APPLICATIONS OF HIGH THROUGHPUT NUCLEOTIDE SEQUENCING

by

JOHANNES WAAGE, M. SC.

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in
Computational Biology and Bioinformatics
at
The PhD School of Science, Faculty of Science,
University of Copenhagen, Denmark.

Committee:
Assoc. Prof. JEPPE VINTHER, Chair, Uni. Of Copenhagen
Prof. ANDERS GORM PEDERSEN, Technical University of Denmark
Assoc. Prof. RICHARD SANDBERG, Karolinska Institutet, Sweden

Supervisors:
Prof. ALBIN SANDELIN (Principal), University of Copenhagen
Prof. BO PORSE, University of Copenhagen

THE BIOINFORMATICS CENTRE, DEPARTMENT OF BIOLOGY
AND
BIOTECH RESEARCH AND INNOVATION CENTRE
UNIVERSITY OF COPENHAGEN

September 1st, 201
“Biology has at least 50 more interesting years.”

- JAMES WATSON, 1989
Preface

This work is presented as the requirement for obtaining the PhD degree at the Bioinformatics Centre, Department of Biology, Faculty of Science, University of Copenhagen. The work was carried out under the supervision of, in no particular order, associate professor Albin Sandelin (Bioinformatics) and professor Bo Porse (Biotech Research and Innovation Centre) from medio 2010 to medio 2013. A portion of the work leading to article I was carried out from medio 2008 to medio 2010 during my master’s thesis. The work was in part funded by the European Research Council (EU FP7 framework programme/ERC grant agreement 204135), and in part by the faculty of science, Uni. Of Copenhagen.
Table of Contents

Preface ................................................................................................................................. 5
English Summary .................................................................................................................. 7
Dansk Resumé ....................................................................................................................... 8
Acknowledgements .............................................................................................................. 9
Author Publications Included In This Thesis ................................................................. 10
Author Publications Not Included In This Thesis ......................................................... 11
Abbreviations ..................................................................................................................... 12
Introduction ....................................................................................................................... 13
Introduction to articles ........................................................................................................ 16
Sequencing and mapping issues ....................................................................................... 20
RNA-sequencing ................................................................................................................ 22
  Milestones and history ...................................................................................................... 22
  Normalization issues ........................................................................................................ 25
  Isoform de-convolution and splice class detection ......................................................... 27
Coding Potential Prediction ............................................................................................... 31
Processing and analysis of CAGE-data ........................................................................... 32
Chromatin Immunoprecipitation coupled with sequencing .......................................... 36
  Milestones and history .................................................................................................... 36
  Peakfinding and normalization ...................................................................................... 37
  Motif discovery .............................................................................................................. 38
Concluding Remarks ......................................................................................................... 40
Original Articles ............................................................................................................... 42
  Article I ........................................................................................................................... 43
  Article II ......................................................................................................................... 63
  Article III ....................................................................................................................... 66
  Article IV ....................................................................................................................... 97
References ......................................................................................................................... 110
Supplemental Figures ....................................................................................................... 116
English Summary

The recent advent of high throughput sequencing of nucleic acids (RNA and DNA) has vastly expanded research into the functional and structural biology of the genome of all living organisms (and even a few dead ones). With this enormous and exponential growth in biological data generation come equally large demands in data handling, analysis and interpretation, perhaps defining the modern challenge of the computational biologist of the post-genomic era.

The first part of this thesis consists of a general introduction to the history, common terms and challenges of next generation sequencing, focusing on oft encountered problems in data processing, such as quality assurance, mapping, normalization, visualization, and interpretation.

Presented in the second part are scientific endeavors representing solutions to problems of two sub-genres of next generation sequencing.

For the first flavor, RNA-sequencing, a study of the effects on alternative RNA splicing of KO of the nonsense mediated RNA decay system in Mus, using digital gene expression and a custom-built exon-exon junction mapping pipeline is presented (article I). Evolved from this work, a Bioconductor package, spliceR, for classifying alternative splicing events and coding potential of isoforms from full isoform deconvolution software, such as Cufflinks (article II), is presented. Finally, a study using 5'-end RNA-seq for alternative promoter detection between healthy patients and patients with acute promyelocytic leukemia is presented (article III).

For the second flavor, DNA-seq, a study presenting genome wide profiling of transcription factor CEBP/A in liver cells undergoing regeneration after partial hepatectomy (article IV) is included.
Dansk Resumé

De nye nukleinsyresekventeringsteknologiers komme har i høj grad udvidet horisonterne og mulighederne inden for forskning i funktionel og strukturel biologi af alle levende (og et par få uddøde) organisms genoms. Med denne enorme og eksponentielle vækst af biologisk data opstår ligeledes store krav til behandling, analyse og fortolkning af data, og således defineres den fremtidige udfordring for bioinformatikkeren i den post-genomiske æra.

I den første del af denne afhandling præsenteres en general introduktion til næste-generationssekvenseringens historie, terminologi og udfordringer, med fokus på typiske problemer indenfor datahåndtering, såsom kvalitetssikring, mapping (placering af sekventerede aflæsningsfragmenter på kromosomet), normalisering og fortolkning.

I den anden del præsenteres 4 videnskabelige arbejder der forsøger at løse nuværende problemer indenfor to undertyper af sekvensering.


For den anden type, DNA-sekvensering, præsenteres et studie omhandlende en fuld genomisk profilering af transkriptionsfaktoren CEBP/As binding i murine leverceller der undergår regenerering efter partiel hepatektomi.
Acknowledgements

A sincere thank-you to my bosses Bo and Albin for being not just that when things get rough.

Thank you to both the Porse group and the whole of Upper Binf for good times!
Author Publications Included In This Thesis

Sorted by date (oldest first). Underlined articles are included in this thesis.


*= shared first author


<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3SS</td>
<td>Alternative 3’ Splice Site</td>
</tr>
<tr>
<td>A5SS</td>
<td>Alternative 5’ Splice Site</td>
</tr>
<tr>
<td>AFE</td>
<td>Alternative First Exon</td>
</tr>
<tr>
<td>ALE</td>
<td>Alternative Last Exon</td>
</tr>
<tr>
<td>AML</td>
<td>Acute Myeloid Leukemia</td>
</tr>
<tr>
<td>APL</td>
<td>Acute Promyelocytic Leukemia</td>
</tr>
<tr>
<td>AS</td>
<td>Alternative Splicing</td>
</tr>
<tr>
<td>ATSS</td>
<td>Alternative Transcription Start Site</td>
</tr>
<tr>
<td>ATTS</td>
<td>Alternative Transcription Termination Site</td>
</tr>
<tr>
<td>BMM</td>
<td>Bone Marrow-derived Macrophages</td>
</tr>
<tr>
<td>CAGE</td>
<td>Cap Analysis of Gene Expression</td>
</tr>
<tr>
<td>ChIP</td>
<td>Chromatin Immunoprecipitation</td>
</tr>
<tr>
<td>ENCODE</td>
<td>ENCyclepana Of DNA Elements</td>
</tr>
<tr>
<td>FANTOM</td>
<td>Functional ANnotation Of the Mammalian genome</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments per Kilobase per Million mapped</td>
</tr>
<tr>
<td>IC</td>
<td>Information Content</td>
</tr>
<tr>
<td>MES</td>
<td>Multiple Exon Skipping</td>
</tr>
<tr>
<td>MX</td>
<td>Mutually Exclusive exon</td>
</tr>
<tr>
<td>NMD</td>
<td>Nonsense Mediated Decay</td>
</tr>
<tr>
<td>ORF</td>
<td>Open Reading Frame</td>
</tr>
<tr>
<td>PTC</td>
<td>Premature Termination Codon</td>
</tr>
<tr>
<td>PWM</td>
<td>Position Weight Matrix</td>
</tr>
<tr>
<td>RPKM</td>
<td>Reads per Kilobase per Million mapped</td>
</tr>
<tr>
<td>SES</td>
<td>Single Exon Skipping</td>
</tr>
<tr>
<td>TMM</td>
<td>Trimmed Mean of M-values</td>
</tr>
<tr>
<td>TPM</td>
<td>Tags Per Million mapped</td>
</tr>
<tr>
<td>TSS</td>
<td>Transcription Start Site</td>
</tr>
</tbody>
</table>
Introduction

The history of nucleic acid sequencing started with the complete sequencing of the RNA gene of a bacteriophage coat protein in 1972 (1), based on the method of incorporating dye-coupled chain-terminating nucleotides, published by Frederick Sanger a few years later (2). An impressive feat at that time, this deduction of the full 387 nucleotides of sequence marked perhaps the earliest beginning of the era of the genome – an era which is currently experiencing its renaissance in the 21th century with the advent of the next generation of nucleotide sequencing.

From the early 1970s to the late 1990s, technological advances in sequencing were significant but somewhat slow compared to the evolution of the microprocessor (driving many other scientific advances), and sequencing of large full genome projects required large machine-parks and extensive collaborative efforts. Following the first fully sequenced genome, the 5386 nucleotides long DNA of the phiX174 phage (3) (and later the first genome to be fully synthesized and assembled in vitro (4)), several small genomes of prokaryotes were sequenced, an by the late eighties, an international consortium of more than 74 laboratories and 600 scientists began a 7-year effort to sequence the first eukaryotic and large genome of *Saccharomyces cerevisiae*. Completed in 1996 (5), this 12-megabase genome was a huge step forward in terms of sheer size, but the 3000-megabase human genome was still a daunting, distant and unfeasible task. But the foundation was laid, and from the yeast genome project (and other early sequencing projects) sprang a plethora of technological and computational disciplines and methods. And perhaps just as important, from these early sequencing projects rose the cross-disciplinary field of bioinformatics, combining sequence analysis with mathematical and statistical models of machine learning, allowing for sequence alignment, gene prediction, structure prediction, and so on. An iconic example of an early bioinformatics algorithm, BLAST, published in 1990 (6), was perhaps the foundation of fast computational analysis of biological sequence.

The obvious culmination of the first era of sequencing was the complete deduction of the human genomic sequence, an effort first planned in the early 80s,
initiated in 1991, and completed in 2001. A multi-party race to finish, private company Celera intended to achieve the completion of both sequencing and assembly first, raising fears that genes and genomic sequence could be patented and not available on the open domain for research. Initial efforts to assemble the genome from the public sequencing consortium were unsuccessful, but on June 22, 2000 GigAssembler (7), a assembler written by then graduate student at University of Santa Cruz in California (UCSC) Jim Kent completed the job just days before Celera, and a few weeks later, the first draft of the human genome was made available at the newborn human genome browser at the UCSC (8), releasing the genome into the public domain.

Concurrent with the human genome project, still based on the Sanger sequencing methods invented in the 1960s, companies were trying to reinvent nucleotide sequencing. Several full-fledged automated high-throughput sequencers appeared on the scene around 2005, and underwent several iterative improvements to both speed, accuracy and cost over the next 8 years. At time of writing, such machines are relatively widespread, and used in a number of research disciplines. Omicsmaps.com, a user-updated map of sequencers in academia, show 23 machines in Denmark alone.

Three platforms currently dominate the market, based on different approaches for deducing nucleotide sequence in parallel: Roches/454s pyrosequencing is based on fixing DNA to micro-beads, followed by amplification and sequencing, and currently supports read lengths up to 1000 nt, it's main advantage. Applied Biosystems SOLiD is based on ligation and two-base encoding (colorspace), and has high accuracy as its primary advantage. Illuminas Solexa platform is based on sequencing-by-synthesis, and has the highest throughput and the lowest cost of the three (9). As of 2013, Illumina by far has the largest marketshare (10), and all of the work included in this thesis is based on data from this platform. The different sequencing platforms will not be covered more in detail, but for a comprehensive comparison and overview, including advantages and disadvantages, please refer to (9).
Much of the work presented here was initiated between 2008 and 2010, a time where the field of high-throughput sequencing and its sub-disciplines, such as RNA-seq and ChIP-seq, were in their nascency. Working with the data presented an array of interesting (and some not so interesting) problems with few ready-made tools and pipelines available, necessitating a healthy dose of experimentation. Fortunately, the bioinformatics discipline borrows a lot of its mentality and affinity for open-source from computer science in general, and from early on, researchers and programmers were sharing experiences, approaches and source code online.
Introduction to articles

The articles presented herein represent the main work of this dissertation, as well as some additional work initiated during my master thesis (article I). Here, a brief motivation, introduction to the subject matter as well as an overview of my contributions and approaches to each work is presented. A more detailed treatise on the methodologies used follows in each section of the general introduction, referencing to each article as needed.

Obviously, the bioinformatic method cannot be wholly separated from the underlying biology, but it should be noted that the aim of this thesis is not to carry the reader through the different biological systems that are the focus point of the articles included (although a brief introduction is given beforehand), but rather to investigate the history, techniques, challenges and solutions in current appliances of next-generation sequencing, and selected examples on how these were used in the included works.
In eukaryotes, most multi-exon genes undergo alternative splicing, the process of regulated shuffling of genomic coding entities (exons) in a combinatorial manner, creating several distinct mRNA molecules (and thereby protein products) from a single gene unit, and this process is believed to be a significant contributor to complexity of higher organisms. Such flexibility in any biological system often carries with it a level of noise, and in this case, a potentially detrimental kind.

NMD or nonsense mediated decay is a general mRNA pathway in eukaryotes that senses aberrantly spliced transcripts, or transcripts that would otherwise lead to potentially dysfunctional protein products, and removes these by means of nucleolysis and degradation.

For this paper, originally planned in early 2008, (and at the dawn of RNA-seq technology) we sought to take advantage of the new digital platform to carry out the first genome-wide and non-discriminative analysis of the effects of NMD KO (facilitated by a murine KO of the essential NMD-factor Upf2) in an mammalian system on the global splicing pattern as a whole. The study gave rise to several bioinformatic challenges: dealing with the nascent RNA-seq data as a whole; creating an artificial exhaustive genome of exon-exon splice junctions facilitating deconvolution of full isoforms based on junction evidence (this was before paired-end sequencing was widespread), predicting coding potential of each transcript, and downstream interpretative analyses, including splice class classification, motif search, and sequence conservation assessment. A challenge for this data was to meaningfully describe, quantify and visualize “noisy” transcripts rescued by NMD KO, without setting inhibitory filter thresholds that would otherwise remove these. Perhaps in some sense a boon for this project, using biological replicates for sequencing was not feasible cost-wise, and post-processing with replication would potentially have eliminated some of the “stochastic” noisy splicing.

Overall, the study was successful, and we found that NMD KO upregulates a large population of non-productive mRNAs through splice factor deregulation, and
extended the understanding of NMD from an mRNA quality pathway to a regulator of several layers of gene expression. The study was published in Genome Biology in 2012 (11).

Paper II: a package for splice class detection – under preparation (formatted as a Bioinformatics Application Note) (2013)

Our NMD-paper produced a more-or-less ready to run pipeline (RAINMAN) for the community to run on their own RNA-seq data. But by 2012, paired-end RNA-sequencing was the norm, and serious advances had been made in the field of full-length isoform deconvolution based on robust statistical methods, including both expectation–maximization and bayesian methods. One such tool published was Cufflinks, part of the popular Tuxido suite for RNA-seq analysis; Cufflinks were among many other studies used for the assembly of RNA-seq data in the large scale ENCODE study of ultimo 2012.

As such, RAINMAN was outdated, and Kristoffer Knudsen (master student in the Sandelin lab) and I set out to write a new tool for splice class detection and coding potential prediction based on full length isoforms. From this emerged spliceR, an R and Bioconductor package that allows for user-friendly annotation of assembled RNA-seq data from Cufflinks (or any other full-length assembler), and outputs a range of useful data, including isoform fraction switch scores and genomic locations of differentially spliced elements for downstream analyses. spliceR was accepted into the Bioconductor 2.13 release in October ’13 (12), and the manuscript is currently under preparation.


A different application for RNA-seq is CAGE-seq, or Cap Analysis of Gene Expression. This method effectively captures small sequence fragments from the 5’ end of capped RNA molecules, which are in turn sequenced, and processed to generate a genome wide map of promoter usage. In humans (and other higher organisms), genes
often have more than one transcription start site, allowing for a diversification of the resulting protein product, similar to and in concert with alternative splicing. For this study, we took advantage of in-house cell-sorting facility and expertise combined with access to samples from patients with acute promyelocytic leukemia (APL). The data for this paper was particularly challenging, as isolatable material from the (liquid) tumors was scarce, resulting in low input material for the method. This generated a lot of technical noise, and necessitated rigorous filtering methods. After processing, we found promoters with upregulated usage in APL, which were often located downstream of annotated promoters, while promoters supporting the annotated longest transcript variants were commonly downregulated in cancer. This manuscript has been prepared, and is awaiting experimental validation of selected targets.

Paper IV: Transcription factor profiling in liver regeneration – Genome Research (2013)

Another major advance for biology brought about by the high-throughput-era was the coupling of traditional transcription factor binding site profiling by ChIP to sequencing, a field previously limited by the microarray platform’s limited and pre-selected number of probes. In this study we presented one the first genome wide studies of temporal transcription factor dynamics – the binding patterns of key hepatocyte factors CEBPA and CEBPB during liver regeneration following partial hepatectomy in a mouse model.

Combined with DNA polymerase II binding data, as well as large array of in-silico binding predictions of other transcription factors, we described a circuitry of temporal codes, binding classes, and mutual exclusive motifs. Major challenges for this study were many and included handling and processing (at that time nascent) ChIP-seq data, peak calling, filtering and clustering techniques using unsupervised approaches. In addition, considerable effort went into generating a high information / low redundancy set of transcription factor motifs from three large repositories (JASPAR, Transfac and UniProbe).
Sequencing and mapping issues

A number of issues are shared between sequencing platforms and appliances pertaining to read length, read quality and trimming and mapping. Although not one of the most exciting issues of sequence analyses, this short section will briefly cover the most common challenges, and how some of these were overcome in the presented works.

Sequence data generated by current platforms is exported in the FASTQ format, a sequence format that combines the raw sequence format FASTA with the quality data. Each platform has its own idiosyncrasies in terms of sequencing quality; the Illumina platform, for instance, generally has a low quality towards the edges of the flow cell, in the very beginning of the sequencing run and trailing off towards the end of the run (data not shown). Most common workflows include a read quality visualization and trimming step for ensuring optimal mapping. In addition, this step can include the removal of sequencing adapter sequences, and identical reads arising from PCR amplification artifacts. Separated from the analytic and interpretive side, a large amount of time is spent readying and "massaging" data in bioinformatics. For article I, this was done manually in Perl and Python. For article III and IV, this was done using FastX-tools (13) and FastQC (14).

Perhaps the pivotal step in sequence processing is the mapping of short sequenced fragments back to the relevant genome, a field where significant progress has been made to accommodate the much larger amounts of data since the early days of BLAST (6). The seed-and-extend approach of BLAST is still applicable in some cases, but modern mappers applying more efficient forms of genome indexing have significantly increased the speed of mapping, a necessity when dealing with millions of reads. Except for the ELAND mapper proprietary to the Illumina pipeline, MAQ was one the earliest attempts at a fast short-read mapper (15), using a hash table of the genome a look-up index, incorporating the quality scores into the mapping algorithm. A bit on the slow side (typically requiring > 1000 CPU hours for larger sequencing projects on workstations at the time of publishing), a big stride forward was made with
Bowtie (16), utilizing the Burrows-Wheeler transform for encoding the reference genome. This transform, published by computer company HP in 1994 as a lossless compression algorithm, proved to be an ingenious way of compressing the genomic index to a size manageable by most workstations (<2GB RAM), but at the same time allowing for very rapid lookup of sequences. The bowtie suite of tools has later been expanded with an RNA-seq assembler and isoform deconvolver (described later), and is at the time of writing still the most widely used short read mapper in research (based on number of citations).
RNA-sequencing

Like the first fully sequenced biological polymer in 1972, the first two works included in this theses is based on sequencing of RNA. This section will briefly cover the history and important landmarks over the last few years of RNA-sequencing by next generation technologies, and will subsequently dig deeper into common problems and issues with the presented works as main focus.

Milestones and history

RNA-seq was introduced to the research community as a group of 5 papers coming out at approximately the same time, but the first significant paper using RNA-seq and paving the way for digital gene expression was by Mortazavi et al. in 2008. An important benchmark paper, their results showed high inter-sample fidelity between technical replicates (later confirmed in (17), arguing that technical replicates are per se not needed in RNA-seq), and, using in-vitro synthesized transcripts as spike-ins, high sensitivity and dynamic range, establishing the method as a serious competitor to the previously ubiquitous microarray-based gene expression platform (18). This work also established the first simple normalization and quantification methods and metrics, and showed, perhaps surprisingly, a correlation with arrays with a R^2 at around only 0.7. More recent advances in data processing and analysis have increased this correlation to around 0.8 (19). Whether the remaining difference owes to the shortcomings of the array platform, the immaturity of RNA-seq processing, or a combination of the two, is not yet readily discernible.
Later the same year, the first global study of alternative splicing using RNA-seq was published in Nature by the Burge lab (20). Surprisingly, this study found that more than 90% of all human genes undergo alternative splicing. Classifying eight different types of alternative events, authors found strong tissue-specific and highly conserved “switch-like” splicing events. To accomplish this extensive mRNA isoform discovery, reads were mapped to an artificial custom genome based on a transcript database of all possible exon-exon junctions, in turn allowing capture of all possible splice events (between the annotated exons from the selected repositories, at least). Until recent advances in full isoform deconvolution, this approach was popular, and was also employed in a modified version by the RAINMAN pipeline presented in paper I (as described below).

Much progress followed in the following years, not all relating to digital gene expression and alternative splicing. RNA-seq found other interesting usages including sequencing of non-coding RNA-species, assessment of allele-specific expression and detection of fusion genes (one of the first works on this is (21)). Technical advances were also made, including single cell RNA-seq (22) and strand-specific RNA-seq (23). Similar advances were made on the software side (some will be discussed later).
including software packages for normalization, differential calling, and visualization, and reference-free assembly (they are too numerous to be referenced here).

Perhaps the pinnacle of research in genome wide transcription profiles was the publication of the ENCODE consortium's large scale sequencing of the transcriptome of 15 human cell lines in 2012 (24). This study validated the previous claim of almost all human multi-exon genes undergoing alternative splicing, and showed that at least 60% of the full human genome is transcribed (but not necessarily translated), perhaps the final nail in the coffin of "junk-DNA". Using Tophats/Cufflinks for isoform assembly (discussed below), ENCODE made several interesting findings – perhaps one of the more curious ones showed, that as the number of annotated isoforms increase for a given gene, the fraction of the most expressed isoform (the major isoform) decreases (signifying biological importance of multiple isoforms, Error! Reference source not found.a), but that most genes reach a plateau at 10-12 isoforms (perhaps signifying a rate-limiting or energy-wise prohibitive value of expressing too many differentially spliced isoforms, Error! Reference source not found.b). This is a good example of a bias-free finding facilitated by sequencing, which would have been impossible just 10 years ago with microarray-technology.

In the relatively short life of RNA-seq, the field has been undergoing rapid advances, and methods are ever evolving. This sometimes translates to older methods being called obsolete by newer, as will be evident of this follows in the section about isoform deconvolution.
RNA-seq normalization. Left: RNA-seq data from Upf2 KO mouse liver (data from article I) after RPKM normalization. Red dots represent the 50 most expressed genes in WT, green and pink dots represent 50 random sampled genes from the second 2nd and 4th bin of an 8-bin quantization of expression, respectively. Note how neither subset straddles the black line of no DE. Right: same data after quantile-normalization, one of many normalization methods available. Note: this figure is for illustrative purposes, and does not relate to method used in paper I.

Normalization issues

A pervasive RNA-seq challenge is the problem of normalization, an issue that has been apparent in almost all data this author has been in touch with. In this context, normalization is understood as the action of adjusting count values proxy of transcript or gene expression in a way that adjusts for unwanted (often technical) bias, and makes comparisons between and within samples meaningful. Normalization methods can roughly be divided into two groups - the first group, intra-sample methods, normalize each given data point (transcript or gene count) using a specific approach. One of the earliest proposed intra-sample methods was that of RPKM (Reads Per Kilobase of transcript length per Million mapped reads), simply normalizing coverage of a given transcript to the transcript length and the sequencing depth (18). Even though the RPKM method has been shown to have some pitfalls in relation to gene length bias (in particular in the over-estimation of lowly expressed transcripts)(25,26), it is still a popular method and readily gives a assessment of the relative levels of transcripts intra-
sample-wise. Another method, borrowed from microarray analysis, is quantile normalization.

RNA-seq doesn’t disclose any information about the total RNA levels in the cell. This have sometimes been dubbed the “sequencing real estate” problem and stems from the notion that in some experiments, as a given sample is sequenced at a given depth, larger groups of RNA species can be more or less abundant in only one sample. After the RPKM normalization method is applied, ranges of expression can be skewed towards or away from this sample (over- or undersampling), although this does not represent an actual difference in expression (but rather a difference in absolute RNA abundances), and this can be potentially dangerous in downstream significance calling. An example of a skewed RNA-seq expression profile is given in figure 2. Here, the highly expressed range of genes is skewed towards the WT, resulting in undersampling of the remaining expression space. This phenomenon is somewhat akin to issues of probe saturation found in microarrays, but additionally gives rise to more peculiar distributions of inter-sample expression comparisons, where difference in absolute abundances in certain ranges of expression may produce a local skew - e.g. “banana-like” scatterplots (data not shown). The second group of normalization methods, inter-sample methods, generally calculates normalization factors to be applied equally to all data-points for a given sample to account for differences in the true abundance of RNA. The TMM method, used in paper I, and part of the EdgeR Bioconductor package (27), assumes that technical variation can be modeled using the poisson distribution, and biological variation using the negative binomial distribution, a assumption that also holds true for other seq-data, including ChIP-seq. Using this variance as weights, TMM simply proposes an inter-sample normalization procedure by calculating a normalization factor based on the trimmed mean of log expression ratios, under the assumption that most genes are not differentially expressed (28).
Since the publication of paper I, several methods have been proposed in literature regarding RNA-seq normalization, and it’s now evident that this issue is just as pertinent for RNA-seq as it was for microarray. The results from one recent comparison study including several methods (29) confirmed earlier suspicions that the RPKM (and by extension, the FPKM metric used by Cufflinks) metric is a bad choice due to gene length bias issues, and found DESeq (30) and TMM normalization to be most robust (Figure 3). A combination of RPKM and TMM normalization, as was applied in paper I, wasn’t tested, but is expected to inherit some of the strengths and weaknesses of both methods.

**Isoform de-convolution and splice class detection**

Initial approaches in detecting alternative splicing was based on creation of artificial genomes, consisting of all possible exon-exon junctions generated from an mRNA repository of choice. For article I, we developed our own pipeline based on this approach, RAINMAN. This method, primarily motivated by single-end sequencing being the available data, and by its relative simplicity, has a few inherent pitfalls. Firstly, each alternative splicing event is only detected based on co-occurrence of two exons, and the inter-dependence between multiple events cannot be determined. In our case (article I), the observation that more than 95% of junctions mapped to concurrent exons of annotated, productive and coding mRNAs, provided

---

**Figure 3:** A comparison inter- and intrasample normalization methods for RNA-seq. TC = total count, UQ = upper quartile, Med = median, DESeq = normalization by the R-package of the same name, TMM = trimmed mean of M-values (edgeR), Q = quantile, RPKM = reads per kilobase per million mapped, RawCount = raw read counts (no normalization). Different normalization methods produce widely different results depending on their approach, and whether they’re intra- (Q and RPKM) or inter-sample (the rest). Adapted from Dillies et al., 2012
the argument that although we couldn’t detect multiple splice events within the same messenger, those would statistically be relatively rare \( (0.05^2 = 0.0025) \), assuming no dependency between events. We dared this normally dangerous assumption, as noisy splicing was thought to happen in a random fashion.

Under this assumption, the remaining exons of the longest supported annotated isoform from a array of public repositories (Refseq, UCSC knownGene, Ensembl, etc...) was attached to each exon of the respective junction, and thus the full length isoform inferred, and, later, its coding potential predicted. This exhaustive exon-exon junction catalog also provided a simple but suitable platform for inferring alternative splicing events, and the alterations of NMD KO on AS. We chose to detect the classes of AS recognized in literature as the minimum informative set, namely single and multiple exon-skipping (SES/MES), alternative 5’ and 3’ splice sites (A5SS/A3SS), alternative first and last exons (AFE, ALE) and mutually exclusive exons (MX). A graphical overview of these classes is given in Supplemental figure 1. The detection algorithm was detected on an iterative approach – in the first step, for each given splice junction, junctions with similar left or right exonic edges (or coordinates) were isolated. Next, the connections and supporting exons of these junctions were included in the model, and depending on the trace back to the original junction, and the inclusion or exclusion of supportive or inhibitve junctions and exons, the splice class was deduced. An example is given in Figure 4 for two splice classes. For (A), two junctions (contJunc1 and contJunc2) with identical left and right coordinates with

![Figure 4: Classification of alternative splicing events based on exon-exon junction evidence. Refer to text for details.](image-url)
the single junction testJunc, and with evidence supporting only one exon between those two junctions, classifies as SES. For (B), a junction (testJunc), sharing right coordinates with another junction (juncShareRight) and left coordinates with an exon, which itself has support for an additional exon sharing left coordinates, but with different right coordinates, from which a junction (juncShareRight) has identical right coordinates with juncShareRight, classifies as A5SS. Similar logic was applied for the remaining splice classes. It has to be noted, that although desirable, exon read coverage was not included in the classification, as time did not permit this feature to be implemented.

As a nicety and convenient functionality, visualization of splice junctions color coded by splice class for export to genome browser was included in the RAINMAN software. Figure 5 shows a classical example exon-exon junction coverage of the NMD-regulated splice factor Srsf7, color coded by splicing class – in NMD KO tissues, isoforms including a PTC-containing exon (marked in yellow), normally spliced in by the splicing factor itself in a auto-regulatory manner, are rescued.

For the work presented, this platform was sufficient and successful in detecting a marked phenotype of unproductive PTC+ transcripts and a global alteration in alternative splicing in the two NMD KO systems utilized.

Figure 5: Screenshot from the UCSC genome browser, showing bonemarrow-derived macrophage (BMM) RNA-seq coverage in Upf2 WT vs KO of the NMD-regulated splice factor Srsf7 (Sfsr7). Green junctions represent alternative splice sites and red junctions indicate single exon skipping. Numbers above junctions including or skipping the yellow auto-regulatory STOP-containing exon are relative expression values.
Around 2010 perhaps, paired-end sequencing was becoming the norm in RNA-seq. Here, both ends of the library insert are sequenced and can be mapped to the genome. This information, combined with the known insert size as well as exon coverage / junction-evidence, enables modern RNA-seq assemblers to attempt full-length transcript de-convolution. Perhaps the most widely used and best cited is the Cufflinks assembler (31), part of the ubiquitous Tuxedo-suite from the hand of Cole Trapnell. Cufflinks achieves de-convolution by a series of simple steps (Figure 6). After initial mapping of reads to the genome, and de novo splice junction detection (A), all mutually incompatible splice fragments are reduced to the parsimonious number of possible paths through the splice graph (B, C). For genes with several splice events, this combinatorial approach generates a large number of potentially false positive isoforms, and this initial step is similar to the junction-based approach described above and presented in paper I, where “canonical”, annotated exons are simply concatenated to each junction. In perhaps the pivotal step, each putative isoforms compatibility with the evidence is determined, using exon coverage and the fragment length distribution (D) (either determined the mate distance in pre-annotated single exon genes or given by the user), setting the prior

Figure 6: RNA Deconvolution by Cufflinks. Refer to main text for details. Adapted from Trapnell et al., 2010
abundances of each transcript. Finally, a minimum cost probability function is maximized, finding the final set of correct parsimonious isoforms, and abundances are reported (E). This maximizing approach vastly improves isoform inference (31).

Other approaches exist to RNA-seq assembly (for a recent review, see (32)), but the Tuxedo suite has proved to be robust in benchmarks, and was among others used as assembler in the ENCODE project, and for our splice class detection software spliceR (paper II), implemented in R and available in the Bioconductor project, we choose to build directly on Cufflinks-assembled transcripts (although other full-length assemblers are supported). As several earlier efforts in splice detection from RNA-seq data has been made, the main aim for spliceR was easily interoperability with R, Bioconductor and Cufflinks. Another major point was complete classification of all splice classes, as well as no requirement for pre-annotated splice event or transcript data. In addition, the option to import the genomic locations of elements spliced in/out for each alternative splicing type was included. This should be a powerful tool for downstream analysis of systems where global alternative splicing is altered, including sequence analysis and motif finding, and phylogenetic analyses. At the time of writing, three collaboration projects take advantage of spliceR, and we hope to see a good amount of interest in the package, when Bioconductor 2.13 is released.

Coding Potential Prediction

Coding potential prediction of mRNAs is in itself not a super-tough problem, but a few different approaches exist. PhyloCSF (33) and CPC (34) have an evolutionary approach, using alignments of potential coding regions combined with machine learning approaches to scan for the frequency of synonymous and conservative substitutions, scoring for coding potential. Coding potential prediction. CPAT (35) bases its score on both the length of the longest ORF in all reading frames, as well as other sequence linguistic features.

The approach chosen in article I to predict coding potential and position of (potentially premature) STOP codons was based on the concept, that in the NMD KO, events representing aberrant splicing are independent (as exemplified by 95% of
junctions mapping to canonical junctions), and thus annotated ORFs can be applied to transcripts generated by exon concatenation and scanned trough. The approach is outlined in Supplemental figure 2.

For spliceR (article II), the same approach was chosen, scanning full isoforms with the most upstream compatible ORF, if any. Cufflinks, however, often finds many novel genes and isoforms, and spliceR in general only annotates 30–35% of found isoforms with ORF and STOP positions, and heuristic approaches, e.g. as presented in CPAT, would be a desirable addition to the package.

Processing and analysis of CAGE-data

An entirely different flavor of RNA-seq is CAGE-seq, or Cap Analysis of Gene Expression combined with high-throughput sequencing. CAGE is based on the sequencing of 5' ends of RNA, and when these short sequences, referred to as "tags", are mapped to back to the genome, a genome-wide map of transcription start sites (TSSs) can be inferred, and the number of times each TSS is represented can be taken as a proxy of that transcript's expression (36). This technique was originally devised as a technological platform for the FANTOM project for genome-wide profiling of promoters (36). Obviously, regular RNA-seq intrinsically has the power to map 5' transcription start sites, but with lower precision, both due to lower coverage as well as coverage bias – all parts of the mRNA is not covered equally, and in some cases, the 5'UTR may not be covered at all.

For paper III, blood was taken from human patients diagnosed with acute promyelocytic leukemia (APL, a cancer of the myeloid lineage of blood cells), as well as from healthy patients, and thus very little biological material was available from the purified and cell-sorted samples. To facilitate a promoterome analysis, we applied a variant of the CAGE technique dubbed nanoCAGE, which allows for small library input sizes (37,38). This was the first time this technique had been applied to human clinical samples, and served as a proof-of-concept for the method in this setting.

In this low-input material scenario, PCR-amplification bias was expected to be a significant issue, and after initial observations of the data, we had some doubts to
whether any meaningful biological signals were to be deducted. Applying rigorous filtering and thresholding, however, resulted in an interesting phenotype - a switch in transcript-length preference between WT and cancer cells: After initial read quality filtering, and mapping to the genome, the following four filters were applied:

1. PCR-bias: Removal of all tag “towers” with a width of one read (in normal CAGE, this could be dangerous, removing several meaningful TSS clusters, but manual inspection of the nanoCAGE data showed, that these towers cropped up in a more random fashion.)

2. High-pass expression filter: in CAGE, the analog to RNA-seqs RPKM is tags per million (TPM), but without the gene length normalization. To further weed out technical noise, an expression of 2 TPM was required.
3. Variance: as some heterogeneity in patients with APL is expected due to unique cancer genotypes, some of the variance between the cancer samples may not be due to noise. None the less, variance obfuscates downstream analyses and significance calling, and clusters with a coefficient of variance above 1 were removed.

4. Exon noise filtering: doing “birds-eye” analysis of the data, an amount of “exon-painting”, or low-coverage reads spanning exonic areas, were evident, possibly due to capture of degradation products. To avoid clusters being called in these regions, clusters were required to be expressed more than 1.5 times the average exonic coverage for the relevant gene, for some analyses. Variance filtering also contributed to countering this problem.

Figure 7: For nanoCAGE data, rigid filtering is required to weed out noise - lowly expressed and narrow clusters are generally not associated with pre-annotated TSSs, motivating filtering on these parameters. (Distance to nearest annotated TSS on the x-axis).
Some of these filters, and the motivation for the thresholds may seem arbitrary. Figure 7 represents three variables, TPM (color coding), width of clusters (Y) and distance from closest annotated UCSC knownGene transcription start site. Although an intrinsic strength of the CAGE technique is promoter discovery, many promoters detected in this experiment represent existing pre-annotated promoters, and the distance to the nearest TSS can be used as a quality measure of different filtering thresholds. From the plot, distance to nearest TSS is much tighter for higher expressed clusters, and for wider clusters, legitimizing filters for both expression and width.

In CAGE, tag clustering and filtering is just one component of the data analysis, for which many parts are similar to those of RNA-seq, including gene ontology analysis, overlap with other datasets, network analysis, etc. One specific analysis of paper IV was specific to CAGE, however, and shows well how the ever-growing amount of public data can often be taken advantage of. To deduce more about the regulatory neighborhood of our putative TSSs (lacking any transcription factor ChIP-seq data for these samples), we instead generalized the question by downloading all *Mus* transcription factor binding sites from the ENCODE project (across cell lines), comprising more than 125 data sets in replicates. The distance distribution from putative TSSs to the nearest transcription factor binding site was then measured and visualized. Although no mechanistic inference can be made directly, as tissues and sample conditions are not comparable, this is an example of large-scale data integration in genomics.
Chromatin Immunoprecipitation coupled with sequencing

The work done in the fourth article presented in this thesis is based largely on data from sequencing of DNA-fragments bound by certain transcription factors (or ChIP-seq), and this section will focus on common challenges in analyzing ChIP-seq data and how these were countered in the presented work.

Milestones and history

The predecessor of ChIP-seq is ChIP-chip or ChIP-on-chip, using chromatin pull downs coupled with DNA micro arrays. Before the first papers using high throughput sequencing were published in '07/'08, this was the method of choice for assaying global transcription factor binding profiles, but obviously had its limits in the same way that microarray analyses for gene expression is limited, most apparent being the restriction to a pre-defined set probes. As such, researchers could only probe a very small fraction of a mammalian genome, resulting in arrays covering, for instance, only known promoters, or exonic areas (39).

Harnessing the (theoretically) non-discriminatory power of sequencing, a natural evolution of the method was given, and the first papers utilizing ChIP-seq was published in Nature and Cell in 2007. In the first, profiling the global binding of transcription factor STAT1 in interferon-treated versus normal HeLa cells, more than 40,000 putative binding regions were presented with, compared to todays standards, relatively few mapped reads (40), and this paper established a methodological and statistical framework for signal processing of ChIP-seq data, that is still in use today. The second paper presented a genome-wide map of histone lysine and arginine methylations, providing new insights in epigenetic regulation that was not readily obtainable with older technologies, and likewise defining much of the data analysis terminology (41).

Since then, a slew of papers have been published using ChIP-seq, and the largest effort so far, similar to RNA-seq, is most likely that of the ENCODE project, integrating data from thousands of experiments and sequencing runs into a combined
regulatory lexicon, describing the transcriptional, regulatory, and epigenetic landscape of the human genome (24,42–44).

Peakfinding and normalization

Perhaps the primary challenge in analyzing ChIP-seq data of transcription factors (as well as histone modifications) is the detection of peaks, i.e. DNA regions that can statistically be induced to be bound by the protein of interest, compared to the background, defined either by the signal-to-noise ratio in the sample, by a mock IgG sample, or both. Since the introduction of next-gen sequencing, developing peak finder software have been an active pursuit in genomics, and to date, there is more than 20 published attempts at solving this problem. Peak finders generally differ in the signal scanning process, statistical models used, and in sensitivity for detecting different regions (e.g. broad histone marks vs. narrow transcription factor peaks).

As mentioned above, a statistical framework based on modeling the background genomic coverage (often referred to as lambda) using a Poisson distribution, and calculating thresholds and significance levels based on this was
presented in the first effort on ChIP-seq, as was subsequently carried through in one of
the earliest most popular peak finders, MACS (45). MACS refined peak calling by
introducing a local lambda (defined as either 1, 2 or 10k nt around the peak) to correct
for local biases, as well as the employment of a shift size modeling step, trying to
deduce the DNA library insert size for more accurate peak definition.

For article VI, MACS and uSeq (a peak caller with similar inner workings),
were employed for our sample pool size of 16 (two TFs, 8 time points). An example of
the binding landscape and peak definition is given in Figure 8. When quantifying
binding intensity for each sample, a common reference for inter-sample comparison is
required (akin to gene or transcript ids for RNA-seq). As each sample was peak-called
separately, the common reference was simply defined as the area under all overlapping
peaks for all samples – analog to the 1-dimensional “shadow” simultaneously cast by all
samples on the genome. Having these consensus peaks, regular downstream
bioinformatic analyses could follow, including clustering, gene-association, network
analysis, and ...

Motif discovery

Discovering transcription factor binding motifs has been a pursuit in genomics
for some time, and data from ChIP-on-chip and ChIP-seq methods is a natural
catalyst for this. Three large repositories containing binding motifs exist, and were the
basis of the combined high-confidence reduced motif bank generated for article IV:
JASPAR, open-source, and containing a small, non-redundant and literature-backed
set, primarily based on the SELEX method (46); TRANSFAC (47), a commercial set
of binding motifs, but with a limited open access set; and UniPROBE (48), an open
access library based on protein-binding microarray data. Our purpose was to reduce
the data from these three sources into something smaller, more manageable and of
higher quality and lower redundancy. Aggregation of all available models resulted in
939 position weight matrices (PWMs), and the smaller, final high-confidence set was
generated as follows:

1. Removal of duplicates
2. End-trimming of models, removing all subsequent nucleotides with an information content (IC) less than 0.5 bits.

3. Removal of all models with a total IC of less than 8 bits.

4. Scoring of all models against each other using the spearman correlation coefficient, and subsequent hierarchical clustering.

5. After inspection of the distance tree, a threshold of ~500 clusters were chosen (figure 9) based on biologically meaningful clusters, and within each cluster, the motifs were aligned using Smith-Waterman local algorithm. The consensus alignment motif was extracted, and subsequently end-trimmed and filtered as in step 2 and 3.

This approach resulted in a little less than 500 high confidence motifs, which were the basis of the binding motif profiling used in article IV.
Concluding Remarks

In the 14 years that has passed so far of the 21st century, genomics can truly be said to have taken centre stage, in no small part due to the advent of high throughput sequencing. The genomic era, perhaps starting in 1972 with the first determination of the sequence of a gene, and perhaps ending with the final draft of the human genome, is now over, and the post-genomic era has begun. But as easy as it is to fling around clichés about the marvels of this time of cutting edge research, evenly hard is it perhaps to pinpoint defining key research discoveries, and this is an inherent challenge in modern day genomics. As the amount of data grows larger and larger, research has a tendency to move away from hypothesis-driven manipulation of simpler systems to genome-wide, organism-wide profiling of any biological information system.

The articles presented in this thesis all present attempts to solve problems in contemporary biology by applying the methodology of quantification, classification and statistics on data from high throughput sequencers.

They were all challenging in the sense that no clear-cut hypothesis existed a priori, in least in terms of analyzing the sequence data. This required an exploration-heavy approach, which when most frustrating (and it will inevitably be at times) could be likened to a “fishing expedition” and when most giving, an unbeatable feeling of actually finding the needle in the haystack. I found, through trial-and-error and an increasing confidence in dealing with big data, that a certain amount of systematization and consistence in data handling and analysis should be a priority in bioinformatics projects. A wide variety of tools and pipelines exist, and even though some general approaches are agreed on in an implicit manner in the field, few standard operating procedures exist (one example of such is the ENCODEs complete guideline for ChIP-seq studies, from lab-bench to computer (49)) – the data is only as good as the tools you use to analyze it with! The Bioconductor project is a good example of an inter-operable, common (and open source) platform for genomics, and with spliceR
(presented in article II), we hoped to present a valuable addition to that project. And as applications and tools mature with the field as a whole, more standards will evolve.

As discussed in the introduction, sequencing has become an accepted and widely used method in most laboratories that do some sort of genomic research, and a lot more robust and well-documented tools and pipelines have been published since. It is this author's observation, that the gap between researchers from a classical biological background (and with few skills in computer science), and researchers with a background in computer science is decreasing, and most likely, future researchers in genomics will be equally well trained in both disciplines. "Big Data" is not a phenomenon restricted to genome biology, and undergraduate programs in science (and other disciplines) are aware of this.

The future in sequencing is promising, especially as the clinical domain starts to accept the new possibilities fully, in part also facilitated by decreasing costs and improved opportunities for running hundreds of samples by the means of multiplexing. With the increased processing and normalization quality of RNA-seq, this platform is expected to completely replace microarrays, with its non-discriminative approach, offering de-novo transcript detection, characterization of alternative splicing, etc. And as reads get longer, and transcript assemblers more advanced, we will come closer to solving the isoform problem fully. For ChiP-seq, the advantages over ChIP-on-chip are just as apparent, and even though large-scale efforts are increasing our knowledge of the regulatory landscape repeatedly, luckily there's a few thousand more transcription factors in the human genome, who hasn't been described yet, and just as many cell types! Combined with profiling of epigenetic marks, genome resequencing and all the other wonderful appliances of sequencing, the unraveling of the complex layers of life goes perhaps a little faster.
Article I

Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns
Mammalian tissues defective in nonsense-mediated mRNA decay display highly aberrant splicing patterns

Joachim Weischenfeldt, Johannes Waage, Geng Tian, Jing Zhao, Inge Damgaard, Janus Schou Jakobsen, Karsten Kristiansen, Anders Krogh, Jun Wang and Bo T Porse

Abstract

**Background:** Nonsense-mediated mRNA decay (NMD) affects the outcome of alternative splicing by degrading mRNA isoforms with premature termination codons. Splicing regulators constitute important NMD targets; however, the extent to which loss of NMD causes extensive deregulation of alternative splicing has not previously been assayed in a global, unbiased manner. Here, we combine mouse genetics and RNA-seq to provide the first in vivo analysis of the global impact of NMD on splicing patterns in two primary mouse tissues ablated for the NMD factor UPF2.

**Results:** We developed a bioinformatic pipeline that maps RNA-seq data to a combinatorial exon database, predicts NMD-susceptibility for mRNA isoforms and calculates the distribution of major splice isoform classes. We present a catalog of NMD-regulated alternative splicing events, showing that isoforms of 30% of all expressed genes are upregulated in NMD-deficient cells and that NMD targets all major splicing classes. Importantly, NMD-dependent effects are not restricted to premature termination codon+ isoforms but also involve an abundance of splicing events that do not generate premature termination codons. Supporting their functional importance, the latter events are associated with high intronic conservation.

**Conclusions:** Our data demonstrate that NMD regulates alternative splicing outcomes through an intricate web of splicing regulators and that its loss leads to the deregulation of a panoply of splicing events, providing novel insights into its role in core- and tissue-specific regulation of gene expression. Thus, our study extends the importance of NMD from an mRNA quality pathway to a regulator of several layers of gene expression.

Background

Alternative splicing (AS) involves the selective inclusion and exclusion of exons from a nascent pre-mRNA that results in various combinations of mature mRNAs with different coding potential and thus protein sequence [1]. Importantly, it has recently been estimated that nearly 95% of all multi-exon genes in the mammalian cell undergo AS [2,3], suggesting a pivotal role for AS in regulating and expanding the repertoire of isoforms expressed. By examining ESTs, it has been proposed that one-third of all AS isoforms contain a premature termination codon (PTC) [4], and these are expected to be targeted for degradation by nonsense-mediated mRNA decay (NMD). NMD is an mRNA quality control mechanism, and the primary function of NMD was initially thought to be in removal of aberrant transcripts arising from mutations or faulty transcription, mRNA processing or translation, but it is now evident that NMD impacts on both diverse physiological processes [5-7] as well as pathophysiological conditions (reviewed in [8]). The conserved core components of the NMD pathway are the UPF1, UPF2 and UPF3A/B proteins, and mutations or depletion of these factors inactivate NMD [9,10]. In mammalian cells, PTCs are distinguished from normal stop codons by their position relative to a downstream exon-exon junction, which is marked by the deposition of the exon junction complex
It has been generally established that for a stop codon to be recognized by the NMD apparatus, it must be situated at least 50 nucleotides upstream of an exon-exon boundary (the 50 nucleotides rule) [12]. Thus, nearly all naturally occurring eukaryotic stop codons are found downstream of the last intron, thereby rendering them immune to NMD. Although recent data have demonstrated that the proximity of the poly(A)-binding protein (PABP) to the PTC is inversely correlated with the efficiency of NMD [13,14], the 50 nucleotides rule applies to almost all studied mammalian transcripts, taking heed of a few noted exceptions [15,16]. Mechanistically, AS can utilize NMD to selectively degrade transcripts by the selective inclusion of a PTC-containing (PTC+) exon or exclusion of an exon, resulting in a PTC- downstream exon. This coupling, initially discovered for serine/arginine-rich (SR) proteins in Caenorhabditis elegans [17], has been coined regulated unproductive splicing and translation (RUST) or AS coupled to NMD (AS-NMD) [4,18]. Intriguingly, proteins involved in splicing processes utilize AS-NMD to autoregulate their own synthesis through a negative feedback loop. The most well characterized splicing activators, the SR proteins, bind to cis elements in the pre-mRNA, usually stimulating the inclusion of an exon. The SR proteins have been shown to utilize AS-NMD in a negative feedback loop to activate the inclusion of a PTC+ exon (PTC upon inclusion) in their own pre-mRNA, thus resulting in NMD [18-21]. The other major class of splice regulators, the heterogeneous nuclear ribonucleoproteins (hnRNPs), are a class of RNA binding proteins with roles in mRNA splicing, export and translation [22,23]. The hnRNPs often, but not always, bind to splice silencer elements and repress splicing at nearby splice sites. Splicing repressors, such as hnRNPs, use AS-NMD to repress the inclusion of a coding exon in their own pre-mRNA that leads to an out-of-frame skipping event, consequently inducing a downstream PTC and thus NMD (PTC upon exclusion). Moreover, AS-NMD is also used to cross-regulate expression of other splice factors, as described elegantly for PTBP1 and PTBP2 [24].

AS is regulated by the selective recruitment of splice regulators to pre-mRNAs. It is well established that splicing activators (such as SR proteins) compete with splicing repressors (such as hnRNPs) for binding to splice sites in an antagonistic manner, where the relative concentration of the two classes regulates the level of AS [25,26]. Thus, the fate of an alternative exon is usually decided by the antagonism between SR proteins and hnRNPs and their concentration and activity (reviewed in [27]). Due to the autoregulatory feedback loop employed by splice regulators, modulating NMD could potentially have widespread effects on the concentration of splicing activators and repressors and thus AS and AS-NMD.

Despite the potential of AS-NMD, it is presently not known to what extent this pathway regulates the transcriptome on a global scale, and how AS homeostasis is affected by NMD. A major problem in discovering the full spectrum of PTC+ transcripts is that EST and cDNA repositories are biased against these isoforms due to their unstable nature in normal cells. Additionally, most studies have primarily focused on microarrays to query the transcriptome upon muting NMD [18,19,28], and the newer studies using sequencing [29] have focused on single cassette exon events, thus disregarding many other physiologically important splice classes such as mutually exclusive exons and alternative 5' and 3' splice site usage. Last but not least, the transcriptomic consequences of genetically modulating NMD and thereby AS in the mammalian organism are largely unanswered.

In the present study, we have performed RNA-seq on two different tissues from a Upf2 conditional knock-out (KO) mouse line. Thus, in addition to providing significantly novel insights into common and tissue-specific functions of NMD, our study represents the first comprehensive and unbiased transcriptome analysis of adult genetically modified mice. To facilitate a high-resolution analysis of all possible exon-exon combinations, we have generated a bioinformatic pipeline, named RAINMAN (Rnaseq-based Isoform detection and NMD ANalysis pipeline; available to the scientific community), that maps reads to a combinatorial database that incorporates both known and in silico predicted exon-exon junctions. The pipeline predicts NMD susceptibility based on junction evidence and groups AS events into seven major splice isoform classes. Using this approach, our results reveal an unprecedented increase in AS upon ablating UPF2, and by inference NMD (although small additive effects from UPF3A/B interaction and deregulation of UPF1 phosphorylation cannot be wholly excluded from analysis, we juxtapose UPF2 and NMD ablation for all purposes in this work) and show that a high proportion of these upregulated AS events are not predicted to contain a PTC. Hence, we find that only 50% of the increase in AS is directly due to stabilization of PTC+ isoforms. Our data demonstrate that muting NMD results in deregulated levels of core splice regulators at both the mRNA and protein level, and further suggests that this contributes to the deregulation of general AS upon ablating NMD. Hence, our data support a model where NMD ablation leads to deregulated levels of core splicing factors that ultimately lead to aberrant global splicing, implying a novel intricate interplay between the NMD machinery and the splicing factors. Finally, our data analysis generates the first insights into...
a putative functional role of the PTC and its surrounding regions through analysis of their conservation patterns.

**Results**

A major task in analyzing AS by RNA-seq is to explore and quantify the differential representation of various splice classes and the protein-coding potential of the mapped reads. To this end, we have developed RAINMAN, a streamlined bioinformatic pipeline available to the scientific community that maps RNA-seq reads to a comprehensive combinatorial database of exon-exon junctions for unique junction discovery, PTC detection and identification of splice isoforms. We used this pipeline to investigate the complexity of the transcriptome and global role of AS-NMD in mammalian cells, by taking advantage of our Upf2 conditional KO mouse, which we previously used to demonstrate the in vivo importance of NMD [7,30]. To explore the effect of NMD on global splicing and to generate and validate an attractive bioinformatic pipeline to study AS and NMD, we chose to analyze the transcriptomes of two different mammalian organ systems with distinct phenotypes upon UPF2 deletion. In one end of the spectrum, we analyzed liver, wherein removal of UPF2 results in failure in liver metabolism and a high mortality rate [30], and in the other, we analyzed bone marrow-derived macrophages (BMMs). These macrophages are generated in vitro from murine bone marrow cells, and Upf2 deleted BMMs are completely devoid of NMD activity but nevertheless show no morphological or functional phenotype compared to wild-type (WT) controls [7].

**Splice isoform inference and PTC detection**

In order to obtain biological material, we first generated mice in which the NMD core factor Upf2 could be selectively inactivated in liver or in BMMs using our previously reported strategy (see Materials and methods). We performed whole transcriptome sequencing (single-end) on poly(A)-purified RNA, isolated from poly-IC injected Upf2fl/fl; Mx1Cre and Upf2fl/fl livers (termed from now on 'Liver KO' and 'Liver WT', respectively) and Upf2fl/fl; LysMCre BMMs and Upf2fl/fl BMMs (henceforth 'BMM KO' and 'BMM WT', respectively). In order to minimize biological variation, we generated libraries from pools of poly(A)-purified RNA derived from three individual animals. Due to the underrepresented nature of NMD-susceptible transcripts in EST and cDNA databases [4], we generated a comprehensive combinatorial database of exon-exon junctions in the murine genome, using sequences from exon models annotated in different repositories (RefSeq, Ensembl, UCSC Known Genes, GENSCAN and Exoniphy; Supplemental Materials in Additional file 1 and Figure S1B in Additional file 2). To discover junctions between exons not previously recorded in the murine transcriptome, we also employed TopHat [31], a de novo mapping algorithm, to map reads to junctions between unannotated exons (Table S1 in Additional file 1). To utilize reads that cover three or more exons, thus aligning to two or more exon-exon junctions, and not initially mapped to our combinatorial database or by TopHat, we incorporated an extra read truncation step (Figure S1B in Additional file 2), trimming reads in steps of 10 bp and remapping, allowing us to recover an additional 4 to 7% of total reads. As a result, 84% of reads mapped to the genome or transcriptome, and out of these, 28% mapped to splice junctions (Figure 1a), corresponding to a combined total of 323,474 unique splice junctions across our two tissues and two genotypes. With a minimum requirement of three reads to a splice junction, minimizing sequencing and mapping artifacts, we mapped approximately 150,000 unique junctions. Of note, we found that the de novo mapper TopHat contributed significantly to our transcriptome set, with 11 to 16% of all discovered junctions uniquely defined by TopHat (Table S1 in Additional file 1). Mapping of the KO samples benefited the most from these TopHat predicted splice junctions, suggesting that the curated transcript repositories are biased against the TopHat-predicted junctions due to their NMD susceptibility. Indeed, TopHat identified 20% and 24% of the PTC+ splice junctions above our minimum read cutoff, compared to only 4% and 11% of PTC- splice junctions (junctions that do result in the generation of a PTC, in BMM and liver samples, respectively) [7,32]. These data demonstrate that at least a quarter of NMD susceptible transcripts are not present in the normal murine repositories. This has implications for future splice isoform detection, since many transcripts that are below detectable levels under physiological conditions, and hence are absent in the repositories, will not be detected under conditions that could favor their presence, for example, perturbed or disease states. Thus, taking advantage of our combinatorial database that includes a pipeline for splice isoform detection and PTC prediction is likely to considerably increase the number of isoforms detected under conditions that favor increased AS.

In the next step of RAINMAN, the pipeline calculates the distribution of seven different splice isoform classes and predicts PTCs (see Supplementary Methods in Additional file 1 for details), and all data are combined for easy visualization and data mining. An illustrative example of a genome browser output from the mapping steps is shown in Figure 1b, with the conditional Upf2
KO gene deleted for exons 2 and 3 in the KO sample. The UCSC-based visualization includes both evidence of reads mapped to junctions and to exons, and, in this case, demonstrates increased AS in the liver KO compared to WT.

We mapped a total of 150,000 unique junctions with approximately 100,000 and 130,000 junctions in BMM and liver tissues, respectively (minimum of three reads to a junction; Figure 2a, top, and data not shown). Tallying up, we found that 6,256 and 7,997 unique genes harbored 14,056 and 25,534 upregulated junction events in BMM and liver, respectively, an average of 2.3 and 3.2 upregulated junction events per gene (Figure 2a), demonstrating that loss of UPF2, and by inference NMD, leads to a substantial deregulation of splicing.

Ablating NMD results in significant upregulation of PTC+ junctions

It has been shown that a stop codon is generally recognized as a PTC by the NMD machinery, if it is situated at least 50 nucleotides upstream of an exon-exon junction, termed the 50 nucleotides rule. We assayed and verified this distance requirement by simply plotting the position of RefSeq gene model stop codons that reside in the penultimate exon (Figure S2 in Additional file 3), observing that the vast majority of these stops indeed fall precisely within 50 nucleotides of the final intron boundary, and thereby avoiding the elicitation of NMD for those transcripts. Incorporating this distance metric in our PTC prediction algorithm, we calculated the fraction of regulated PTC+ junctions as a function of UPF2

KO gene deleted for exons 2 and 3 in the KO sample.
Ablation. As expected, this analysis demonstrated a highly significant upregulation of PTC+ junctions in the KO samples (Figure 2a, all junctions versus PTC+ junctions, P-value \(1 \times 10^{-16}\), Chi-square). In total, approximately 44% of predicted PTC+ junctions were upregulated in BMMs and Liver KO (Figure 2a), amounting to 16% (BMM) and 28% (liver) of expressed genes containing at least one NMD-susceptible splice junction regulated more than two-fold.

Among the upregulated PTC+ splice junctions, 516 were shared between the two tissues, and these corresponded to 448 unique genes (Figure 2a; Table S3 in Additional file 4), which we thus termed core NMD targets. The group includes well-known NMD targets such as Smg5, Hsf1, and Zcchc6 as well as many known NMD-susceptible splicing factors [6,18,19]. Indeed, Gene Ontology (GO) analysis demonstrated that the highest ranked cluster contained genes involved in mRNA processing and splicing (P-value of \(9.0 \times 10^{-18}\), Bonferroni corrected; see the full list of GO terms and genes in Table S3 in Additional file 4). All classical SR proteins have been shown to utilize AS-NMD to autoregulate their own synthesis [20], and our finding that common splicing factor junctions are upregulated upon
muting NMD is in agreement with earlier studies [7,18,19,32]. We also determined the GO terms associated with genes containing upregulated PTC+ junctions unique to either liver or BMMs (Figure 2b). In the BMM-specific set, genes involved in G-protein coupled receptor function were enriched (P-value 1.5 × 10^{-4}, Bonferroni corrected; Table S4 in Additional file 5), whereas the liver specific set was strongly associated with mitochondrion, among others (P-value 9.7 × 10^{-56}, Bonferroni corrected; Table S4 in Additional file 5).

In summary, we find that 43 to 44% of all predicted PTC+ junctions are upregulated in Upf2-ablated tissues and that 516 of these junctions are common in BMM and liver, several of which are well-known NMD targets. Moreover, we show that NMD is predicted to downregulate isoforms of 16 to 28% of all expressed genes (excluding the well-described regulation of splicing factors from the analysis). Moreover, NMD regulates genes involved in mitochondria and G-protein-coupled receptor functions in liver and BMM, respectively, suggesting that NMD also serves tissue-specific functions.

Loss of NMD leads to the selective stabilization of alternative splicing events

We next compared the impact of abrating NMD on canonical versus AS in more detail. Here, a canonical junction is defined as a splicing junction between two consecutive exons of the longest RefSeq isoform for each gene, and AS junctions are thus all other combinations. In total, our pipeline detected approximately 10,000 unique AS junctions in BMMs compared to approximately 35,000 in the liver (Figure 3). A significant challenge in analyzing expression changes upon NMD ablation is to distinguish primary from secondary effects. Using the sensitive PTC-detection in RAINMAN, we found that close to 50% of AS was predicted to generate a PTC (Figure 3a, bottom), which is in stark contrast to canonical splicing junctions (Figure 3a, top). This suggests that NMD directly degrades 16% (0.47 × 0.35) of all AS in BMMs and 18% in liver (0.45 × 0.39), resulting in more than a two-fold reduction of the involved isoforms.

The BMM and liver samples demonstrated essentially similar fractions of up- and downregulated AS junctions (Figure 3a). However, the number of unique AS junctions were higher in liver samples, even after calibrating for the number of mapped junction reads (and thereby sequencing depth; Figure 3b, top). To investigate whether this was a characteristic feature of the organs or whether it was due primarily to muting of NMD, we analyzed the AS complexity per gene (Figure 3b). The number of mapped junction reads was first normalized to BMM WT (Figure 1a) by stochastically removing reads from BMM KO, liver WT and liver KO, to exclude any skewing due to differences in sequencing depth. Interestingly, these data demonstrate an inherently different AS profile between the two tissues. Whereas BMM WT and KO showed similar proportions of AS, measured as the correlation between the number of AS junctions and the total number of junctions per gene (Figure 3b, top), the proportion of AS of the liver KO sample increased considerably more compared to both BMM KO and liver WT (Figure 3b, top). This strongly suggests that the liver has a high level of AS relative to BMM, both normally and through ablation of NMD.

Finally, we also compared the proportion of PTC+ junctions, measured as the number of PTC+ junctions as a function of total splicing for each gene (Figure 3b, bottom). The PTC+ junctions followed the same trend as for AS junctions, with a steeper relative increase in PTC+ junctions in liver KO compared to liver WT and BMM. From linear interpolation, we found that, on average, 27% of all spliced junctions in a given gene in the liver are predicted to result in a PTC (liver KO slope of 0.27; Figure 3b).

In conclusion, these data demonstrate that approximately 17% of all AS is downregulated via NMD more than two-fold. We furthermore show that liver is characterized by a high degree of AS compared to BMMs and that close to one-third of all uniquely spliced junctions in the liver are predicted to elicit a PTC, thus giving the first in vivo evidence of the pervasive effect and importance of AS-NMD in the mammalian organism.

Splice isoform classes are differently affected by NMD ablation

As described above, we detected a significant increase in AS upon abrating UPF2. It has, however, not previously been described to what extent NMD affects different splicing classes. Transcriptome-wide studies where NMD has been muted have primarily looked at single cassette exon skipping (ES) events [28,32], thus disregarding many physiologically relevant AS events. To this end, we implemented an AS isoform classification module that allowed us to infer the degree to which all major splice isoform classes were affected by NMD. Hence, our splice isoform inference pipeline assessed the distribution of splice junctions to seven major groups of AS events, namely single exon skipping (SES) and multiple exon skipping (MES), alternative 5’ splice site (A5SS) and alternative 3’ splice site (A3SS), mutually exclusive exons (MXE), alternative first exon (AFE) and alternative last exon (ALE) (see Table 1, Supplementary Materials in Additional file 1 and Table S6 in Additional file 6 for pipeline performance). We narrowed our downstream splice isoform analysis to those that demonstrated a change in
'percent spliced in' (PSI) between KO and WT (ΔPSI) higher than 20%. PSI is calculated from the ratio of junctions supporting a given feature (for example, the inclusion of an exon) versus junctions supporting the reciprocal event (for example, the skipping of the same exon), and the ΔPSI is thus a metric of how much the inclusion changes upon NMD ablation (see Supplementary Methods in Additional file 1 for further details).

We found that approximately 40% of all ES events with a ΔPSI higher than 20% (inclusion) or lower than -20% (exclusion) in the KO tissues were upregulated due to stabilization of a PTC+ isoform (Table 1). Next, we subdivided ES events into PTC upon inclusion and exclusion events. An example of a PTC upon exclusion event that is stabilized upon NMD ablation in both BMM (ΔPSI = -61%) and liver (ΔPSI = -63%) is Mgea5, which was also found upregulated in immortalized mouse embryonic fibroblasts (MEFs) ablated for NMD [32]. We validated this Mgea5 isoform among others by RT-PCR (Figure 4b). In the liver, 55% of all single exon inclusion events generated a PTC, whereas 40% were found in the BMM (compare PTC upon inclusion to total inclusion events for single ES in Table 1). Tmem183a is an example of a highly skipped PTC upon inclusion that was found stabilized in both BMM (ΔPSI = 51%), liver (ΔPSI = 54%) and immortalized MEFs ablated for NMD [32], and was similarly validated by RT-PCR (Figure 4b).
Interestingly, for alternative usage of A5SSs and A3SSs, we saw an even higher proportion of PTC+ events, especially in the liver (compare the fourth and fifth rows in BMM versus liver tissue in Table 1), suggesting a previously unrecognized importance for these splice events in AS-NMD. The ribosomal protein Rps12 is an example of a gene that generates a PTC+ isoform as a result of A5SS in both BMM (ΔPSI = 50%) and liver (ΔPSI = 55%), and likewise for Smg5 in BMM (ΔPSI = 37%) and liver (ΔPSI = 50%) as a result of A3SS (Table 1).

Using stringent criteria (Supplementary Methods in Additional file 1), we detected 44 MXE events in the BMM KO, but more than 10 times as many in the more splice-prone liver KO tissue (Table 1). Out of the detected MXE events, few were predicted to elicit a PTC+ isoform, possibly due to the regulated nature of a mutually exclusive splice event, which in most instances would not be predicted to undergo NMD. An exception is the pyruvate kinase gene Pkm2, which encodes two different enzymatic active isoforms due to a single MXE event [33]. In the KO samples, and this was particularly true for BMMs, aberrant splicing resulted in inclusion of both mutually exclusive exons, resulting in a PTC+ isoform (validation in Figure 4b and schematic in Figure 4d). This may indicate an important role for NMD in ensuring the exclusive incorporation of exons in MXE events.

Finally, we quantified the number of isoforms with an increase in AFE or ALE. From transcriptome data alone, it is often impossible to determine whether an AFE is the result of alternative promoter usage or AS between two mutually exclusive first exons and the second exon. Nevertheless, AFE results in alternative transcripts with potential PTCs. Analysis of AFE showed that more than one-third of the detected AFE events are predicted to generate a PTC+ isoform (Table 1). For ALE categorization, we required AS to two mutually exclusive 3’ terminal exons, and PTC+ prediction is thus not meaningful. Hence, changes in levels of ALE are therefore most likely secondary effects of altered splicing upon UPF2 ablation. A functional mechanism for ALE is the regulated inclusion of microRNA target sites in the 3’ UTR, and this potentially adds another layer of complexity when studying conditions where splicing is perturbed, such as by removing NMD.

To test the accuracy of our splice isoform detection algorithm, we chose 50 RAINMAN-predicted AS events (ΔPSI > 20% (inclusion), or < -20% (exclusion)) for RT-PCR on independent liver and BMM material. We validated AS events that are not predicted to result in a PTC+ isoform (Figure 4b, top row), isoforms stabilized in the KO due to inclusion of a PTC (Figure 4b, middle row) and isoforms that elicit a PTC due to a skipping event (Figure 4b, bottom row). Out of the 50 tested AS events, we were able to validate 49 of these (98%), and we therefore conclude that our splice isoform detection algorithm is highly accurate in predicting different splice isoform events.

To validate the inferred expression changes and to assess inter sample-replicate variability, we performed quantitative PCR experiments on biological replicates. This analysis showed a strong correlation with the

Table 1 Splice events and PTC+ classification

<table>
<thead>
<tr>
<th>Tissue</th>
<th>Splice event</th>
<th>Total events</th>
<th>PTC+ events</th>
<th>Percentage PTC+</th>
<th>Exclusion events</th>
<th>PTC upon exclusion</th>
<th>Inclusion events</th>
<th>PTC upon inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMM</td>
<td>Total ES</td>
<td>730</td>
<td>281</td>
<td>38%</td>
<td>428</td>
<td>201</td>
<td>302</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Single ES</td>
<td>511</td>
<td>131</td>
<td>26%</td>
<td>310</td>
<td>148</td>
<td>201</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>Multiple ES</td>
<td>219</td>
<td>66</td>
<td>30%</td>
<td>118</td>
<td>53</td>
<td>101</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A5SS</td>
<td>130</td>
<td>68</td>
<td>52%</td>
<td>118</td>
<td>53</td>
<td>101</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A3SS</td>
<td>238</td>
<td>110</td>
<td>46%</td>
<td>94</td>
<td>44</td>
<td>34</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MXE</td>
<td>44</td>
<td>20</td>
<td>46%</td>
<td>13</td>
<td>6</td>
<td>7</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AFE</td>
<td>130</td>
<td>68</td>
<td>52%</td>
<td>118</td>
<td>53</td>
<td>101</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALE</td>
<td>49</td>
<td>NA</td>
<td>NA</td>
<td>49</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Liver</td>
<td>Total ES</td>
<td>3,102</td>
<td>1,285</td>
<td>41%</td>
<td>1,926</td>
<td>965</td>
<td>1,176</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>Single ES</td>
<td>1,505</td>
<td>932</td>
<td>59%</td>
<td>932</td>
<td>531</td>
<td>573</td>
<td>320</td>
</tr>
<tr>
<td></td>
<td>Multiple ES</td>
<td>1,597</td>
<td>994</td>
<td>59%</td>
<td>994</td>
<td>434</td>
<td>603</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A5SS</td>
<td>449</td>
<td>263</td>
<td>59%</td>
<td>199</td>
<td>112</td>
<td>87</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>A3SS</td>
<td>654</td>
<td>370</td>
<td>57%</td>
<td>370</td>
<td>203</td>
<td>167</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>MXE</td>
<td>475</td>
<td>66</td>
<td>14%</td>
<td>475</td>
<td>66</td>
<td>409</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>AFE</td>
<td>172</td>
<td>66</td>
<td>38%</td>
<td>172</td>
<td>66</td>
<td>106</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ALE</td>
<td>143</td>
<td>NA</td>
<td>NA</td>
<td>143</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Splice isoform classes are shown for both single and multiple exon skipping (ES), alternative 5’ splice site (A5SS), alternative 3’ splice site (A3SS), mutually exclusive exons (MXE), alternative first exon (AFE) and alternative last exon (ALE). For ALE, no events were classified as PTC+, since the class relies on splicing to two different last exons. Here, a ΔPSI > 20% was required for all classes. NA, not applicable.
Figure 4 Splice isoform classes are differentially affected by loss of UPF2. (a) Schematic drawing of the main isoform classes detected in our pipeline. (b) RT-PCR validation of 25 splicing events predicted from our pipeline (see Table S9 in Additional file 12 for a list of the 49 validated events out of 50 tested). Top: normal AS events. Middle: PTC upon inclusion events. Bottom: PTC upon exclusion by ES. (c) Western blotting from two different liver pairs, showing UPF2 (rabbit α-UPF2), SRp55, SRp40 and SRp30 (mouse α-mAb104) and β-actin (rabbit α-actin). The asterisk denotes the truncated UPF2 isoform found in cells ablated for NMD. (d) MXE splicing of Pkm2 and the inclusion of both mutually exclusive exons in the KO sample.
RNA-seq inferred expression changes (R = 0.975, Pearson’s correlation; Figure S3A in Additional file 7) and a very low sample replicate variability (Figure S3B in Additional file 7). In addition, we assessed our ability to correctly infer PTC+ isoforms by cloning and sequencing the full-length isoforms of Srsf9, a gene with a PTC-upon inclusion event (∆PSI = -40% in liver; isoform not detectable in BMMs). The predicted PTC+ isoform was precisely recapitulated in vivo (Figure S3C in Additional file 7), further validating the usefulness of our approach.

**Loss of NMD leads to deregulation of PTC- isoforms through breakdown of splice factor homeostasis**

We have demonstrated that only 50% of upregulated AS junctions are predicted to be the direct result of stabilized PTC+ isoforms, suggesting that a major proportion of the increased AS is the indirect result of perturbed NMD (Figure 3a). As we have shown here, the liver has a high degree of AS (Figures 1 to 3), and it has been found to have a prominent divergent expression pattern of SR proteins and hnRNPs compared to other mammalian tissues [34]. Hence, we hypothesize that this could make the liver particularly susceptible to changes in splicing patterns upon ablation of NMD.

Splicing factors are known to utilize AS-NMD to autoregulate their own synthesis in a negative feedback loop, through the selective inclusion or exclusion of PTC+ exons [4,18]. Indeed, our RNA-seq analysis also identifies a range of these factors among the core NMD targets (Figure 2; Table S3 in Additional file 4). However, as most of the NMD sensitive splicing factor isoforms are predicted to encode truncated proteins, the rescue of these upon NMD ablation is not expected to affect AS per se (unless they encode proteins with dominant negative properties). Previous microarray-based studies have not addressed the formal possibility that the stabilization of PTC+ isoforms upon loss of NMD could occur against the backdrop of changed levels of the canonical isoforms capable of encoding the full-length form of proteins such as splicing regulators. To test this possibility, we scrutinized our datasets for changes in the expression levels of 25 canonical mRNA isoforms capable of encoding full-length splicing regulators by correcting the changes in the gene expression ratios between KO and WT samples with the fraction contributed by the PTC+ isoforms (Table S5 in Additional file 1). Strikingly, for the 19 splice factors for which we have evidence for AS-NMD in the liver, 10 displayed a > 1.5-fold deregulation, with 5 displaying increased (Sfrs3, Sfrs4, Tra2a, Srsf16, Hnrnpl) and 5 decreased (Tra2b, Hnrnpj, Hnrnph3, Hnrnpq, Ptbp2) expression of the canonical mRNA isoforms. Moreover, western blot analysis revealed that SRp30 and SRp40 protein levels were elevated in KO liver, demonstrating that the changes in canonical mRNA isoform levels for at least some splicing factors are indeed translated into full-length protein.

Collectively, these findings support a model where the increase in global AS upon loss of NMD is at least in part caused by a massive deregulation of key splicing factors, and further help to explain the PTC- AS fraction deregulated upon muting NMD. A highly aberrant level of splicing regulators would likely tilt the normal balance between activators and repressors and deregulate a large subset of exons [4,18].

**Ablating NMD leads to skipping of highly conserved cassette exons**

SES represents the best characterized AS event, and is a frequent PTC-generating mechanism. RAINMAN classified 40% and 60% of all SES as PTC+ events in the BMM and liver KO samples (Table 1 - compare PTC upon exclusion with total exclusion events, and PTC upon inclusion with total inclusion events). Out of the PTC upon inclusion events, a total of 26 events were upregulated in both KO tissues (∆PSI > 20%) and 51 PTC upon exclusion events were similarly upregulated in both KO tissues (ΔPSI < -20%; manually curated lists are included in Table S7 in Additional file 8 and Table S8 in Additional file 9). In support of an important physiological function, almost all of the genes harboring these highly upregulated PTC+ isoforms were also upregulated at the gene level in KO samples (Table S7 in Additional file 8 and Table S8 in Additional file 9), suggesting that stabilization of the PTC+ isoforms led to an appreciable increase in total mRNA levels in KO tissues. It should be noted that most of the splicing factors and ribosomal proteins known to use AS-NMD were indeed upregulated in both tissues, but had ΔPSI > -20% or < 20% in BMMs (data not shown). Hence, the high frequency of PTC+ splice events suggests that these are regulated AS events. It is generally found that AS exons are more conserved than constitutive exons, especially in the flanking introns [35]. The core members of the SR protein family and hnRNPs have been proposed to use AS-NMD in an autoregulatory loop, by binding to highly conserved regions, termed ultraconserved elements (UCEs), to elicit the PTC+ splice event [18,20]. Upon inhibition of NMD in cell lines, it has been shown that many PTC+ exons are surrounded by high intronic conservation [18,19]. As described above, our sensitive approach demonstrated a high proportion of PTC- AS events upregulated in KO tissues, and we therefore examined the conservation of PTC+ and PTC- single exon exclusion (ΔPSI < -20%) and inclusion (ΔPSI > 20%) events in the KO samples compared to the unregulated skipping events. The latter control group

**Collectively, these findings support a model where the increase in global AS upon loss of NMD is at least in part caused by a massive deregulation of key splicing factors, and further help to explain the PTC- AS fraction deregulated upon muting NMD. A highly aberrant level of splicing regulators would likely tilt the normal balance between activators and repressors and deregulate a large subset of exons [4,18].**
contains exons that undergo AS but are not differently skipped between WT and KO (-20% < ΔPSI ≤ 20%). We first examined exclusion events, and found that flanking introns of the skipped exons are significantly more conserved compared to unregulated skipping events (Figure 5a; P-value < 2.2 × 10^{-16}, Komogorov-Smirnov test), and this was even more pronounced in BMMs (Figure S4 in Additional file 10). Interestingly, flanking introns of PTC+ upon exclusion events were less conserved than PTC- exclusion events, but displayed a significant higher exonic conservation (P-value < 2.2 × 10^{-16}, Komogorov-Smirnov test). We next considered exons that demonstrated increased inclusion in KO samples. In these cases, flanking introns were also significantly more highly conserved compared to skipping events not regulated by NMD (Figure 5b). Again, PTC- inclusion events were more highly conserved compared to PTC+ events in the flanking introns.

In summary, these results demonstrate that 40 to 60% of upregulated SES events are predicted to undergo NMD, depending on tissue type, and that these splice events are highly conserved. Importantly, PTC- inclusion and exclusion events displayed higher intronic conservation compared to PTC+ events, strongly suggesting that the UCE-containing splice regulator isoforms (which are all PTC+) are not the primary reason for the high intronic conservation surrounding skipped exons. Moreover, removal of UCE elements from the analysis did not alter the intronic conservation profiles of either PTC- or PTC+ events (data not shown). Thus, the high conservation for both included and excluded exons appears to be a general attribute of highly regulated cassette exons (ΔPSI > 20%) that are affected directly and indirectly by NMD. Thus, the deregulation of PTC- events is therefore consistent with a mechanism where deregulated levels of splice regulators in the KO samples markedly affect regulated splicing of their cognate target exons. This again implies that ablating Upf2, and by inference NMD, has important secondary effects on splicing factor homeostasis and that this impacts widely on global AS.

**Muting NMD leads to increase in splicing by-products**

We have demonstrated that removing UPF2, and by inference NMD, causes a dramatic increase in AS, in part by stabilizing PTC+ AS isoforms. Apart from AS-NMD, NMD has been implicated in removal of genomic

---

**Figure 5** Introns flanking regulated exons are highly conserved. (a, b) Mean per position phastCon conservation scores around SES events are shown for exclusion events upregulated in the KO sample (a), and for inclusion events upregulated in the KO sample (b). Shown are PTC+ exclusion/inclusion events (red line), PTC- exclusion/inclusion events (green line) and unregulated skipping events (grey line). Yellow lines are scores for all mm9 RefSeq exons and 75 bp into surrounding introns. Data shown are for liver. Exclusion events: PTC+, 439; PTC-, 251. Inclusion events: PTC+, 64; PTC-, 162. Unregulated events: 3,494. Mm9 RefSeq exons: 274,281. See Figure S4 in Additional file 10 for a graph with BMM data. Numbers on the x-axis indicate nucleotide intervals - 25 and 75 nucleotides for exons and flanking introns, respectively. Curves represent a cubic smoothing spline fitted to data.
noise and splice errors that might otherwise lead to a
panoply of spurious isoforms [7,36-38]. The extent to
which NMD destroys such splice by-products has not,
however, been studied on a transcriptome-wide basis in
mammalian tissues. We therefore examined the relative
expression level of PTC+ junctions, measured as the
PTC+ fraction (PTC-generating junctions/All other
junctions) against the junction expression level (Figure
6), binned by the RNA-seq metric RPKM (reads per kb
of gene model per Mb of mapped reads, refer to [39]).
These data show that the PTC+ fraction is enriched in
liver compared to BMM over the entire expression pro-
file, further emphasizing that the liver is characterized
by an increased abundance of PTC+ isoforms compared
to BMM. Importantly, there is a distinct overrepresenta-
tion of PTC+ isoforms in the lower expression range for
both liver (Figure 6 - compare dark red and light red for
liver KO and WT, respectively) and BMM (Figure 6 -
compare dark blue with light blue for BMM KO and
WT, respectively) relative to more highly expressed iso-
forms. These data thus confirm and quantify, on a glo-
bal scale, another major role for NMD in removing low-
abundance isoforms such as the products of erroneous
splicing.

PTC exons are distinct from normal stop codon exons
We have shown that NMD-affected SES events demon-
strate a high conservation score and that this is not due
to UCE-containing SR proteins and hnRNPs. The ques-
tion then arises to what extent the PTC+ exon itself is
aberrant compared to other expressed exons, and in
particular the normal stop codon-containing exon. We
thus examined the conservation of the exonic sequence
surrounding the PTC in upregulated PTC+ exons, as
evidenced by upregulated PTC+ junctions. To normalize
for normal exonic conservation variation and for biases
in tissue-specific exon conservation, we subtracted ran-
dom exonic sequences expressed in the same samples.
Interestingly, we found that PTC+ exons had a markedly
different conservation profile compared to normal stop
codon exons (Figure 7a; for comparison, see unnorma-
lized conservation score in Figure S5 in Additional file
11). Hence, exonic sequence conservation increased
towards the PTC, and displayed only a minor drop in
relative conservation score in nucleotides following the
PTC. Importantly, the PTC and surrounding nucleotides
displayed an overall higher conservation score relative to
random exons, and this was particularly striking for
PTCs in BMM. Notably, removal of the UCE-containing

![Figure 6 Low-expressed junctions are enriched in Upf2 KO samples](http://genomeweb.com/2012/13/5/R35)
Figure 7 Conservation of regulated PTCs and surrounding exons (a) Mean per-position phastCon scores are shown, centered on the PTC, for upregulated junctions in liver and BMM. To visualize conservation around PTCs in comparison to normal exonic areas, phastCon scores from a random sample (n = 4,000) of BMM and liver-expressed exons were subtracted from either sample. For normal STOPs, phastCon-scores from random RefSeq exons (n = 4,000) were subtracted. Normal STOPs are based on all RefSeq transcript models. Ranges of scores do not extend into introns, and may be shorter than 100 bp for individual PTCs. For PTC+ junctions, a KO/WT fold change of 2 was required. BMM PTC+ positions: 884. Liver PTC+ positions: 3,091. Normal RefSeq STOP positions: 23,231. (b) Distribution of stop codons: for intronic, intergenic and exonic bins, all mm9 trinucleotides in all three reading frames were sampled. RefSeq STOPs represent normal stop codons for all 21,470 RefSeq transcript models (for genes with multiple models, the longest was used). For BMM (n = 497), PTC-inducing junctions were required to have a log2(KO/WT) fold change > 2, and a minimum of 5 reads for both genotypes summed. For liver (n = 670), PTC-inducing junctions were required to have a log2(KO/WT) fold change > 2, and a minimum of 10 reads for both genotypes summed. Fisher’s exact test was used to test for significance.
core splicing factor genes did not influence the conservation profile in liver or BMM (data not shown). The BMMs are characterized by a more stringent splice pattern, whereas the liver has a high degree of AS and increased PTC+ junctions expressed at low levels (Figures 2 to 4 and 6), pointing to an increased fraction of random splice byproducts in the liver. Hence, the increased PTC+ exonic conservation in BMMs compared to the liver is likely due to a higher proportion of functionally relevant PTCs in BMMs, found in various vertebrates. Moreover, the relatively high conservation score at and after the PTC suggests that these sequences may possibly harbor un-recognized cis elements that could assist in recognizing the PTC.

Finally, we also gauged whether PTCs associated with highly regulated PTC-inducing junctions displayed any preference for the identity of the stop codon (Figure 7b). Surprisingly, whereas the frequencies of normal canonical UAA, UAG and UGA RefSeq stop codons corresponded closely to the distribution observed in exonic regions, regulated PTCs had a marked and highly significant preference for UGA in both tissues. These findings suggest that the NMD elicting PTCs have been under selective pressure in order to adapt to the regulatory needs of AS-NMD.

Discussion
Here, we present the first RNA-seq study from genetically modified adult mice accompanied by a thorough comparative study of the role of NMD in two distinct murine tissues. Moreover, this is the first analysis to date of the global impact of NMD on all major classes of AS in untransformed mammalian cells.

To facilitate a detailed analysis of AS, we have generated a streamlined bioinformatic pipeline, RAINMAN, available to the scientific community, that maps RNA-seq reads to a comprehensive combinatorial exon-exon junction database to obtain maximum AS isoform information. This method allowed us to map 28% of all mappable reads to junctions (all samples combined; Figure 1) with a total discovery of 150,000 unique junctions, thus giving us high AS information. The junction data are further processed in the pipeline to predict NMD susceptibility and thus coding potential with high accuracy at single junction resolution. Finally, seven major splice isoform classes are inferred and processed for multiple comparison purposes.

Loss of NMD leads to a pronounced deregulation of alternative splicing
Using this pipeline, we have shown that ablating Upf2, and by inference NMD, impacts on several layers of AS in both BMMs and liver. By examining genes that harbor the exact same spliced junctions predicted to elicit a PTC in both tissues, our data confirmed that genes involved in mRNA processing are highly overrepresented among common NMD targets, as previously shown (Table S3 in Additional file 4). From our data, we estimate that close to one-third of all junction events supporting AS result in PTC+ and NMD-sensitive transcripts in the liver (Figure 3b), which is in agreement with earlier predictions [4]. Previous computational analyses of dbEST and SWISS-PROT have found that 8 to 12% of human genes are predicted to be targets of NMD [4,40], but these repositories inherently underestimate the true number of NMD-sensitive isoforms. Our data demonstrated that 16% and 28% of expressed genes in BMM and liver, respectively, were predicted to undergo NMD (upregulated more than two-fold in KO) of one or more isoforms. These numbers expand upon and underscore the global importance of NMD and in particular AS-NMD. Nearly all detected PTC+ events could be attributed to AS, and we found that approximately 50% of all AS in the KO was in fact predicted to elicit a PTC (Figure 3a), with the liver as the tissue with the most AS (Figure 3b). This complements previous findings showing that the liver has one of the highest levels of AS compared to other organs [2,34]. The high level of AS in the liver was also reflected in the fraction of low-level PTC+ junctions (Figure 6), most likely consisting of splice errors. Thus, our data make a strong case for NMD playing an important role in removing ‘splicing’ noise also in untransformed mammalian tissues.

In contrast to the splice error-generated PTC+ isoforms, our analysis of highly regulated splice classes (ΔPSI > 20% or ΔPSI < -20%), revealed that approximately 40% of all highly skipped exons were degraded by NMD (Table 1). These ES events were further subdivided into SES and MES events. Whereas three-quarters of all ES events in BMMs were SES, the proportion of SES and MES in liver was approximately equal. We speculate that the increased MES proportion reflects the more promiscuous splicing pattern in the liver, thus leading to novel AS between exons not normally spliced together.

Importantly, our analysis of conservation patterns in SES revealed a high conservation score of NMD-regulated skipping exons and their flanking introns compared to that of unregulated events. This suggests that events deregulated in the absence of NMD are under tight control in normal NMD-proficient cells and that AS-NMD plays a crucial role in controlling the expression of a vast number of genes.

The splice isoform inference analysis expands the repertoire of NMD targets
From our splice isoform inference, we found a high proportion of PTC+ events in most classes with the notable
exception of MXEs (ALE events were not included in PTC classification). Due to the regulatory nature of MXEs, in which two coding exons compete for the selective inclusion, it is not surprising that this group is underrepresented in the PTC+ events. One interesting example was Pkm2, in which a well-characterized MXE event either dictates whether the adult form (M1 isoform containing exon 9) or the embryonic isoform (M2 containing exon 10) is produced. In many tumors, the embryonic M2 isoform is switched on, leading to increased aerobic glycolysis [33]. However, we found that ablating UPF2, and by inference NMD, led to inclusion of both mutually exclusive exons with a resulting downstream PTC (Figure 4b, d), suggesting an interesting function for NMD in controlling MXE. We also provide the first report of a surprisingly high frequency of PTC+ A5SS and A3SS events, suggesting that the alternative usage of splice acceptor or donor sites is often utilized to regulate expression by AS-NMD. Another possibility could be that these events are more error-prone - for example, by the use of a strong and weak splice site (the alternative splice site), leading to a less stringent AS than between two strong splice sites such as SES. Nevertheless, our study significantly expands the known repertoire of NMD targets in all classes of AS, many previously uninvestigated on a global scale.

Common and tissue-specific functions of NMD
The impact of NMD on transcriptome composition has up to now only been studied in whole organisms or individual cell lines, which has therefore precluded any insights into conserved versus tissue-specific functions of NMD. Focusing first on core NMD targets, our stringent comparative analysis found 50 genes with a common highly upregulated PTC upon exon exclusion in BMMs and liver (Table S8 in Additional file 9). Almost all of these genes were upregulated in both tissues, suggesting a functional role for AS-NMD in regulating the abundance of transcripts with full coding potential from these genes (see below). Several of the known splicing factors previously shown to utilize AS-NMD in an auto-regulatory loop, such as Ptbp1, demonstrated PTC upon exclusion in both tissues, but below ΔPSI of 20% in BMMs. Besides SR proteins and hnRNPs, most of the genes found upregulated are unknown targets for AS-NMD, such as Soat2 and Acat2. The latter gene is involved in esterification of cholesterol, and is known to be regulated in a tissue-specific manner, with a particularly high expression level in liver and intestine and to a lesser extent in macrophages [41]. Here, we found that this gene was upregulated 8.4-fold in BMMs and 4.2-fold in liver upon muting NMD, suggesting an important regulatory function for AS-NMD in the synthesis of cholesterol esters.

We found 26 genes with a common PTC upon inclusion event (ΔPSI > 20%) in both BMM and liver KO (Table S7 in Additional file 8). An example of a gene with a PTC upon inclusion isoform is Nktr, which is exclusively expressed in and required for natural killer (NK) cells. We found that inhibiting NMD caused a high upregulation of a PTC+ isoform and a concomitant 2.4- to 3.9-fold upregulation at the gene level. It has been shown that aberrant isoforms of the Nktr gene are present in cells not expressing Nktr [42]. Here, we show that NMD mutes Nktr in BMMs and liver (and also in NMD ablated transformed MEFs [32]), suggesting that AS-NMD serves an important regulatory function in regulating the functional expression of this NK cell-specific protein. Interestingly, Nktr is involved in NK cell activity but contains RS repeats and a cyclophilin-domain found in several splicing factors, and this could confer on the protein properties sufficient to facilitate regulation of its own expression through AS-NMD. Thus, it could be that several other proteins not directly involved in splicing have gained RNA-binding properties that would allow them to autoregulate their own synthesis by AS-NMD.

Apart from the core NMD targets, our analysis also yielded the first insights into potential tissue-specific effects of NMD. Through GO analysis of parent genes of PTC+ junctions that were exclusively upregulated in either the BMM or liver datasets, we could show a tissue-specific role for NMD in the regulation of G-protein-coupled receptors and mitochondria function, respectively. Interestingly, these GO classes mirror the main biological functions of monocytes/macrophages and liver tissue, that is, in immunological reactions and energy metabolism, respectively, and we predict that NMD-dependent transcriptome analysis in other organs would uncover tissue-specific NMD targets in pathways of particular importance for the organ in question.

NMD controls the expression of a network of splicing factors
In contrast to the well-characterized role of AS-NMD in the autoregulation of individual splicing factors, the global consequences of its disruption on broad splicing patterns have not been studied in detail. We were therefore intrigued by finding that approximately 50% of the upregulated AS events in the NMD-deficient tissues were devoid of PTCs, suggesting that they were indirect targets of NMD and that the importance of NMD in AS extends well beyond AS-NMD.

Using conventional microarray-based gene expression analysis, we have previously shown that several splice factors were upregulated in both UPF2-deficient BMMs and liver [30,43]. However, these studies could not discriminate between canonical mRNA isoforms and
stabilized PTC+ isoforms, which is important as many of the latter would be predicted to encode truncated and thereby (most likely) non-functional proteins. Using our RNA-seq data, and correcting for the presence of the PTC+ isoforms, we can now show deregulated expression of the canonical mRNA isoform (encoding the full-length protein) for a prominent number of splice factors, a finding that was also corroborated by changes in protein levels for two of these. These findings suggest that the prominent mechanism of PTC-containing events, at least in part, can be explained by changes in the protein levels of splicing factors. This is further supported by the observation that the intronic regions surrounding deregulated PTC-exons are highly conserved, suggesting that they are subjected to splice factor-mediated regulation.

Our data therefore extend the importance of NMD to the regulation of PTC-deficient mRNA isoforms through the deregulation of splice factor levels. The mechanisms (s) by which this occurs will be the subject of future studies. They also provide a clue as to why the liver is more affected by the loss of NMD than BMMS, as the former organ displays much higher levels of AS and is therefore predicted to be more sensitive to alterations in splicing factor levels.

Finally, and most importantly, these findings expand the functional impact of NMD on transcriptome behavior to also include substantial indirect effects on PTC-splicing events and highlights the crucial importance of this pathway as a gatekeeper of transcriptome integrity.

A role for the stop codon and its flanking regions in regulating NMD

The importance of PTC identity and its flanking regions in the regulation of NMD is a relatively unexplored area and has not previously been addressed on a global scale. Here we were able to show that regions flanking regulated PTCs display a divergent and increased level of conservation when compared to both canonical stop codons and exonic regions, suggesting that they have been under selective pressure. Similarly, the distribution of regulated PTCs was markedly different from that of normal termination events with a strong preference for the NMD machinery to use UGA. This is interesting in light of recent findings showing that sequences (including stop codons) having positive impact on stop codon read-through lead to a reduction of NMD through the removal of UPF1 from the 3’ UTR through a translation-dependent mechanism [44]. These findings may suggest that the PTC and its surrounding sequences have been under selective pressure to optimize NMD efficiency in a process involving translational read-through, which in turn may be used by the cell for regulatory purposes.

Conclusions

By developing and applying a robust bioinformatic pipeline mediating a high-resolution study of AS and its dynamics, this whole-transcriptome analysis of two NMD-deficient primary mouse tissues provides a comprehensive quantification of the impact of NMD on untransformed mammalian transcriptomes, providing crucial novel insights into its role in both core and tissue-specific regulation of gene expression, significantly extending the importance of NMD from an mRNA quality pathway to a regulator of several layers of gene expression. Thus, in addition to removing low levels of splicing ‘errors’, and destabilizing targets of AS-NMD, our analysis reveals the potentially crucial importance of NMD in maintaining splicing homeostasis. In particular, the observed deregulation of full-length splicing regulators suggests that the NMD pathway controls their expression in a manner distinct from direct AS-NMD, and that their deregulation is an important contributor to the global deregulation of AS that we observe in NMD-deficient tissues.

Finally, we have provided a reference catalogue of NMD-regulated AS events as well as an open source tool, RAINMAN, facilitating the process of splice isoform inference and PTC prediction from RNA-seq reads. Future transcriptome studies in other organs and organisms will further reveal the impact of NMD on splicing patterns and how this pathway modulates biological read-out in a tissue- and pathway-specific manner.

Materials and methods

Mice

For the selective deletion of Upf2 in liver and BMM we used our conditional floxed Upf2 line and the LysMCre and MxICre driver lines as described previously [7]. Upf2 is recombined during macrophage differentiation, and the mature BMMS are devoid of any functional UPF2 protein as well as NMD activity. To rescue the Upf2fl/fl; MxICre from the previously reported hematopoietic lethality, Upf2fl/fl; MxICre and Upf2fl/fl were transplanted with wild-type bone marrow cells prior to poly-I:C injection. As described previously, liver was harvested 3 weeks after Upf2 recombination to avoid any indirect effects from the poly-I:C treatment [30]. All mouse work was performed according to national and international guidelines and approved by the Danish Animal Ethical Committee. This study was approved by the review board at the Faculty of Health, University of Copenhagen (LT-P0658).

cDNA synthesis for RNA-seq

Total RNA was harvested in Trizol (Invitrogen; Carlsbad, CA, USA from Upf2fl/fl; LysMCre and Upf2fl/fl-derived
male BMMs, grown as previously described [7]. Briefly, bone marrow cells were grown in vitro in the presence of macrophage colony-stimulating factor-conditioned medium for 7 days, giving essentially a 100% pure macrophage population. Total RNA from liver was harvested in Trizol 21 days after injection of poly I·C from Upf2fl/fl; Mx1Cre and Upf2fl/fl; Mx1Cre. We pooled 50 μg RNA from each of three age-matched males to get a total of 150 μg RNA, which was subjected to two rounds of mRNA purification by hybridizing to oligo(dT) beads (Dynabeads mRNA purification kit, Invitrogen). The resulting mRNA (1 μg) was then used as template to prepare cDNA. Double-stranded cDNA was essentially prepared as described by the manufacturer (Superscript Double-Stranded cDNA Synthesis kit, Invitrogen), using 10 μM random hexamers to prime first strand synthesis. Finally, the double-stranded cDNA was purified using QiaQuick PCR columns (Qiagen, Hilden, Germany) followed by phenol-chloroform extraction. The quality of the cDNA was verified using a Bioanalyzer.

RT-PCR and western blotting
Total RNA was purified using Trizol and 0.5 μg RNA was used to synthesize single-stranded cDNA with oligo d(T) primers, using a ProtoScript M-MULV First-Strand Synthesis kit (New England Biolabs, Ipswich, MA, USA) as described by the manufacturer. The cDNA was used in standard PCR reactions using Taq Polymerase (Invitrogen), with primers specific for the exons flanking the alternative exon (see Table S9 in Additional file 1). Sequencing was performed on an Illumina Genome Analyzer II flowcell, generating 75 bp single-end reads. RNA-seq data have been submitted to the NCBI Short Read Archive database with accession number GSE26561.

Statistical methods
The statistical package R was used to calculate Chi-square and Kolmogorov-Smirnov tests.

Database accession
RNA-seq data are available at the NCBI Short Read Archive database (accession number GSE26561). RAINMAN scripts, documentation and complete junction and gene expression lists are available online [47].

Additional material

Additional file 1: Supplementary Information and Supplementary Tables S1, S2 and S5. Supplementary Materials, Methods and References. Supplementary Table S1: number of mapped junctions contributed uniquely by the combinatorial database (Comb DB only) or TopHat (TopHat only) and the number of mapped junctions detected by both the combinatorial database and TopHat (Both). Supplementary Table S2: the contribution of TopHat to the number of junctions predicted to generate a PTC versus all junctions (minimum of three reads per junction). Supplementary Table S5: deregulation of core splice factors. Gene FC indicates the change in mRNA levels for all the isoforms for the particular gene between KO and WT.

Additional file 2: Supplementary Figure S1. UCSC Genome browser output of Pion gene and schematic of the RAINMAN pipeline with steps for mapping and processing of reads.

Additional file 3: Supplementary Figure S2. Histogram showing distance from normal stop codon to the 3' end of EF-Seq genes with
stop in final exon, and distances to nearest downstream exon-exon junction for genes with stop codons in the second to last exon.

Additional file 4: Supplementary Table S3. Reads per unique junction statistics for all samples, split into junctions discovered by mapping to the combinatorial database versus junctions discovered by TopHat.

Additional file 5: Supplementary Table S4. Table with Gene Ontology terms associated with genes containing upregulated PTC+-junctions that are unique for Upf2 KO liver or BMM.

Additional file 6: Supplementary Table S5. Results from validation by manual inspection of output from isoform class inference.

Additional file 7: Supplementary Figure S5. Validation of expression change inference and isoform inference.

Additional file 8: Supplementary Table S7. PTC upon inclusion isoforms (SES) upregulated in both Upf2 KO liver and BMM (APSI > 20%).

Additional file 9: Supplementary Table S8. PTC upon exclusion isoforms (SES) upregulated in both Upf2 KO liver and BMM (APSI < 20%).

Additional file 10: Supplementary Figure S6. Mean per position phastCon conservation score around single exon skipping events for BMMs. Numbers on x-axis indicate nucleotide intervals - 25 and 75 nucleotides for exons and flanking introns, respectively.

Additional file 11: Supplementary Figure S5. Conservation around upregulated PTCs, with mean per-position phastCon scores centered on the PTC for upregulated junctions in liver and BMMs.

Additional file 12: Supplementary Table S9. List of primers used in RT-PCR validation of splicing events.

Abbreviations
A3SS: alternative 3' splice site; A5SS: alternative 5' splice site; AFE: alternative first exon; ALE: alternative last exon; AS: alternative splicing; AS-NMD: alternative splicing coupled to nonsense-mediated mRNA decay; BMM: bone marrow-derived macrophage; bp: base pair; ES: exon skipping; EST: expressed sequence tag; GO: Gene Ontology; hnRNP: heterogeneous nuclear ribonucleoprotein; KO: knock-out; MEF: mouse embryonic fibroblast; MES: multiple exon skipping; MXE: mutually exclusive exon; NK: natural killer; NMD: nonsense-mediated mRNA decay; PCR: polymerase chain reaction; PSI: percent spliced in; PTC: premature termination codon; PTC-: PTC absence; PTC+: PTC containing; RAINMAN: RNAseq-based isoform detection and NMd sequencing; RPKM: reads per kb of gene model per million mapped reads; SES: single exon skipping; SR: serine/arginine-rich; UCE: ultraconserved element; UTR: untranslated region; WT: wild type.

Acknowledgements
This work was supported by grants from the Danish Natural Science Research Council, the Novo Nordisk foundation, the Lundbeck Foundation and the European Research Council (financial support to J Waage; EU FP7 framework programme/ERC grant agreement 204135).

Author details
1The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, DK2200 Copenhagen, Denmark. 2Biotech Research and Innovation Centre (BRIC), University of Copenhagen, DK-2200 Copenhagen, Denmark. 3Section for Gene Therapy Research, Rigshospitalet, University of Copenhagen, DK-2100 Copenhagen, Denmark. 4The Bioinformatics Centre, University of Copenhagen, DK-2200, Copenhagen, Denmark. 5BGI-Shenzhen, Shenzhen 518083, China. 6Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark.

Authors’ contributions
JoWe carried out experiments, drafted the manuscript and conceived the study. JoNa carried out the bioinformatic analysis and helped to draft the manuscript. GT, LZ, JuWa, and KK facilitated sequencing. ID and JSJ assisted with experimental work. AK participated in study design. BP helped to draft the manuscript and conceived the study. All authors have read and approved the manuscript for publication.

Competing interests
The authors declare that they have no competing or conflicting interests.

Received: 28 November 2011 Revised: 6 May 2012 Accepted: 24 May 2012 Published: 24 May 2012

References


47. RAINMAN (RnaSeq-based isoform detection and NMD Analysis pipeline). http://people.bioinf.uki.de/jwagie/RAINMAN/.

Submit your next manuscript to BioMed Central and take full advantage of:

• Convenient online submission
• Thorough peer review
• No space constraints or color figure charges
• Immediate publication on acceptance
• Inclusion in PubMed, CAS, Scopus and Google Scholar
• Research which is freely available for redistribution

Submit your manuscript at www.biomedcentral.com/submit
spliceR: An R package for classification of alternative splicing and prediction of coding potential from RNA-seq data
Gene expression

spliceR: An R package for classification of alternative splicing and prediction of coding potential from RNA-seq data

Kristoffer Knudsen1,2,3,4, Bo Torben Porse1,2,3, Albin Sandelin2,4,* and Johannes Waage1,2,3,4,*

1The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, DK2200 Copenhagen, Denmark
2Biotech Research and Innovation Centre (BRIC), University of Copenhagen, DK-2200 Copenhagen, Denmark
3 The Danish Stem Cell Centre (DanStem) Faculty of Health Sciences, University of Copenhagen, DK2200 Copenhagen Denmark
4The Bioinformatics Centre, University of Copenhagen, DK2200, Copenhagen, Denmark

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: With the advent of increasing depth and decreasing costs in digital gene expression technologies exemplified by RNA-sequencing, researchers are now able to profile the transcriptome with unprecedented detail. These advances not only allow for precise approximation of gene expression levels, but also for characterization of alternative isoform usage/switching between samples. Recent software improvements in full transcript deconvolution prompted us to develop spliceR, an R package for classification of alternative splicing. spliceR labels isoforms based on fully assembled transcripts, detecting single- and multiple exon skipping, alternative donor or acceptor sites, intron retention, alternative first or last exon usage, and mutually exclusive exon events. Alongside, event spliced-in/out values are calculated for effective post-filtering, and genomic coordinates of differentially spliced elements are annotated for downstream sequence analysis. Furthermore, spliceR has the option to predict the coding potential and thereby the nonsense mediated decay (NMD) sensitivity of transcripts based on stop codon position.

Availability and Implementation: spliceR is implemented as an R package, is freely available from https://github.com/splicer-tool/splicer, and has been submitted to the Bioconductor repository.

Contact: johannes.waage@gmail.com

Alternative splicing is an important part of the multi-layered process of RNA processing, elevating the potential number of unique products with orders of magnitude. More than 95% of all human genes undergo alternative splicing, and this is thought be a key element in driving the phenotypical complexity of mammals (Pan et al., 2008). Recent advances in sequencing technology of RNA (RNA-seq), combined with modern RNA-seq transcript assemblers, such as Cufflinks (Roberts et al., 2011), now allows for describing the diverse RNA landscape with high resolution.

Here we present an easy-to-use tool, spliceR, which builds upon common RNA-seq assembly workflows by annotating multiple transcripts from the same gene with alternative splicing classes. For each gene entity, each alternative transcript is classified against either the gene’s hypothetical pre-mRNA (based on all isoforms) or against the gene’s most expressed transcript. Additionally, spliceR allows for the characterization of the protein coding potential of each transcript by translating the full exonic sequence of each transcript with supported annotated open reading frames (ORFs).

spliceR is fully based on object types found in the Bioconductor (Gentleman et al., 2004) project, such as GRanges, allowing for full flexibility and modularity, and has been submitted to the Bioconductor repository.

Classification of alternative splicing: spliceR takes full-length transcript information either from Cufflinks, or from a data generated by any RNA-seq assembler that outputs fully deconvoluted transcripts. spliceR supports a number of filters, letting the user choose to classify alternative splicing only from those transcripts that have passed set of qualifiers, either given by cufflinks (transcript and/or gene model confidence), or based on NMD-sensitivity or expression thresholds.

In some scenarios, especially when looking at changes between samples where splicing patterns are expected to change, users may opt to configure spliceR to use the most expressed transcript as the reference transcript instead of the theoretical pre-RNA. Based on this comparison spliceR outputs a complete classification of splicing events as seen in figure 1a. Furthermore, percent spliced in (PSI) values are calculated for each of the transcripts, representing the percentage of the total parent gene expression originating from this transcript. Finally, delta-PSI values (dPSI) allow researchers to assess changes in splicing (i.e. isoform switching) between samples.
The output of spliceR also facilitates various downstream analyses, including filtering on transcripts that have major changes between samples, filtering for specific splicing classes, or sequence analysis on elements that are spliced in or out between samples. The latter could include detection of enriched motifs in or surrounding such elements, or identification of protein domains that are spliced in/out. spliceR facilitates this type of analysis by outputting the genomic coordinates of each alternatively spliced element.

Visualization: To analyze global trends in splicing, spliceR can produce a range of Venn diagrams, showing the overlap of splicing events between samples for either a specific type of alternative splicing or for all events. An example is shown in figure 1b.

Analysis of coding potential: For assessment of coding potential, spliceR initially retrieves the genomic exon sequence using BSGenome objects, easily downloadable from Bioconductor directly in R. Next, ORF annotation is retrieved from the UCSC Genome Browser repository from either Refseq, Known Gene or Ensembl. Alternatively, one can generate a custom ORF-table. Finally, the RNA sequences of input transcripts are assembled, and if a compatible ORF exists, translated, and positional data about the stop codon, including distance to final exon-exon junction, is recorded and returned to the user. Based on the generally accepted 50 nt rule in the NMD literature (Weischenfeldt et al., 2012), transcripts are marked NMD-sensitive if the stop codon falls more than 50 nt upstream of the final exon-exon junction, although this setting is user-configurable.

Conclusion: We present a Bioconductor package spliceR, which harnesses the power of current RNA-seq and assembly technologies, and provides a full overview of alternative splicing events and protein coding potential of transcripts. spliceR is flexible and easy integrated in existing workflows, supporting input and output of standard Bioconductor data types. To our knowledge, spliceR is the only software that facilitates comprehensive splice class detection based on full-length transcripts.

ACKNOWLEDGEMENTS

Funding: This work was supported in part through a center grant from the Novo Nordisk Foundation (The Novo Nordisk Foundation Section for Stem Cell Biology in Human Disease).

Conflict of Interest: The authors have nothing to declare.

REFERENCES


Visconte, V. et al. (2012) SF3B1 haploinsufficiency leads to formation of ring sideroblasts in myelodysplastic syndromes. Blood, 120, 3173–86.

Article III

Genome-wide mapping of transcription start sites reveals deregulation of full-length transcripts in acute promyelocytic leukemia
Genome-wide mapping of transcription start sites reveals deregulation of full-length transcripts in acute promyelocytic leukemia.

Johannes Waage1,2,3,4, Mette Boyd1,2, Kim Theilgaard-Mönch2,3,5, Helena Mora-Jensen6, Piero Carninci7, Nicolas Rapin1,2,3,4, Berit Lilje1,2, Peter Hokland8, Niels Borregaard6, Bo Torben Porse2,3,4, Albin Sandelin1,2,*

1 The Bioinformatics Centre, Department of Biology, University of Copenhagen, Copenhagen, Denmark; 2 Biotech Research and Innovation Centre (BRIC), University of Copenhagen, Copenhagen, Denmark; 3 The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; 4 The Danish Stem Cell Centre (DanStem) Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; 5 The Department of Hematology, Skanes University Hospital, Lund University, Sweden, 6 The Granulocyte Research Laboratory, Department of Hematology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark; 7 RIKEN Center for Life Science Technologies, Yokohama, Japan; 8 Department of Hematology, Århus University Hospital, Århus, Denmark

* Corresponding author
Abstract

Transcription start sites are the focal points of transcriptional regulation, where information from regulatory elements is integrated to stabilize initiation of transcription. In humans, most genes have more than one transcription start site, and these often exhibit different tissue specificity, serving as distinct regulatory frameworks for the same gene. Usage of such promoters can also result in differential gene function manifested on the protein level. Alternative promoter usage is increased in several disease states, including cancer, but its impact in most disease states is uncharacterized.

Most methods for genome-wide detection of transcription start sites cannot be applied to rare cell samples or highly purified cells because they need high amounts of RNA. In this study, we have applied the nanoCAGE method to create a genome-wide map of transcription start site usage in highly purified cancer cells from patients with acute promyelocytic leukemia, and corresponding normal bone marrow cells (i.e. promyelocytes) purified from healthy subjects, using as little as 50 ng total RNA.

We show that the nanoCAGE method gives similar results as experiments made with microarrays in terms of expression on gene level, but in addition allows for the identification of alternative promoter usage in cancer cells. We identify 2,162 putative promoters that are significantly differentially regulated between APL and controls. Interestingly, promoters whose usage is upregulated in APL have an increased propensity to be located within genes, downstream of annotated promoters. Conversely, promoters supporting annotated longest transcript variants are commonly downregulated in cancer. We show several examples of genes with upregulated alternative promoters located downstream, many of which confer protein domain loss that could contribute to leukemic transformation and maintenance. Moreover, we show that these downstream promoters likely have different regulatory cues than cancer-specific promoters corresponding to the longer RNA isoforms.
Introduction

Messenger RNA (mRNA) heterogeneity is a hallmark of higher eukaryotes, contributing to the plethora of tissue- and cell specific patterns of protein expression and function. The underlying mechanisms contributing to the final selection of mRNA isoforms include alternative promoter usage, alternative splicing, and alternative termination (reviewed in refs [1–3]), potentially generating several isoforms from a single gene.

There are several examples where alternative isoform usage can affect the protein product of a gene, typically by excluding exons that code for important protein residues, or domains (for a review, see [4]). Alternative promoters are particularly interesting generators of isoform diversity, since they confer additional regulatory inputs to the same gene. Indeed, there are many examples of alternative promoters that confer different tissue specificity within the same gene. An example is the Dlgap1 gene in mouse which has at least four promoters that are specific for different brain tissues [5]. There are indications that alternative promoter usage is increased in disease states, and particularly cancer[1].

Thus, mapping the promoter usage genome-wide in specific cell states, including disease, is important. There are several methods for high-throughput discovery of promoter usage, ranging from chromatin immunoprecipitation (ChIP) of key components of the preinitiation complex, to sequencing RNA 5’ ends [6]. RNA-based methods offer base-pair resolution, and are typically based on selection of cap structures at mature 5’ ends of RNAs followed by sequencing of the first 20-30 nt of the transcript from the 5’ end. The two most commonly used methods are Cap Analysis of Gene Expression (CAGE) [7] and TSS Seq [8]. While these methods have high sensitivity and specificity, they also require RNA from a substantial number of cells (typically >1 µg total RNA of high quality). This makes it hard to analyze rare cell types or highly purified cells from clinical tissue samples. The nanoCAGE method, based on template switching instead of cap trapping, is an alternative which can measure promoter usage with as little as 10ng total RNA, at the cost of lower
sensitivity [9,10]. Previous studies have used the nanoCAGE method to investigate the promoter landscape of cultured hepatocellular carcinoma cells [10,11], but it has not yet been applied to clinical samples. As a proof of concept, we applied the nanoCAGE method to comprehensively identify promoters specifically used in highly purified cells from patients with acute promyelocytic leukemia (APL), compared to corresponding purified controls, namely promyelocytes.

Acute myeloid leukemia (AML) encompasses a variety of clonal disorders, whose malignant populations do not respond normally to regulatory cues and have lost the ability to differentiate into fully mature blood cells [12]. APL represents a genetically defined subclass according to the current WHO classification of AML, and comprises 5-10% of all AMLs. APL is characterized by a block in terminal neutrophil differentiation and the accumulation of “leukemic” promyelocytes (APL blast cells) in the bone marrow and blood [13,14]. Over 98% of APLs harbor a t(15;17) translocation, which juxtaposes the promyelocytic leukemia gene (PML) and the retinoid acid receptor alpha gene (RARA) resulting in the expression of the PML-RARA fusion protein[14]. In normal cells, RARA forms heterodimers with the retinoid X receptor (RXR) that binds the retinoic acid responsive elements (RARE) of their target gene promoters. In the absence of retinoic acid (RA), RARA interacts with histone deacetylase (HDAC)-containing co-repressor complexes and represses transcription, whereas binding of RA triggers a conformational switch resulting in recruitment of coactivator complexes and subsequently, activation of transcription [15]. In APL cells, PML-RARA forms homodimers, multimeric complexes, and heterodimers with RXR that all bind to RARE and repress transcription at physiological levels of RA through recruitment of corepressor complexes containing HDACs, DNA methyltransferases and polycomb complexes [16–18]. At pharmacological levels, however, RA reverses PML-RARA mediated repression of RARA target genes resulting in terminal neutrophil differentiation of APL cells.

Despite the tremendous progress in our understanding of how PML-RARA regulates transcription at the molecular level, little is currently known to what extent
PML-RARA changes the TSS landscape in APL. Alternative promoter usage has been associated with cancer; hence, elucidating the promoter usage landscape in APL is an important aspect of the regulatory state of the disease.

In the present study we used APL as a cancer model to investigate changes of TSS usage associated with malignant transformation. For this we compare the TSS landscapes of highly purified leukemic promyelocytes (iAPL blasts) from APL patients with those of normal early promyelocytes (EPMs) purified from healthy subjects, using as little as 50 ng total RNA.

We find that the cancer cells tend to use downstream alternative promoters much more often than normal cells, thereby producing shorter transcripts, which lie at a median distance of ~6400 nt from the TSS generating the longest pre-mRNA. We demonstrate that usage of these downstream alternative promoters often leads to transcripts lacking the potential to code for important protein domains that could contribute to the cancer phenotype. Conversely, the usage of canonical TSSs (corresponding to full-length genes) for genes in repair pathways is downregulated in cancer cells. Furthermore, using transcription factor ChIP-seq data from the ENCODE project, we demonstrate that the two categories of putative TSSs have distinct regulatory profiles. We also report 90 cases of non-coding RNA promoters, in particular snoRNAs, that are upregulated in the APL cells.

**Materials and Methods**

**Sample preparation**

Bone marrow was aspirated from the posterior iliac crest of healthy donors and patients with newly diagnosed APL after informed consent had been given according to guidelines established by the Danish National Committee on Health Research Ethics, and cell sorted by fluorescence activated flow cytometry (FACS). Detailed FACS protocols are described in [19], and the FACS strategy is depicted in Figure 2.

RNA was isolated from sorted cells using the RNeasy Micro Kit (Qiagen, Valencia, CA, USA) according to manufacturer’s instructions. Briefly, BM populations were sorted once and then resorted directly into 350 µl RLT lysis buffer/β-
Mercaptoethanol (Sigma-Aldrich). Tubes containing sorted cells were vortexed immediately to lyse all cells, then snap-frozen on dry ice and stored at -80°C before mRNA was purified according to the manufacturer’s instructions (RNeasy Micro Kit, Qiagen, Valencia, CA, USA).

**NanoCAGE library preparation**

NanoCAGE was performed as described in [9], with 50 ng total RNA for all samples. RNA was sequenced on the Illumina Genome Analyzer IIX in biological duplicates for both EPM and APL samples, with an input of 2 nM in 10 uL per sample, following standard protocol.

**NanoCAGE tag mapping**

Reads were mapped to the NCBI GRCh37/hg19 human reference genome using Bowtie v. 0.12.8 [20] with standard parameters. See suppl. Figure S1 for mapping results.

**Tag clustering**

Initial consensus clusters were created, consisting of overlapping and neighboring tags for all samples merged, followed by tag quantification for each sample in the consensus clusters. Clusters were end-trimmed for low tag content, reducing each tag cluster to the minimum width required for at least 80% of tags to remain in the cluster, and clusters of 1 nt width were removed to eliminate PCR artifacts. Tag counts were quantified as TPM (tags per million mapped reads), and clusters with less than 2 TPM in the highest sample were removed. Clusters with high inter-replicate variance were removed using a coefficient of variance-threshold of 1, resulting in a final set of 7,458 clusters.

**Exon noise filtering**

To further solidify our distal promoter discovery, the per nt coverage of a given TC was divided by the average per nt coverage across exons for the respective gene (minus exons containing any tag clusters), and a ratio of minimum 1.5 was required.
Statistical testing for differential expression

To test for differential promoter usage between APL and EPM, we utilized edgeR v. 3.13 [21], using standard settings, with FDR = 0.05, and resulting in 1,437 clusters downregulated, and 725 clusters upregulated in APL.

Promoter classification

Promoters (tag clusters defined above) were assigned to the closest UCSC knownGene transcriptional unit, and were partitioned into four categories: Full-length canonical (within 300 nt from the most upstream annotated TSS of the closest gene), proximal (from 300 nt to within 1/10 of the total gene length), distal (from 1/10 to the 3’-end of the gene), and intergenic. Ambiguous promoters (due to e.g. overlapping gene units) were discarded.

Protein domain loss analysis

Transcript regions lost in due to usage of alternative promoters located within genes and upregulated in APL were intersected with protein domain mappings from the SUPERFAMILY repository [22], reporting both full and partial overlaps.

Regulatory neighborhood analysis

The full ENCODE transcription factor ChIP-seq binding site set [23] was downloaded from the UCSC Genome Browser, and the distance from each TC to the nearest binding site was determined for each TF. For a select subset of TFs, the distance from each TC to the full TF binding site set was measured.

Data availability

Raw sequence data as well as the full TSS set is available at the NCBI Gene Expression Omnibus and Short Read Archive under accession no. GSE46561.

Results

Figure 1 gives a full overview of the analysis flow, starting from FACS-sorting of cells, followed by microarray and nanoCAGE experiments and concluded by computational analysis and comparisons with existing datasets.
Analysis of cancer phenotypes is often hampered by the lack of appropriate normal controls, i.e. normal cells at a similar stage of differentiation. We have previously established that the closest normal counterparts of APL blast cells in terms of expression are early promyelocytes (EPM) cells [24]. Therefore, a FACS cell sorting strategy was developed to purify these two cell types from bone marrow samples, as described in Figure 2: reanalysis of double-sorted populations demonstrated a purity between 90%-100% for all populations. Importantly, by comparing APL blast to normal EPMs, we are therefore able to determine cancer specific transcriptional changes as opposed to those arising from differences in degree of differentiation.

Following cell sorting, for each patient, RNA was extracted and used to measure gene expression and promoter usage by nanoCAGE. Sequenced nanoCAGE reads were mapped to the human genome as described in Methods (suppl. Figure S1A). Using pooled data, neighboring tags were aggregated into tag clusters (TCs), whose expression from each library was quantified by the normalized number of tags within the cluster on the same strand (expressed as tags per million mapped reads, TPM). For simplicity, we will refer to these clusters interchangeably as TSSs or promoters in the text, as in [5]. Following initial filtering, the resulting TCs were found to have a distance distribution showing an increase of clusters close to previously annotated TSSs, and a decrease of clusters with similar distance to TSSs as random genomic locations (suppl. Figure 1B).

Visual inspection of raw mapped data showed a tendency towards low intensity tag clusters mapping uniformly over exons. These might represent cases where the reverse transcriptase failed to reach the real 5' end, or the capture of partially degraded (and possibly recapped) mRNAs, as discussed in [10]. We also noted cases of PCR clonal expansion, typically only present in one of the replicates; these are likely attributed to the low amount of starting RNA and the high number of PCR cycles employed in the NanoCAGE method. To take these issues into account, we employed additional filtering on clusters, requiring low inter-replicate variance, as well as constraints on cluster widths and expression levels. This produced a final set of 7,458
TCs. Of these, 1,398 overlapped with annotated RefSeq TSSs, 1,795 overlapped with GenCode V14 TSSs, and 5,663 were un-annotated.

Next, we identified differentially expressed promoters using edgeR [21] and found 725 promoters to be upregulated and 1,437 promoters to be downregulated in APL vs. control (Figure 3A) (P<0.05, FDR-adjusted).

To validate our findings in terms of expression, we compared our set of differentially expressed promoters with previously published microarray-based gene expression data [25]. Genes assigned as significantly downregulated by microarrays (n=2,666, P < 0.05, FDR-adjusted) overlap the corresponding promoters sets from CAGE significantly (n=563, P=1.34e-22, hypergeometric test) (Figure 3B, top). Similarly, for all genes found upregulated in APL, compared to normal counterpart cells (EPMs), by microarray (n=855, P<0.05, FDR-adjusted), we found a less marked, but still significant overlap (n=165, P=2.14e-25, hypergeometric test) with promoters upregulated in APL by nanoCAGE (Figure 3B, bottom). The combined microarray data was based on samples from 37 APL patients. Owing to the sheer difference in sample numbers between our APL nanoCAGE (2) and the array studies, some precaution is necessary when comparing expression values between platforms, and smaller overlap fractions were expected.

Next, we investigated the location of differentially expressed promoters in relation to annotated genes. We partitioned all promoters across both samples into bins of location relative to nearest gene body, grouping 16% as full-length canonical (here defined to be within -150/150 bp of longest UCSC transcript TSS on the same strand), 7% as intragenic proximal (from 150 bp to 1/10 th of the total gene length), 43% as intragenic distal (placed somewhere along the remaining gene body), and 33% as intergenic (figure 3, grey bars).

Interestingly, promoters upregulated in APL were overrepresented in the category of TCs located downstream, distal of annotated TSSs (P<0.01, hypergeometric test); conversely, downregulated promoters were overrepresented in the bin of TCs overlapping canonical TSSs (P < 0.01, hypergeometric test) (figure 3).
The distal upregulated promoters, by definition located between 1/10 of the gene length and the transcription termination site, were not placed uniformly along the gene, peaking around a median distance from the TSS of ~6400 nt (suppl. Figure S2). In summary, TSSs preferentially used in APL are more often downstream alternative promoters compared to promoters preferentially used in the EPM control samples.

Comparing to gene set signatures collected from the Molecular Signature Database (http://www.broadinstitute.org/gsea/msigdb/index.jsp) [26], we observed a significant overlap between our upregulated distal promoters, and genes overexpressed in APLs (P = 2.3e-12, hypergeometric test) [27], chronic myelogenous leukemia (CML) (P = 2.44e-15, hypergeometric test) [28], as well as a number of other cancers, indicating that promoters of oncogenic and cancer maintaining genes, expressed in APL, produce shorter transcripts (Figure 4, overlay). Similarly, we observed a significant overlap between down-regulated canonical promoters, and gene sets for DNA repair pathways, including tumor suppressors CHEK2 (P<0.01, hypergeometric test) and ATM (P<0.01, hypergeometric) both involved in activation of the DNA damage checkpoint, leading to repair, cell cycle arrest, and/or apoptosis (Figure 4, overlay) [29], further validating the observation.

Since degraded RNA has potential to be captured by the nanoCAGE protocol, heightened RNA degradation activity in the cancer cells could potentially cause some of these observations. We reasoned that if this were the case, then we would expect the nanoCAGE tags to be uniformly distributed over the exons, and not congregate to distinct and reproducible clusters over replicates. Even though our initial intra-replicate filtering strategy would catch most of these exon tags, we further refined our filter, requiring that the downstream, cancer-specific tag clusters of interest should have higher tag coverage than expected over the exons of the respective gene (see Methods). This filtering resulted in a list of 40 high confidence downstream promoters upregulated in APLs (the full list is given in suppl. table T1), suitable for further downstream functional analysis. Two examples of such are given in Figure 5. Both of these are close to annotated alternative promoters (further validating the approach),
but here we give evidence that the alternative promoters are used preferentially in the APL cells compared to controls. Figure 5A shows an APL-specific alternative promoter, which generates a shorter transcript version of ADAMTSL2. This is placed downstream of several protein-coding exons, and usage of the alternative promoter will produce a transcript which cannot encode the signal peptide present in the longer variant of the RNA, as well as several regions in which mutations have been shown to result in increased amounts of available TGF-beta [30]. TGF-beta, known to inhibit proliferation of hematopoietic precursors, is normally found highly expressed in APLs [31], and the selective downstream promoter usage of ADAMTSL2 could contribute to this effect. Figure 5B shows a downstream alternative promoter in the GPR56 gene, a GPCR family receptor recently shown to play a role in both hematopoietic stem cell as well as leukemic stem cell maintenance [32]. In this case, the alternative promoter is not predicted to give a different protein product, as it is located in an extended 5' UTR. An overview of 24 selected genes with downstream promoters upregulated in APL is presented in Figure 6, indicating novel promoters (15/24) and previously annotated alternative promoters (9/24). To further investigate the functional effects of downstream promoter usage, we extrapolated the lost mRNA exons to protein domains. Several genes exhibited domain loss, which could be implicated in an oncogenic phenotype, including: USP13, involved in deubiquitination and implied in the stabilization of p53 [33], loses its proteinase domains. ADAMTSL2, in addition to the regulation of TGF-beta described above, looses its TSP-1 type 1 repeat domains (see Figure 5A), previously shown to be implicated in induction of apoptosis [34] and inhibition of angiogenesis [35]. CTNNA3, a key player in cellular adhesion, loses its cadherin associating alpha-catenin domains. A full list of domains lost is given in suppl. table T2.

To further assess the regulatory differences between the promoter classes, we investigated the binding of transcription factors in the genomic neighborhood of each TSS class. To facilitate this, we measured the distance from each TC to the nearest transcription factor binding site, based on ChIP-seq peaks from ENCODE (all cells)
We observed that transcriptional inhibitor members of the E2F family were associated closer with APL-downregulated full-length promoters, while early response transcription factors JUN and FOS (together forming the AP1 complex, and both expressed in APL cells measured by nanoCAGE (data not shown)) were closer associated with upregulated downstream promoters (figure 7). This indicates that some regulatory cues differ between the downstream and canonical promoters, even within AML cells.

A strength of nanoCAGE is that it is not limited to a pre-defined gene set, but can potentially capture any RNA species in the cell, including ncRNAs. To categorize putative non-coding TSSs, we extracted all non-coding RNA identifiers from RefSeq: of these, 90 and 84 were up- and downregulated (figure 8A), respectively, with a significantly larger fraction of non-coding RNAs being upregulated compare to non-regulated (figure 8B). Suppl. table T3 lists the most upregulated ncRNAs. SNORD114-1 has previously been found upregulated APL in a PML-RAR-alpha context by microarray and ChIP-seq analyses, conferring cell growth.[36] Similarly, miR-21, classically described as an “oncomir” [37], and with many tumor-suppressor targets, is upregulated. Curiously, MEG3 is high on the list of upregulated ncRNA TSSs, normally thought to be a tumor suppressor not normally expressed in tumors [38] and has previously been show to be hypermethylated in AML and myeloplastic syndrome (MDS). Microarray data confirm this upregulation (7 times upregulated in APL). The high expression of MEG3 in APL may point to a new and as yet undescribed role in cancer pathology.

Discussion

In this paper we present the first investigation of promoter usage in APL cancer cells, compared to their corresponding normal counterpart, both FACS-purified from human bone marrow. The nanoCAGE method allows for analyses of promoters with these rare cells – in our case with as little as 50 ng total RNA. By applying stringent filtering and statistical analyses, we identify 2,162 TSS clusters that are significantly changed between the two states. On the gene level, the findings correspond well with
microarray studies, but the method can pinpoint novel promoters in the set, which array approaches cannot. Surprisingly, we see a shift of promoter usage: APL cells are prone to use promoters that are downstream of canonical TSSs, downregulating usage of full-length canonical transcripts. To further refine these findings, we employed a stringent filter on the TC-to-exon-coverage, which resulted in a list of 40 high-confidence APL-specific promoters - followed by an extrapolation of mRNA to protein domains, we demonstrated several examples of domain loss with clear implications to protein function in the cancer cells, several of which can be linked mechanistically with a cancer phenotype. Furthermore, we noted that the regulatory landscape around these downstream promoters is different to those of full-length transcripts, even if the full-length transcripts are preferentially expressed in APL. In particular, AP1 transcription factors are preferentially bound closer. A caveat with these results is that the ENCODE ChIP-seq data is based on multiple cell lines. On the other hand, we could in many cases see corresponding support of expression of the TF in question by nanoCAGE, which argues that the observed regulatory event should also be relevant for the APL cells.

In summary, our findings indicate that alternative promoter usage is potentially important yet often overlooked biological feature of APL cells. While these results are indicative of functional impact, it is unclear what the causality is: we do not at present know if the observed shorter isoforms are a result of disease progression or play a causal role in malignant transformation. To address such questions, it would be necessary to either knock down or overexpress the shorter variants and observe the resulting phenotypes.

Regardless of cause and effect, these types of cancer-specific alternative promoters could potentially be used as biomarkers for APL, since they have additional power compared to microarray expression profiles, which cannot detect the placement of alternative promoters. However, for a comprehensive delineation of APL-specific biomarkers it would then be necessary to expand the study to cover multiple APL and normal cancer states, as well as biological variation over many patients.
Acknowledgments

The authors thank the National High-throughput DNA Sequencing Centre of Copenhagen, Denmark for collaborative assistance with sequencing.

This study was supported by grants to AS from the Danish Cancer Society, The Lundbeck Foundation, the Novo Nordisk Foundation and the European Research Council (FP7/2007–2013/ERC grant agreement 204135). Work in the Porse lab was supported by grants from the Danish Cancer Society, The Danish Research Council for Strategic Research and through a center grant from the Novo Nordisk Foundation (The Novo Nordisk Foundation Section for Stem Cell Biology in Human Disease).

Author contributions

JW and AS carried out the computational analysis, MB and PC carried out laboratory work. HMJ and KT collected cell samples and carried out cell sorting. NR provided microarray data and analysis. BL contributed with domain loss analysis. PH and NB contributed with clinical sample collection. BTP and AS supervised the project. JW, BTP and AS wrote the paper.

References


times and a unique clustering of S-phase cells in patients with acute promyelocytic leukemia [see comments]; 1037–1048.


Figure 1: Pipeline overview

Flowchart of the full data analysis process presented in this work. Tag cluster data sets are depicted in green, experimental and computational steps to produce data in blue, and analysis modules in red.
Figure 2: FACS sorting strategy for the purification of APL blasts and their corresponding normal bone marrow population.

Mononuclear cells (MNCs) were purified from BM aspirates of healthy subjects and APL patients. MNCs were stained with a cocktail of fluorochrome-conjugated MoAbs and the DNA dye 7AAD allowing for sorting of immunophenotypically identical normal early promyelocytes (EPM) and APL blast as described previously [19]. Purified EPM and APL blasts were defined as Lin-CD34lo/negCD15+ cells. Wright Giemsa stains of sorted EPM and APL blasts demonstrate a typical promyelocyte morphology.
Figure 3: Characteristics of gene expression in APL by nanoCAGE

(A) MA-plot for normalized nanoCAGE-seq data, showing the combined sample expression on the x-axis, and the sample fold change on the y-axis. Blue data points represent candidate TSSs which are significantly different between APL cells and control (with FDR <= 0.05), giving 1,437 downregulated and 725 upregulated genes, respectively. (B) Venn diagrams comparing the overlap of TSSs of genes down- or upregulated in APL using microarrays and nanoCAGE.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADAMTS12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AKAP9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ARGLU1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GCNT1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CTNNA3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ELF1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ERG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ESYT2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLG1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPR56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSPA8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFT88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIAA0226L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KIF24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PZP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROBO3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUVBL1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SNX20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ST5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TPM4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USP13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>USP53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VPS13B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ZFR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- APL promoter, supported
- UCSC longest transcript
- APL promoter, novel
Figure 4: (Previous page) Classification of candidate transcription start sites

Barplot of genomic location classifications of candidate TSSs for pooled samples (grey bars), or the subsets which are not changing (green), upregulated in APL (blue), or downregulated in APL (red). For these subcategories, we investigated their locations: overlapping known TSSs (canonical), novel TSSs within genes either proximal or distal to the TSS, or intragenic (See Methods for details).

TSSs downregulated (FDR<0.05, N=827) in APL are overrepresented in canonical promoters, while TSSs upregulated in APL (FDR<0.05, N=364) are overrepresented in intragenic promoters, supporting shorter transcripts of known genes. Overlay boxes show gene sets from other studies enriched for either downregulated canonical promoters (pink), or upregulated distal promoters (teal).
Figure 5: Examples of AML-specific alternative promoters

Two examples of AML-specific alternative promoters identified in the study, at the ADAMTSL2(A) and GPR56(B) loci. Top panel shows UCSC genes from the UCSCS genome browser: grey panels show normalized mean TPM counts over replicates for AML and NanoCAGE tags on the Y axis and the genomic location on the X axis. For ADAMTSL2, the mRNA stretches corresponding to protein domains lost by alternative promoter usage are indicated in colored boxes.
Figure 6: Visual representation of 24 genes with downstream promoters upregulated in APL.

Gene models visualized are based on longest canonical UCSC gene models, and are normalized to have the same lengths. Thick boxes indicate protein-coding exons. Orange arrows indicate the TSSs inferred from the gene models. Red arrows indicate AML-specific promoters not previously annotated by either UCSC Known Gene, RefSeq or Ensembl. Green arrows indicate AML-specific promoters that are supported by aforementioned databases. Note that the large majority of downstream promoters are within or between coding exons.
Figure 7: Distance distributions from transcription factor binding sites to candidate TSSs

Distribution of selected ENCODE transcription factor ChIP peaks (pooled over all samples) in relation to TSS categories. For each TSS in all three categories, the distance to the nearest respective TF was logged, and the total distance distribution is shown. *** indicate $p < 0.001$, * indicates $p < 0.05$, NS indicates non-significance.
Figure 8: Characteristics of expression of non-coding RNAs

A) MA-plot of non-coding RNAs. Blue dots indicate ncRNAs upregulated in APL (p < 0.05, n = 90), red dots indicate ncRNAs downregulated in APL (p < 0.05, n=84). B) Compared to the total number of genes deregulated in APL, ncRNAs are overrepresented in upregulated genes.
Figure S1: A; Vigourous filtering ensures a reduced but high confidence set of candidate promoters.

“Reads mapped” indicate reads mapped to mm9, after removing PhiX-specific sequences. Common tag clusters are defined as overlapping reads, based on all samples pooled. The filtering steps done to reach 5,748 candidate TSS’s are described in the text.

B: Violin plots of distance to nearest RefSeq TSS of all (pink) and filtered (blue) TSSs. The blue line indicates the mean of the distance of random genomic locations to nearest RefSeq TSS.
Figure S2: Density plot of distance in nucleotides from distal upregulated promoters to the TSS of the longest refseq gene model.

The green line indicates the median.
Figure S3: Distance boxplots from each tag cluster of a given category to the closest ENCODE TF binding site.

[Figure too large for inclusion. Downloadable at https://www.dropbox.com/s/2e50t51xnntpx5r/Figure_S3.pdf]
Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries
Temporal mapping of CEBPA and CEBPB binding during liver regeneration reveals dynamic occupancy and specific regulatory codes for homeostatic and cell cycle gene batteries

Janus Schou Jakobsen,1,2,3,7 Johannes Waage,1,2,3,4 Nicolas Rapin,1,2,3,4 Hanne Cathrine Bisgaard,5 Fin Stolze Larsen,6 and Bo Torben Porse1,2,3,7

1The Finsen Laboratory, Rigshospitalet, Faculty of Health Sciences, 2Biotech Research and Innovation Centre (BRIC), 3The Danish Stem Cell Centre (DanStem), Faculty of Health Sciences, 4The Bioinformatics Centre, 5Department of Cellular and Molecular Medicine, Faculty of Health Sciences, University of Copenhagen, DK-2100 Copenhagen, Denmark; 6Department of Hepatology, Rigshospitalet, DK2200 Copenhagen, Denmark

Dynamic shifts in transcription factor binding are central to the regulation of biological processes by allowing rapid changes in gene transcription. However, very few genome-wide studies have examined how transcription factor occupancy is coordinated temporally in vivo in higher animals. Here, we quantified the genome-wide binding patterns of two key hepatocyte transcription factors, CEBPA and CEBPB (also known as C/EBPalpha and C/EBPbeta), at multiple time points during the highly dynamic process of liver regeneration elicited by partial hepatectomy in mouse. Combining these profiles with RNA polymerase II binding data, we find three temporal classes of transcription factor binding to be associated with distinct sets of regulated genes involved in the acute phase response, metabolic/homeostatic functions, or cell cycle progression. Moreover, we demonstrate a previously unrecognized early phase of homeostatic gene expression prior to S-phase entry. By analyzing the three classes of CEBP bound regions, we uncovered mutually exclusive sets of sequence motifs, suggesting temporal codes of CEBP recruitment by differential cobinding with other factors. These findings were validated by sequential ChIP experiments involving a panel of central transcription factors and/or by comparison to external ChIP-seq data. Our quantitative investigation not only provides in vivo evidence for the involvement of many new factors in liver regeneration but also points to similarities in the circuitries regulating self-renewal of differentiated cells. Taken together, our work emphasizes the power of global temporal analyses of transcription factor occupancy to elucidate mechanisms regulating dynamic biological processes in complex higher organisms.

Mammalian liver regeneration is a well-studied process, in which the large majority of mature hepatocytes rapidly and in a highly synchronized manner re-enter the cell cycle upon injury (Fausto et al. 2006; Michalopoulos 2007; Malato et al. 2011). Serial transplantation of liver tissue has demonstrated a very high “repopulating” capacity of hepatocytes (Overturf et al. 1997), which lends hope to using these cells in regenerative medicine. The ability of mature liver cells to proliferate is reminiscent of specific, differentiated cells of the immune system (naïve T cells, B cells) that are kept in quiescence until exposed to specific stimuli (Glynn et al. 2000; Yusuf and Fruman 2003; Feng et al. 2008). However, it remains open whether similar programs control the proliferation of hepatocytes and immune cells. In the liver, a number of studies have mapped temporal changes in mRNA levels during regeneration (e.g., White et al. 2005). This, coupled with functional studies, has led to the identification of several TFs involved in the regenerative response (for review, see Kurinna and Barton 2011). Still, knowledge about how these factors are coordinated temporally throughout the regenerative process is limited.

CEBPA (C/EBPalpha) and CEBPB (C/EBPbeta) are two key hepatocyte TFs known to have divergent roles in liver function and regeneration. The two factors belong to the same basic region leucine zipper-family (bZIP), and several studies have shown that they bind the same core DNA sequence, acting as either homo- or
heterodimers (Diehl and Yang 1994; Rana et al. 1995; Osada et al. 1996). While CEBPA is highly expressed in the quiescent condition (before injury) and regulates many metabolic liver genes, CEBPB is up-regulated during liver regrowth and is required for a full regenerative response (Greenbaum et al. 1995; Wang et al. 1995).

In many tissues, CEBPA is observed to be an anti-proliferative factor facilitating differentiation, while CEBPB has been found to be either pro- or anti-proliferative in different settings (Porse et al. 2001; Nerlov 2007). In the skin, the two factors appear to act redundantly to limit epidermal stem cell activity (Lopez et al. 2009).

In the current study, we have examined dynamic TF binding during liver regeneration in the mouse by performing a time course of ChIP-seq experiments to map and quantify binding of CEBPA and CEBPB on a genome-wide scale at a high level of temporal resolution. We find that CEBPA and CEBPB generally occupy the same positions in the genome of hepatocytes in vivo, but they do so with quantitatively divergent temporal patterns during regeneration.

To further dissect the dynamic transcriptional network behind liver regeneration, we interrogated the temporally defined groups of CEBP-bound elements for regulatory properties, both with respect to sequence composition and differential expression of associated genes.

Results

A time course of CEBPA and CEBPB in vivo ChIP experiments reveals three distinct patterns of binding

Liver regeneration in rodents has been studied in detail using partial hepatectomy, in which three of five lobes of the liver are resected. The hepatocytes in the remaining liver lobes undergo up to two cell cycles during the first week of liver regeneration, hereby reestablishing presurgical liver mass (for reviews, see Fausto et al. 2006; Michalopoulos 2007). To examine the binding dynamics of CEBPA and CEBPB during liver regeneration, we harvested regenerating liver tissue at eight time points (0, 3, 8, 16, 24, 36, 48, and 168 h; Methods). These time points cover the quiescent state (G0), several stages of the first cell cycle growth phase (G1), the G1–S-phase transition at 36 h, a later time point at 48 h, as well as the terminal phase point of 168 h (Fig. 1A; Matsuo et al. 2003; Fausto et al. 2006). Livers from five mice for each time point were subjected to ChIP using specific CEBPA or CEBPB antibodies (Methods) (Fig. 1B). We confirmed antibody specificity by doing ChIP in mice deficient for Cebpa or Cebpb or by including epitope blocking peptides (Supplemental Fig. S1). To minimize experimental variation, we pooled DNA precipitated from each of the five
mice in equal amounts and sequenced the combined material. We mapped the sequences to the mouse genome (mm9) and determined base pair coverage after internal normalization to the total mapped read count for each antibody (Supplemental Methods; Supplemental Fig. S2; Supplemental Table S1). Genomic regions bound by either CEBPA or CEBPB were identified with the Useq peak-finder algorithm using a mock immunoprecipitation (IP) control (Supplemental Methods; Supplemental Table S2). We validated IP consistency of two series of three independent ChIPs at four different CEBP bound genomic locations (Supplemental Fig. S3). For further validation, we performed de novo motif searches on several data sets and a conservatism score analysis centering on CEBP motifs, all of which confirmed the quality of our ChIP-seq data (Supplemental Fig. S4).

Next, CEBPA or CEBPB peaks for each time point were used to construct a sum list of all enriched regions. Four examples of enriched regions located just upstream of gene loci are shown in Figure 1, depicting genomic coverage for the CEBPA and CEBPB ChIP time series (Fig. 1C).

After applying a stringent filtering regimen, we identified a total of 11,314 high-confidence regions bound by CEBPA or CEBPB, many of which show highly dynamic occupancy during the course of regeneration (for full peak calling numbers and region coverage, see Supplemental Tables S2, S3; for filtering, see Supplemental Methods). We assembled temporal binding profiles for all bound regions based on maximal genomic coverage. As illustrated in Figure 1 for the four regions mentioned above, distinct profile patterns can be observed (Fig. 1D).

The high temporal resolution of the obtained binding profiles allowed us to query for groups of putative cis-regulatory regions with similar occupancy dynamics. By hierarchical clustering, we identified three prevalent clusters (Fig. 2A), containing roughly equal numbers of bound regions (3549, 2818, and 3034 for the A, B, and C clusters, respectively). Two clusters (A and B) were defined by strong CEBPB binding, with maxima either at the 3- and 36-h (A) or at the 3-h (B) time points, which is also evident in a summed profile of temporal coverage (Fig. 2B). As opposed to the A and B clusters, the C cluster was characterized by a much more dynamic occupancy pattern, which can be observed in the coverage (Fig. 2B).
Regulation of liver regeneration

clusters, the C cluster was characterized by a high degree of temporal correspondence between the CEBPA and CEBPB binding levels, with maxima at the quiescent state 0-h time point and at 24 h (Fig. 2A,B). The detected robust binding of CEBPA at the 16- and 24-h time points was unexpected, as previous studies have reported decreased protein levels for this TF throughout the regenerative process (Greenbaum et al. 1995, 1998). To clarify this, we examined the protein levels for CEBPA and CEBPB until the first round of replication (time points, 0–36 h) by Western blotting (Fig. 2C). This highlighted a clear correlation between the CEBPA to CEBPB protein level ratios and the observed binding patterns. Specifically, we detected high levels of CEBPB at the 3- and 36-h time points, high CEBPA levels at the quiescent state (0 h), and a return of CEBPA predominance at 16- and 24-h time points (Fig. 2C,D).

The three binding pattern clusters are associated with specific functional classes of genes

The temporally distinct CEBP binding patterns of putative cis-regulatory regions belonging to the three clusters suggest divergent regulatory roles through the regenerative process. To address this possibility, we investigated whether each cluster was associated with specific sets of differentially expressed genes. To this end, we performed ChIP-seq experiments with liver tissue from the eight time points outlined above with an antibody specific to RNA polymerase II (POL2). In contrast to examining steady-state mRNA levels, this allowed us to directly measure the transcriptional activity at any given time point as POL2 binding to each gene body (Supplemental Methods; Sandoval et al. 2004). To pinpoint genes regulated by CEBPA and/or CEBPB, a single putative target gene was assigned to each bound region, based on proximity to the transcription start site (TSS) of neighboring genes (Supplemental Methods). Genes associated with regions belonging to one of the three binding profile clusters (A, B, or C cluster genes) were inspected for differential expression (for a full target gene list, see Supplemental Table S4).

For an initial overview, we counted differentially expressed genes from each cluster, defined as genes with a change in POL2 gene body coverage above twofold, comparing each time point to 0 h (Fig. 3A). This revealed that all clusters are associated with prominent gene expression changes at the 3-h time point. The C cluster was associated with a large group of down-regulated genes at 3 h and, furthermore, displayed up-regulation of many genes at 24 h. In contrast, the A and B clusters mostly exhibit down-regulation at the 24-h time point and up-regulation early in the time course (3 and 8 h), as well as a resurgence at 36 h (mainly the A cluster).

Next, we focused the analysis on the 0-, 3-, 24-, and 36-h time points, as these displayed the most prominent TF binding differences between the three clusters and therefore are likely to represent the most distinct regulatory states. Groups of genes, up- or down-regulated from one time point to the next (i.e., from 0–3 h, 3–24 h, and 24–36 h), were subjected to gene ontology (GO) analysis with the online DAVID tool (Supplemental Methods) (Huang da et al. 2009). As summarized in Table 1, we observed extensive differences in functional classes (GO terms; Biological Process) of differentially expressed genes associated with the three binding patterns (for full GO analysis lists, see Supplemental Table S5).

At the 0- to 3-h transition, the A cluster displays almost equal numbers of genes up- and down-regulated (554 vs. 682) (Table 1). In contrast, the B and C clusters showed mainly decreased target gene activity, with 610 and 709 down-regulated genes (63% and 70% of total regulated genes, respectively). The down-regulated genes associated with both B and C regions are annotated with lipid metabolism or oxidation-reduction terms, while the activated genes are associated with acute phase or inflammatory response. This shift of gene activity is expected, as the liver shifts from a quiescent, homeostatic state to an acute stage as a response to injury. By inspection of shared B and C target gene loci, we found several examples of B and C peaks in close proximity to each other. These often bind CEBPs at mutually exclusive time points, as exemplified in Figure 3B (lower panels). This observation suggests that CEBP complexes, possibly via interaction with specific co-activators or repressors at B or C regions, could have opposing actions on transcription, leading to timed fine-tuning of gene activity.

Prominent gene expression changes from 3 to 24 h include a significant down-regulation of A cluster genes with the term “transcriptional regulatory activity,” a shift that is paralleled by a decrease of A cluster CEBPA and CEBPB occupancy. This may suggest that the expression of these regulatory factors is dependent on CEBP binding. Another finding was the up-regulation of genes targeted by C cluster regulatory regions (889 of 1209 differentially...
Table 1. Target gene GO categories define distinct regulatory roles for the A, B, and C binding clusters

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time point 0–3 h</th>
<th>P-value</th>
<th>Time point 3–24 h</th>
<th>P-value</th>
<th>Time point 24–36 h</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>UP</td>
<td>Hexose metabolic process</td>
<td>0.255</td>
<td>UP</td>
<td>Purine nucleotide binding</td>
<td>0.082</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Programmed cell death</td>
<td>0.267</td>
<td>562</td>
<td>Adenylyl nucleotide binding</td>
<td>0.086</td>
</tr>
<tr>
<td>DOWN</td>
<td>Amine biosynthetic process</td>
<td>0.012</td>
<td>DOWN</td>
<td>Transcription regulator activity</td>
<td>2.6 × 10^{-6}</td>
<td>DOWN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cellular amino acid biosynthetic process</td>
<td>0.014</td>
<td>749</td>
<td>DNA binding</td>
<td>2.2 × 10^{-4}</td>
</tr>
<tr>
<td>B</td>
<td>UP</td>
<td>Regulation of inflammatory response</td>
<td>0.016</td>
<td>UP</td>
<td>Solute:cation symporter activity</td>
<td>0.277</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cofactor metabolic process</td>
<td>0.063</td>
<td>518</td>
<td>Symporter activity</td>
<td>0.285</td>
</tr>
<tr>
<td>DOWN</td>
<td>Organic ether metabolic process</td>
<td>0.007</td>
<td>DOWN</td>
<td>Transcription regulator activity</td>
<td>0.036</td>
<td>DOWN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Triglyceride metabolic process</td>
<td>0.009</td>
<td>460</td>
<td>Lipase activity</td>
<td>0.521</td>
</tr>
<tr>
<td>C</td>
<td>UP</td>
<td>Acute inflammatory response</td>
<td>0.008</td>
<td>UP</td>
<td>Oxidation reduction</td>
<td>2.4 × 10^{-4}</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Response to wounding</td>
<td>0.031</td>
<td>889</td>
<td>Steroid metabolic process</td>
<td>1.0 × 10^{-3}</td>
</tr>
<tr>
<td>DOWN</td>
<td>Oxidation reduction</td>
<td>5.8 × 10^{-8}</td>
<td>DOWN</td>
<td>Positive regulation of macromolecule biosynthetic process</td>
<td>0.064</td>
<td>DOWN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lipid biosynthetic process</td>
<td>8.9 × 10^{-5}</td>
<td>320</td>
<td>Positive regulation of nitrogen compound metabolic process</td>
<td>0.064</td>
</tr>
</tbody>
</table>

The two top hits of GO biological process categories enriched are shown. Genes assigned to a CEBP binding cluster were included in the UP and DOWN bins when exceeding a fold change of transcriptional activity above 1.5. Changes are measured as POL2 binding levels relative to a preceding time point as indicated. P-values are Benjamini corrected for multiple testing. For a complete GO term list, see Supplemental Table S5.
expressed genes). Enriched C cluster GO terms were consistent with a resurge of homeostatic and metabolic gene activity before S-phase entry (Table 1). One notable example is Cps1, carboxylmethylphosphate synthetase 1, which encodes a rate-limiting enzyme in the urea-cycle (Jones 1965) and displays a pronounced up-regulation at the 24-h time point (Fig. 3B, upper left panel).

At the 36-h time point, the A cluster associated genes displayed a strong bias toward increased transcriptional activity, with 545 up-regulated against only 73 down-regulated, in contrast to the C group target genes with 177 up and 196 down. In the large set of up-regulated A cluster genes, we found a strong overrepresentation of cell cycle GO terms (Table 1). An example of a cell cycle regulator gene (Mdc1) targeted by an A cluster enhancer is shown in Figure 3 (Fig. 3B, upper right panel). The B cluster shows association with relatively few regulated genes at this time point (272 up and 77 down).

These data collectively show the three clusters to be associated with defined sets of genes with distinct biological functions and timing of expression.

Motif analysis identifies specific and mutually exclusive sets of regulatory code

The targeting of distinct sets of genes and the distinct binding patterns suggest the presence of divergent regulatory features in the regions constituting the three temporal clusters. Combinatorial binding of multiple TFs has been proposed to explain dynamic changes in occupancy and gene regulation during development (e.g., Zinzen et al. 2009).

To examine this possibility in our system, we carried out an analysis of the collective set of sequences of each binding cluster (A, B, or C) versus a background set. We counted instances of known TF cognate sequences to generate probability scores using the ASAP software (Marstrand et al. 2008). In order to attain comprehensive coverage with minimal redundancy, the analysis was performed using a condensate of three publicly available databases of TF binding sequences or position weight matrices (PWMs) (Supplemental Methods; Supplemental Table S6). Probability Z-scores for all 246 condensed PWMs with representation above a threshold were used for hierarchical clustering to examine the general differences among the three clusters (Fig. 4A). Key examples are shown with full Z-score information (Fig. 4B).

As expected, we found CEBP motifs to be among the most enriched for all of the binding profile clusters (Fig. 4A,B; Supplemental Table S6). From the hierarchical clustering, it was evident that the majority of motifs found in the A set were shared with the B set, while a small group of motifs was found to have robust overrepresentation in all the three clusters (Fig. 4A). Prominent among these were motifs recognized by CREB and HNF4A (Fig. 4A,B), two TFs known to have a wide set of target genes in hepatocytes and central roles in liver metabolism and development (Costa et al. 2003; Montminy et al. 2004; Bolotin et al. 2010).

Notably, many motifs displayed a clear mutually exclusive pattern of overrepresentation (Fig. 4A). Specifically, many motifs found in the A and B clusters of cis-regulatory elements were underrepresented in the C cluster and vice versa. Binding sequences of the A and B regions belong to TFs related to stress response, proliferation, or cell cycle regulation, such as E2F, EGR1 and MYC (c-myc), the hypoxia response factor HIF1A, the metal response factor MTF1, and Kruppel-like factors (KLFs) (Fig. 4B; Supplemental Table S6; Lichtlen and Schaffner 2001; Blais and Dynlacht 2004; Liao et al. 2004; Eilers and Eisenman 2008; Majumdar et al. 2010; McConnell and Yang 2010; Zwang et al. 2011). The C regions, on the other hand, contain matches to PWMs of liver-specific TFs—such as FOXA2, ONECUT1 (previously known as HNF6), and HNF1A (Costa et al. 2003; Guillaumond et al. 2010)—or factors found to be associated with a quiescent state (G0), e.g., MAFB and FOXOA3A (Gree and Brunet 2005; Sarazin et al. 2009). Additionally, several different SOX motifs were identified as strongly overrepresented within this cluster (Fig. 4B; Supplemental Table S6).

Overall, these observations suggest significant differences in the regulatory mechanics of the A/B clusters versus the C cluster of putative cis-regulatory elements. Moreover, the data point to sets of specific, dynamic TF binding partners of CEBPs, defining a temporal regulatory code.

Multi-level support of temporal cis-regulatory code predictions

To assess if the TFs predicted to interact with A or C cluster regions are present in the liver, we examined their expression levels (POL2 gene body reads) against all genes (Fig. 4C; Supplemental Table S7). This revealed that the majority are highly expressed, being represented by either direct matches or top protein family members (e.g., Klf10, -13, -15 or Sox13, -15, -18). Strikingly, Cebp, Cebp, and Maf are at the top of the list (rank positions of 107, 21, and 61, respectively).

We further examined the indicated differences of the three clusters by interrogating their genomic position relative to the nearest TSS (Fig. 4D). The three clusters clearly diverge in this respect, as the C cluster regions are distinct from A (or B) regions by showing no proximity to TSSs. Several studies have shown differences in preferred genomic positioning of TFs (e.g., Gerstein et al. 2012). This suggests, based solely on the genomic difference in position, that C cluster regions recruit other cofactoring factors than the A (or B) regions. Specifically, E2Fs and MYC have been shown to bind at positions similar to the A regions, supporting the relevance of enrichment for E2F and MYC cognate sequences in the A cluster (e.g., Eilers and Eisenman 2008).

Next, we took advantage of published mouse ChIP-seq data sets (Chen et al. 2008; Hoffman et al. 2010; Laudadio et al. 2012) to test if a panel of factors bound to A versus C regions as predicted by our computational analysis. We find our analysis to be supported as the E2F1, KLF4, and MYC factor bound regions overlap significantly more with A regions, while FOXA2 and ONECUT1 preferentially bind C regions (Fig. 4E; Supplemental Methods).

Finally, we tested co-occupancy of CEBPs and several putative cofactors by performing sequential ChIP (Methods). This shows that the factor E2F3 binds to the two A regions Slbp and Cbxs simultaneously with CEBP factors, but none of the examined C regions, in accordance with our predictions. In contrast, the C cluster-associated factors ONECUT1, HNF1A, and MAFB interact with several C regions also occupied by CEBPs (Fig. 4F). Noticeably, a number of target regions are shared among the three C region factors.

Two modes of transcriptional regulation by EGR1

EGR1 has been shown to be essential for a timely regenerative response in the mouse liver (Liao et al. 2004) and represents a classical "immediate early TF" which is up-regulated upon a growth stimulus (for review, see Thiel and Cibelli 2002). Recently, it was found to be part of a growth-signal discriminatory circuit together with TP53 (Zwang et al. 2011). Canonical EGR1 cognate sequences are among the most highly enriched in the "cell cycle" or "proliferation" set of CEBP bound regulatory regions (the A cluster), together with E2F and MYC motifs (Supplemental Table S6).
To investigate the role of EGR1 in liver regeneration in vivo, we performed ChIP-seq experiments with EGR1-specific antibodies at the 24- and 36-h time points. Unexpectedly, we observed a preference of EGR1 for the C set over the A set regions (Fig. 5A). Accordingly, A and EGR1 peak summits were generally positioned further apart than C and EGR1 peaks (Fig. 5B). Moreover, a GO analysis of putative target genes suggested that EGR1 may be involved in all aspects of liver regeneration, as acute phase genes, metabolic genes, and cell cycle genes are found proximal to EGR1 bound regions (Supplemental Table S9; data not shown).

Figure 4. The CEBP temporal binding patterns display two sets of cis-regulatory code. (A) Representation of motif frequencies in each temporal cluster. Hierarchical clustering based on Z-scores of 210 position weight matrices (PWMs). Each row indicates a specific motif (PWM), while columns represent each binding cluster (A, B, or C). Color scale indicates representation relative to background set (overrepresented, dark blue; underrepresented, dark gray; no difference, white). Z-score scale is cut off at 6.40 for clarity. Example PWMs are shown in black text. (B) Selected transcription factors with clear binding sequence overrepresentation in A and B clusters (upper six), all clusters (middle three), or only the C cluster (lower seven). Clusters are denoted by color. (C) Expression level frequency distribution (POL2 gene body read coverage) of all genes (reads per kilobase, sum of eight time points, normalized, above 0.5) with candida transcription factors indicated. (D) Distances from A, B, and C cluster peak summits or random positions to the most proximal RefSeq transcription start site (TSS). (E) External ChIP-seq data peak regions (yellow bars) showing overlaps with CEBP A and C cluster regions (white numbers indicate proportions) and P-value of hypergeometric test (Fisher’s one-tailed) of overlap similarity. (F) Sequential ChIP (reChIP) assessing co-occupancy at A and C cluster regions. Anti-CEBP was used as first-round antibody (recognizing both CEBPA and CEBPB), with second-round IgG or antibody against indicated TFs. Enrichments are normalized to IgG levels. N = 2–5. Error bars, SEM. (*) P < 0.05; (**) P < 0.01, t-test versus IgG enrichments. For gene names, see Supplemental Table S10.
By manual inspection of tightly overlapping EGR1 and C regions bound by CEBPs, we found many peaks positioned precisely at a CEBP sequence but lacking EGR1 cognate sequence (Fig. 5C, bottom panels). Conversely, EGR1 peaks that possessed EGR1 cognate sequences did not overlap exactly with the CEBP peaks, and these peaks were consistently associated with A cluster regions (Fig. 5C, top panels). Sequences of target regions with positions of CEBP and EGR1 PWM hits and logos of used PWMs can be found in Supplemental Figure S5.

To test if EGR1 depends on CEBPs for interacting with DNA at C regions (lacking EGR1 sequences), we performed EGR1-specific ChIP with liver tissue from wild-type and Cebpb knockout mice. Our results show that EGR1 binding depends significantly more on CEBPB at C regions than at A regions, suggesting that EGR1 could bind DNA in two modes, indirectly (via CEBPs) at C regions and directly at A regions (Fig. 5D).

Differential sets of transcriptional regulators targeted by the three clusters of cis-regulatory regions

To further explore the transcriptional network of liver regeneration, we focused on the 120 most highly expressed CEBP target genes annotated with the “transcriptional regulator, DNA-dependent” GO term (Supplemental Table S8). These represent as targets of either A, B, or C cluster regions or of any combination hereof (Fig. 6A).

We find that the large majority of regulators (110 of 120) are targeted by either the A or B cluster regions, while the C cluster targets much fewer (38), suggesting a massive shift at the transcriptional network level upon transition from a quiescent to a regenerative state. Moreover, the injury response clusters (A/B) also impinge on many of the regulators targeted at the quiescent stage (29 of 38 bound by the C cluster), hinting at tight integration of the genetic programs at successive stages of the complex regenerative process.

A number of genes are associated with a high number of CEBP bound regions belonging to two or three clusters (Fig. 6A), suggesting a complex cis-regulatory structure. This may suggest that they are important components, or “hubs,” of the network. Prominent examples are the genes for the bHLH DNA-binding dominant repressor, ID2, as well as the cell cycle regulators JUN and EGR1 (Fig. 6A,B). Also Cebpa and Cebpb themselves are targets of multiple bound regions, as are Mafb, Hnf4a, and several Klf genes (Fig. 6A).

Several genes of regulatory factors are targeted by regions of which they may be coregulators based on the binding motif analysis (Figs. 4B, 6A). This can be interpreted as network feed-forward
or auto-regulatory loops (Fig. 6C). Examples include genes encoding E2Fs and KLFs for the A and B cluster regions and FOX factors for the C regions. Hnf4a is a shared target of regions from all three clusters, and the HNF4A binding sequence is also enriched in all clusters (Fig. 6A; Supplemental Tables S6, S7). These regulatory loops may be involved in adding robustness to the network as suggested for other transcriptional hierarchies (Lee et al. 2002).

We detect the cholesterol Nr1h3 (also known as Lxr) and bile acid Nr1h4 (also known as Fxr) metabolism regulatory genes to be A and C cluster targets, respectively (Fig. 6A; Supplemental Table S8; Kalaany and Mangelsdorf 2006). We also find the gene encoding RXRA, the obligate binding partner of both TFs, to be targeted by multiple CEBP bound elements. This suggests that these factors may be interconnected with the CEBPs in regulation of cholesterol and bile metabolism during liver regeneration.

Finally, we note that almost all core members of the circadian clock system (Zhang and Kay 2010) turn out to be putative CEBP targets (Fig. 6A; Supplemental Table S8), e.g., the core TF genes Clock and Bmal1 (also known as Arntl), as well as the downstream effectors encoded by Per1/2/3 and Cry1/2, and secondary component genes Rora, Rorc, Nr1d2 (also known as Rev-erb beta), Dbp, and Nrfl3 (also known as E4bp4). This points to a high level of cross-talk between circadian clock and regenerative response regulation.
Discussion

The multi-time point global quantification of TF occupancy enabled us to study the temporal dynamics of regulation in vivo. We find that the two central transcriptional regulators CEBPA and CEBPB interact with a large group of cis-regulatory elements in a highly dynamic manner through the progressive phases of liver regeneration. This group of genomic regions can be subdivided by CEBP binding dynamics into three clusters with distinctly enriched sets of sequence motifs, pointing to molecular mechanisms governing differential CEBP binding. Guided by this observation, we validate the predicted binding preference of several key co-binding TFs with sequential ChIP and published ChIP-seq data.

An important aspect of our study is the integration of observed TF binding dynamics and expression of putative target genes. We find that each of the three binding clusters is associated with distinct sets of regulated target genes, i.e., acute phase genes, metabolic/homeostatic genes, and cell cycle–related genes (Table 1). This demonstrates the possibility of linking dynamic TF occupancy patterns with the shifting phases of the regenerative program (Fig. 7A). Moreover, the concomitant resurge of CEBP binding to the C cluster regions and metabolic gene activity at 24 h provides evidence for a previously uncharacterized, early “homeostatic” phase of liver regeneration, taking place before the first round of replication.

The observed shifting phases of CEBP binding is closely temporally coupled to the dynamic ratio of CEBPA to CEBPB (Fig. 2C,D). One possible interpretation is that this ratio is involved in dictating enhancer occupancy by CEBP complexes. As such, a high ratio of CEBPA to CEBPB would lead to binding of C cluster cis-regulatory elements enhancing metabolic and repressing acute phase response genes. Conversely, a low ratio would direct binding of B cluster elements that repress metabolic or activate acute phase genes or of A cluster elements that activate cell cycle genes. This differential recruitment of CEBPs could be explained by interaction with distinct sets of TFs, depending on CEBP complex composition. This model is summarized in Figure 7B. Our observations and model contradict the conventional view that CEBPA levels are low while CEBPB levels are high throughout the regenerative process (Greenbaum et al. 1995, 1998), and emphasizes the requirement for a high degree of temporal resolution in studies of dynamic biological processes.

Many cognate sequences associated with TFs not previously known to be involved in liver regeneration were found to be overrepresented in the CEBP bound regions (Supplemental Table S6). In the C cluster regions, motifs of several SOX factors were abundant (Fig. 4B). Recent findings implicate SOX factors in hepatocyte differentiation and stem cell biology (Duan et al. 2010; Furuyama et al. 2011). Our data find Sox18, -13, and -15 to be highly expressed (Fig. 4C; Supplemental Table S7), but exactly which SOX factors cooperate with CEBPs in the liver remains to be determined. In the A cluster, we found overrepresented KLF factor motifs, and several highly expressed Klf6 (e.g., Klf10, -13, and -15) are targets of CEBP A cluster enhancers (Figs. 4B,C, 6A; Supplemental Table S8). Liver functions were recently identified for Klf15 and Klf10 (Guillaumond et al. 2010; Takashima et al. 2010). Moreover, KLF factors are central for self-renewal programs in embryonic stem cells (Takahashi and Yamanaka 2006; Jiang et al. 2008), which suggests that they may play similar roles in the self-renewal circuitry of fully differentiated hepatocytes, very likely as both binding partners and targets of CEBPs.

An unexpected observation of this study was that EGR1 binding preferentially colocalizes with the C cluster regions (Fig. 5A), which is almost depleted of EGR1 cognate sequence (Fig. 4B; Supplemental Table S6). Our data support a model of indirect or “assisted” EGR1 binding to explain this observation (Fig. 5C,D). Only one example of indirect EGR1 binding via CEBPB has been published (Zhang et al. 2003), but our genome-wide observations suggest that EGR1 targeting of metabolic gene promoters (C cluster regions) via CEBPs could be a general phenomenon in the liver. EGR1 has also been found to be important for interpretation of mitogenic signals (Zwang et al. 2011) and targets A cluster regions in this study (Fig. 5A,D; Supplemental Table S9). Hence, the observed CEBPA-to-CEBPB shift in the CEBP pool could regulate EGR1 recruitment to A or C cluster cis-regulatory elements and thus ensure a tight temporal control of EGR1 action in line with the progressive phases of liver regeneration.

In essence, liver regeneration upon partial hepatectomy is compensatory growth, as the cells proliferating are fully differentiated hepatocytes. Previous studies have shown that liver regeneration does not involve loss of hepatocyte differentiation state (e.g., Malato et al. 2011; for reviews, see Fausto et al. 2006; Michalopoulos 2007). This is consistent with our finding that many hepatocyte-specific genes, e.g., involved in metabolic functions, are up-regulated rather than down-regulated. (Fig. 3B; Table

**Figure 7.** Protein level ratios of CEBPA versus CEBPB may define a dynamic transcriptional switch. (A) Schematic diagram showing the divergent binding patterns and biological roles of the A, B, and C cluster cis-regulatory elements through the first cell cycle of liver regeneration. A previously uncharacterized early resurge of metabolic/homeostatic genes, associated with C cluster regions with binding peaking at 24 h, was observed. (B) Model of a transcriptional switch centered on the relative ratio of CEBPA and CEBPB determining the composition of the CEBP complex pool (homo- or heterodimers). Gene activating or repressive actions are indicated, as well as sets of transcription factors found to be associated with A or C putative enhancers. Metabolic genes are induced and acute phase genes repressed by binding in a CEBPA-high setting (C regions), while the opposite is true for the CEBPB-high setting (B regions). A-type regions are only targeted by CEBPs when the CEBPB form is abundant.
1. Only few examples are known of differentiated cells capable of hepatocyte-like “self-renewal” in the adult mammalian organism. It has recently been shown that the quiescence of naïve T cells is actively enforced by a balance between FOXP1 and FOXO1, allowing swift expansion upon external cues (Feng et al. 2011). Similarly, deficiency of just two TFs leads long-term, non-tumorigenic expansion capacity to mature macrophages (Aziz et al. 2009). These two factors are MAF (c-maf) and MAFK, both of which we find to be putative CEBP targets in the liver (Fig. 6A; Supplemental Table S8). MAFB is among the most highly expressed TFs in the quiescent condition and is strongly down-regulated from the 8-h time point (Supplemental Tables S4, S8). Furthermore, we find MAFB and MAFK binding sequences enriched in the quiescence/metabolism-associated C cluster of enhancers, predominantly bound by CEBPs at the 0- and 24-h time points. In the proliferation-associated A cluster, we find KLF and MYC sequences (Fig. 4B; Supplemental Table S6); KLF4 and MYC were demonstrated to be required for expansion of the differentiated macrophages (Aziz et al. 2009). Moreover, CEBPA is a key determinant of macrophage as well as liver cell differentiation (e.g., Feng et al. 2008). These observations may suggest that similar transcriptional networks centered on a CEBP/MAF axis are enforcing quiescence in both fully differentiated hepatocytes and macrophages. Understanding this circuitry has the potential to be of use in regenerative medicine as a step toward manipulation of hepatocytes or other fully differentiated cells for therapeutic purposes.

In conclusion, the present work shows how time-resolved analysis of two core TFs of liver regeneration can reveal specific aspects of the regulation operating at distinct phases during the process. In a broader sense, our approach of analyzing dynamic TF occupancy globally in vivo should be applicable to other experimental systems and holds promise to aid in the elucidation of complex transcriptional networks in higher vertebrates.

Methods

Mouse work

The partial hepatectomy was performed on wild-type mice (C57BL/6j, male, 7 wk old) by removing three of five liver lobes according to the method described earlier (Thoren et al. 2010). Mice were culled after varying amounts of time covering the regrowth phase of liver regeneration (0, 3, 8, 16, 24, 36, 48, and 168 h) (Fig. 1A,B). Livers from five individual mice for each time point were harvested, generating 40 samples; directly snap-frozen in liquid nitrogen; and stored at −80 °C. Animal experiments conformed to institutional as well as Danish national guidelines.

Chromatin immunoprecipitation

After thawing, tissue samples were homogenized by douncing (loose pestle, Wheaton 15-mL douncer) in cold PBS, and were cross-linked for 10 min in 1% formaldehyde using a rotator. Chromatin was fragmented by sonication (Bioruptor, Diagenode). After thawing, tissue samples were homogenized by douncing (loose pestle, Wheaton 15-mL douncer) in cold PBS, and were cross-linked for 10 min in 1% formaldehyde using a rotator. Chromatin was fragmented by sonication (Bioruptor, Diagenode). 

Mouse work

The partial hepatectomy was performed on wild-type mice (C57BL/6j, male, 7 wk old) by removing three of five liver lobes according to the method described earlier (Thoren et al. 2010). Mice were culled after varying amounts of time covering the regrowth phase of liver regeneration (0, 3, 8, 16, 24, 36, 48, and 168 h) (Fig. 1A,B). Livers from five individual mice for each time point were harvested, generating 40 samples; directly snap-frozen in liquid nitrogen; and stored at −80 °C. Animal experiments conformed to institutional as well as Danish national guidelines.

Chromatin immunoprecipitation

After thawing, tissue samples were homogenized by douncing (loose pestle, Wheaton 15-mL douncer) in cold PBS, and were cross-linked for 10 min in 1% formaldehyde using a rotator. Chromatin was fragmented by sonication (Bioruptor, Diagenode). The 40 samples were used in Chip-seq experiments with antibodies against CEBPA, CEBPB, EGR1 (Santa Cruz: sc-61, sc-150, sc-110x), RNA-POL2 subunit B1 antibody (AC-055-100, Diagenode), and Mock IgG (Sigma I8140). Chip was performed according to the method described earlier (Sandmann et al. 2006a). Sequential Chip was done according to the method described previously (Truax and Greer 2012), with the modification of cross-linking antibody–Protein A beads (www.neb.com). Santa Cruz antibodies as above were used for sequential CEBPA-CEBPB Chip (Supplemental Fig. S6). An in-house CEBP antibody (JSF1052) recognizing both CEBPA and CEBPB was used in the first round in sequential Chip with antibodies against ONECUT1 (H1N6), HNF1A, E2F3, MAFB, or IgG for the second round (Santa Cruz: sc-13505x, sc-8966x, sc-878x; Novus Bio: nb600-266). All sequential Chip enrichments were normalized to IgG ratios. Enrichment was validated by qPCR (ABI Prism 7000 or Roche Lightcyler 480), using ratios of a positive detector primer set versus a negative (Mannist, Chr_12_desert1 or Sf12). All primer sets are listed in Supplemental Table S10.

Clustering of CEBPA and CEBPB peaks

Consensus regions for CEBPA and CEBPB peaks for all time points were defined by merging overlapping regions between sets, requiring a called peak in at least one sample, producing a total of 87,049 consensus regions. Subsequently, coverage, defined as the maximal coverage level for each region, was determined using the normalized counts (Supplemental Material). Filtering was applied, requiring normalized coverage of at least one sample for any given region, reducing the set to 11,314. Hierarchical clustering was performed in MeV (http://www.tm4.org/mev/) (Saeed et al. 2006), and three predominant clusters were identified: A (3449 regions), B (2818), and C (3034). The remaining 2013 regions were excluded from further analysis. Coverage reads were summed for each time point/cluster, normalized, and used for displaying sum coverage tracks (Fig. 2B). Consensus region coverage data can be found in Supplemental Table S3.

Data access

The timeseries Chip-seq data generated for this work have been deposited in the NCBI Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) and are accessible through GEO Series accession number GSE42321.

Acknowledgments

We thank Claus Nerlov and Agnes Zay for Cebpb knockout mouse livers, Mie Poulsen and Bjerg Krog for expert help in performing the partial hepatectomy, and Thomas Sandmann, Federico De Masi, and members of the Pose laboratory for critical reading of the manuscript. This study was supported by grants from The Novo Foundation and the Danish Cancer Society.

References


References


Supplemental Figures

Supplemental figure 1: A graphical overview of splice classes detected by RAINMAN and spliceR. In addition, spliceR detects intron retention events, not shown here.
Supplemental figure 2: Strategy for predicting coding potential of transcripts based on exon-exon junction evidence. In A, short sequenced reads are mapped to an artificial genome consisting of all possible exon-exon combinations for each gene. In B and C, each adjoining exon from the longest supported refseq (or another repository) isoform is concatenated to the exons supported by the junction. In D, if an annotated CDS exists, the putative isoform is translated and the position of the STOP codon is stored. In E, if no CDS exists, translate in all frames, choosing a combination of longest supported polypeptide and most upstream ORF.
“Formula for breakthroughs in research: Take young researchers, put them together in virtual seclusion, give them an unprecedented degree of freedom and turn up the pressure by fostering competition.”

- JAMES WATSON, 1989