RESEARCH REPORT

The tissue-specific transcriptomic landscape of the mid-gestational mouse embryo

Martin Werber1, Lars Wittler1, Bernd Timmermann2, Phillip Grote1,.* and Bernhard G. Herrmann1,3,*

ABSTRACT
Differential gene expression is a prerequisite for the formation of multiple cell types from the fertilized egg during embryogenesis. Understanding the gene regulatory networks controlling cellular differentiation requires the identification of crucial differentially expressed control genes and, ideally, the determination of the complete transcriptomes of each individual cell type. Here, we have analyzed the transcriptomes of six major tissues dissected from mid-gestational (TS12) mouse embryos. Approximately one billion reads derived by RNA-seq analysis provided extended transcript lengths, novel first exons and alternative transcripts of known genes. We have identified 1375 genes showing tissue-specific expression, providing gene signatures for each of the six tissues. In addition, we have identified 1403 novel putative long noncoding RNA gene loci, 439 of which show differential expression. Our analysis provides the first complete transcriptome data for the mouse embryo. It offers a rich data source for the analysis of individual genes and gene regulatory networks controlling mid-gestational development.

KEY WORDS: RNA-seq, Embryogenesis, Long noncoding RNA, Mouse, Tissue-specific expression, Transcriptome

INTRODUCTION
The mouse is a well-established model organism for the investigation of mammalian development and human disease. The elucidation of the mouse genome sequence has provided access to any gene locus in the genome. Various approaches for gene function analysis have been described, making any gene amenable to loss-of-function studies. However, annotation of the transcriptome is still far from being finalized as the transcriptome differs from cell type to cell type and changes rapidly during development. Large-scale analyses of the mouse transcriptome have assessed temporal changes of 19,000 coding genes in the early mouse embryo [embryonic day (E) 6.5 to E9.0, corresponding to Theiler stage (TS) 9-14] using a custom microarray platform (Mitiku and Baker, 2007). In different approaches, single genes have been assayed for expression at various embryonic stages or in different tissues using in situ hybridization techniques (Diez-Roux et al., 2011; Gray et al., 2004). Whole-mount expression data have been obtained for 1912 genes of various functions by our group (http://mamep.molgen.mpg.de), and for 1561 transcription factors and TF-related genes (http://embry.jp) analyzed in midgestational embryos (E8.5 to E11.5; reviewed by Geffers et al., 2012). All of these datasets provide invaluable information pertaining to mainly protein-coding genes. However, none of these datasets has provided an accurate representation of the transcriptome of the mouse embryo. In particular, earlier studies using expression profiling did not cover the complete set of protein coding genes or alternative transcripts or noncoding RNA genes. The latter have come into focus in recent years, as noncoding genes are assumed to play important roles in gene regulation (for reviews, see Pauli et al., 2011; Rinn and Chang, 2012).

Among the highly diverse class of noncoding genes, long noncoding RNAs (lncRNAs) are thought to influence transcription by a wide range of mechanisms. For example, lncRNAs can interact with chromatin-modifying protein complexes involved in gene activation or silencing, and thus may be directly involved in epigenetic gene control (for a review, see Ulitsky and Bartel, 2013). Recent data generated from human cell lines by the ENCODE (Encyclopaedia of DNA Elements) Project Consortium provided a comprehensive view of the abundance of lncRNA genes in humans (Derrien et al., 2012; Dunham et al., 2012). However, despite recent functional analysis of a few lncRNA genes, not much is known about the functional role of lncRNAs in tissue differentiation and organogenesis in the mouse embryo (Grote et al., 2013; Li et al., 2013; Sauvageau et al., 2013).

RESULTS AND DISCUSSION
RNA deep sequencing of six tissues from the TS12 mouse embryo
We have determined the transcriptome of Theiler stage 12 (corresponding to E8.25, 3-6 somites) mouse embryos. To obtain a global view of the transcriptomic landscapes of different tissues at single-nucleotide resolution, we micro-dissected five prominent tissues comprising most of the embryo mass: the mesoderm caudal to the somites – predominantly containing presomitic and lateral mesoderm (caudal end mesoderm: CEM); the somites; the heart; the head (comprising the future brain, including the rhombencephalon and head mesoderm); and the presumptive spinal cord (PSC), as well as the remainder (carcass; extra-embryonic tissue was largely removed) from ten embryos (Fig. 1A,B). We generated cDNA from total RNA of each tissue and sequenced it on an Illumina HiSeq 2000 (RNA-seq).

We obtained 100 base paired-end reads and mapped approximately one billion reads in total against the latest mouse genome build mm10 (GRCm38) (supplementary material Table S1). The mapping results, coverage, and splice junction mappings are visualized and accessible via the UCSC genome browser (http://overview.molgen.mpg.de/ucsc.html; Fig. 1C). In each sample, the vast majority of reads mapped to coding exons (69.4-73.4%) and UTRs (19.6-21.1%), while the remaining reads mapped within introns (4.5-7.8%) or intergenic (1.3-2.0%) regions (supplementary material Table S2).

We calculated normalized transcript levels for each of the six samples on the basis of all annotated ENSEMBL transcripts with cufflinks, taking into account the default parameters for the alignment status (status=OK) and significance thresholds for FDR.
(5%; significant=yes). To validate the accuracy of normalized transcript levels obtained from RNA-seq data, we compared relative transcript levels of 22 housekeeping genes obtained by quantitative real-time PCR (qPCR) to normalized transcript levels calculated from RNA-seq data and found that they correlate strongly ($r=0.87$, p=1.12E-07; Fig. 1D). To confirm that the RNA-seq data accurately reflect the tissue specificity of gene expression, we compared the expression data derived by RNA-seq with whole-mount in situ hybridization data for ten marker genes (Fig. 1E,F). The tissue-specific expression was well reflected by RNA-seq data. In addition, we compared RNA-seq data to quantitative RNA measurements by qPCR on an independent, biological replicate (Fig. 1F,G). The values correlated well between the two methods. Taken together, our transcriptome dataset provides accurate quantitative information of transcript levels in each of the six tissues.

For a comparison of the number of genes expressed in each tissue, we focused on fully read-covered transcripts derived from genes annotated in ENSEMBL (cufflinks alignment status (status=OK)). Using this approach, we detected expression of 12,864 genes in the CEM, 13,384 in the head, 13,477 in the heart, 12,738 in the PSC, 12,674 genes in the somites and 14,158 in the carcass sample. Combining results over all sampled tissues, we detected significant transcription of 14,158 genes representing 38% of all 37,224 annotated genes.

We identified 1375 genes that are differentially expressed between the six tissue samples using cufflinks and cuffdiff with default settings (Fig. 2A and supplementary material Table S3). To extract co-expressed gene groups from this set, we applied hierarchical clustering to gene-normalized (Z-score-transformed) data for all tissues using Genesis (Sturn et al., 2002). We identified a specific gene signature for each of the six tissues. The CEM (52), head (85), PSC (37) and somites (68) revealed relatively small tissue-specific gene sets, while the heart (442) and the carcass sample (389) displayed significantly more specifically expressed genes (supplementary material Table S4). In addition to many key developmental genes, these clusters contain a number of genes that are not well annotated, and thus significantly extend the gene catalogues for studying fundamental developmental processes.

We visualized similarities in global expression and correlation between samples by applying Self Organizing Maps clustering and mosaic image graphing (GEDI-maps, Gene Expression Dynamics Inspector) (Eichler et al., 2003). The GEDI-maps illustrate that the differentially expressed gene sets are largely complementary between the six tissues, but also show significant overlap between head and PSC sharing neural tissue, between CEM and PSC sharing common progenitors, between CEM and one of its derivatives, the somites, and between the heart and carcass (Fig. 2B).

In order to characterize the functional significance of the tissue-specific gene signatures, we applied gene ontology enrichment analysis using DAVID for three categories: molecular function, biological process and cellular component (GOTERM_MF_FAT, GOTERM_BP_FAT, GOTERM_CC_FAT) (Huang et al., 2009). The top GO terms (cut off: FDR $\leq 0.05$), ranked by q-value, reflect the varied differentiation stages of the tissue samples (Fig. 2C).

**Alternative splicing and extended UTRs**

Alternative splicing greatly increases the variety of functionally distinct proteins in the cell. With the RNA-seq method employed here, the splicing of primary RNA can be observed by mapping of sequence read clusters to two genomic regions separated by intervening DNA fragments (introns). We compiled a unified database of splice junctions from all samples in our dataset using Tophat (Trapnell et al., 2009) and a second aligner, STAR (Dobin et al., 2013), and used this database with Tophat to prime additional junction alignments.

We evaluated the precision of splice junction mapping by comparing our mappings against annotated splice sites from public
reference annotations (PRA) (ENSEMBL v.69). The majority of the splice junctions (75% on average) mapped exactly to the PRA positions, demonstrating that our approach reliably detects splice junctions (Fig. 3A). One quarter of the splice junctions fall into one of the following two groups: partially in agreement with PRA (partially novel), whereby either splice donor or splice acceptor site are different (17.4%); and completely novel (6.7%).

We determined the transcript isoforms from the six tissue samples and assigned them to seven transcript classes (Fig. 3B). In each tissue, the number of novel transcript isoforms exceeds the number of transcripts, which perfectly match PRA datasets. We classified alternative splice junctions and exon coverage with respect to the gene structure found in the PRA, into the following classes with the software ASprofile: exon skipping (SKIP), alternative transcript start or termination site (TSS, TTS), and alternative exon (AE) (Florea et al., 2013). The vast majority of novel events comprise alternative or extended first and last exons, most likely reflecting frequent tissue-specific differences in the choice of TSS and TTS [Fig. 3C and http://dx.doi.org/10.6084/m9.figshare.993872 (see also http://overview.molgen.mpg.de/Table_alternative_exon_usage.zip)]. To evaluate whether the novel splicing events giving rise to alternative transcripts are embryo specific (3.9% also found in ENCODE). These data illustrate the excessive use of alternative promoters and alternative exons in embryonic tissue.

Novel long noncoding transcripts
We extended our analysis to include the class of long noncoding RNAs (lncRNAs). Out of 2240 lncRNA transcripts annotated in ENSEMBL, 1378 (61.5%) show significant expression in at least one of the six tissues, and 225 are differentially expressed (using cuffdiff; FDR 5%; Fig. 4B and supplementary material Table S5).

Based on comparison to the ENSEMBL (v.69) reference annotation, we additionally identified 4196 novel intergenic transcripts that could be condensed to 2438 loci with a minimum length of 200 bases. We characterized this heterogeneous set of transcripts in several steps. We identified 1143 (27.2%) transcripts with significant similarity to a protein using BLASTx against the NCBI non-redundant protein database (e-value threshold: e-30) (Fig. 4A). From the remaining 3053 transcripts we identified 449 (10.7%) as putative miRNAs in miRBASE (Griffiths-Jones, 2010). A comparison against RFAM revealed 474 (11.3%) transcripts with similarity to other types of small noncoding RNAs. Thus, excluding known classes of coding and noncoding RNA, the remaining 2130 (50.7%) transcripts represent high-quality putative novel lncRNAs. They comprise 2082 multi-exon transcripts belonging to 1403 expressed gene loci. Among these, 439 are differentially expressed between the six embryonic tissues (Fig. 4B and supplementary material Table S5). Thus, through this analysis we
substantially extend the current annotation of lncRNAs in the mouse and provide the first set of differentially expressed lncRNAs for the mouse embryo.

Divergent coding-noncoding gene pairs (CNPs)

Visual inspection of putative novel lncRNAs in our transcriptome dataset suggested that many are transcribed in close proximity to protein-coding genes. Similar observations have already been described for mouse ES cells and human datasets (Khalil et al., 2009; Sigova et al., 2013). We identified 392 divergent coding-noncoding gene pairs (supplementary material Table S6). The majority are very close to each other, within less than 200 bp (Fig. 4C).

Divergent CNPs are likely to share the same promoter. We compiled the expression patterns for each gene pair across all samples and calculated the Spearman rank correlation coefficient as a measure of similarity. Using $r \geq 0.5$ as a cut-off, we identified a set of 150 CNPs, which show significantly correlated expression across tissue samples (Fig. 4D).

Alternative TSS and novel transcripts supported by H3K4me3 marks

Histone modifications reflect chromatin states and transcriptional activity of genomic regions. In particular, the tri-methylation of lysine 4 on histone H3 (H3K4me3) has been shown to reliably mark the transcriptional start sites of active and poised genes (Guenther et al., 2007). We reasoned that this mark would also provide support for low expression and novel transcripts, alternative first exons and novel long noncoding RNAs. Therefore, we determined the H3K4me3 histone modification marks on a genome-wide scale in whole TS12 embryos using ChIP-seq.

We mapped 13.3 million reads using bowtie and determined the regions significantly enriched for H3K4me3 enrichment ($\gamma$-axis of RNA-seq track is 1000x coverage). We detected 29,262 such regions ranging in size between 50 bp (lower detection threshold) and 3233 bp (mean=1004 bp; $P$-value cut-off 1.00E-05). Of these, 14,686 (50.2%) colocalized with TSS annotated in ENSEMBL (v.69), confirming that they reliably predict TSS on a genome-wide scale (supplementary material Table S7; data not shown). We found 1241 regions marking a distant (median=6.3 kb) novel TSS. An example is shown on Fig. 3D. Another 6442 H3K4me3 regions map inside gene bodies. Of these, 1703 mark putative enhancers identified in multiple mouse cell lines and tissues (Shen et al., 2012). We found 6893 H3K4me3 enrichment peaks outside of transcribed regions. Of those, 1187 overlapped with enhancers and 5706 regions could not be related to TSS, gene bodies or enhancers.

The H3K4me3 marks support the localization of the TSS of 888 putative novel lncRNAs, as illustrated by an example on Fig. 4E. Our datasets significantly extend the number of lncRNA loci expressed in the mouse embryo and, in particular, the number of lncRNAs showing tissue-specific or differential expression. These data also aid in the identification of mouse orthologues of human lncRNAs involved in disease.

MATERIALS AND METHODS

Embryo preparation, RNA isolation and in situ hybridization

Mouse embryos of the C57BL/6J strain were dissected at E8.25 from their mothers and staged according to Theiler (Theiler, 1989). Embryos at TS12 (3- to 6-somite stage) were treated with Dispase (BD Biosciences) for 5 min at room temperature to facilitate the separation of the five tissues. Microdissection was performed in PBS using a flame-sharpened tungsten needle under a stereomicroscope. Tissue samples were pooled from ten embryos, which were lyed (RLT +2-mercaptoethanol, Qiagen), and RNA was isolated using the RNeasy Micro Kit (Qiagen) according to the manufacturer’s protocol. The protocol for the whole-mount in situ hybridization and the probe sequence can be found on the MAMEP website (http://mamep.molgen.mpg.de/protocol.html).
Whole embryos were stage matched (see above) and ten embryos pooled for H3K4me3 analysis (Abcam, ab8580) ChIP, which was performed as described previously (Grote et al., 2013). Libraries were prepared according to the ChIP-Seq DNA Sample Prep Kit (Illumina). 13.3 million 36-base single reads were mapped with bowtie against the mouse mm10 (GRCm38) reference assembly. Significantly enriched regions for H3K4me3 were determined with Model-based analysis of ChIP-seq (MACS) (Zhang et al., 2008). Overlaps with TSS and putative enhancer regions were calculated using BEDTools (Quinlan and Hall, 2010).

RNA-seq library preparation and sequencing
Sequencing libraries were prepared from total RNA (300 ng) according to the TruSeq RNA Sample Prep (Ver. A) (Illumina) library protocol and analyzed on a HiSeq 2000 genome analyzer with 100 base paired-end read setting. The paired-end libraries had an average fragment size of 310 bp; the paired-end sequences were 100 and 98 bases in length.

Transcriptome assembly and differential expression analysis with RNA-seq
The reads were mapped against the mouse mm10 (GRCm38) reference assembly using two different alignment programs, Tophat (with option –b2–very-sensitive) (Trapnell et al., 2009) and STAR (Dobin et al., 2013), using ENSEMBL reference annotation for guided splice junction mapping. To exploit advantages of reference guided assembly without compromising the detection of novel transcripts, the mapped reads were assembled into transcripts using cufflinks with and without the RABT-guided assembly option (Trapnell et al., 2009). The resulting tissue-specific transcriptomes were merged using cuffcompare and custom PERL scripts re-assigning the gene identifiers. For each transcriptome the expression strength of individual transcripts and genes, and differential expression between tissues was estimated with cuffdiff.

Quantification of housekeeping genes by real-time PCR
For the biological replicates, five embryos were dissected and RNA isolated as described above. The following genes were used as housekeeping genes: Prmt1, Srm, Csnk2b, Ube2m, Cd81, Atp6v0c, Rab1, Pgd, Rtn4, Erp29, Psma7, Gpx4, Lypla2, Rad9, B2m, Add1, Ext3, Usp11, Cd3eap, 1810031K17Rik (Cnddp1 – Mouse Genome Informatics), Kif1c and Mrsa8. Predesigned TaqMan probe and primer sets were purchased (Life Technologies) and quantitative PCR was carried out on a StepOnePlus system (Life Technologies) and quantitative PCR was carried out on a StepOnePlus system (Life Technologies), as suggested by the manufacturer.

Accession number
The accession number at the ENA for the RNA-seq and the ChIP-seq data is PRJEB4513.

Acknowledgements
We thank Tracie Pennimpede and Arica Beisaw for comments on the manuscript. We are grateful to Sonia Paturej for help with preparing the sequencing libraries.

Competing interests
The authors declare no competing financial interests.

Author contributions
P.G., M.W. and L.W. conceived the project, L.W. and P.G. carried out experiments, B.T. supervised the sequencing, and M.W. performed all bioinformatics analyses. B.G.H., P.G. and M.W. designed the figures and wrote the manuscript.

Funding
This work was supported by the Max-Planck Society.

Supplementary material
Supplementary material available online at http://dev.biologists.org/lookup/suppl/doi:10.1242/dev.105858/-/DC1

References

RESEARCH REPORT


