Composite transcriptome assembly of RNA-seq data in a sheep model for delayed bone healing

Marten Jäger1,2†, Claus-Eric Ott1†, Johannes Grünhagen1, Jochen Hecht2,3, Hanna Schell2,4, Stefan Mundlos1,2,3*, Georg N Duda2,4, Peter N Robinson1,2,3* and Jasmin Lienau2,4

Abstract

Background: The sheep is an important model organism for many types of medically relevant research, but molecular genetic experiments in the sheep have been limited by the lack of knowledge about ovine gene sequences.

Results: Prior to our study, mRNA sequences for only 1,556 partial or complete ovine genes were publicly available. Therefore, we developed a composite de novo transcriptome assembly method for next-generation sequence data to combine known ovine mRNA and EST sequences, mRNA sequences from mouse and cow, and sequences assembled de novo from short read RNA-Seq data into a composite reference transcriptome, and identified transcripts from over 12 thousand previously undescribed ovine genes. Gene expression analysis based on these data revealed substantially different expression profiles in standard versus delayed bone healing in an ovine tibial osteotomy model. Hundreds of transcripts were differentially expressed between standard and delayed healing and between the time points of the standard and delayed healing groups. We used the sheep sequences to design quantitative RT-PCR assays with which we validated the differential expression of 26 genes that had been identified by RNA-seq analysis. A number of clusters of characteristic expression profiles could be identified, some of which showed striking differences between the standard and delayed healing groups. Gene Ontology (GO) analysis showed that the differentially expressed genes were enriched in terms including extracellular matrix, cartilage development, contractile fiber, and chemokine activity.

Conclusions: Our results provide a first atlas of gene expression profiles and differentially expressed genes in standard and delayed bone healing in a large-animal model and provide a number of clues as to the shifts in gene expression that underlie delayed bone healing. In the course of our study, we identified transcripts of 13,987 ovine genes, including 12,431 genes for which no sequence information was previously available. This information will provide a basis for future molecular research involving the sheep as a model organism.

Background

The sheep is an established model organism for medically relevant research in cardiology [1,2], reproductive medicine [3], respiratory medicine [4,5] and many other fields. The sheep is particularly important in the field of orthopedics, because the dimensions of ovine long bones allow the use of implants designed for application in humans, and the bone mineral composition as well as the metabolic and remodeling rates are similar to those in humans [6-10].

Although bone usually heals spontaneously, failures in bone healing remain an important medical and research challenge. Bone healing is a highly complex regenerative process that is directed by a series of cytokines and growth factors and leads to restoration of skeletal integrity. Despite advances in the field of orthopedic research, our understanding of the molecular mechanisms involved in standard and impaired healing is still limited, and delayed unions and non-unions are still major clinical problems.

A tibial osteotomy healing model in sheep is well established in our laboratory, and its biological and
mechanical characteristics have been previously published [11-17]. The model has been used to simulate standard healing by means of the application of a rigid external fixator following osteotomy as well as delayed healing by application of a rotationally unstable external fixator.

To date, molecular research in the sheep has been hampered by the paucity of information about the sheep genome and gene sequences. Therefore, a systematic comparison of the expression characteristics of the transcriptome between a standard and impaired healing osteotomy has not yet been performed. Massively parallel DNA sequencing platforms, widely referred to as "next-generation sequencing" (NGS), are changing the playing field in biomedical research by enabling the comprehensive and relatively inexpensive analysis of genomes and transcriptomes [18-22]. NGS technologies have opened the door to genome scale experiments in organisms that lack comprehensive genome or transcriptome information, making it possible to assemble novel transcripts and identify differential regulation in a single experiment [23,24].

We have previously used EST sequencing to investigate genes differentially expressed in the course of standard bone healing in sheep [25]. In this study, we compare the gene expression profiles of standard and delayed bone healing in the sheep by means of next-generation sequencing and computational analysis of the sheep transcriptome to identify transcripts of 13,987 ovine genes, for 12,431 of which no mRNA sequence was previously available. We identified characteristic clusters of gene expression including several with striking differences between standard and delayed healing, some of which have known roles in the extracellular matrix and skeletal development.

Results
Sheep osteotomy model
A total of 63 Merino mix sheep (2.5 years old) with a mean weight of 72 kg (± 10 kg) received a standardized mid-shaft tibial osteotomy (3 mm gap) stabilized with a monolateral external fixator. Group I (n = 31) received a rigid external fixator, which has been previously shown to support standard healing [15,16]. Group II (n = 32) was treated with a mechanically critical external fixator, which allowed free rotation through highly stiff stainless-steel tube set in two tapered roller bearings, thereby producing a distinct delayed healing [17] (Figure 1A). The standard healing group stabilized with the rigid external fixator showed uneventful healing with complete bony bridging by day 42 or 63. In contrast, the delayed healing group stabilized with the rotationally unstable external fixator did not reach union by day 63 (Figure 1B). Calcified histology at day 14 and 21 revealed newly formed bone in the periosteal callus that was covered by a layer of osteoid seam in the control group. Furthermore, in regions of intramembranous bone formation periosteally, mineralization of deposited osteoid was visible (Figure 1C). In contrast, in the delayed healing group, an increased amount of osteoid covering the newly formed bone and less mineralization of deposited osteoid was found at day 14 (Figure 1D).

Massively parallel sequencing
To date, the use of the sheep to investigate the genetic correlates of bone healing has been limited owing to the lack of ovine genome and transcriptome sequences. The purpose of the current project was therefore to generate an initial sheep transcriptome using next-generation technologies and to perform an exploratory analysis of differential gene expression between standard and delayed healing in the sheep.

We developed a bioinformatics workflow (Figure 2) that would take advantage of ovine sequences where possible and otherwise use homologous sequences from mouse and cow for mapping short reads. We analyzed one flowcell with eight lanes of samples at different time points of standard and delayed healing. Pooled samples were obtained from 5-6 animals each at four different time points (7, 11, 14, 21 days). Each sample was run in a single lane of the flowcell, resulting in 18-27 million 76 bp reads per lane corresponding to 9-14 million unique reads per lane with a total of 177 million reads including 69 million unique reads (Table 1).

De novo transcriptome assembly
Prior to de novo assembly the reads from all lanes of the flowcell were pooled and duplicate reads were removed to generate a list with each different sequence represented only once. The best mean read score for a single unique read was preserved. The list of unique reads was then trimmed by removing base calls from the 3’ end of the read with Phred (quality) score [26] of 2 or less. A total of 4,599 ovine mRNA genbank entries (corresponding to 1,556 genes) and 325,596 ovine ESTs were downloaded from the NCBI database using the E-Utils [27]. NCBI reference sequences containing one or more ambiguous bases (‘N’) were removed from further analysis, such that 4,363 mRNA and 294,896 EST sequence files were available for the assembly. De novo assembly using the mRNAs, the ESTs, and the uniquified and trimmed reads was then performed using Velvet [28]. Velvet output a total of 830,469 contigs with an average length of 134 bp. Oases is a de novo transcriptome assembler designed to produce extended contigs from short read sequencing technologies in the absence of any genomic assembly. It clusters the contigs from a preliminary assembly by Velvet into small groups called...
loci and uses a de Bruijn graph-based algorithm to construct transcript isoforms [29]. The contigs produced by Velvet were postprocessed using Oases yielding 85,555 loci (gene predictions) comprising a total of 117,594 extended contigs with an average length of 1,374 bp. 56,298 loci exceeded the minimum length threshold of 150 bp and were included in further analysis (Table 2). For each of the 56,298 loci, the contig with the highest Oases confidence score was chosen for further analysis. The average length of these contigs was 956 bp.

BLAST was used to identify the gene models by comparing the contig sequences to available sheep, cow, and mouse mRNA sequences. The resulting hits (limited to only the one best matching sequence per query sequence) were filtered for matches with significant E-value according to the species being compared and percentage of identical matches meeting the requirements shown in Table 3. In comparison to *Ovis aries*, more reference sequences are available for *Bos taurus* including various splice variants. Therefore, a higher identity cut-off was used to identify cow ortholog transcript models. 22,117 contigs were annotated to a gene symbol using this pipeline. The average length of the annotated contigs was 1,662 bp, and the longest single annotated contig was 21,746 bp long and annotated to dystonin (*DST*). There were 34,181 contigs which could not be assigned to any mouse, cow or sheep transcript, and these were excluded from further analysis.

**Read mapping**

The *de novo* transcriptome was then combined with all available 4,599 ovine mRNA sequence files as well as all 43,102 bovine mRNA sequence files from NCBI for mapping of the short reads with Bowtie [30]. The Bowtie mapping algorithm was used to map short reads with a seed length of 18 and a maximum of three allowed mismatches in the seed. In each lane each read was mapped to a single gene. On average, about 75% of the short reads could be mapped (Table 4). The use of *Bos taurus* transcripts for the mapping substantially increased the number of mappable reads (Figure 3).

**Evaluation of differential gene expression**

Each of the lanes corresponded to a pooled sample of 5-6 animals at one time point. As one lane per condition was available, differential expression between conditions was evaluated using the Audic-Claverie method [31,32] in order to enable exploratory analysis. The raw counts of the reads mapped as described above were used for the Audic-Claverie analysis. In addition, RPKM analysis was used to estimate the fold change. Gene expression was compared between the time point 7 days
and those at 11, 14, and 21 days for both standard and delayed healing. In addition, each of the individual time points was compared between the standard and the delayed healing groups. A gene was considered to be differentially expressed if the Audic-Claverie $p$-value was $< 10^{-15}$ and the fold change of the normalized (RPKM) expression values was at least 2 in either direction (see also Additional file 1: Supplemental Figure S2). There were 5 genes differentially expressed between standard and delayed healing at day 7, 173 at day 11, 59 at day
Table 1 Read counts

| lane | reads | unique reads |
|------|-------|-------------|
| 1    | 22,145,090 | 11,682,018 |
| 2    | 23,356,043 | 12,420,275 |
| 3    | 27,617,415 | 14,663,522 |
| 4    | 20,234,612 | 10,168,941 |
| 5    | 24,793,840 | 12,944,693 |
| 6    | 18,894,344 | 9,796,924  |
| 7    | 18,788,777 | 10,383,698 |
| 8    | 21,722,529 | 12,302,346 |
| all  | 177,552,650| 69,019,744 |

Lanes 1, 3, 5, and 7 correspond to standard bone healing on days 7, 11, 14, and 21, and lanes 2, 4, 6, and 8 to delayed healing on days 7, 11, 14, and 21.

14, and 109 at day 21. Within the time course of standard healing, there were 177 genes differentially expressed between day 11 and 7, 265 between day 14 and 7, and 318 between day 21 and 7. Within the time course of delayed healing, there were 136 genes differentially expressed between day 11 and 7, 139 between day 14 and 7, and 259 between day 21 and 7. This corresponded to a total of 884 distinct genes showing differential expression in at least one comparison. Q-PCR analysis of the pooled samples for 26 selected genes confirmed the analysis of the RPKM values (Additional file 1: Supplemental Tables S1-S4 and Figure S3).

Clustering and GO analysis
Hierarchical clustering was performed to identify groups of genes with similar expression profiles. A total of 13 clusters were identified by visual inspection. Figure 4 shows the expression profile for standard and delayed healing groups. The heat plot on the left displays the expression patterns for all genes. Separate plots are shown for each cluster with the average and standard deviations of the RPKM expression values. There were clusters with relatively minor differences between the standard and delayed healing groups, and several others with marked differences between the groups. Model-based gene set (MGSA) Gene Ontology analysis [33] was performed for each of the clusters, and up to three GO terms with a marginal probability of at least 50% and the highest number of annotated genes are shown (Figure 4). Additionally, MGSA was performed for the entire set of 884 differentially expressed genes (Table 5).

The GO terms include terms such as extracellular matrix and chemokine activity with well known roles in skeletal biology and bone healing. 24 of the differentially expressed genes were annotated to striated muscle contraction, and 45 to contractile fiber. The role of α-smooth muscle actin fibroblasts in the contraction of skin wounds is well known; smooth-muscle actin expressing connective tissue cells have also been shown to take part in fracture healing [34], and our results could be a reflection of this phenomenon. Therefore, we investigated all differentially expressed genes annotated to contractile fiber (n = 45). Most of these genes are localized in clusters E and G (Figure 4) and were more highly expressed in standard healing at day 11 and 14, and more highly expressed in the delayed healing group at day 21, consistent with a delay in the regulation of these genes (Figure 5A). Several of the genes annotated to striated muscle contraction and contractile fiber overlap with genes annotated to calcium ion binding (n = 59). These genes were also significantly more highly expressed in standard healing at day 11 and day 14 (Figure 5B) where clear signs of mineralization were shown by histology (Figure 1C). In contrast, differentially

Table 2 Results of the de novo assembly using Velvet and Oases

| Source | Description | number |
|--------|-------------|--------|
| Velvet | Contigs     | 830,469|
|        | Average contig length (bp) | 134 |
|        | Oases       | 117,594|
|        | Extended contigs | 134 |
|        | loci        | 85,555 |
|        | loci > 150 bp | 56,298 |
|        | loci (quality filtered + annotated) | 22,117 |
|        | Number of unique mappable sheep genes | 13,546 |
|        | Average annotated contig length (bp) | 1,662 |
| NCBI   | ovine genbank mRNA entries | 4,599 |
|        | ovine genes with known mRNA sequence | 1,556 |
|        | ovine EST entries | 325,596 |
|        | bovine genbank mRNA entries | 43,102 |
|        | bovine genes with known mRNA sequence | 16,052 |
| Assembly | Total sheep genes with known mRNA sequence | 13,987 |

Following all quality control and filtering steps, there were a total of 263,325 mappable genes. For 21,865 of these genes, positive counts were detected in all 8 lanes, and these genes were used for the further analysis of differential expression. A gene with known mRNA sequence refers to a gene with a gene symbol for which at least one mRNA sequence was found. Note that mRNA sequence entries assigned to hypothetical genes were not included, and that multiple sequence entries were found for some genes. A unique mappable sheep gene refers to a set of one or more Oases loci that could be mapped to a unique gene symbol via BLASTing to sheep, cow, or mouse sequences. The total number of sheep genes with known RNA sequence is derived from the union of de novo assembled genes and previously sequenced genes, 1,115 of which overlapped.

Table 3 Threshold settings used for homolog mapping using Blastx/n

| homolog mapping | identical matches | expect value |
|-----------------|-------------------|--------------|
| Blastx (mouse)  | 80%               | 1-50         |
| Blastn (sheep)  | 90%               | 1-50         |
| Blastn (cow)    | 97%               | 1-50         |

Only Blast matches fulfilling both conditions were considered as true orthologs.

Table 3 Threshold settings used for homolog mapping using Blastx/n

| homolog mapping | identical matches | expect value |
|-----------------|-------------------|--------------|
| Blastx (mouse)  | 80%               | 1-50         |
| Blastn (sheep)  | 90%               | 1-50         |
| Blastn (cow)    | 97%               | 1-50         |
Table 4 Read counts for the 8 lanes as mapped with Bowtie

| lane | reads        | mapped reads | %    | annot | unannot | de novo | Cow (%) | de novo | annnot | unannot |
|------|--------------|--------------|------|-------|---------|---------|---------|---------|--------|---------|
| 1    | 22,145,090   | 17,127,532   | 77.3 | 11.8  | 0.1     | 51.2    | 35.4    | 1.5     |        |         |
| 2    | 23,356,043   | 17,964,339   | 76.9 | 11.5  | 0.1     | 51.4    | 35.5    | 1.5     |        |         |
| 3    | 27,617,415   | 17,964,339   | 76.3 | 10.8  | 0.4     | 48.7    | 38.6    | 1.5     |        |         |
| 4    | 20,234,612   | 14,980,730   | 74.0 | 10.5  | 0.1     | 47.6    | 40.4    | 1.3     |        |         |
| 5    | 24,793,840   | 18,367,757   | 74.1 | 12.1  | 0.2     | 48.4    | 37.6    | 1.7     |        |         |
| 6    | 18,894,344   | 13,810,552   | 73.1 | 11.0  | 0.1     | 48.6    | 38.7    | 1.5     |        |         |
| 7    | 18,788,777   | 13,599,338   | 73.1 | 12.1  | 0.2     | 49.6    | 36.7    | 1.4     |        |         |
| 8    | 21,722,529   | 14,276,174   | 65.7 | 11.8  | 0.2     | 48.2    | 38.2    | 1.5     |        |         |

The sections entitled Sheep (Ovis aries) and Cow (Bos taurus) indicate the percentage of the corresponding target sequences for the mapped reads for each of the eight lanes. See also Figure 3.

**Figure 3** Composite reference transcriptome assembly and assignment of short read sequences

**A** De novo transcriptome assembly was performed with Velvet and Oases on the basis of mRNA sequence entries from GenBank for sheep and cow as well as the de novo assembled contigs from this study. All sequences were used as targets to map the short reads. The pie chart shows the relative proportions of the sequence entries from each of the sources used for the mapping (OAR = Ovis aries, BTA = Bos taurus). **B** The distribution of the targets that were matched by bowtie for short read mapping are shown. Most reads mapped to the de novo transcriptome assembly, but it was possible to map a substantial number of additional reads by use of the Ovis aries and Bos taurus mRNA sequences. Table 4 displays the exact counts for each lane.
expressed genes that were annotated to cartilage development \((n = 10)\) showed a characteristic upregulation in the latter two time points of delayed healing (Figure 5C). These genes are a subgroup of genes that are annotated to the GO term skeletal system development, which was significant in the MGSA analysis of cluster F (Figure 4).

Especially during the early phase of bone healing, fundamental processes such as inflammation are strongly influenced by the mechanical conditions under which bone healing takes place [35,36]. In our experiment, genes annotated to response to wounding \((n = 84)\) were consistently more highly expressed in the standard
healing group (Figure 5D). This term is a parent term of inflammatory response (cluster I) and related to defense response (cluster K and L). In agreement with the biological observation that the inflammatory phase of bone healing takes place in the first few days [37] most genes annotated to these terms showed the highest expression at day 7 both in standard and in delayed healing. In a similar fashion genes annotated to cell division (n = 29) showed the highest expression levels at day 7 (cluster I). These genes were consistently more highly expressed in the delayed healing group at the later time points (Figure 5E). Genes involved in complement activation were significantly overrepresented in all 884 differentially expressed genes (Table 5) and were from day 11 on significantly more highly expressed in the standard healing group (Figure 5F).

### Discussion

Most bony injuries heal without problems, but there are several conditions under which enhancement of the repair process would be of great benefit to ensure the rapid restoration of skeletal function. Animal models are essential for investigating the different molecular processes underlying bone healing. Several studies have identified molecular differences between standard healing or non-unions [38-40], but these studies have been performed in small-animal models. Currently, investigative tools are still limited for analyses in large-animal models such as sheep and relatively little research has been performed into sheep genetics. However, the sheep model is critical for medical applications because the size of the bone, the loading, and the time to healing are comparable to human fracture healing [6].

A virtual sheep genome has been constructed by mapping ovine contigs obtained by 454 sequencing onto bovine sequences that had been rearranged in sheep order [41-44]. At present, however, only low coverage genomic data is available. The sequences presented in this work will represent a valuable and complementary resource to current efforts to sequence the sheep genome.

In previous studies, we have focused on investigating the expression of specific genes during mechanically induced delayed healing in the sheep osteotomy model compared to standard healing [12,13]. The present study is the first conducting a systematic comparison of the expression characteristics of the transcriptome between a standard and impaired healing osteotomy in a large-animal model. A limitation of our study is the fact that only a single pooled sample could be investigated for each condition. The Audic-Claverie test allows an estimation of the statistical significance of observed differences in the counts of genes that are interpreted as differential expression, but may tend to overestimate significance. For this reason, we applied a stringent P-value cutoff (a gene was considered differentially expressed with a Benjamini-Hochberg corrected P-value of p < 10^{-15} and a fold-change of at least 2-fold). With this proviso, our study has identified a large number of differentially expressed genes corresponding to biological categories that are thought to be most relevant for bone healing. For instance, transcriptome-wide analyses revealed that about 9% (81/884) of the genes found to be differentially expressed during bone healing are annotated to extracellular matrix. Some of these ECM genes are typically found in cartilage. Semi-rigid fixation associated with delayed healing results in a larger cartilage component of the callus, which persisted longer [11]. In agreement with this observation, our study showed higher expression of genes related to cartilage formation after mechanically critical fixation with higher instability of the bone fragments. Cytokines play important roles during bone healing and were shown to be significantly overrepresented in the MGSA analysis. We additionally identified a large set of genes annotated to striated muscle contraction and contractile fiber that displayed a characteristic shift in delayed bone healing. Genes from these categories have not previously been known to be differentially expressed in bone healing. More research will be required to identify the cell types within the fracture callus that express these genes and to elucidate their functional role. The fact that many genes whose proteins are involved in binding calcium ions are differentially expressed provides leads as to the molecular correlates of the differential mineralization observed in delayed bone healing.

### Conclusions

RNA-Seq is an approach to expression profiling based on next-generation sequencing technologies, whereby a sample of RNA is converted to a library of cDNA.
Figure 5 Differentially expression of genes annotated to selected GO-terms. **A** contractile fiber (GO:0043292), **B** calcium ion binding (GO:0005509), **C** cartilage development (GO:0051216), **D** response to wounding (GO:0009611), **E** cell division (GO:0051301), and **F** complement activation (GO:0006956). The normalized RPKM values of all significantly differentially expressed genes annotated to the indicated GO terms from standard healing were divided by the corresponding RPKM values from delayed healing. Positive values indicate higher expression levels in standard healing, negative values higher expression levels in delayed healing. A t-test against the null-hypothesis that there is no difference between the healing groups was performed (**p < 0.01, **p < 0.01**).
fragments attached to adaptors. Individual molecules, with or without amplification, are then sequenced in a high throughput fashion [45]. For model organisms such as the sheep for which relatively few gene sequences have been previously published, RNA-seq allows combined identification of previously unsequenced transcripts together with transcriptome analysis. Prior to this study, partial or complete mRNA sequences corresponding to 1,556 ovine genes were publicly available. In the course of this project, partial or complete transcript sequences were generated for 13,987 ovine genes, corresponding to a nearly nine-fold increase in the number of sheep genes with publicly available sequence information. A FASTA file with sequences of the longest available transcript for each of the 13,987 ovine genes is available as Additional File 2. The short reads have been deposited in NCBI’s short read archive.

Methods

Surgical procedure

All animal experiments were conducted following national regulations for the care and use of laboratory animals and approved by the local legal representative (Landesamt für Gesundheit und Soziales Berlin: G0127/07, G0172/04). Surgery was performed as described previously [12,17]. The newly generated tissues were harvested at days 7, 11, 14 and 21 after surgery. For all time points the sample size was \( n = 6 \) for both groups, except for day 21 (group I, \( n = 5 \); group II, \( n = 6 \)). In the 7 day groups, the tissue formed in the gap was harvested under general anesthesia and the animals were sacrificed 14 days after osteotomy, for comparative analysis of healing after hematoma harvesting which will be reported in another study. In the other groups, tissue harvesting occurred after euthanasia and involved removal of newly formed tissue at the osteotomy site (within the gap and along the periosteal/endosteal surfaces of the bone fragments). Tissue was placed in an RNA stabilization reagent (RNAlater; Qiagen, Hilden, Germany) for storage at -80°C.

RNA extraction

Total RNA was isolated from the tissues using the RNeasy Maxi Kit (Qiagen) according to the manufacturer’s instructions. The concentration of each RNA sample was determined spectrophotometrically and the integrity of all RNA samples was monitored on agarose gels.

Histology

For histological examination, the callus regions of the explanted tibiae from additional animals of the standard and delayed healing groups euthanized at the day 14 (\( n = 4 \) each) were sectioned into 3 mm slices in the frontal plane. For calcified histology, histological slices were dehydrated with alcohol and xylol, embedded in methylmetacrylate (Technovit 9100 NEU, Heraeus Kulzer, Germany), cut into 6 µm-thick sections and stained with Movat Pentachrome and Safranin Orange/von Kossa.

Library preparation and Massively parallel sequencing

In each experimental group total RNA of all samples per time point was pooled prior to library preparation. All libraries were prepared using the mRNA-Seq sample prep Kit (Illumina, San Diego, CA USA) according to the manufacturer’s instruction. Clusters were generated with Illumina’s v4 Single Read Cluster Generation Kit. Each library was loaded onto one lane of the flow cell at 7 pM concentration. The flow cell was then sequenced on a Genome Analyzer IIx (Illumina) for 76 cycles with v4 sequencing kits following the standard protocol and using SCS v2.6 software.

Expression profiling and Analysis of Differential Expression

The mapped read counts for each gene were normalized for RNA length and for the total read number in the lane according to reads per kilobase of exon model per million mapped reads (RPKM), which facilitates comparison of transcript levels between samples [45]. The Audic-Claverie method was used to estimate differential expression between standard and delayed bone healing at the same time point as well as for the comparison of different time points within each group. The Audic-Claverie method, which was originally developed for SAGE data, is based on the assumption that the counts of each gene in each of two libraries under comparison follow the same unknown Poisson distribution, and thus allows an estimation of differential expression based on single measurements for two conditions [31]. Fold changes were calculated after quantile normalization of the RPKM values (Additional file 1: Supplemental Figure S1).

Gene Ontology Analysis

Gene Ontology [46] annotations for the 21,865 sheep gene models were obtained by mapping the gene symbols for the sheep gene models to human gene symbols (\( n = 15,343 \)) using bioMart [47]. GO annotations were available for 13,785 of these genes. Model-based gene set analysis (MGSA) was used to perform Gene Ontology analysis. MGSA analyzes all GO terms at once by embedding them in a Bayesian network, in which gene response is modeled as a function of the activation of the GO terms and probabilistic inference is used to identify the active categories [33]. Analysis was performed using the Ontologizer [48].
Clustering
For each gene determined to be differentially expressed, normalized RPKM values were transformed to a mean of zero and a standard deviation of one separately for the standard and delayed healing groups (four time points each). The normalized values were then combined into vectors of 8 values per gene. Hierarchical clustering was performed to group genes according to similarity in pattern of gene expression [49].

Short read sequences
The data from the experiments described in this work are available from the NCBI Sequence Read Archive at http://www.ncbi.nlm.nih.gov/sra under the accession number SRA020182.

Quantitative PCR (Q-PCR)
After transcription of 1 μg RNA into cDNA (RevertAid H minus cDNA Synthesis Kit, Fermentas) quantitative RT-PCR was performed in MicroAmp optical 384-well plates on ABI Prism 7900 Sequence Detection System in a total volume of 12 μl in each well containing 6 μl of Power SYBR Green PCR Master Mix (Applied Biosystems), 5 μl cDNA (in a 1:50 dilution) and 1 μl primers (0.2 μmol each). For some RNA samples with lower concentrations, higher volumes were reverse transcribed, and the cDNA dilutions were adapted accordingly prior to Q-PCR analysis. For validation of the target genes obtained from RNA-seq equal amounts of cDNA were pooled for each group and time point. Primer pairs were designed to span exon-exon junctions. Primer sequences can be obtained upon request. All samples were run in triplicates in separate tubes to permit the quantification of the target genes’ mRNA expression relative to the mean expression of GTPB1, HDAC6 and SNRPN, i.e. three stably and highly expressed genes obtained from the RNA-seq data. Q-PCR result data was exported from the SDS 2.3 software (Applied Biosystems) and further analyzed as described previously [50].

Additional material

Additional file 1: Supplementary Information A PDF file with Supplementary Figures S1 - S3 and Supplementary Table S1 - S4.
Additional file 2: Ovis aries composite transcriptome A FASTA file containing the sequence of the longest transcript of each of the 13,987 ovine gene models.

Acknowledgements
This study was supported by a grant from the German Research Foundation (DFG SFB 760) and by an additional grant to the Berlin-Brandenburg Center for Regenerative Therapies (BMBF project number 0313911). The authors would like to thank Mrs. Catrin Janetzki, Mrs. Norma Schulz and Mrs. Camilla Bergmann for excellent technical assistance.

References
1. Shothi R, Zaretski A, Cohen E, Engel A, Bar-E Y: The sheep as a model for coronary artery bypass surgery. Lab Anim 2004, 38(2):149-157.
2. Matsuzaki K, Morita M, Hamamoto H, Noma M, Robb JD, Gilespie MJ, Gorman JH, Gorman RC: Elimination of Ischemic Mitral Regurgitation Does Not Alter Long-Term Left Ventricular Remodeling in the Ovine Model. Ann Thorac Surg 2010, 90(3):788-794.
3. Yamagami T, Yoshimatsu R, Matsumoto T, Anzai H, Yoshizawa M, Fukuy K, Nishimoto T: Reproductive ability after uterine artery embolization in a sheep model: observation over 2 seasons. Reprod Sci 2010, 17(1):569-584.
4. Schmalstieg FC, Keeney SE, Rudloff HE, Palkowetz KH, Cevallos M, Zhou X, Cox RA, Hawkins HK, Traber DL, Zwischenberger JB: Arteminogenous CO2 removal improves survival compared to high frequency percussive and low tidal volume ventilation in a smoke/burn sheep acute respiratory distress syndrome model. Ann Thorac Surg 2007, 246(3):512-21, discussion 521-3.
5. Zhou X, Wang D, Ballard-Croft CK, Simon SR, Lee HM, Zwischenberger JB: A tetracycline analog improves acute respiratory distress syndrome survival in an ovine model. Ann Thorac Surg 2010, 90(2):419-426.
6. Sturmer KM, Schuchardt W: [New aspects of closed intramedullary nailing and narrow cavity reaming in animal experiments. I. The tibia of the sheep, as a model for intramedullary nailing (author’s transl)]. Unfallheilkunde 1980, 83(7):341-5.
7. Newman E, Turner AS, Wark JD: The potential of sheep for the study of osteopenia: current status and comparison with other animal models. Bone 1995, 16(4 Suppl):277S-284S.
8. den Boer FC, Patka P, Bakker FC, Wippermann BW, van Lingen A, Vink GQ, Boshuizen K, Haarman HJ: New segmental long bone defect model in sheep: quantitative analysis of healing with dual energy x-ray absorptiometry. J Orthop Res 1999, 17(5):654-660.
9. Ravaglioli A, Krajewski A, Celotti GC, Piancastelli A, Bacchin B, Montanari L, Zama G, Giombi L: Mineral evolution of bone. Biomaterials 1996, 17(6):617-622.
10. Reichen JC, Saffzadeh S, Wollschleger ME, Epari DR, Schtz MA, Duda GN, Schell H, van Griensven M, Riedl H, Hutmacher DW: The challenge of establishing preclinical models for segmental bone defect research. Biomaterials 2009, 30(12):2149-2163.
11. Epari DR, Schell H, Bail HJ, Duda GN: Instability prolongs the chondral phase during bone healing in sheep. Bone 2006, 38(6):864-870.
12. Lienau J, Schmidt-Bleek K, Peters A, Haschke F, Duda GN, Pecka C, Bail HJ, Schutzte N, Jakob F, Schell H: Differential regulation of blood vessel formation between standard and delayed bone healing. J Orthop Res 2009, 27(9):1131-40.
13. Lienau J, Schmidt-Bleek K, Peters A, Weber H, Bail HJ, Duda GN, Pecka C, Schell H: Insight into the molecular pathophysiology of delayed bone healing in a sheep model. Tissue Eng Part A 2010, 16:191-9.
14. Peters A, Schell H, Bail HJ, Hannemann M, Schumant T, Duda GN, Lienau J: Standard bone healing stages occur during delayed bone healing, albeit with a different temporal onset and spatial distribution of callus tissues. Histol Histopathol 2010, 25(9):1149-1162.
15. Lienau J, Schell H, Duda GN, Seebeck P, Muchow S, Bail HJ: Initial
vascularization and tissue differentiation are influenced by fixation
stability. J Orthop Res 2005, 23(3):639-45.
16. Schell H, Epaü DR, Kass J, Bragulla H, Bail HJ, Duda GN: The course of
bone healing is influenced by the initial shear fixation stability. J Orthop
Res 2005, 23(3):1022-8.
17. Schell H, Thompson MS, Bail HJ, Hoffmann JE, Schill A, Duda GN, Lienau J:
Mechanical induction of critically delayed bone healing in sheep:
radiological and biomechanical results. J Biomech 2008, 41(14):3066-72.
18. Shendure J, Ji H: Next-generation DNA sequencing. Nat Biotechnol
2008, 26(10):1168-75.
19. Metzker ML: Sequencing technologies - the next generation. Nat Rev
Genet 2010, 11:31-46.
20. Wetterborn A, Ameur A, Feuk L, Gyllenstein U, Cavailer L: Identification of
novel exons and transcribed regions by chimpanzee transcriptome
sequencing. Genome Biol 2010, 11(7):R78.
21. Costa V, Angelini C, Fois ID, Coccodrillo A: Uncovering the complexity of
transcriptomes with RNA-Seq. J Biomed Biotechnol 2010, 2010:853916.
22. Cirilli ET, Singh A, Shianna KV, Ge D, Smith JP, Maia JM, Heienlz EL,
Goedert JJ, Goldstein DB, Center for HIV/AIDS Vaccine Immunology (CHAVI):
Screening the human exome: a comparison of whole genome and
whole transcriptome sequencing. Genome Biol 2010, 11(S5):R57.
23. Baele F, Schmeits W, Cremers C, Burt W, Klaunmann H, Weeth A, Schulz
TW, Hildebrandt T: Into the unknown: expression profiling without
genome sequence information in CHO by next generation sequencing.
Nucleic Acids Res 2010, 38:3999-4010.
24. Sun C, Li Y, Wu Q, Luo H, Sun Y, Song J, Lui EMK, Chen S: De novo
sequencing and analysis of the American ginseng root transcriptome
using a GS FLX Titanium platform to discover putative genes involved in
ginsenoside biosynthesis. BMC Genomics 2010, 11:262.
25. Hecht J, Kühl H, Haas SA, Bauer S, Poustka AJ, Lienau J, Schell H, Steige AC,
Seitz V, Reinhardt R, Duda GN, Mundlos S, Robinson PN: Gene
identification and analysis of transcripts differentially regulated in
fracture healing by EST sequencing in the domestic sheep. BMC
Genomics 2006, 7:172.
26. Cock PJ: Fileds: CJ, Goto N, Heuer ML, Rice PM: The Sanger FASTQ
file format with sequences with quality scores, and the Solexa/Illumina FASTQ
variants. Nucleic Acids Res 2010, 38(6):1767-1771.
27. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K,
Chetvernin V, Church DM, Dicuccio M, Federhen S, Feolo M, Geer LY,
Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL,
Madden TL, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J,
Panchenko AR, Phan I, Tebko CK, Tatusova TA, Wagner L, Weng Z, Wang
Y, Wilbur WJ, Yankulov K, Zhang D, Zhang J, Zhou H, Zhu S: Database
resources of the National Center for Biotechnology Information.
Nucleic Acids Res 2010, 38(Database):1-16.
28. Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly
using de Bruijn graphs. Genome Res 2008, 18(5):821-829.
29. Schulz M, Zerbino D: Oases - De novo transcriptome assembler for very
short reads. Tech. rep., MPI for Molecular Genomics, European Bioinformatics
Institute (EMBL-EBI), UC Santa Cruz 2010.
30. Langmead B, Trapnell C, Pop M, Salzberg SL: Ultrafast and memory-
efficient alignment of short DNA sequences to the human genome.
Genome Biol 2009, 10(3):R25.
31. Audic S, Claverie JM: The significance of digital gene expression profiles.
Genome Res 1997, 7(10):986-995.
32. Tino P: Basic properties and information theory of Audic-Claverie statistic
for analyzing cDNA arrays. BMC Bioinformatics 2009, 10:310.
33. Bauer S, Gagneur J, Robinson PN: GOing Bayesian: model-based gene set
analysis of genome-scale data. Nucleic Acids Res 2010, 38:3523-3532.
34. Kinner B, Gerstenfeld LC, Einhorn TA: Spectr M: Expression of smooth
muscle actin in connective tissue cells participating in fracture healing
in a murine model. Bone 2002, 30(5):736-745.
35. Schmidt-Beek K, Schell H, Kolar P, Pfiff M, Perka C, Buttgereit F, Duda G,
Lienau J: Cellular composition of the initial fracture hematoma
compared to a muscle hematoma: a study in sheep. J Orthop Res 2009,
27(9):1147-1151.
36. Kolar P, Schmidt-Beek K, Schell H, Gaber T, Toben D, Schmidmaier G,
Perka C, Buttgereit F, Duda GN: The Early Fracture Hematoma and Its
Potential Role in Fracture Healing. Tissue Eng Part B Rev 2010, 16:427-434.
37. Shapiro F: Bone development and its relation to fracture repair. The role of
mesenchymal osteoblasts and surface osteoblasts. Eur Cell Mater 2008,
15-33-76.
38. Le AK, Mclau T, Hu D, Helms JA: Molecular aspects of healing in
stabilized and non-stabilized fractures. J Orthop Res 2001, 19:76-84.
39. Meyer J R A, Meyer MH, Tenholder M, Wondracek S, Wasserman R,
Panges G: Gene expression in older rats with delayed union of femoral
fractures. J Bone Joint Surg Am 2003, 85-A(7):1243-54.
40. Nikura T, Hak DJ, Reddy AH: Global gene profiling reveals a
downregulation of BMP gene expression in experimental atrophic
nonunions compared to standard healing fractures. J Orthop Res 2006,
24(7):1463-71.
41. Dalynryp BP, Kirkness EF, Nefedorov M, McWilliam S, Ratnakumar W,
Zhaow S, Shetty J, Maddox JF, O’Grady M, Nicholas F, Crawford A,
Smith T, de Jong PJ, McEwan J, Oddy VH, Cockett NE: Consortium ISG: Using
comparative genomics to rerorid the human genome sequence into a
virtual sheep genome. Genome Bio 2007, 8(7):R152.
42. Tino P, Audic S, Claverie JM: The significance of digital gene expression profiles.
Genome Res 1997, 7(10):986-995.
43. Ashburner M, Ball CA, Blake DJA, Botstein D, Butler H, Cherry JM, Davis AP,
Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A,
Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene
Ontology: tool for the unification of biology. The Gene Ontology
Consortium. Nat Genet 2000, 25:25-29.
44. The International Sheep Genomics Consortium, Archibald AL, Cockett NE,
Dalynryp BP, Faraut T, Kijas JW, Maddox JF, McEwan J, Hutton Oddy V,
Raadsma HW, Wade C, Wang J, Wang W, Xun Y: The sheep genome
reference sequence: a work in progress. Anim Genet 2010, 41(5):449-543.
45. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and
quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008,
5(7):621-628.
46. The International Sheep Genomics Consortium, Archibald AL, Cockett NE,
Dalynryp BP, Faraut T, Kijas JW, Maddox JF, McEwan J, Hutton Oddy V,
Raadsma HW, Wade C, Wang J, Wang W, Xun Y: The sheep genome
reference sequence: a work in progress. Anim Genet 2010, 41(5):449-543.
47. Kasprzyk A, Williams BA, McCue K, Schaeffer L, Wold B: Mapping and
quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 2008,
5(7):621-628.
48. Ashburner M, Ball CA, Blake DJA, Botstein D, Butler H, Cherry JM, Davis AP,
Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A,
Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: Gene
Ontology: tool for the unification of biology. The Gene Ontology
Consortium. Nat Genet 2000, 25:25-29.
49. dei Medley GC, Ballester B, Holland R, London D, Thorisson G,
Kasprzyk A: BioMart-biological queries made easy. BMC Genomics 2009,
10:22.
50. Bauer S, Grossmann S, Vingron M, Robinson PN: Ontologizer 2.0-a
multifunctional tool for GO term enrichment analysis and data
exploration. Bioinformatics 2008, 24(14):1650-1651.
51. Eisen MB, Spellman PT, Brown PO, Botstein D: Cluster analysis and display
of genome-wide expression patterns. Proc Natl Acad Sci USA 1998,
95(25):14863-14868.
52. Ott CE, Bauer S, Manke T, Ahrens S, Rodelsperger C, Grünhagen J, Kornak U,
Duda G, Mundlos S, Robinson PN: Promiscuous and depolarization-
induced immediate-early response genes are induced by mechanical
strain of osteoblasts. J Bone Miner Res 2009, 24(7):1274-1262.

Cite this article as: Jäger et al: Composite transcriptome assembly of
RNA-seq data in a sheep model for delayed bone healing. BMC
Genomics 2011 12:158.