# Discovering the BMP Signaling Pathway by using Motif Enrichment and Conservation

Nanning G. de Jong

Information and Communication Theory Group, Delft University of Technology, The Netherlands

This research identifies new functions of known transcription factors related to osteoblast differentiation. It reveals parts of the osteoblast differentiation pathway by using an integrated approach of *in vitro*, *in vivo* and *in silico* techniques. Perturbation experiments are performed in MC3T3 cells with *Bone Morphogenetic Protein 2 (*BMP2). The differently expressed genes whose expression correlates to the BMP2 concentration in the medium are identified as potential direct target genes of the BMP2 signaling pathway. Then known transcription factors are identified with enriched binding sites in the upstream genomic region of these direct target genes. These are the transcription factors *TCFAP2a*, *ROAZ* and *SNAI1*. This identification is biologically validated with RNAi perturbation experiments on these transcription factors.

The use of known transcription factors is complemented by using the same clusters of differently expressed genes to discover conserved binding site motifs with the program PhyME.

Contact: n.g.dejong@student.tudelft.nl

*Keywords*

Osteoblast differentiation | Transcription Factors | Motif Scanning | Motif Discovery | Motif Enrichment | Motif Conservation |

*Table of Contents*



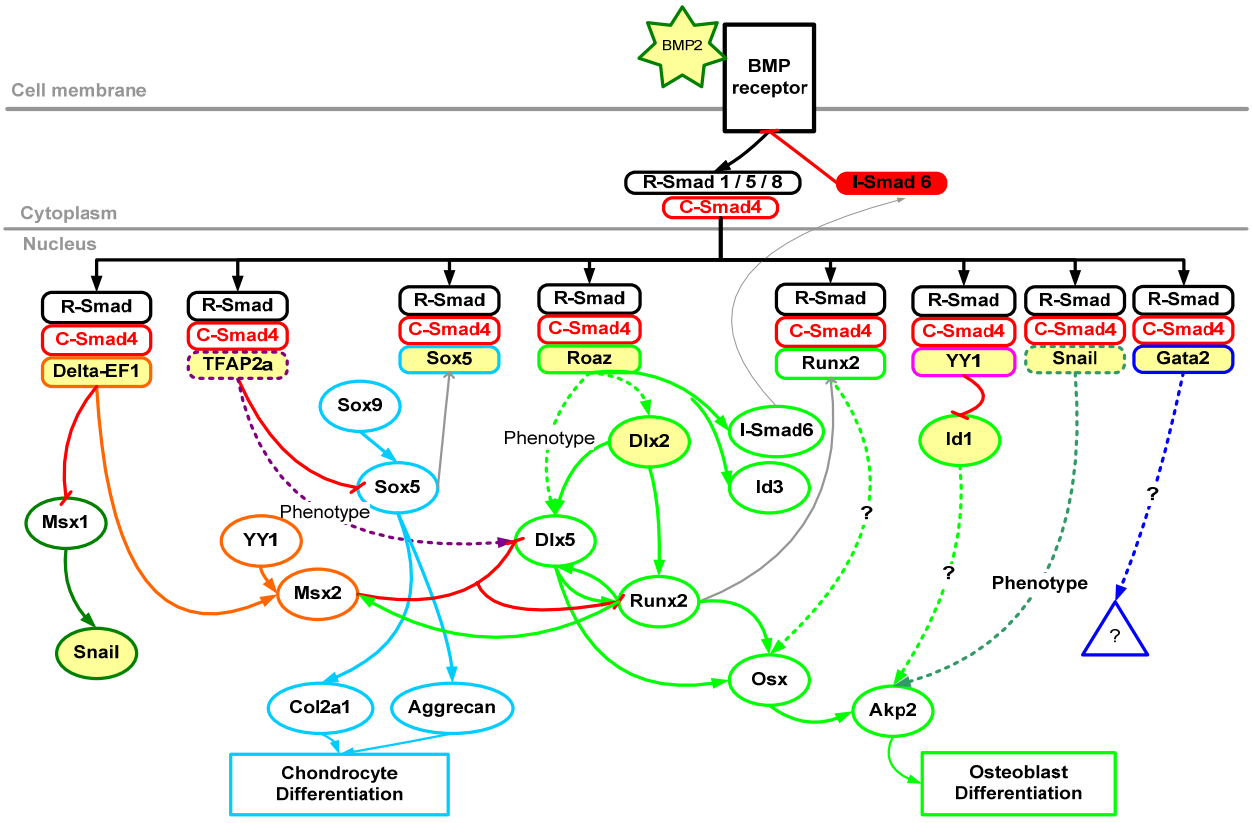



**Figure 1: Proposed BMP signaling pathway. The R-Smads are activated upon binding of BMP2 on the receptor. The R-Smads can then form a complex with C-Smad4 and be transported to the nucleus. This complex binds several transcription factors such as Delta-EF1, which transcribe or inhibit other genes downstream and lead towards chondrocyte or osteoblast differentiation. Dotted lines were not described in literature before, but proposed in this research. The dotted lines with the word "phenotype" have an osteoblast differentiation phenotype in RNAi knockdown experiments. The transcription factors or genes that have high rankings in our results are highlighted in yellow.**

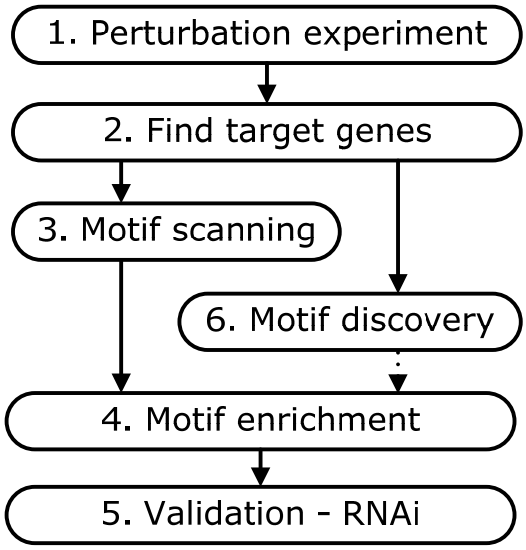



**Figure 2: Schematic of the general approach**

## 1. Introduction

More than 50% of women above 50 suffer from osteoporosis and have a high risk of obtaining a bone fracture due to reduced bone mass. The cause of osteoporosis lies in the imbalance between osteoblast-mediated bone formation and osteoclast-mediated bone resorption [1].

Osteoblasts are specialized cells that synthesize bone proteins which are essential for the formation of the extra-cellular matrix that is subsequently mineralized. Bone marrow contains pluripotent mesenchymal stem cells that can differentiate towards osteoblasts (bone forming cells), but also towards adipocytes (fat cells), myocytes (muscle cells), chondrocytes (cartilage cells), and osteoclasts (bone resorbing cells). The choice of lineage depends on the availability of different signal molecules, like growth factors, morphogens and transcription factors. [2].

A signaling molecule that is known to stimulate osteoblast differentiation is *bone morphogenetic protein 2* (*BMP2*). This molecule activates the developmental pathway towards osteoblasts and simultaneously inhibits the pathways towards the other cell types [3]. It is clear that many regulatory steps are needed to ensure that pluripotent stem cells will differentiate towards osteoblasts and not to other cell types. Different transcription factors are involved in this process, however their exact roles and the genes they target still needs to be elucidated [4].

The signal transduction of extra cellular *BMP2* from the cell membrane to the nucleus is well known and is illustrated in the upper part of Figure 1. Outside the cell, *BMP2* first binds the *BMP* receptor which activates the intracellular part of the receptor. The activated receptor then phospohorylates the *Receptor-Smad* proteins, i.e. *R-Smad1*, *R-Smad5* and / or *R-Smad8*. These activated Receptor-Smads bind with the *Co-Smad* (*C-Smad4*) upon which this complex is transported into the nucleus [5].

In the nucleus, the *Smad* complex may associate with other transcription factors (named co-factors) to allow the induction of different target genes. Which genes are targetted may depend on which co-factors are available. Some of the known target genes of the BMP signaling pathway, such as *distal-less homeobox 2 and 5 (Dlx2* and *Dlx5)*, *activate runt related transcription factor 2* (*Runx2)*, which is the key transcription factor in the osteoblast differentiation route [6]. *Runx2* is known to stimulate the transcription factor *Osterix* (*Osx*), which in turn stimulates *Alkaline Phosphatase* (*Akp2*) [5]. Alkaline Phosphatase is a key bio marker used to identify early osteoblast cells [7]. A measure for the presence of more mature osteoblasts in a culture is the amount of mineralization of the cells.

*Smads* are involved in a wide variety of signal transductions and have different functions. The *Inhibitor-Smad* proteins, i.e. *I-Smad6* and *I-Smad7*, inhibit the phosporylation of the *R-Smads* by the receptor [4] (See Figure 1). Besides *BMP*s, the signalling molecule *transforming growth factor beta 1 (TGF-β1)* also uses *Smads* to transduce its signal [8]. Therefore, co-factors may play an important role to determine a different set of target genes of smad-mediated signalling for different cellular contexts.

This paper presents an approach that reveals parts of the osteoblast differentiation pathway by using an integrated approach of *in vitro*, *in vivo* and *in silico* techniques. The approach focuses on first identifying the direct target genes of *BMP2* signaling. Their upstream promoters are then analyzed to determine which known transcription factors as well as unknown binding sites may be involved in their regulation. The approach consists of six steps which are depicted in Figure 2.

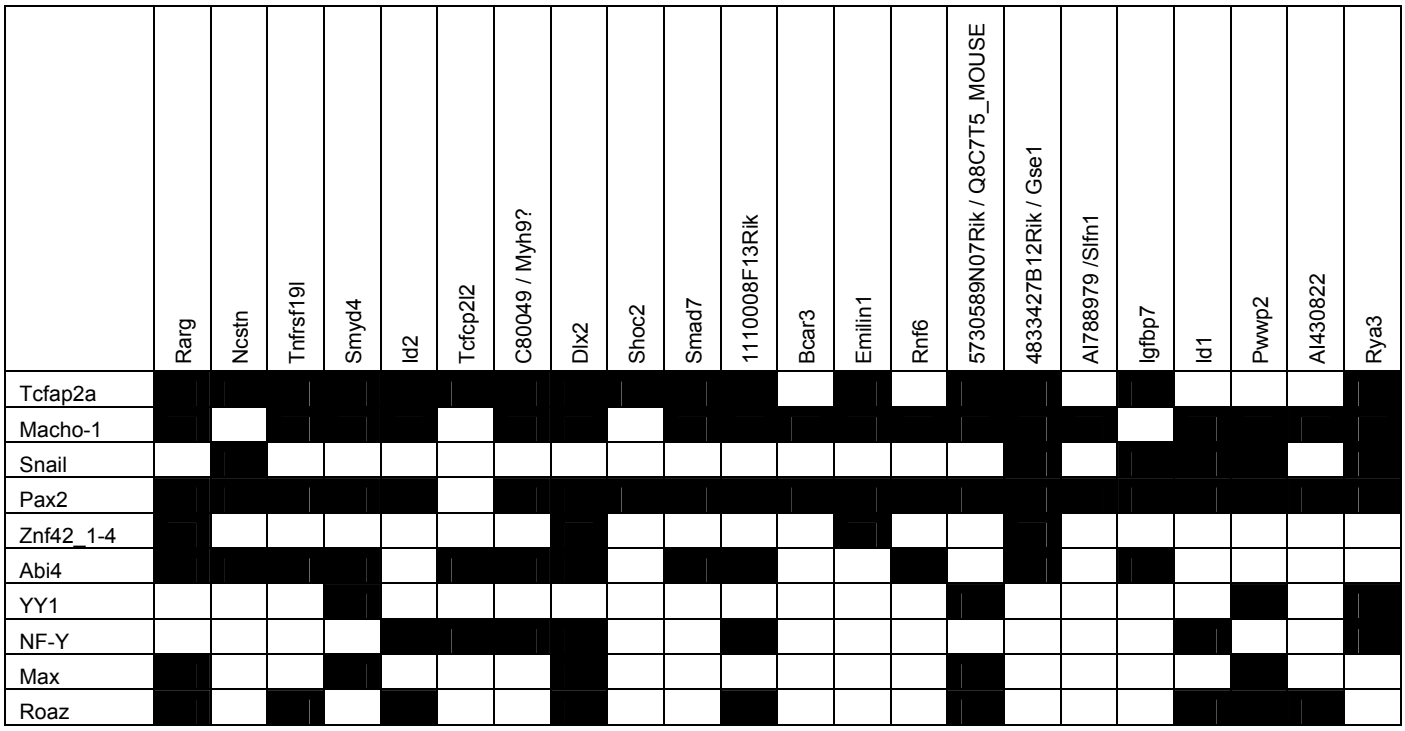
The first step is to measure the gene expression levels over time during *BMP2*-induced osteoblast differentiation using DNA micro arrays. The second step is to analyze the DNA expression levels and to find the direct target genes of *BMP* signaling. The third step is to determine which transcription factors with known binding site motifs can bind the upstream promoters of the direct target genes. The fourth step is to calculate the enrichment of these transcription factors for this specific group of target genes.

The highly enriched transcription factors are in the fifth step validated by using RNAi experiments. The effects of these known transcription factors on osteoblast differentiation are established when knockdowns of these transcription factors show an osteoblast phenotype.

The sixth step is to discover new binding sites in the promoter regions of the target genes. This is done by using the motif discovery tool PhyME. This can then be followed by the fourth and fifth step again.

Application of this approach on *BMP2*-induced osteoblast differentiation in the MC3T3 cell-line

**Table 1: The cluster of BMP target genes (columns) and the occurrence of enriched transcription factor binding sites (rows). The black cells in the table depict the genes for which the number of TFBS exceeds a certain threshold for a certain upstream length (see Table 2 for these values). These genes are counted as a hit in the hypergeometric test.**




resulted in the discovery of a novel role in osteoblast differentiation of the transcription factors *Snail homolog 1* (*Snai1*), *zinc finger protein 423* (*Zfp423* or *ROAZ*) and *transcription factor AP-2 alpha* (*Tcfap2a*).

# 2. Results

## 2.1. Direct target genes of the BMP2 signal

The first goal of this research is to find direct target genes of the *BMP* signal transduction route. Therefore a time-course DNA microarray experiment was performed on MC3T3 cells that differentiate towards osteoblasts under the influence of *BMP2*. Besides mRNA level, also the activity levels of the signal molecule *BMP2* in the culture, the alkaline phosphatase levels as well as the amount of deposited mineralization was measured in this experiment. For more details concerning the microarray dataset see Methods section 4.1.

Based on these measurements a group of 22 genes (see Table 1) was selected. The mRNA levels of these genes showed the highest (anti)-correlation with the measured activity levels of BMP2 in the experiment (see Methods section 4.6). These genes react without time lapse on the *BMP2* pulses and are therefore assumed to be the direct target genes of the signal transduction route. Several of these genes, i.e. *inhibitor of DNA binding 1* and *2* (*Id1* and *Id2)*, *Dlx2* and I-*Smad7*, are known to play a role in osteoblast differentiation. Their known interactions as far as they are relevant to osteoblast differentiation are also depicted in Figure 1 [4] [9] [10] [11].

## 2.2. Enrichment of known motifs in the cluster of direct target genes

The promoter regions of the direct target genes were then analyzed to find transcription factors that are enriched for this set of target genes. The upstream promoter regions were scanned (see Methods section 4.3) with the motifs of known transcription factors to find binding sites in these sequences (denoted as TFBS). The enrichment score of the occurrence of these TFBS in the promoters of the target genes was calculated with the hypergeometric test. The coupling of a transcription factor to a gene, however, depends on the length of the considered promoter region and the amount of binding sites. We followed a procedure in which the threshold for the minimum number of binding sites ($T^b$) and the length of the promoter region ($T^l$) length were optimized for each transcription factor separately. The configuration for which resulted in the lowest uncorrected p-value possible has been considered the enrichment score of that transcription factor for the group of genes in consideration. See Methods section 4.4 for details on this procedure. It should be noted that this minimum of p-values is no longer a p-value itself, but represents a score for the enrichment of this transcription factor to this cluster of genes. The top 10 TFBS with the highest enrichment are presented in Table 2.

**Table 2: Top ranking transcription factors found for the cluster of BMP target genes. The second column shows the enrichment score upon which the ranking is based. The third and the fourth column show the number of genes in the cluster that are marked as a hit to calculate the enrichment score with the hypergeometric test (Methods 4.4). For this cluster there are 22 genes in total (foreground) and about 22.000 genes in the genome (background). The fifth and sixth columns show the number of binding sites and upstream length at the minimal uncorrected p-value (Chapter 2.2). The last five columns show with a "+" if references to these topics were found in literature. The Macho-1 transcription factor is found in Sea Squirts (*Ciona intestinalis)* and ABI4 in plant life (*Arabidopsis thaliana).* Therefore no references in literature were found related to bone formation. The NF-Y transcription factor is related to the ERK pathway and is a very common for eukaryotes.**

| Transcription Factor (corresponding to TFBS) | Enrichment Score (minimal uncorrected p-value) | Foreground hits (in the cluster) | Background hits (in the genome) | ($T^b$) Number of Binding Sites at min. p-value | ($T^l$) Upstream length (nt) at min. p-value | BMP (relation in literature) | SMAD (relation in literature) | Bone (relation in literature) | TGF-B (relation in literature) | Bone Phenotype in Knockout//Knockdown |
|---|---|---|---|---|---|---|---|---|---|---|
| TCFAP2A | 1,76E-08 | 16 | 4812 | 5 | 480 | + | | + | + | + |
| Macho-1 | 4,56E-05 | 18 | 10723 | 9 | 3760 | | | | | |
| Snai1 (Snail) | 2,27E-04 | 6 | 1163 | 7 | 3590 | + | + | | + | |
| Pax2 | 2,71E-04 | 21 | 16836 | 1 | 600 | + | + | | | |
| ZNF42_1-4 (MZF1) | 4,35E-04 | 4 | 460 | 27 | 4440 | | | + | | |
| ABI4 | 4,56E-04 | 12 | 5690 | 1 | 270 | | | | | |
| YY1 | 9,33E-04 | 4 | 564 | 7 | 610 | + | + | + | | + |
| NF-Y | 9,50E-04 | 7 | 2125 | 1 | 430 | | | + | | |
| MAX | 1,48E-03 | 5 | 1102 | 2 | 720 | | + | | + | |
| Roaz | 1,52E-03 | 9 | 3780 | 1 | 3670 | + | + | + | | + |

Table 2 shows that the optimal number of binding sites, $T^b$ , and promoter lengths, $T^l$, varies significantly for the different transcription factors. For instance a variety in the minimal amount of binding sites, $T^b$ between 1 and 27 can be observed. The upstream promoter length, $T^l$ , varies between 270 and 4440 nucleotides for different transcription factors. Note that using fixed values for these parameters, as most other tools do, would be an arbitrary choice and will result in substantially different rankings for each choice. Some transcription factors need to have Transcription Factor Binding sites (TFBS) close to the transcription start to be functional, while others may need multiple binding sites distributed over longer promoter regions. Our method adapts to the conditions of each transcription factor.

## 2.3. Enrichment of motif pairs in the cluster of direct target genes

A similar scanning and enrichment calculation method was performed for combinations of transcription factors (Table 3). It is known from literature that transcription factors can cooperate to transcribe target genes specifically or block each other to prevent that [12]. The binding sites of these transcription factors should then be both part of the promoter region of a gene.

## 2.4. Literature Validation

As a first validation, a literature study was done on the found transcription factors. Many references to *BMP*, *Smads*, Bone formation or *TGF-β* (a related pathway that also uses *Smads*) were found for many top ranking transcription factors, as can be seen in the second part of Table 2 and Table 4. Four different pairs of transcription factors were previously described in literature. This is shown in the last column of Table 3. Some of these literature study results are discussed in more detail hereafter.

*ROAZ*

Hata *et al.* show that, in *Xenopus*, *BMP2* induces the association of *ROAZ* (synonym: Zfp423) with *Smad1* and *Smad4* and that this complex then synergistically binds to the *BMP2* response element in the promoter of target genes [13]. In the murine C2C12 cell-line, Ku *et al.* show through several experiments that *ROAZ* is also a co-factor for *BMP4*-induced Smad-mediated induction of *I-Smad6* [14]. For example, Ku *et al.* show with mutations in the promoter of *Smad6* that the binding site for *ROAZ* and the *Smads* are essential for induction of *I-Smad6* and that over expression of *ROAZ* leads to decreased levels of *ALP*. *I-Smad6* antagonizes the *R-Smads1, 5* and *8* and therefore inhibits the *BMP* signal. The activation of *I-Smad6* by *ROAZ* under the influence of *BMP* is therefore a negative feedback

**Table 3: Top ranking combinations of motifs found with BMP response cluster**

| Motif combination | | Enrichment Score (minimal uncorrected p-value) | Foreground hits (in the cluster) | Background hits (in the genome) | ($T^b$) Number of Binding Sites at min. p-value | ($T^l$) Upstream length (nt) at min. p-value | Combination found in literature |
|---|---|---|---|---|---|---|---|
| Snai1 | YY1 | 2,62E-06 | 8 | 1223 | 5 | 4000 | + |
| Snai1 | deltaEF1 | 3,27E-06 | 12 | 3516 | 3 | 3270 | + |
| deltaEF1 | Roaz | 4,88E-06 | 8 | 1330 | 1 | 4310 | |
| NF-Y | Sox5 | 5,42E-06 | 7 | 936 | 1 | 2450 | |
| TFAP2A | Snail | 6,39E-06 | 5 | 344 | 8 | 4010 | + |
| GATA2 | ZNF42_1-4 | 6,74E-06 | 4 | 157 | 6 | 2340 | + |
| YY1 | Macho-1 | 7,75E-06 | 19 | 10901 | 9 | 4550 | |
| Gata1 | ZNF42_1-4 | 7,82E-06 | 4 | 163 | 8 | 2220 | |
| c-ETS | Macho-1 | 1,10E-05 | 19 | 11126 | 9 | 4780 | |
| MYB.ph3 | NF-Y | 1,62E-05 | 7 | 1107 | 1 | 2510 | |

**Table 4: Literature results of Transcription Factors from Table 3 that are not discussed in Table 2. Sox-5 is related to cartilage formation. C-ETS is known to bind Runx2. MYB.ph3 is found in plant life (*Petunia hybrida*)**

| | BMP | SMAD | Bone | TGF-B | Bone Phenotype Knockout |
|---|---|---|---|---|---|
| deltaEF1 | + | + | + | | + |
| Sox-5 | + | | + | | + |
| GATA1 | + | + | | | |
| GATA2 | + | + | | | |
| c-ETS | + | + | + | + | |
| MYB.ph3 | | | | | |

loop.

A study by Karaulanov *et al.* on the *BMP*-response element in the promoters of Xenopus *BMP* targets found a highly conserved and functional motif that consists of *Smad* binding elements (SBE) in close proximity to a (*R)OAZ* binding element (OBE) [15] They found this SBE-OBE motif in the *Id* proteins and in inhibitory *Smads* (both types of genes are also present in our set of target genes).

Ku *et al.* also found that *inhibitor of DNA binding 3 (Id3)* is another direct target of *ROAZ* [14] in murine C2C12 cell lines. It was shown by Maede *et al.* that *Id1* / *Id3* heterozygous knock out mice suppress BMP-induced bone formation in vivo [9]. These results indicate that *Id1* and *Id3* promote bone formation in vivo and that *ROAZ* might also induce bone formation.

In summary, *ROAZ* is known to be a functional co-factor for *BMP*-induced *Smad*-mediated bone formation and targets *Id* proteins and inhibitor *Smads*.

*YY1*

Kurisaki *et al.* show with GST pull-down experiments that endogenous *Yin Yang 1 (YY1)* interacts strongly with the *C-Smad4* and in lesser extend with R-Smad1 [3]. They also show that *YY1* does not interfere with the binding of *R-Smad1* and *C-Smad4*, but it does reduce the affinity of this *Smad*-complex for their DNA binding sites. As a result *YY1* inhibits the induction of direct *BMP* target genes, such as *Id1*.

Through over expression as well as knockdown of *YY1* in C2C12 cells Kurisaki *et al.* indeed show that higher levels of *YY1* lead to lower levels of *ALP*.

Hence, *YY1* is known to be an inhibiting co-factor of BMP-induced *Smad*-mediated bone formation and is known to target *Id1*.

*Delta-EF1*

*Zinc finger E-box binding homeobox 1 (Delta-EF1 or ZEB-1)* was shown by Postigo to form a transcriptional complex with *Smads* [16]. In murine C2C12 cells, he showed that over-expression of *Delta-EF1* resulted in higher levels of *BMP2*-induced ALP. This finding is supported by the fact that *Delta-EF1* homozygote knockout mice were shown to have multiple skeletal malformations [17].

In addition, two known direct targets of *Delta-EF1* are known to be important for bone formation. Firstly, Lazarova *et al.* found that the activity of the murine *osteocalcin 2* promoter is modulated by *Delta-EF1*. Osteocalcin is the most abundant osteoblast-specific non-collagenous protein [18].

Secondly, *Delta-EF1* also binds to two sites within the vitamin D3 receptor promoter and activates the transcription of this receptor in a cell-specific manner [19]. A knockout of this receptor in mice resulted in impaired bone formation [20].

Thus, Delta-EF1 is known to be a functional co-factor of BMP-induced Smad-mediated bone formation and is known to target osteocalcin and vitamin D3.

*Transcription factor pairs*

The analysis of the promoter of *snail homolog 2 (Snai2)*, a known target of *YY1*, found conserved elements consisting of an *YY1* responsive element, a *TATA* box and a potential *Snail* binding motif in close proximity [21]. Although not functionally confirmed this finding supports a potential role for our top-ranked transcription pair *YY1* and *Snai1*.

Table shows also other pairs of transcription factors that are described in Literarture. *Snai1* and *Delta-EF1 (ZEB1*) are described by Laux *et al*. to be E-cadherin transcriptional repressors [22]. An other transcription factor pair is decribed by Yu *et al.*. They

report about a LTR Enhancer Complex *NF-Y / MZF1* (*ZNF42*) / *GATA-2.* Finally, two other transcription factors, *Snai1* and TCFAP2a, are described in relationship with BMP signaling for neural crest development [23].
In summary, four of our top six transcription factor pairs are reported to be related in literature.

## 2.5. Biological Validation

*TCFAP2a*, *ROAZ* and *SNAI1* were selected for biological validation experiments. These transcription factors all have highly enriched binding sites in the promoter regions our set of BMP-target genes. *ROAZ* is known to regulate BMP4-induced osteoblast differentiation in C2C12 cells, but its role in BMP2-induced MC3T3 cells was unknown. To our knowledge, it was never shown before in literature that either *TCFAP2a* or *Snai1* are involved in the regulation osteoblast differentiation.
The validation was done through siRNA knockdown experiments on MC3T3 stemcells that differentiate toward osteoblasts under the influence of the signal molecule BMP2 (Methods 4.6). Alkaline phosphatase activity was measured as an early osteoblast differentiation marker [5].
The results of these experiments are represented in Figure 3. The first three experiments are controls and should not have any effect on osteoblast differentiation (and ALP formation) and were performed to form the baseline for the rest of the experiments. These three experiments are: Mock (electroporation without any siRNA duplex), Scrambled (random siRNA without any target gene on the genome) and GFP (Green Fluorescent Protein, which is not present on the *M. musculus* genome).
The next two experiments (BMPRII and Smad6) have a known phenotype and are therefore useful as controls. The knockdown of the BMP Receptor II blocks the relay of the BMP signal and limits osteoblast differentiation. The decrease in ALP activity when stimulated with BMP in Figure 3 shows that very well. Smad6 is an inhibitor of the BMP signaling pathway. The ALP activity is therefore

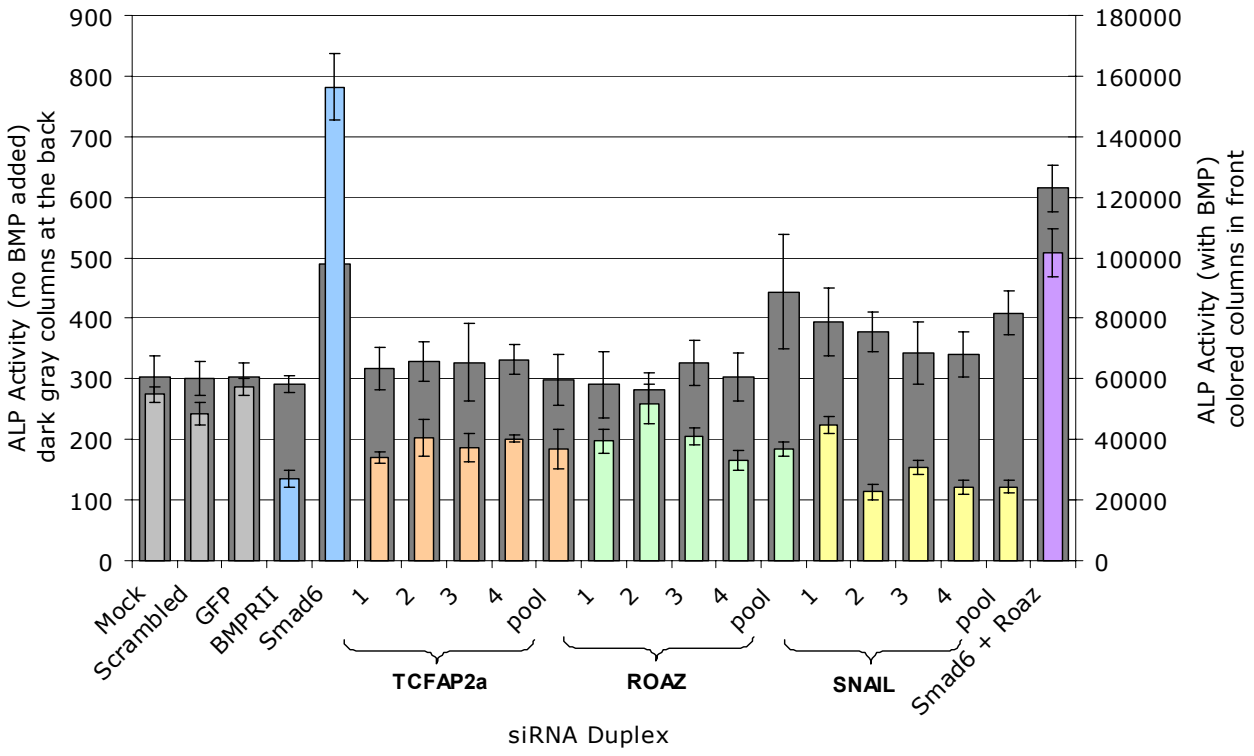



**Figure 3: Biological validation: The ALP activity was measured in MC3T3 cells for days after tranfection with siRNA (see Methods 4.6). The colored bars at the front show the ALP activity in siRNA transfection experiments with 100 ng/ml BMP2 added on day 2. The back row shows the same transfection experiments without added BMP2. From left to right: The first three experiments (Mock, Scrambled and GFP) are control transfections and do not have significant effect on the ALP activity. The knock down of the BMP2 receptor with blocks the BMP2 signal, which decreases the ALP activity (first blue bar). Smad6 is an inhibitor of the BMP2 signal and its knock down increases the ALP activity (second blue bar). Then coupled experiments of 5 bars each are shown for each of the transcription factors TCFAP2a (orange), ROAZ (green) and SNAIL (yellow). For each transcription factor, the set constitutes of 4 different siRNA duplexes and the pool (mix) of these duplexes. All three transcription factors show significant decreases of the ALP activity. The last bar depicts the result of a pool of the Smad6 and ROAZ duplexes. This experiment shows that the ALP activity without added BMP2 is increased compared to the Smad6 duplex by itself. And the ALP activity is decreased with added BMP2. The error bars indicate one standard deviation.**

stimulated, when this gene is targeted with siRNA. This is confirmed in Figure 3.
Then four different siRNA duplexes (plus their pool) were tested for each targeted transcription factor. The results in Figure 3 show that these transcription factors all have an osteoblast differentiation phenotype. The ALP activity decreases with 37% for the *TCFAP2a* #1 duplex when comparing it to the average baseline of Mock, Scrambled and GFP. The *Roaz* #4 and *Snai1* #2 duplexes decrease it with 38% and 58%. The decrease in ALP activity of *Snai1* #2 is even stronger than that of the BMPRII duplex, which has a decrease of 49%.
A final experiment was done with the combination of Smad6 and *Roaz* pool duplexes. A synergy between them can be seen for the background ALP activity (i.e. with no BMP added). Smad6 and the *Roaz* pool both separately increase the background ALP activity as compared to the average baseline of Mock, Scrambled and GFP (no BMP added). This effect is however enhanced when combining these two duplexes as can be seen in the last column of Figure 3. Their effect on ALP activity is, however, counteracted when they are stimulated with BMP.
In summary, we show here that *TCFAP2a*, *ROAZ* (Zfp423) and *SNAIL* (*Snai1*) have a novel positive role in BMP2-induced osteoblast differentiation in murine MC3T3 cells.

## 2.6. Novel PhyME motifs

The last step in our upstream promoter analysis is the discovery of novel binding site motifs using the discovery program PhyME. This is the sixth step in our method (Figure 2) and is described in detail in the method section 4.5. The results are shown in Table 5. New motifs were discovered in the upstream regions of the cluster of BMP2 direct target genes (exp. A). New motifs were also discovered in the upstream regions without the overlapping exons of neighboring genes in this region (exp. B).
Not all discovered motifs are depicted as most of them had a consensus sequence of either AAAAAAAAAA or TTTTTTTTTT. These motifs were considered as false artifacts from the pattern recognition of PhyME and were discarded. These artifacts could have been caused by repetitions of single bases in the genomic data.
Since only 3 interesting motifs were discovered for each experiment, it was not necessary to rank them. However, if more motifs would have been found, we propose to use the calculation of enrichment p-values (as in section 2.2) to rank the motifs on their specificity to the found cluster.
In Table 5, the three most similar JASPAR motifs are also depicted for each of the novel motifs. All novel motifs have at least one highly similar JASPAR motif. If a novel motif would have been found with low similarity, it would be worthwhile to perform follow-up biological Protein-DNA interaction experiments to identify which (novel) proteins may be able to bind them.
The consensus of the three novel motifs of experiment A and B differ slightly as only six unique JASPAR motifs were found. From these six at least two transcription factors, i.e. ID1 and Klf4, are known to be related to osteoblast differentiation [9] [24].

## 3. Discussion

The final result of this research is the new BMP signaling pathway overview of Figure 1. An extensive overview like this, ranging from the BMP receptor to the Akp2 biomarker, was never published before. Also new relationships were discovered with the *in silico* methods in this paper. These new relationships are depicted with dotted lines in the figure.
Obviously, the transcription factors in the pathway in Figure 1 are biased towards the known JASPAR transcription factors, as those factors were used in our *in silico* methods. Nevertheless, our *in silico* methods succeeded in returning a short list of transcription factors from which many are related to BMP, SMAD, bone formation or TGF-B signalling according to literature. Furthermore, all three

**Table 5: Top ranking PhyME motifs compared to JASPAR transcription factor motifs.**

| Exp A: BMP2 cluster - 5000 upstream | | | | Exp B: BMP2 cluster - 5000 upstream minus exons | | | |
|---|---|---|---|---|---|---|---|
| PhyME Motif Consensus | | Similar to JASPAR Transcription factor | JASPAR Similarity Score | PhyME Motif Consensus | | Similar to JASPAR Transcription factor | JASPAR Similarity Score |
| A1 | TGTCTTTCTG | HMG-IY | 93,31% | B1 | TGTCTGTGTT | HMG-IY | 93,52% |
| | | ID1 | 93,21% | | | ID1 | 93,03% |
| | | Pax4 | 92,17% | | | Pax4 | 92,76% |
| A2 | AGGAAAAAGA | Pax4 | 94,23% | B2 | AGAGAAAGAG | Pax4 | 93,91% |
| | | HMG-IY | 93,02% | | | HMG-IY | 93,09% |
| | | Pax5 | 92,44% | | | ID1 | 92,31% |
| A3 | TCCCAGCCTT | Pax4 | 96,42% | B3 | CCTGCCTCCT | Pax4 | 95,45% |
| | | Pax5 | 92,82% | | | Pax5 | 93,21% |
| | | TBP | 89,79% | | | Klf4 | 92,14% |

transcription factors that we validated were found to have a biological osteoblast differentiation phenotype, despite the fact that for two of them no relation to osteoblast differentiation was known before.
The literature validations confirm that our enrichment rankings are highly related to osteoblast differentiation, as 8 of the top 10 single motifs are related to either BMP, SMAD, Bone formation or TGF-B (of the 123 different single transcription factor motifs that we ranked). This confirmation is even more striking when the relationships with osteoblast differentiation were researched for the about 7500 different combinations of transcription factor motif pairs. The top 10 enriched transcription factor pairs consists of 13 different transcription factors and 10 of these are either related to BMP, SMAD, bone formation or TGF-B (see Table). Four motif pairs in this top 10 were reported to be linked to each other in literature before. The other 6 motif pairs that are proposed in this research might also form complexes or cooperate biologically in other ways.
*TCFAP2a*, *ROAZ* and *Snai1* are confirmed with biological experiments to have an osteoblast differentiation phenotype. The biological experiments show phenotypic changes by RNAi knock downs of the candidate TFs. This shows that our method can pin point TFs that are involved in the transcription of direct target genes of a signal molecule.
An advantage of using a database of validated motifs of known transcription factors is that these factors can easily be knocked down, as their genomic sequences are already known. This allows for permutation experiments without the need to discover the protein that is responsible for transcription first. Discovery of novel motifs requires for instance gel-shift techniques to find the corresponding transcription factor protein.
The putative motifs that were discovered with PhyME in table 5 were not sufficiently novel to be worthwhile to investigate further. The putative motif B1 does have a slight similarity to the consensus motif of Smad that was published by Kurisaki *et al.* [3]. The second until seventh nucleotide of the B1 putative motif (GTCTGT) is complementary to the Smad Binding Element CAGACA [25]. It might therefore be possible that we discovered parts of Smad binding sites.
Although we have found novel transcription factors that regulate osteoblast differentiation, transcription factors are not well suited as drug targets as they are difficult to manipulate by extracellular addition of a small molecule. The genes which they transcriptionally regulate can however be successful drug targets. One way to unravel them is to perform RNAi experiments of the transcription factor in combination with microarray measurements.
In this paper we only considered a single application, but our proposed approach can easily be applied to any other cluster (they should not necessarily be direct target genes of a signaling molecule). Our approach could also be applied to unravel the genetic network further downstream in the BMP signaling pathway. Results of our approach successfully applied on a cluster of differently expressed genes in Runx2 knock-out mice are described in Supplement S7.
There exist other databases of validated motifs of known transcription factors like TRANSFAC [26]. Our method could be enhanced with vertebrate motifs from this database. Both JASPAR and TRANSFAC do not include the motifs for Runx2 and SMADs themselves. Their motifs are however known. The detail of these motifs is not as good as in the JASPAR database [3] [14]. The enrichment of combinations of SMAD motifs with the transcription factors from the JASPAR database would be interesting for this pathway.
The transcription factors of plants and insects were not excluded from the used JASPAR database to scan the promoter regions. Some of these transcription factors did show up as enriched in the top10 that we created. Lacking a mammalian bone structure, no relationship between them or their homologs with osteoblast differentiation or SMADS could be found in literature. It is therefore recommended that, when analyzing vertebrate genomes, these ~30 transcription factor be left out of this method. This decreases the possible amount of transcription factor combinations from ~7500 to ~4000 and makes the method more efficient for vertebrate research.
Every highly enriched transcription factor in our research was manually searched for in literature. This is however labor intensive. An automated data miner that searches the PubMed database could give an overview of specific search terms for every transcription factor and therefore enhance the literature validation.

# 4. Methods

## 4.1. Datasets

Three different types of data were used in the experiments, i.e. DNA microarray data, genomic sequences and transcription factor motifs.

*DNA expression data*

In this study, we followed BMP2-induced osteoblast differentiation in MC3T3-E1 cells over time. In one time-series, called the "multiple pulse series", one dose of BMP2 was added at time zero and additional doses were added after 72 hours, after 144 hours and again after 216 hours. Each addition of BMP2 coincided with regular medium refreshments supplied every 72 hours. In the second time-series, denoted as the "single pulse series", BMP2 was added only at time zero and no additional doses of BMP2 were given during subsequent medium refreshments. The gene expression of the multi and single pulse series was then measured with an Affymetrix GeneChip® Mouse Genome 430 2.0 Array

(at t = 0, 1, 2, 3, 4, 8, 16, 32, 56, 72, 73, 74, 75, 76, 80, 144, 145, 146, 147, 148, 152, 176, 200, 224 hours). Between-array normalization was performed using Rosetta Resolver Experiment Builder.

*Genomic sequences*

The -5000 to 0 bp upstream sequences of all genes of the species *Mus musculus* (version 39_36), *Rattus norvegicus* (39_34i) and *Homo sapiens* (39_36a) were acquired from the database Ensembl [27], by using the web interface MartView [28].
MartView was also used to convert Affy IDs from DNA Microarray data into Ensembl Gene IDs and to download ortholog information.

*Transcription Factor Motifs*

All 123 Transcription Factor Motifs (i.e. Position Weigth Matrices, see for an explanation Supplement S2) were downloaded from the JASPAR database [29]. This database contains curated and non-redundant transcription factor DNA-binding preferences.

## 4.2. Cluster selection

Pearson's correlation between the BMP2 activity profile and all probe-set expression profiles was then determined. Probe-sets were ranked according to correlation values. For genes with multiple corresponding probe-sets, only the best ranked probe-set was maintained. The 24 highest ranked genes were selected which corresponds to requiring a minimal correlation of .665 or a maximal p-value of 1E-05. This cluster of immediate-early BMP2 response genes is also called direct BMP-target genes in this paper. See Supplement S1 for a complete overview of the expression profiles of the genes in the cluster.

## 4.3. Motif Scanning

Sequences are scanned with the Perl TFBS module scanning program [30]. A PWM is slid from 5' to 3' along the positive and negative strand of a sequence in increments of one base pair resulting in a score $S_n$ (see Supplements S3). Position *n* on the sequence is marked as a binding site by using a relative threshold *T*.

$$T \leq \frac{100(S_n - S_{\min})}{(S_{\max} - S_{\min})} \qquad (1)$$

Where $S_{min}$ and $S_{max}$ are the absolute minimum and maximum scores of a certain PWM.
For our experiments a threshold T = 90% was chosen. The scanning of the upstream regions of all genes in the genome with the JASPAR motifs resulted in a set of positions of binding sites on these upstream regions.

## 4.4. Enrichment procedure

The hypergeometric test was used to differentiate between binding sites that are common in the upstream regions of all genes (like TATA boxes) and binding sites that are specific for a given cluster. In order to make use of this test it is necessary to define, for a given transcription factor, when a gene is counted as a "hit'". A gene is considered a hit if it has at least $T^b$ binding sites for that transcription factor within $T^l$ nucleotides upstream of transcription start. See Supplement S4 for more details on the hypergeometric test. In Suplement S5 we introduce an alternative way to score the enrichment that operates on the direct output of the scanning (and thus does not need to threshold the scores).
As an upstream region can have multiple binding sites, the p-value was re-calculated for every possible value of $T^b$=[1,2, … ]. Similarly, as the length of the upstream region is unknown, the p-value was re-calculated for different upstream sequence lengths (in increments of 10 bp up to 5000 bp).
Thus for each motif/transcription factor, a matrix is produced with 500 different upstream lengths times the number of all possible thresholds for the number of binding sites. One cell thus contains the p-value of a motif for a certain upstream length and a certain threshold of the number of binding sites**.**
The Enrichment score of a given transcription factor with respect to the cluster of BMP-target genes is determined by taking the minimal p-value of this matrix. This matrix can be visualized as was done for Runx1 in Figure 4.
For the enrichment of combinations of transcription factor binding sites (motif pairs), a gene is considered a hit if it has at least $T^b$ binding sites of both transcription factors within $T^l$ nucleotides upstream of transcription start.

## 4.5. Motif Discovery

The motif discovery tool PhyME [31] was used to find putative motifs of binding sites in the promoter sequences of genes that are presumable regulated by the same transcription factor(s). Our motif discovery experiments were performed on the cluster of BMP-target genes identified in Methods 4.2.
PhyME uses the conservation of binding sites in orthologs to discover motifs. See Suplement S6 for a detailed description of PhyME. Since we are looking for bone developmental transcription factors, the conserved binding sites should be present in all vertebrates. *Homo sapiens*, *Rattus norvegicus* and *Mus musculus* were chosen for this research. The identification of orthologs and the downloading of the sequences is described in Methods section 4.1.
PhyME needs to have the neutral mutation rates of these species compared to a common ancestor as an input. The authors of PhyME suggest the program fastDNAML [32] to determine the mutation rates of the orthologs used. The mutation rates for the used

vertebrate species were taken from the PhyME paper [31] and are depicted in Figure S6.
Once a cluster of genes is selected, the promoter region must be chosen. This is not easy for eukaryotes. The promoter region of the well known osteoblast differentiation gene, Runx2, was reported to be up to 4000 bp upstream of the transcription start site. The promoter length in our experiments was therefore chosen to be 5000 bp upstream of the transcription start site.
Alternatively, motif discovery experiments were performed on -5000 to 0 bp upstream regions where the exons of neighboring genes were removed that overlap with the promoter region under study. This was done because exons are generally better conserved during evolution than introns (except for binding sites). That might result in more false positives of binding sites in the exon sequences.
The desired length of the discovered motifs was chosen to be 10 nucleotides.
The discovered motifs were then compared with the curated JASPAR motifs, using the *"COMPARE custom profile to database profiles"* function of the JASPAR website [29]

## 4.6. Biological experiments: siRNA

*Cell Culture*

The MC3T3-E1 cell line was obtained from Riken institute (cell no. RCB1126) and is cultured in α-minimal essential medium, α-MEM from Bio Whittaker with 10% bovine calf serum (BCS), 2mM L-glutamin, and penicillin-streptomycin (2,000 units/ml Penicillin G and 2 mg/ml Streptomycin) at 37°C in 5%/95% air/$CO_2$ atmosphere. [33]

*siRNA Transfection of MC3T3 by electorporation*

After culturing and harvesting MC3T3-E1 cells (by trypsin) cells were centrifuged and resuspended to a final concentration of $5.0 \cdot 10^6$ cells/ml. For electroporation 400 µl cell suspension was added to a BioRad electroporation cuvet (4mm) and 10 µl 40µM siRNA duplex was added per cuvet. After gentle mixing, electroporation was performed (1000V, 50 µF and infinite resistance) in a Biorad Gene Pulser Xcell total system. Subsequently cells were re-seeded ($1.5 \cdot 10^6$ cells/ml) in 24-well's plates and incubated at 37°C in 5%/95% air/$CO_2$ atmosphere. The cell medium +/- BMP2 and 50 mg/l Ascorbic Acid was refreshed at day 1, day 4 (and day 7). [33]

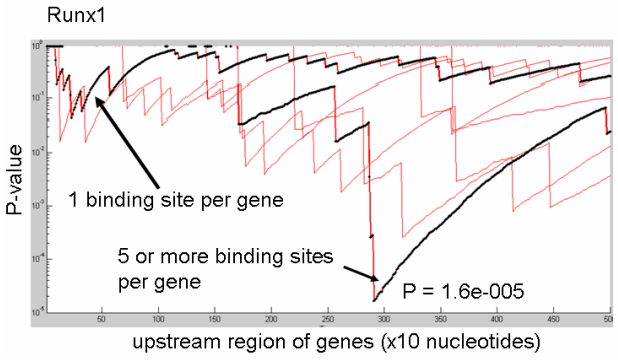



**Figure 4: Enrichment matrix representation: Each line represents the uncorrected p-values of the Runx1 TF for different upstream promoter lengths of a specific threshold of binding sites. The black lines show this for a threshold of one and five (or more) binding sites. The minimum uncorrected p-value is taken as the enrichment score (p = 1.6e-5)**

*Alkaline Phosphatase Activity Measurement (ALP)*

75 µl Lysisbuffer (100mM Potassium phosphate pH 7.8, 0.2% Triton-X) is added to the cells after removal of the medium. It is then incubated for 10 min at 37°C for optimal lysis. 10 µl lysate and 40 µl CDP-star is then added in an optiplate and incubated for 30 minutes in the dark at room temperature (CDP-Star, ready-to-use is an ultra-sensitive chloro-substituted 1,2-dioxetane chemiluminescent substrate for alkaline phosphatase, which exhibits extremely rapid light signal generation, Roche). The luminescence is then measured for 1.0 second with a Wallac Victor Multilabel counter 1420. [33]

*Acidic Phosphatase Activity Measurement*

The activity of alkaline phosphatase was corrected with acidic phosphatase activity to normalize variations in cells numbers. Nitrophenyl phosphate (pNPP) is added to the cell lyaste. pNPP is then converted to p-nitrophenol by cyto-acid phosphatase. The p-nitrophenol product absorbs light at 405 nm, and absorbance at this wave- length is monitored as a measure of cell number [34].
The preparation starts with the addition of 75 µl Lysisbuffer (100mM Potassium phosphate pH 7.8, 0.2% Triton-X) to the cells after removal of the medium. It is then incubated for 10 min at 37°C for optimal lysis. 5 µl lysate plus 100 µl P-nitrophenyl phosphate buffer (Add P-nitrophenyl phosphate tablet from Sigma to 15ml 0.1 M NaAc pH 5.5, 0.1% Triton-X) is put in 96-well's plate and is incubated for 1.5 hour in the dark at 37°C in 5%/95% air/$CO_2$ atmosphere.
The reaction is stopped with 10 µl 1M NaOH and incubated for 10 minutes at room temperature. The absorbance is measured at 405 nm with a Wallac Victor Multilabel counter 1420. [33]

## 5. References

1. Kingsley, L.A., J.M. Chirgwin, and T.A. Guise, *Breaking new ground to build bone.* Proceedings of the National Academy of Sciences, 2007. **104**(26): p. 10753.
2. Valcourt, U. and A. Moustakas, *BMP Signaling in Osteogenesis, Bone Remodeling and Repair.* European Journal of Trauma, 2005. **31**(5): p. 464-479.

3. Kurisaki, K., et al., *Nuclear Factor YY1 Inhibits Transforming Growth Factor ß-and Bone Morphogenetic Protein-Induced Cell Differentiation.* Molecular and Cellular Biology, 2003.
4. Yamaguchi, A., T. Komori, and T. Suda, *Regulation of Osteoblast Differentiation Mediated by Bone Morphogenetic Proteins, Hedgehogs, and Cbfa1*. 2000. p. 393-411.
5. Ryoo, H.M., M.H. Lee, and Y.J. Kim, *Critical molecular switches involved in BMP-2-induced osteogenic differentiation of mesenchymal cells.* Gene, 2005.
6. Lee, M.H., et al., *BMP-2-induced Runx2 Expression Is Mediated by Dlx5, and TGF-{beta} 1 Opposes the BMP-2-induced Osteoblast Differentiation by Suppression of Dlx5 Expression.* Journal of Biological Chemistry, 2003. **278**(36): p. 34387-34394.
7. Kim, Y.J., et al., *Bone Morphogenetic Protein-2-induced Alkaline Phosphatase Expression Is Stimulated by Dlx5 and Repressed by Msx2.* Journal of Biological Chemistry, 2004. **279**(49): p. 50773-50780.
8. Alliston, T., et al., *TGF-beta-induced repression of CBFA1 by Smad3 decreases cbfa1 and osteocalcin expression and inhibits osteoblast differentiation.* The EMBO Journal, 2001. **20**: p. 2254-2272.
9. Maeda, Y., et al., *Inhibitory helix-loop-helix transcription factors Id 1/Id 3 promote bone formation in vivo.* Journal of Cellular Biochemistry, 2004. **93**(2): p. 337-344.
10. Harris, S.E., et al., *Transcriptional regulation of BMP-2 activated genes in osteoblasts using gene expression microarray analysis: role of Dlx2 and Dlx5 transcription factors.* Front Biosci, 2003. **8**: p. s1249-65.
11. Miyazono, K. and K. Miyazawa, *Id: A Target of BMP Signaling*. 2002, American Association for the Advancement of Science.
12. Postigo, A.A., *Regulation of Smad signaling through a differential recruitment of coactivators and corepressors by ZEB proteins.* The EMBO Journal, 2003. **22**(10): p. 2453-2462.
13. Hata, A., et al., *OAZ uses distinct DNA-and protein-binding zinc fingers in separate BMP-Smad and Olf signaling pathways.* Cell, 2000. **100**(2): p. 229-40.
14. Ku, M., et al., *OAZ Regulates Bone Morphogenetic Protein Signaling through Smad 6 Activation.* Journal of Biological Chemistry, 2006. **281**(8): p. 5277.
15. Karaulanov, E., W. Knöchel, and C. Niehrs, *Transcriptional regulation of BMP4 synexpression in transgenic Xenopus.* The EMBO Journal, 2004. **23**: p. 844-856.
16. Postigo, A.A., *Opposing functions of ZEB proteins in the regulation of the TGFbeta/BMP signaling pathway.* The EMBO Journal, 2003. **22**(10): p. 2443-2452.
17. Takagi, T., *DeltaEF1, a zinc finger and homeodomain transcription factor, is required for skeleton patterning in multiple lineages*. 1998.
18. Ducy, P., et al., *Increased bone formation in osteocalcin-deficient mice.* Nature, 1996. **382**: p. 448-452.
19. Lazarova, D.L., M. Bordonaro, and A.C. Sartorelli, *Transcriptional Regulation of the Vitamin D3 Receptor Gene by ZEB 1.* Cell Growth and Differentiation, 2001. **12**(6): p. 319-326.
20. Yoshizawa, T., et al., *Mice lacking the vitamin D receptor exhibit impaired bone formation, uterine hypoplasia and growth retardation after weaning.* Nature Genetics, 1997. **16**: p. 391-396.
21. Morgan, M.J., et al., *YY1 Regulates the Neural Crest-associated slug Gene in Xenopus laevis.* Journal of Biological Chemistry, 2004. **279**(45): p. 46826.
22. Laux, H., et al., *Tumor-associated E-cadherin mutations do not induce Wnt target gene expression, but affect E-cadherin repressors.*
23. Meulemans, D. and M. Bronner-Fraser, *Gene-regulatory interactions in neural crest evolution and development.* Dev Cell, 2004. **7**(3): p. 291-9.
24. King, K.E., et al., *Kruppel-like Factor 4 (KLF4/GKLF) Is a Target of Bone Morphogenetic Proteins and Transforming Growth Factor ß1 in the Regulation of Vascular Smooth Muscle Cell Phenotype.* Journal of Biological Chemistry, 2003. **278**(13): p. 11661-11669.
25. Korchynskyi, O. and P. ten Dijke, *Identification and functional characterization of distinct critically important BMP-specific response elements in the Id1 promoter.* Journal of Biological Chemistry, 2001: p. 111023200.
26. Wingender, E., et al., *TRANSFAC: an integrated system for gene expression regulation.* Nucleic Acids Research, 2000. **28**(1): p. 316-319.
27. Hubbard, T., et al., *Ensembl 2005.* Nucleic Acids Res, 2005. **33**.
28. Kasprzyk, A., et al., *EnsMart: A Generic System for Fast and Flexible Access to Biological Data.* Genome Research, 2004. **14**: p. 160-169.
29. Sandelin, A., *JASPAR: an open-access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research, 2004. **32**(90001): p. 91-94.
30. Lenhard, B. and W.W. Wasserman, *TFBS: Computational framework for transcription factor binding site analysis.* Bioinformatics, 2002. **18**(8): p. 1135-1136.
31. Sinha, S., M. Blanchette, and M. Tompa, *PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences.* BMC Bioinformatics, 2004. **5**(1): p. 170.
32. Olsen, G.J., et al., *fastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood.* Bioinformatics. **10**: p. 41-48.
33. Dechering, K.J., *These protocols were worked on by: Ven - de Laat, J.J.M. van de (Cindy); Assink - van der Laak, J.E.H. (Rianne); Heerkens, R.J.H.*

*(Roland); Dechering, K.J. (Koen). The experiments were executed by: W. Beumer (Wouter).* 2006.
34. Yang, T.T., P. Sinai, and S.R. Kain, *An Acid Phosphatase Assay for Quantifying the Growth of Adherent and Nonadherent Cells.* Analytical Biochemistry, 1996. **241**(1): p. 103-108.
35. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nature Reviews Genetics, 2004. **5**(4): p. 276-287.
36. Vaes, B.L., et al., *Microarray analysis on Runx2-deficient mouse embryos reveals novel Runx2 functions and target genes during intramembranous and endochondral bone formation.* Bone, 2006.
37. Yamashiro, T., et al., *Expression of Runx1,-2 and-3 during tooth, palate and craniofacial bone development.* Mech. Dev, 2002. **119**: p. 107-110.

# 6. SUPPLEMENT DATA

## S1. Expression profiles

Figure S1 shows the expression profiles of the genes that were identified (according the procedure as described in section 4.2) as immediate-early BMP2 response genes.

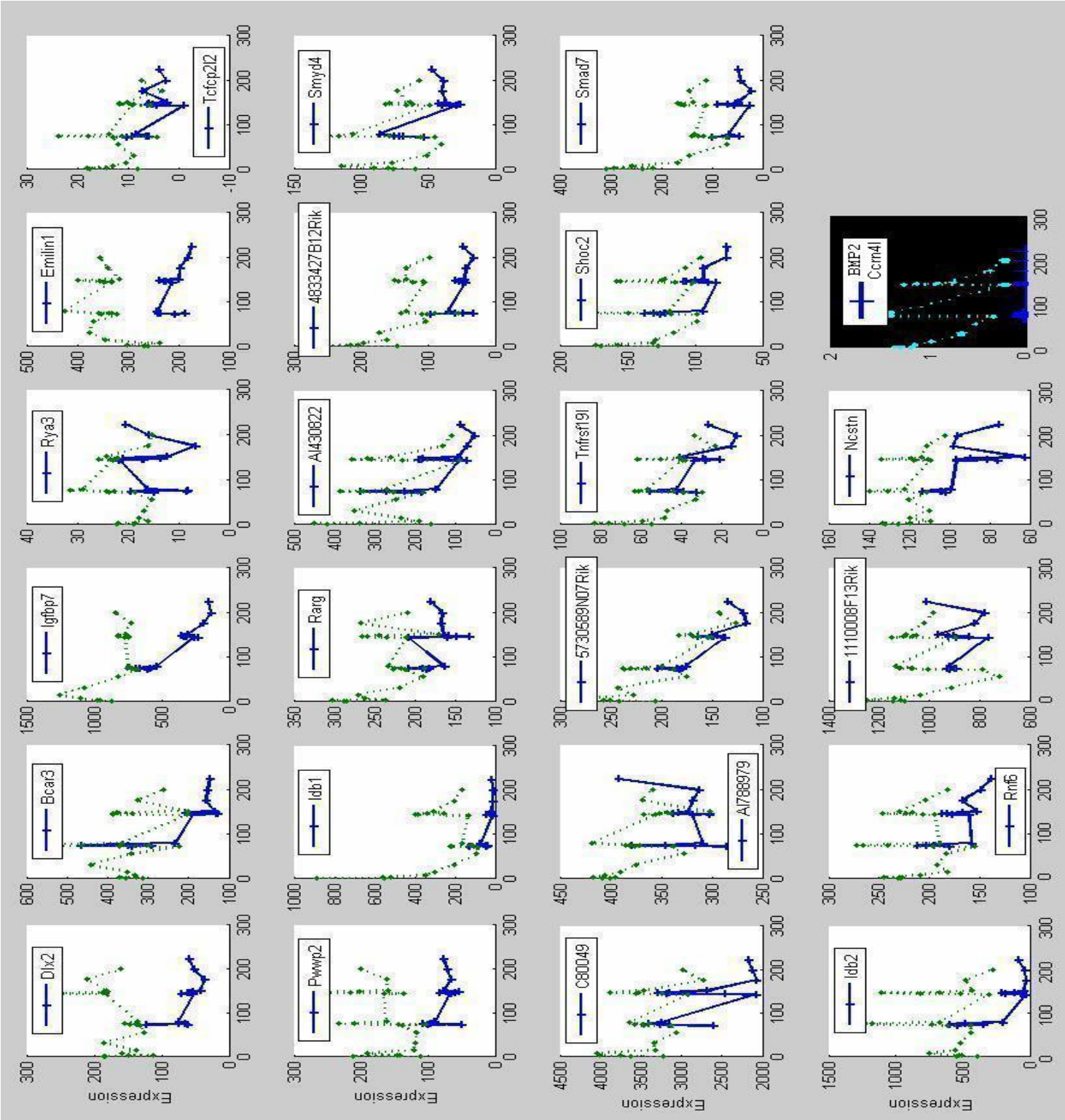



**Figure S1: Gene expression profiles of the immediate-early BMP2 response genes (BMP2 cluster). These expression profiles correlate the best with the BMP2 concentration during the experiment (which is the last graph with the black background).**

## S2. Position Weight Matrix

In order to be able to scan sequences for potential binding sites of a motif, Position Weight Matrices (PWM) need to be created.
Known PWMs can be downloaded from databases like JASPAR [29] or from literature. Putative PWMs are created by motif discovery programs like PhyME in the following way. The first step in creating a PWM is to align all its putative or known binding sites (with length $w$). Then the occurrences $f_{b,i}$ of the bases $b \in (A,C,G,T)$ per position $i$ need to be counted. These occurrences can be stored in a Position Frequency Matrix (PFM) of size $(b,w)$.
Now, the corrected probability $p_{b,i}$ of observing a given base on position $i$ can be calculated using the following formula.

$$p_{b,i} = \frac{f_{b,i} + s(b)}{N + \sum s(b')}, b' \in \{A,C,G,T\} \qquad (2)$$

Where $N$ is the number of aligned binding sites and $s(b)$ is the pseudo count function [35]. The pseudo count is added to correct for small samples and to eliminate null values before log conversion in equation(3. The corrected probabilities from equation (2 can now be converted into a Position Weight Matrix (PWM) or an Information Content Matrix (ICM). The PWM can be used for scanning sequences (see chapter 4.3) and an ICM is used to represent these probabilities in an intuitive way as a motif logo.
A Position Weight Matrix (PWM) can be calculated by dividing the corrected probabilities $p_{b,i}$ of base $b$ in position $I$ (from equation(2) by the expected background probabilities $p_b$ and converting the values to a log-scale [35]. $W_{b,i}$ is then the PWM value of base $b$ in position $i$.

$$W_{b,i} = \log_2 \frac{p_{b,i}}{p_b} \qquad (3)$$

The information content $D_i$ of position $i$ in an ICM is a measure of the nucleotide specificity in that position of the alignment and is a function (eq.(4) of the corrected probability $p_{b,i}$ of base $b$ [35].

$$D_i = 2 + \sum_b p_{b,i} \log_2 p_{b,i} \qquad (4)$$

Figure S3 depicts the logo of an example of a motif, based on equation (4).

## S3. Motif scanning

Sequences can be scanned with PWMs of transcription factor motifs to find potential binding sites. Having the PWM representation of a motif, a PWM score can be calculated for each position on a sequence.

$$S_n = \sum_{i=1}^{w} W_{(n+i),i} \qquad (5)$$

Where $S_n$ is the PWM score of a position n on a sequence with the same length $w$ as the PWM and $W_{(n+i),i}$ is the PWM value of PWM-position $i$ and the base on sequence-position $(n+i)$. During scanning with the Perl TFBS module scanning program, a PWM is slid from 5' to 3' along the positive and negative strand of a sequence in 1-bp increments and is evaluated for each increment with equation(5. Position $n$ on the sequence can be marked as a binding site by using a relative threshold $T$.

## S4. Hypergeometric test

For a given transcription factor, $T^b$ and $T^l$ , the enrichment can be quantified with a p-value, by using the hypergeometric distribution (eq. (6). This calculates the probability that the observed number of hits (or more) will be detected in a random cluster of the same size.

$$p = P(i \geq h) = \sum_{i=h}^{\min(C,H)} \frac{\binom{H}{i}\binom{G-H}{C-i}}{\binom{G}{C}} \qquad (6)$$

Where $G$ is the total number of genes, $C$ is the number of genes in the cluster, $H$ is the number of hits in all genes and $h$ is the number of hits in the cluster genes.

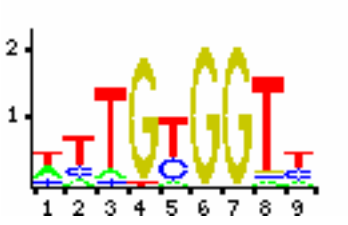



**Figure S3: Example of a motif logo (y-axis is the Information Content $D_i$, x-axis is the position $i$ )**

## S5. Segal Scoring as an alternative for determining TF enrichment

The Segal score (Eq. (7) combines all the PWM scores of a sequence and gives the most weight to the highest PWM scores (e.g. the best binding sites). The PWM score $S_i$ of every position on the sequence is first calculated using equation (5) (see Supplement S3).

$$Segal\_score = \zeta(\log(\frac{1}{n-p+1}\sum_{i=1}^{n-p+1}\exp\{S_i\})) \qquad (7)$$

The sigmoid function ξ (8) scales the log-function between 0 and 1.

$$\zeta(p) = \frac{1}{1+e^{-p}} \qquad (8)$$

The Segal score of a certain transcription factor is calculated for each gene in the genome. The distribution of Segal scores for a cluster of genes can be compared with the distribution of the background (genome), using a t-test (9).

$$T = \frac{\bar{x}-\bar{y}}{s\sqrt{\frac{1}{n}+\frac{1}{m}}} \qquad (9)$$

Runx2 cluster

| Motif name | p-value | Single Motif Enrichment top10 | Combi Motifs Enrichment top10 | Literature BMP | Literature SMAD | Literature Bone | Literature TGF-B | Bone phenotype Knockout |
|---|---|---|---|---|---|---|---|---|
| CFI-USP | 5,37E-04 | + | | | | | | |
| ZNF42_1-4 | 1,08E-03 | + | | | | | | |
| Bapx1 | 6,02E-03 | | | | | | | |
| Macho-1 | 1,01E-02 | | | | | | | |
| NF-Y | 1,24E-02 | | | | | + | | |
| Roaz | 1,27E-02 | | | + | + | + | | |
| NFKB1 | 1,33E-02 | | | | | | | |
| Staf | 1,39E-02 | | | | | | | |
| TFAP2A | 1,40E-02 | | + | + | | + | + | |
| Klf4 | 1,62E-02 | | + | | | | | |

BMP response cluster

| Motifname | p-value | Single Motif Enrichment top10 | Combi Motifs Enrichment top10 | Literature BMP | Literature SMAD | Literature Bone | Literature TGF-B | Bone phenotype Knockout |
|---|---|---|---|---|---|---|---|---|
| 'ABI4' | 3,85E-07 | | | | | | | |
| 'TFAP2A' | 7,90E-07 | + | + | + | | + | + | |
| 'CFI-USP' | 3,36E-05 | | | | | | | |
| 'Roaz' | 2,53E-04 | + | + | + | + | + | | |
| 'ESR1' | 6,78E-04 | | | | | + | | |
| 'Macho-1' | 9,41E-04 | + | + | | | | | |
| 'MAX' | 4,69E-03 | + | | | + | | + | |
| 'TCF11-Ma | 4,76E-03 | | | | | | | |
| 'Mycn' | 5,00E-03 | | | | | | | |
| 'Arnt' | 5,16E-03 | | | | | | | |

**Table S5: Top 10 Segal Scoring for the Runx2 and BPM2 response clusters**

However, a more advanced t-test was used, that assumes that the distributions have unknown and possibly unequal variances. The Satterthwaite's approximation for the effective degrees of freedom was used to solve the Behrens-Fisher problem. This t-test gives a p-value to the equality of the two distributions.
The t-test on the Segal scores (7) is a measure for the specificity of a transcription factor for a certain set of genes (cluster). A general TF, like the TATA-box, should have a similar amount of binding sites in the cluster as in the background set. The distribution of the Segal scores in the cluster is therefore not different from the distribution of the background.
Specific transcription factors for this cluster will have more binding sites than expected from the background and the distribution of the cluster will therefore have a different mean than the background distribution.
The advantage of using this measure is that this score is independent of the length of the sequence. Promoters with different sequence lengths can therefore be compared.

The ranking based on the Segal scores was compared to the enrichment method of the main paper in the third and fourth column of Table S5. It was found that this method has no added value over the enrichment method. Therefore we used the more common approach of the hypergeometric test as enrichment score in our analysis.

## S6. PhyME

*Design assumptions*

PhyME is based on two main assumptions. The first main assumption is that a cluster of genes with similar expression patterns can be regulated by the same transcription factor or the same group of transcription factors. The binding sites of a common transcription factor are then overrepresented in the sequences of this cluster of genes.

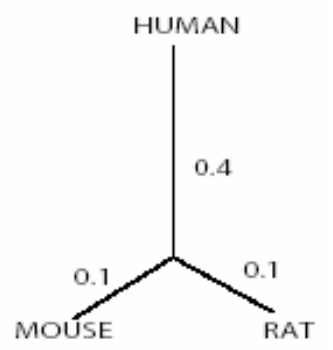



**Figure S6: Phylogenetic tree used [31]; The numbers are mutation rates of each species compared to the hypothetical common ancestor.**

The second main assumption is that the stability of transcription factor binding sites is of high importance during evolution. If these sites were unstable, then critical processes like bone formation in vertebrates could be easily compromised. The assumption is therefore that the mutation rate of binding sites is lower than the rest of the intergenic region. The most critical binding sites should be preserved in all vertebrate species.

*Non-Evolutionary HMM*

The first main assumption about overrepresentation is used by PhyME in a Hidden Markov Model (HMM). The basic HMM in PhyME can be easier explained by excluding the evolutionary model at first. The evolutionary model will be included in the next paragraph.

The Hidden Markov Model is used to generate new sequences based on probabilities. It starts with sampling a nucleotide with a certain emission probability from the background weight matrix $W_b$ (matrix with length 1). Then a choice is made to sample from the background weight matrix again or to start sampling from all the positions in the motif weight matrix $W_m$ (matrix with length ~10). This choice is expressed in the transition probability *p.* The probability of sampling from $W_m$ is $p_m = p$. The probability of sampling from $W_b$ is $p_b = 1 - p$. After the sampling of nucleotides from the motif weight matrix has finished, the model will either choose again between the weight matrices based on the transition probabilities.

This HMM is trained to explain the input sequence as best as possible. The parameters $\theta$ that are being optimized are the motif weight matrix $W_m$ and the transition probability $p$. The background weight matrix is calculated by PhyME based on the nucleotide frequencies in the input sequence $S$. This parameter (also part of $\theta$) is therefore not optimized.

The training of the HMM parameters $\theta$ is done by an Expectation Maximization algorithm. The factor that is being maximized during this training is the likelihood ratio $F$. This ratio measures how much more likely it is that $S$ was generated using the motif weight matrix, than without it ($\theta_b$ just contains the background weight motif).

$$F(S,\theta) = \log \frac{\Pr(S \mid \theta)}{\Pr(S \mid \theta_b)} \qquad (11)$$

Expectation Maximization algorithm is used to optimize the HMM parameters (only $W_m$ and $p$).
This algorithm is only guaranteed to converge to a local optimum. PhyME therefore executes the motif finding step a number of times. A randomly chosen substring from the input sequence is used each time as a seed for the motif weight matrix. The E-M procedure on these seeds is then cut off after a small number of iterations and the seed with the greatest

score *F* is used to run the E-M to convergence.

*Evolutionary HMM*

PhyME uses the second main assumption by searching for small aligned sequences in orthologous genes (genes with high similarity in different species). If critical binding sites are preserved during evolution in multiple species, then they must be part of the aligned sequences of these species. As this research is focused on osteoblast differentiation, these species are chosen from the vertebrate group. For this research they are *Homo sapiens*, *Rattus norvegicus* and *Mus musculus*.
PhyME uses the alignment tool LAGAN to find aligned sequences of multiple species with a minimum length of the specified binding site length and with a minimum degree of similarity.
PhyME then uses the construct Ψ instead of the sequence *S* for the training of the HMM. This construct consists of the complete sequence of the reference species (*Mus musculus*) and the aligned sequences of the orthologs on the positions where that is possible.

PhyME takes the phylogenetic distance between the used species into account. This is done by using the user input of the neutral mutation rates. The following equation is used to describe the evolutionary model and is used in the evolutionary likelihood ratio *F*.

$$\Pr_e(\psi \mid W, k) = \sum_{\alpha \in \sum} W_{k\alpha} \prod_{S_\sigma \in \psi} (\mu_\sigma W_{kS_\sigma} + (1-\mu_\sigma)\delta_{\alpha S_\sigma} \qquad (12)$$

Where $s_\sigma$ is the nucleotide from species σ in alignment Ψ, $\delta_{xy} = 1$ if $x = y$ and 0 otherwise. $\mu_\sigma$ is the neutral mutation probability.
"For the position *k*, one "creates" a base *α* in the ancestor with frequency $W_{k\sigma}$ and each such base is either passed unchanged to the species σ (probability 1 - $\mu_\sigma$) or mutated in species σ with probability $\mu_\sigma$ and a new base selected with a frequency defined again by *W*."

## S7 Runx2 Knock Out cluster

The scanning and enrichment calculation method was applied again on a cluster of differently expressed genes from a Runx2 deficient mouse compared to wild-type to show that the approach is general. These 25 differently expressed genes are assumed to be directly or indirectly regulated by the transcription factor Runx2 (see Table) [36]. Runx2 is known to be a central transcription factor in the osteoblast differentiation pathway. Our method was used to see if other transcription factors also regulate this cluster. A literature search was then performed to see if the transcription factors with the highest enrichment are mentioned in relation with bone growth. The top 10 results of the combination enrichment calculation and the literature search are shown in Table.
The literature results show that many enriched transcription factors are mentioned in literature in relationship with bone development.
Runx2 itself is not part of the JASPAR database. It can be seen, however, that Runx1 is part of the results in Table. Runx1 is highly similar to Runx2 [37] and is also the most enriched single transcription factor in the Runx2 cluster (results shown in Table).

**Table S7a: RUNX2 Knock Out Cluster Top 10: Enrichment of Transcription Factors pairs.**

| TF 1 | TF 2 | Minimal uncorrected p-value | Related to bone in literature: TF 1 | TF 2 | Transcription Factors with relationship to species or tissue (based on literature). |
|---|---|---|---|---|---|
| GABPA | c-ETS | 2,83E-07 | - | + | GABPA: Adipocyte diff |
| GATA3 | Klf4 | 2,24E-06 | + | + | |
| RUNX1 | Broad-complex_4 | 2,35E-06 | + | - | Broad-compl: Insect |
| TCFAP2A | MNB1A | 5,29E-06 | + | - | MNB1A: Plant |
| MYB.ph3 | Prrx2 | 1,40E-05 | - | + | MYB.ph3: Plant |
| ZNF42_5-13 | YY1 | 1,42E-05 | - | + | ZNF42_5-13: Blood |
| RUNX1 | HMG-1 | 1,57E-05 | + | - | HMG-1: Plant |
| TCFAP2A | Dof3 | 1,75E-05 | + | - | Dof3: Plant |
| GATA3 | MafB | 2,36E-05 | + | + | MafB: Cartilage |
| YY1 | MafB | 2,47E-05 | + | + | MafB: Cartilage |

| Rank | Transcripton factor | Related to bone in Literature | Enrichment score Minimal uncorrected p-value | Minimum number of binding sites | Upstream lenght | Foreground hits | Background hits | | | Gene name and ENSEMBL ID |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | RUNX1 | + | 1,66E-05 | 5 | 2910 | 5 | 365 | | Tubb5 | ENSMUSG00000001525 |
| 2 | ZNF42_1-4 | | 6,20E-04 | 3 | 1110 | 22 | 15433 | | Hck | ENSMUSG00000003283 |
| 3 | Arnt-Ahr | | 6,77E-04 | 2 | 4830 | 24 | 18645 | | Knsl8 | ENSMUSG00000003546 |
| 4 | Dof3 | - | 1,08E-03 | 7 | 400 | 4 | 513 | | Clec11a | ENSMUSG00000004473 |
| 5 | TFAP2A | + | 1,32E-03 | 26 | 4830 | 6 | 1409 | | | ENSMUSG00000016995 |
| 6 | CFI-USP | - | 1,42E-03 | 2 | 2770 | 4 | 552 | | Mmp9 | ENSMUSG00000017737 |
| 7 | MAX | | 1,44E-03 | 1 | 950 | 9 | 3240 | | Gtpbp2 | ENSMUSG00000023952 |
| 8 | MafB | + | 1,80E-03 | 3 | 1650 | 15 | 8312 | | Pim1 | ENSMUSG00000024014 |
| 9 | MYC-MAX | | 1,81E-03 | 6 | 3330 | 1 | 2 | | Snf1lk | ENSMUSG00000024042 |
| 10 | SPIB | | 2,09E-03 | 5 | 2230 | 23 | 17966 | | Q8C6A9_MOUSE | ENSMUSG00000024160 |
| | | | | | | | | | Col9a1 | ENSMUSG00000026147 |
| | | | | | | | | | Akp2 | ENSMUSG00000028766 |
| | | | | | | | | | Pogz | ENSMUSG00000029304 |
| | | | | | | | | | Ibsp | ENSMUSG00000029306 |
| | | | | | | | | | Dlx5 | ENSMUSG00000029755 |
| | | | | | | | | | Agc1 | ENSMUSG00000030607 |
| | | | | | | | | | Cfh | ENSMUSG00000033898 |
| | | | | | | | | | Satb2 | ENSMUSG00000038331 |
| | | | | | | | | | | ENSMUSG00000039153 |
| | | | | | | | | | Ndufb10 | ENSMUSG00000040048 |
| | | | | | | | | | BC048355 | ENSMUSG00000040658 |
| | | | | | | | | | | ENSMUSG00000058330 |
| | | | | | | | | | Ppp2r5d | ENSMUSG00000059409 |
| | | | | | | | | | Fdps | ENSMUSG00000059743 |
| | | | | | | | | | Rpl7l1 | ENSMUSG00000063888 |

**Table S7b: Best ranking single transcription factors for the Runx2 KO cluster. The grey cells indicate if a Runx2 KO cluster gene is "hit" by that transcription factor (for the indicated minimum amount of binding sites and upstream length).**