

Extensive transcriptional complexity during hypoxia-regulated expression of the myoglobin gene in cancer

**Anne Bicker¹, Dimo Dietrich², Eva Gleixner^{1,3}, Glen Kristiansen², Thomas A. Gorr^{4,5},
Thomas Hankeln^{1,*}**

¹Institute of Molecular Genetics, Johannes Gutenberg University, J.J. Becherweg 30a,
D55099 Mainz, Germany

²Institute of Pathology, University Hospital Bonn, Sigmund-Freud-Str. 25, 53127 Bonn,
Germany

³present address: Center for Systems Biology, University of Freiburg, 79104 Freiburg,
Germany

⁴Center for Pediatrics and Adolescent Medicine, Clinic IV, Div. of Pediatric Hematology and
Oncology, University Medical Center Freiburg, Mathildenstrasse 1, 79106 Freiburg, Germany

⁵Institute of Veterinary Physiology, Vetsuisse Faculty, University of Zurich,
Winterthurerstrasse 260, 8057 Zurich, Switzerland

*Corresponding author. Institute of Molecular Genetics, Johannes Gutenberg University, J.J.
Becherweg 30a, D55099 Mainz, Germany; Phone: +49 6131 39 23277; Fax: +49 6131 39
24585; Email: hankeln@uni-mainz.de

ABSTRACT

Recently, the ectopic expression of myoglobin (MB) was reported in human epithelial cancer cell lines and breast tumor tissues, where MB expression increased with hypoxia. The better prognosis of MB-positive breast cancer patients suggested that the globin exerts a tumor-suppressive role, possibly by impairing mitochondrial activity in hypoxic breast carcinoma cells. In order to better understand *MB* gene regulation in cancer, we systematically investigated the architecture of the human *MB* gene, its transcripts and promoters. *In silico* analysis of transcriptome data from normal human tissues and cancer cell lines, followed by RACE-PCR verification, revealed seven novel exons in the *MB* gene region, most of which are untranslated exons located 5'-upstream of the coding DNA sequence (CDS). Sixteen novel alternatively spliced *MB* transcripts were detected, most of which predominantly occur in tumor tissue or cell lines. Quantitative RT-PCR analyses of *MB* expression in surgical breast cancer specimen confirmed the preferential usage of a hitherto unknown, tumor-associated *MB* promoter, which was functionally validated by luciferase reporter gene assays. In line with clinical observations of MB upregulation in avascular breast tumors, the novel cancer-associated *MB* splice variants exhibited increased expression in tumor cells subjected to experimental hypoxia. The novel gene regulatory mechanisms unveiled in this study support the idea of a non-canonical role of MB during carcinogenesis.

INTRODUCTION

Myoglobin (MB) was the first oxygen-binding respiratory protein described to appear in extravascular locations (1; 2). It is mainly expressed in skeletal and cardiac muscle of vertebrates at high micro- or millimolar concentrations (3; 4, 5). Due to its high O₂-binding affinity (P₅₀ ~1 Torr), MB supports the intracellular diffusion of O₂ from the sarcolemma to the mitochondria (6; 5). In addition, MB functions in short-term O₂ storage in contracting muscles for a steadier O₂ supply to the mitochondria (4). Upon cellular hypoxia, MB has been described to decrease the level of reactive oxidative stress and to detoxify harmful excess nitric oxide (NO[•]) and nitrite (7; 8; 9; 10).

MB expression was further reported in various human malignant tumors and tumor cell lines that did not originate from myocytic precursors, including acute leukemia, desmoplastic small round cell tumors, breast carcinoma, colon carcinoma and non-small cell lung cancer (11; 12; 13; 14; 15). In example, 40% of invasive breast carcinomas were found to exhibit pronounced MB expression (14). Comparing *MB* levels in breast tumor biopsies and healthy ductal tissue counterparts from the same individuals revealed on average a 350-fold higher abundance of *MB* transcripts in the cancer tissues (14). In a corresponding Kaplan-Meier analysis, the overall survival of 917 primary breast cancer cases was monitored and correlated to their MB and estrogen receptor alpha (ER α) status. Despite the fact that patients with ER α -positive tumors generally have a better outcome, and that mammary carcinomas displayed a tight association between the presence of MB and ER α , we also documented the favorable prognosis of ER α -deficient breast tumors with high MB content (ER-/MB+) relative to ER α -deficient/MB-deficient (ER-/MB-) entities (14). Thus, MB is of prognostic value and might confer tumor suppressive activity *in vivo*.

To investigate the possible function of MB in a tumor context, lentiviral gene transfer was employed to overexpress mouse Mb in the MB-null A549 human lung cancer cell line (16). Resulting experimental tumors expressing ectopic Mb displayed reduced hypoxia,

minimal levels of the O₂-sensitive alpha subunit of the hypoxia inducible factor (HIF-1 α), lower vessel density along with a more differentiated cancer cell phenotype, and largely suppressed metastatic spreading. The authors correlated these beneficial outcomes of Mb overexpression with the reduction of tumor hypoxia (16). However, this approach seems somewhat artificial, since the total amount of Mb in the engineered cancer cells was similar to MB levels in human muscle (i.e. [MB] ~200-300 μ M), while endogenously occurring MB protein levels in breast cancer cells are several hundred-fold lower (14). Therefore, it remains doubtful whether MB in tumors is sufficiently abundant to confer meaningful O₂ supply capacity and to maintain aerobiosis in temporarily hypoxic cancer cells. In line with a negligible role of endogenous MB in O₂ transport in breast tumor cells, high-resolution respirometry on stable *MB* knockdown clones of suspended MDA-MB468 breast cancer cells revealed, when compared to *MB* proficient controls, statistically indistinguishable *in vitro* P₅₀(O₂) values (17). Apparently, endogenous MB in breast tumors and breast cancer cells may thus exert a non-classical molecular function. Along this line, it has been proposed that MB may regulate oxidative/nitrosative stress in cancer cells, e.g. by its ability to scavenge toxic NO[•] under hypoxic conditions (13). Our own experiments using siRNA-driven *MB* silencing in MDA-MB468 breast cancer cells indicated that MB interferes with the capacity of the mitochondria specifically in hypoxic carcinoma cells. MB was additionally found to influence proliferation and motility of these cells, yet in ways not directly related to the facilitated diffusion or storage of O₂ (17).

The attractive hypothesis that MB might have tumor-suppressive properties raised the question for the gene regulatory mechanisms driving the transcription of the globin in cancer cells. MB immunostaining on breast ductal carcinoma in situ (DCIS) entities revealed a gradient of MB expression, which correlated with hypoxia markers (17). Several publications also reported the robust increase of *MB* concentrations by transcriptional upregulation in epithelial cancer cells subjected to hypoxia (13; 14; 15). Analogous to the functional

discrepancies between carcinoma and myocytic MB, this hypoxia-mediated increase in cancer MB expression is also at odds with known regulatory mechanisms of MB in muscles.

According to the current view, hypoxia acts only as an effective regulatory stimulus of MB expression in muscle when it is accompanied by other triggers (such as exercise) (18). This observation is concordant with the fact that the standard muscle MB promoter region apparently lacks canonical HIF-binding *cis*-elements (i.e. hypoxia response elements, HREs) and hypoxia inducibility (19). Recently however, one alternative MB transcript (NM_203377) was reported to originate from a different, non-standard promoter site, which appeared predominantly active in the breast cancer cell line MDA-MB468 (17). This alternative MB mRNA was 300-fold increased over the standard transcript in the breast cancer cell line. Moreover, the alternative mRNA was 2.2-fold induced in cells challenged by prolonged hypoxia, compared to normoxic samples. However, it remained unclear whether the novel transcript occurs also in clinical specimen *in vivo* or in other cell lines. In addition, a novel candidate HRE, located 2.8 kb upstream of the translation start site (TSS) of the MB gene was identified, which enhanced reporter gene transcription by ~43% in hypoxic MDA-MB468 cells. Apparently, this element is therefore only a minor contributor to hypoxic MB induction (17).

In the present study, we systematically characterized the spectrum of MB transcript variants and their initiation sites by transcriptome sequence data mining and experimental verification. We evaluated the expression levels of novel, alternative MB transcripts in various human tissues and cancer cell lines by quantification of RNA-Seq data. To account for mouse as a model organism, we examined conservation of the human and mouse *Mb* gene structures. Expression rates of MB transcript variants were compared in normoxic and hypoxic cultures of a breast and a colon cancer cell line to elucidate how diminished O₂ supply impacts on MB gene regulation in different cancer backgrounds. The biological relevance of cancer-associated, hypoxia-inducible MB transcripts was then validated in breast cancer biopsies, and

the functionality of a novel promoter was studied by dual luciferase reporter assays (DLRAs). To identify functional differences between the cancer-linked *MB* transcript variants, the influence of an alternatively spliced upstream open reading frame (uORF) on globin translation efficiency was examined.

RESULTS

A revised structure of the human *MB* gene

To identify the full spectrum of alternative transcripts from the human *MB* gene, we first mined available *MB* transcriptome data from the NCBI UniGene EST (= expressed sequence tag) database (for a list of representative ESTs supporting alternative *MB* transcripts, see Tab. S1). RT-PCR reactions using cDNA from the breast and colon cancer cell lines MDA-MB468 and DLD-1, respectively, were then performed to verify the *in silico* predicted utilization of *MB* exon sequences (Tab. S2). Amplicons were sequenced to assemble a list of novel exons and splice variants of the human *MB* gene (Fig. S1, Fig. 1). EST and RT-PCR data from cancer cells revealed six novel exons located within the 5'UTR (= untranslated region) of the human *MB* gene and one alternative noncoding start exon located between the first and the second CDS-containing exon (Fig. 1). Considering all known and newly annotated exons, the genomic sequence of human *MB* now encompasses 31796 bp (coordinates in genome build 37/hg19 are: q12.3 Chr 22: 36,002,811-36,034,607), thus about doubling the size of the current human *MB* gene annotation in the NCBI database (Fig. S1, Fig. 1). Exons were numbered in 5'>3'-direction, tagging untranslated exons with a "u" and exons that encode at least part of the CDS, with a "c". The three long-known CDS-containing exons 1 to 3 were thus renamed as exons 9c, 11c and 12c, respectively (20).

From the newly detected exons, we were able to infer 16 novel splice variants in addition to three previously annotated ones (Fig. S1). Variant 2 corresponds to the NCBI-

annotated transcript NM_005368.2 (AceView transcript b) and the first published *MB* structure (20). Variant 13 equals NM_203377.1 (AceView transcript d; also see (17)) and variant 5 has been annotated as NM_203378.1 (AceView transcript c). According to EST data, these previously annotated transcripts are the most common ones. The CDS-encoding exons 11c and 12c were present in all known and novel splice forms, and most variants also contained exon 9c with the canonical start codon. Thus, 9 out of 19 transcript variants encode the standard MB protein as found in muscle, whereas all others translate into either no or a truncated MB protein. All introns are equipped with consensus splice sites.

The integrity of the novel protein encoding *MB* transcripts was confirmed via 5' and 3' RACE, using total RNA from the MDA-MB468 cell line. In the case of transcript variants 9, 10 and 11, the sequences resulting from 5'RACE reactions suggested the first exon (exon 5u) to start 25 bp upstream of the *MB* transcript version annotated by AceView. The splice variant 13, starting with exon 4u, was identified to contain an additional 15 bp as compared to the AceView annotation. We further approved the integrity of variant 19, starting at exon 1u, via 5' RACE (Fig. S1).

Phylogenetic conservation of the *MB* gene

A search for evolutionary conserved regions within the *MB* gene was conducted to reveal stretches of most important biological function. BLASTN pairwise local alignments of human *MB* and the respective gene regions from other species (Tab. S3) showed that exon 9c displays the highest sequence conservation. This conservation ranged from 100% to 76% identity to *MB* of chimp and chicken, respectively. The other CDS-encoding exons 11c and 12c were also highly conserved in mammals and birds with percent identities ranging from 99% to 70%. Regarding the noncoding 5'UTR, the alignments showed a gradual exon conservation between vertebrate species (Tab. S3). Exons located proximal to the CDS of the

globin (such as 8u) tended to show higher conservation than distant exons, suggesting a lower selection pressure on the 5' region of the *MB* gene in evolution.

Due to the importance of mouse (*Mus musculus*) as the prime mammalian model organism for biomedical research, a comparison of human and mouse *Mb* gene structures (coordinates in mouse genome build 38/mm10 are: qD3 Chr15:77,015,487-77,050,668) for synteny was of particular interest. For this, all mouse ESTs available on NCBI-Unigene were annotated in terms of exon presence and tissue type. Equivalent to the human *MB* gene, mouse *Mb* exons were renamed numerically in 5'>3' orientation. Two novel upstream exons and four additional splice variants were detected in mouse (Fig. S2b). However, conservation amongst vertebrates turned out to be very limited: all CDS-containing exons and human exon 8u, equivalent to exon 4u in mice (Fig. S2b) showed the (expected) sequence conservation, whereas all other human and mouse *Mb* upstream noncoding exons were not alignable and are thus not conserved between the two taxa (Fig. S2a). A reason might be that ~56% of the genomic area encompassing the 5'UTR of human *MB* consist of repetitive elements, such as Alu and L1 elements typical for primate genomes (Fig. S3).

Tissue-specific expression of *MB* transcript variants

To identify tissue-specific *MB* splicing patterns, all ESTs (n = 972) were assigned to their tissues of origin (Fig. S1). Expectedly, the standard *MB* transcript (variant 2), represented by 557 ESTs, was most prominently seen in skeletal and heart muscle, and to a much lesser extent in smooth muscle tissues. At least 8 of the 16 novel transcripts were also expressed in muscle tissues, but are only represented by very few EST reads (n = 13). Fifteen EST entries, corresponding to 8 *MB* transcript variants, were observed in datasets from a large variety of non-muscular tissues, including liver, lung, prostate, testis and sciatic nerve (Fig. S1). In cancer cell lines, EST analysis (n = 8 reads) and RT-PCR experiments demonstrated the

presence of 12 out of the 19 *MB* splice variants. As suggested by our previous study (17) and now confirmed by EST data, transcript variant 13 appeared to be one of the major transcripts expressed in cancer cell lines. In addition, three other protein-coding *MB* transcripts were found in the breast and colon cancer cell lines, i.e. variants 9 (AceView transcript a), 10 and 11 (AceView transcript f), which all have a common transcriptional start at exon 5u (Fig. 1). The *MB*-coding variant 19 was only detectable in MDA-MB468 cells, but missing in DLD-1 cells.

To infer relative expression profiles of different body tissues and cancer cell lines, we additionally analyzed 41 RNA-Seq datasets from *MB* expressing human tissues and cancer cell lines (Tab. S4). Sequence reads were mapped against the human *MB* CDS-containing exons and the 5'UTR starting exons, quantified and normalized. Transcription of *MB*-CDS encoding exons was on average 25-fold and 333-fold higher in heart muscle than in normal breast and colon tissue, respectively (Fig. 2). We noted that the mRNA expression measured in breast tissues appeared higher than expected from protein studies (14), and might therefore be partly due to contamination with myocytes. In the cancer cell lines MDA-MB468, MCF-7, LNCaP and DLD-1, average CDS-exon expression was about 115-fold lower than in heart tissue (Fig. 2).

In addition, great differences were noted in the utilization frequency of the alternative 5'UTR starting exons among the investigated datasets. As expected from our EST analyses, starting exons 8u and 9u are predominantly used in muscle and heart, whereas upstream exons 4u and 5u are distinctly underrepresented in non-cancerous tissues (Fig. 2). For instance, in human heart those transcripts starting at exon 9u are about 800 times more abundant than those starting with exon 5u (Fig. 2). In contrast, cancer cell lines such as MB468, DLD-1 and LNCaP as well as prostate tumor biopsies only express *MB* transcripts starting at upstream exons 4u and 5u, yielding mRNA variants 9, 10, 11 and 13 (Fig. S1). Since there were no

reads mapping to the start exons 1u, 2u and 10u in all analyzed transcriptomes, we do not expect their associated *MB* transcripts to typically appear in cancer cells.

Quantification of *MB* transcript variants in cancer cell lines

In silico analyses of RNA-Seq datasets (Tab. S4) and analyses of cDNAs from MDA-MB468 and DLD-1 cells using exon junction-spanning primers to specifically distinguish between splice variants revealed that the protein-coding *MB* mRNA variants 9, 10, 11 and 13 are present in cancer cell lines. Therefore, these mRNAs might be crucial for the expression of functional MB in tumor tissues. Based on this idea, we quantified the expression of these *MB* transcripts in MDA-MB468 cells and in DLD-1 cells via qRT-PCR. Compared to the standard muscle transcript (variant 2), normoxic MDA-MB468 cells produced 374 to 708 fold higher quantities of the alternative *MB* transcript variants 9, 10 and 11 (Fig. 3). Even transcript variant 13 is significantly more strongly expressed than variant 2 in the breast cancer cell line, although it accounts for only 2% of all transcripts in MDA-MB468 cells. Since the amount of splice variants 9, 10 and 11 greatly exceeded that of variant 2 also in the colon cancer cell line DLD-1, these mRNAs appear to dominate the transcriptional profile of *MB* in a more widespread fashion in epithelial cancer cells.

Hypoxia induction of *MB* transcripts in cancer cell lines

Our previous work pointed to tissue hypoxia as a contributing stimulus of MB expression in DCIS entities (17). We also noted that transcripts corresponding to the cancer-associated *MB* splice variant 13 (starting with exon 4u) were 2.2-fold upregulated in MDA-MB468 cells subjected to 1% O₂ for 72h (17). Equivalent to those experiments, we now compared the abundance of novel, protein-coding cancer-associated *MB* splice forms among the cDNA samples of MDA-MB468 cells and DLD-1 cells raised under hypoxic (72h, 1% O₂) and normoxic (72h, air) conditions via qRT-PCR. Relative to normoxic oxygen supply, transcripts

variants 9, 10 and 11 showed a statistically significant 3- to 3.5-fold increase in MDA-MB468 and a 1.5- to 2.1-fold up-regulated expression in DLD-1 cells in response to oxygen deprivation (Fig. 3). As a result, splice variants 9, 10 and 11 (all transcribed from exon 5u) most likely underlie the hypoxia-mediated increase of *MB* in breast and colon cancer cell lines (see (17) for Western Blot evidence).

Detection of *MB* transcript variants in breast cancer tissues

To prove the relevance of alternative *MB* transcripts in clinically relevant human breast cancer specimens, we also investigated biopsies of 21 breast cancer patients. Following surgical resection of the tumors, tissue specimens were first tested for *MB* positivity by immunostaining. After RNA extraction and cDNA synthesis, the four protein-encoding *MB* splice forms most commonly found in cancer cell lines and the standard transcript variant 2 were quantified by qRT-PCRs in each tumor sample. In line with our *in vitro* data of breast and colon cancer cell lines, the *MB* transcripts 9, 10 and 11 also represented the dominant *MB* mRNA fraction in the breast cancer cases (Fig. 4a). Compared to the sum of variants 9, 10 and 11, expression of the *MB* standard transcript variant 2 was on average 14-fold suppressed in tumors. The vast majority of the cancer cases expressed variant 9 at maximal quantities, followed by variant 11 and variant 10. In conclusion, the promoter upstream to their common start exon 5u is the most active driver of *MB* transcription in breast tumors, while the muscle-associated standard promoter most likely plays only a minor role. The *MB* splice variant 13, previously described in the breast cancer cell line MDA-MB468 (17), was weakly expressed in tumor specimens. Compared to the ~ 26-fold more abundantly present variants 9, 10 and 11 (Fig. 4a), variant 13 is probably of minor importance in an *in vivo* context.

Hypoxia inducibility of *MB* transcripts in breast cancer tissues

We have shown so far that the novel *MB* transcript variants 9, 10 and 11 are inducible in hypoxic cancer cell lines. In addition, variant 13 was earlier reported to be upregulable in breast cancer cells upon challenge with 1% O₂ (17). In order to correlate the expression of *MB* transcripts with the degree of hypoxia in tumor tissues, we further quantified mRNA levels of the hypoxia-specific biomarker vascular endothelial growth factor A (*VEGFA*) by qRT-PCR. A significant positive Pearson's correlation between *VEGFA* mRNA increase upon hypoxia stimuli and enhanced expression of the *MB* variants 9, 10, 11 ($p < 0.001$) and variant 13 ($p < 0.001$) under hypoxia was found. In addition, the hypoxia induction of *MB* variant 13 correlated strongly and positively with the hypoxia response of the variants 9, 10 and 11 (Pearson, $p < 0.001$). However, no significant correlation was found between the response of *MB* variant 2 to hypoxia and the hypoxia inducibility of *VEGFA* ($p = 0.7$).

In addition, the degree of hypoxic stress was also inferred by performing carbonic anhydrase IX (CAIX) antibody staining in tumor tissue sections adjacent to the ones used for RNA preparation, as the HIF-1 α -2 α target gene *CAIX* is a widely used marker of local tissue hypoxia in solid malignancies. Most tumors with strong *MB* transcription were also CAIX positive (Fig. 4b), with few cases showing a discrepant result (i.e. sample D). For a biological interpretation, one should take into account that hypoxia is considered a local phenomenon, where totally different O₂ levels may occur in adjacent tissue sections.

Identification of a novel, active *MB* promoter in cancer cells

As described above, in the cancer cell lines MDA-MB 468 and DLD-1 the *MB* mRNA variants 9, 10 and 11 are all transcribed from the same start site, the 5' end of exon 5u. To characterize the novel gene regulatory region upstream of exon 5u, we utilized the program Gene2Promoter to identify a 638 bp sequence fragment (Genomatix promotor identification number GXP_2244649; see Fig. 5) as a candidate active promoter sequence. Next, different

regions of the predicted promoter sequence were cloned into luciferase reporter vectors to investigate their transcriptional activity in transfected MDA-MB468 cells. The quantified light units of each plasmid were normalized to light intensities of empty vector controls. The largest candidate fragment (692 bp), ranging from 501 bp upstream of the to 191 bp downstream of the transcription start site, was active in breast cancer cells with a 23-fold increase in relative light units (RLUs), as compared to the control (Fig. 5). This fragment fully encompassed the Gene2Promoter-predicted GXP_2244649 sequence. To further characterize transcriptional activities along the promoter region, progressive deletions of the 692 bp region from its 5' and 3' ends were made and compared in their luciferase readouts to the full-length fragment. Regions containing up to 179 bp upstream to the transcription start site were as active as the complete 692 bp fragment in MDA-MB468 cells (i.e. ~50-fold increased reporter gene activity), while DNA sections containing only 108 bp upstream of the transcription start site showed much weaker promoter activity (Fig. 5). At the 3' end, DNA sections that ended 25 bp downstream of the transcription start site conferred the strongest reporter gene activation (i.e. ~40-fold increase of RLUs). In contrast, constructs that contained only up to 17 bp downstream of the transcription start site triggered only a 22-fold increase in RLUs (Fig. 5). In result, the most active promoter region upstream of exon 5u consists of a 204 bp fragment that at least starts 179 bp upstream and ends 25 bp downstream of the transcription start site. A ConSite search for common cis-acting promoter elements revealed a CCAAT-box motif (NF-Y) proximate (57 bp upstream) to the transcription start site (see Fig. 5) that contains a perfectly matching pentanucleotide motif. Due to its position within the 204 bp fragment, this site might be relevant for promoting the transcription of cancer-associated *MB* mRNAs.

An additional open reading frame within the human *MB* gene

Among the protein-coding *MB* transcripts, variant 11 is unique in containing exon 7u, which is located on an Alu element. This exon encodes a potential 78 bp uORF, starting at its first basepair. uORFs usually do not encode functional proteins but control the translation efficiency of the following protein coding sequence by facilitating ribosome dissociation (21; 22). To test the functionality of the uORF in exon 7u, pGL3 control reporter assay plasmids were constructed that contain either the full uORF or a point-mutated version (c.-3067A>T, resulting in TTG instead of ATG as a start codon). Neither the uORF, nor its point-mutated version contained a second, additional ATG. To mirror the natural degree of ribosome reinitiation as occurring on the *MB* mRNA variant 11, the exact basepair composition that stretches between the uORF and the *MB*-ORF (=open reading frame) was added to both plasmids upstream to their luciferase-ATGs. Both constructs were co-transfected with pGL4.74 vectors in MDA-MB468 cells and quantified by DLRA. Since the unmutated construct containing the uORF (ATG) conferred a 34% decrease in reporter gene translation efficiency relative to the mutated uORF (TTG) construct (Fig. 6), the candidate uORF is indeed able to significantly impair the ribosome attachment upon translation of the *MB* CDS in transcript variant 11.

DISCUSSION

Although MB is traditionally regarded as one of the best-characterized proteins in biochemistry, advances in recent years have revealed several novel, additional functions of this protein in NO[•] and ROS (=reactive oxygen species) metabolisms, thereby extending its classic role in O₂ supply to muscle cells (7; 8; 5; 23). In parallel it was shown that beyond its primary localization in cardiac myocytes, skeletal and smooth muscle (3; 4; 24; 5), MB is expressed at several “ectopic” sites. Importantly, these additional expression sites include

breast cancer, non-small cell lung cancer, colon and ovary carcinomas, desmoplastic small round cell tumors and leukemic bone marrow, as well as derived breast, colon and osteosarcoma cancer cell lines (11; 12; 13; 14; 15), and also luminal cells of healthy breast tissue (14). From a zoological perspective, these data may not be entirely surprising since, e.g. in the common carp, small amounts of the globin were detected in various non-muscle tissues, including liver, gill, kidney and brain (25). The biological function of ectopic *MB* expression, however, has remained elusive up till now. In the present study, data mining of publicly available RNA-seq datasets have confirmed transcription of *MB* in several healthy and tumorous human tissues and cancer cell lines at mRNA levels corresponding to roughly 1/100 of the amount present in heart muscle (Fig. 2). Most importantly, detailed analyses of *MB* transcription revealed that the organization, splicing patterns and promoter usage of the *MB* gene are also much more complicated and variable than hitherto thought.

Unexpected complexity of the *MB* gene and its transcripts

Our EST and RNA-Seq data analyses revealed 11 alternative *MB* protein-coding transcripts in human tissues, the majority of which has been hitherto undescribed (Fig. S1). An additional 8 transcripts, producing aberrant or truncated *MB* versions, could be detected. However, their frameshift-induced premature stop codons in the *MB*-CDS are positioned distantly upstream of the distal 3' splice site, making it likely that RNA-decay mechanisms (reviewed in: 26) degrade these transcripts prior to translation. This bewildering complexity of *MB* transcript variants is due to the presence of several novel exons within a substantially enlarged *MB* gene region of 31.8 kb instead of its former size of 16.6 kb. The *MB* gene thus illustrates results of the ENCODE project, stating that roughly 75 % of the human genome is primarily transcribed (27). ENCODE cell lines were reported to express on average sets of 10 to 12 mRNA isoforms per gene, which agrees well with the expression of 9 and 6 alternative *MB* transcripts (out of 19 splice forms in total) in human skeletal and heart muscle, respectively. In cancer

cell lines, 12 out of the 19 *MB* mRNA variants were detected (Fig. S1). Three of those variants, all hitherto undescribed, were most frequently observed in breast cancer entities, and one breast and one colon cancer cell line (Fig. 3 and 4a). Thus, *MB* expression also adheres to the ENCODE rule that mostly one isoform per gene is predominantly expressed upon a certain cellular condition, and that 75% of all genes produce at least 2 major transcripts (27).

The obvious question for the adaptive functional value of multiple *MB* splice forms cannot be fully answered to date, but the observed transcriptional complexity is clearly contradictory to a “minimalistic expression strategy” (27; 28). As a possible example for the functionality of upstream exons we found that the alternatively spliced cassette-exon 7u is able to decrease *MB* translation efficiency via its part encoding a uORF (Fig. 6). Due to the fact that the mRNA section between the uORF and the downstream *MB*-ORF is rather short (49 bp) and does not contain internal ribosome entry sites (IRES), we speculate that ribosome re-initiation is likely to be hindered to some extent when exon 7u is spliced in (21; 29; 30; 22). While the recorded effect on translation appears rather moderate *in vitro*, a non-linearity between *MB* protein and mRNA levels in a breast cancer cell line, possibly indicating post-transcriptional effects, has been reported previously (13). Notably, cassette-exon 7u is part of an AluSc family element that was inserted into the *MB* gene, illustrating the potentially adaptive effect of retrotransposons as modulators of gene regulation during primate evolution (31; 32; 33). Various primate-associated transposons such as Alu and L1 elements have invaded the 5' part of the human *MB* gene region (Fig. S3). Not surprisingly therefore, the novel *MB* upstream exons are only conserved throughout higher primate orders (Tab. S3) and lack orthologues in the mouse *Mb* gene, which displays less variability on the genomic level and in its splice patterns. Studying transcriptional regulation of ectopic *Mb* expression in the mouse may therefore be of limited value.

Most important to understand the gene regulatory mechanisms driving classic and ectopic *MB* expression is the fact that the alternative usage of 5' upstream exons implies the

existence of different promoters for the *MB* gene. For instance, *MB* transcript variants 2 and 5 are the most abundant mRNAs in heart and skeletal muscle (Fig. 2), but only variant 2 starts downstream of the classical TATA box motif (20). Interestingly, despite extensive work on this promoter region, which has led e.g. to the identification of functionally relevant MEF2, Sp1 and NFAT transcription factor binding sites (34), the presence of an alternative transcriptional start site 5' of starting exon 8u has escaped notice so far, although it originates between the two NRE-elements and therefore covers the conventional promoter region. For the most active *MB* transcriptional start site in cancer cells, our reporter assays confirmed a 204 bp promoter section upstream of exon 5u to confer high transcriptional activity in the *MB* expressing breast cancer cell line MDA-MB468. This section contains an integer CCAAT-box (also known as NF-Y), a motif that is often used to facilitate the transcription of genes overrepresented in cancer (35). Further molecular studies are necessary to characterize the relevant gene regulatory elements of this tumor-associated *MB* promoter beyond mere bioinformatic predictions.

Cancer-associated human *MB* splice variants are hypoxia-regulated

The existence of additional promoter regions in the *MB* gene may clearly have the adaptive value of facilitating gene expression under special circumstances. From the viewpoint of tumor biology, it is an important finding that the *MB* transcripts initiated by the novel exon 5u promoter (variants 9, 10, 11) are distinctly upregulated by hypoxia, a frequent microenvironmental influence determining tumor physiology and malignancy (36; 37; see discussion below). A 1.5-3.5 fold induction of those *MB* mRNAs was demonstrated *in vitro* in the breast cancer cell line MDA-MB468 and in DLD-1 colon cancer cells upon experimental hypoxia. In breast cancer biopsies, *MB* transcriptional upregulation was paralleled by increased mRNA expression of the hypoxia-responsive *VEGFA* gene and on the protein level by positive immunostaining of the hypoxia-marker CAIX, suggesting that *MB* upregulation at

lowered oxygen levels also occurs *in vivo*. In line with these observations, we had previously reported (17) that another minor *MB* transcript (variant 13) also showed hypoxia-sensitivity (at 1% O₂) in MDA-MB468 breast cancer cells (but not in colon-derived DLD-1). Also, *MB* immunostaining distribution correlated with local tissue hypoxia in many cases of DCIS cancer entities, which originate in the breasts' avascular milk ducts and completely rely on diffusional O₂ and nutrient supply, thus becoming more and more hypoxic towards their center during growth (17).

The degree of hypoxia-responsivity of the *MB* gene in the classic context of muscle cells has been a debated issue (38; 39; 40; 41; 42). Recent work showed that while the Mb content of the contracting heart of mice increased under moderate hypoxia (10% O₂), mouse Mb remained unchanged in sedated skeletal muscle (18). However, *MB* transcription in skeletal muscle could be induced by exercise via Ca²⁺ ion release to activate the Calcineurin/NFAT pathway, with an increased Calcineurin sensitivity upon hypoxia (18). Thus, at least in skeletal muscle, *MB* regulation does not respond to an isolated hypoxic stimulus, but requires a combination of exercise and hypoxia (reviewed in 34; 43). In line with this finding, the muscle-associated standard *MB* promoter was reported to lack canonical binding sites for the hypoxia-responsive master transcription factors HIF-1 α -2 α (19; 18). In contrast, we have recently identified a candidate HRE within the upstream region of the *MB* gene that conferred a moderate 43% hypoxia induction to a downstream luciferase reporter gene (17). It is therefore probable that this HRE along with other, hitherto uncharacterized enhancer sites drives the hypoxia-inducible transcription of *MB* mRNA variants in cancer cells. The recruitment of novel upstream exons and promoters in the primate lineage thus has facilitated evolution of an additional regulatory potential, which enables *MB* expression under hypoxic cellular stress. This potential appears to be specifically realized in a tumor context. It remains unclear, however, why a minor usage of the non-standard *MB* splice forms is also

recorded in non-cancerous tissues, as observed in healthy breast, skeletal and heart muscle (comp. Fig. 2).

A non-standard molecular function of MB in hypoxic tumor tissues?

At current, the biochemical function of ectopically expressed MB in mammals is unclear. MB protein levels in breast tumor cells as well as in healthy breast tissue appear much too low to account for a respiratory function, suggesting a non-standard function of MB in tumor cells (44; 17). This is even more likely, as work in myocytes has revealed important additional roles of MB, e.g. in compensating oxidative and nitrosative stress (7; 8; 9; 10). In hypoxic tumor areas, an induction of (low-level) MB expression might interfere with the homeostasis between formation and sequestration of ROS and RNS (=reactive nitrogen species), and subsequent biochemical studies will have to address these functional scenarios. Most importantly, MB-positive tumors reflect the luminal, better differentiated carcinoma phenotype along with an ameliorated prognosis for the patient, possibly due to advanced cell differentiation, less aggressiveness and reduced metastatic spread (16; 13; 14). Recent experiments on *MB* knockdown MDA-MB468 breast cancer cells indeed indicated that MB - by unknown mechanisms - impairs mitochondrial activity under hypoxia (17). MB therefore harbors the potential of acting as a tumor-suppressor, and its hypoxia-responsive upregulation may partly explain its beneficial effect on breast tumor patients' survival.

MATERIALS AND METHODS

Bioinformatic analyses of the *MB* gene region

To elucidate the human and mouse *MB/Mb* gene structures, their corresponding EST files were downloaded from NCBI UniGene

(<http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?UGID=908333&TAXID=9606&SEARCH=mb>) and aligned to the genomic *MB* sequence using the NCBI Spidey software (45). Novel

exon regions and splice variants were portrayed with GenePalette 1.2 (46). All new splice variants were cross-checked with the transcripts available at NCBI AceView (<http://www.ncbi.nlm.nih.gov/iebr/research/acembly/>) and annotated, if a matching sequence has been found. We applied Gene2Promoter (www.genomatix.de) to predict the promoter sequences of novel 5'UTR splice variants (47). Common *cis*-acting promoter elements (TATA, CCAAT and GC boxes) were searched within the experimentally validated cancer-active *MB* promoter (Fig. 5) with ConSite (<http://www.phylofoot.org/consite/>), choosing a minimum specificity of 12 bits and a TF cutoff of 82%. All detected elements were then checked for the integrity of their essential core motif sites. The revised gene structures of human *MB* and mouse *Mb* were depicted with SeqBuilder (Lasergene Sequence Analysis Software, DNASTAR Inc., USA). The positions of repeats relative to the newly identified exons were determined by RepeatMasker (<http://www.repeatmasker.org>; 48) and outlined on the human *MB* gene by SeqBuilder. All novel *MB* splice variants were inspected for encoded ORFs with EditSeq (Lasergene Sequence Analysis Software, DNASTAR Inc., USA). Additionally, NCBI-BLASTX (49) searches were performed to reveal whether novel exons and exon combinations had protein coding potential. The phylogenetic conservation of newly found human *MB* exons was assessed in NCBI-BLASTN searches against the respective genes (plus a flanking 50kb frame) of *Pan troglodytes* (2.1.4), *Macaca mulatta* (051212), *Callithrix jacchus* (3.2), *Canis familiaris* (3.1), *Bos taurus* (UMD3.1), *Mus musculus* (GRCm38.p1), and *Gallus gallus* (4.0). Only hits with an e-value of $E < 10^{-10}$ were listed.

For an extended *de novo* detection of *MB* in different human tissues, high-throughput RNA-Seq datasets of different human tissues and cancer cell lines were downloaded from the NCBI Sequence Read Archive (SRA; see Tab. S4). All datasets were mapped against the genomic sequence of the *MB* gene (Chromosome 22 of the human genome built 37) by applying the Tophat program package (<http://tophat.cbcb.umd.edu/>) with a maximum of 2 mismatches (default settings). To quantify *MB* exon usage in the different transcriptomes

(Tab. S4), reads mapping to the coordinates of each *MB* transcript start exons (4u, 5u, 8u, 9u) and CDS-containing exons (9c, 11c and 12c) were counted and normalized relative to the according exon and dataset size (RPKM values; 50). Dependent of the availability of tissue datasets on the SRA, average ratios were calculated via Excel 2003 (Microsoft) if multiple datasets per tissue type were used.

Characterization of *MB* splice variants in cancer cell lines

The breast cancer line MDA-MB468 and the colorectal adenocarcinoma cell line DLD-1 were cultured in DMEM/Hams-F12 with stable glutamine (PAA), supplemented with 10% FBS Gold (PAA) and 1% Pen/Strep (PAA). For normoxic culture conditions, cells were raised in an IG150 incubator (Jouan) at 5% CO₂ in an H₂O-saturated atmosphere at 37°C. For experimental hypoxia, cells were raised in a CB 53 incubator (Binder) in an H₂O-saturated atmosphere at 37°C, 1% O₂ and 5% CO₂.

To experimentally verify novel splice variants, total RNA was extracted from cell lines by applying the RNeasy Mini Kit (Qiagen), including a DNase I digestion. First-strand cDNA synthesis was carried out on 1 µg of total RNA using the Superscript III RT- Kit (Invitrogen). 5'/3' RACE reactions were performed on RNA from MDA-MB468 cells using the Gene Racer Kit (Invitrogen). RT-PCRs on cDNA were conducted with the TrueStart Taq DNA Polymerase kit (Fermentas) in a peqSTAR 96 Universal Gradient Thermocycler (Peqlab) or T3 Thermocycler (Biometra), according to the manufacturer's instructions. Annealing temperatures were chosen in accordance with the melting temperatures of the primer sets. PCR amplicons were purified by the High Pure clean-up kit (Roche) and either sequenced directly or cloned into pGem-T Easy vectors (Promega). After transformation in DH10B-cells and singularization of clones, plasmids were purified with the GeneJET Plasmid Miniprep Kit (Fermentas). Sanger-based sequencing reactions were carried out by a commercial service (StarSeq). Sequence reads of novel *MB* splice forms, such as *MB*

sequences derived from 5'/3' RACE reactions were submitted to the EMBL database (<http://www.ebi.ac.uk/>).

Quantitative real-time reverse-transcriptase PCR (qRT-PCR)

The qRT-PCR reactions were performed on cDNA samples from hypoxic or normoxic MDA-MB468 and DLD-1 cells. The primer combinations used for each qRT-PCR reaction were chosen to specifically hybridize to only one *MB* splice variant of interest (see Fig. 3; Tab. S2). The Power SYBR Green PCR Mastermix (Applied Biosystems) was used at a total volume of 10 μ l with annealing temperatures of 57°C in an ABI 7500 real-time cycler (Applied Biosystems).

For absolute quantification of samples with unknown cDNA content, we applied the standard curve approach with serial tenfold dilutions of standard plasmids containing a matching amplicon as an insert. As a positive control for cellular response to hypoxic conditions, transcripts of the *VEGFA* gene were quantified in hypoxic and normoxic samples (51). Triplicate assays were measured in each run, with a total of n=3 experiments for each amplicon assay. Average ratios and standard deviations were calculated via Excel 2003 (Microsoft). A two-sided Students' T-test with an error value of 5% ($\alpha=0.05$) was performed to infer statistical significance.

Detection of *MB* splice variants in cancer biopsies

MB expression was analyzed in tumor and normal adjacent tissues obtained from surgical specimens from 21 primary breast cancer patients. Patient material was collected at the University Hospital of Bonn between 2007 and 2012. The study was approved by the institutional review board. Matched formalin-fixed and paraffin-embedded tissue (FFPET) and fresh-frozen tissues from the same patients were analyzed. *MB* and CAIX immunostaining of 3- μ m FFPET sections was performed using the LabVision Autostainer

480S system (Thermo Scientific). The PT-Module was used for dewaxing and epitope retrieval (pH 6,0 at 99°C for 20 min). The following antibodies and dilutions were used: rabbit monoclonal MB antibody EP3081Y, (Abcam, Cambridge, UK), dilution 1:150; rabbit polyclonal CAIX antibody (Abcam), dilution 1:300.

Total RNA was isolated from the fresh frozen tumor and normal adjacent tissues by means of the RNeasy Kit (Qiagen), following the Animal Tissue protocol. As described above, the first-strand cDNA synthesis was conducted with the Superscript III RT- Kit (Invitrogen), but on 500 ng total RNA. The resulting cDNA was diluted 1:4 and quantified with the Qubit ssDNA Assay Kit (Life Technologies). QRT-PCRs were performed as described above, with the FastStart Universal SYBR Green Master (ROX) Mix (Roche) in duplicate assays. Absolute copy numbers of *MB* transcripts per 10 ng cDNA were determined using the standard curve approach. For statistical correlations, parametric Pearson tests were performed with SPSS, Version 21 (IBM SPSS Statistics). Samples were sorted according to the overall *MB* level detected in the biopsied tumors.

Functional analysis of a candidate uORF and a novel *MB* promoter via dual luciferase reporter assays

To investigate the functionality of the candidate uORF and the novel promoter, the corresponding sequence regions were amplified via PCR on human genomic placenta tissue DNA and cloned into luciferase reporter gene vectors (pGL3-control vectors, Promega). For PCR amplification of uORF and mutated uORF (c.-3067A>T) constructs, all forward and reverse primers were equipped with a 5'-end HindIII and NcoI site, respectively (Tab. S2). Forward and reverse primers used for the amplification of candidate promoter sections were engineered to carry 5' MluI and HindIII recognition sites, respectively (Tab. S2). PCR reactions were prepared as described above. Annealing temperatures were in accordance with

the melting temperatures of each primerset (see Tab. S2). Obtained PCR products were digested with the restriction enzymes that matched their flanking recognition sites. pGL3-control vectors (Promega) for cloning the uORF constructs and candidate promoter sections were digested with respective restriction enzymes (Fermentas). All vectors were then dephosphorylated at their 5'-ends by Antarctic Phosphatase (NEB). PCR products were ligated into the modified vectors using T4 ligase (Fermentas) and transformed as described above. We also generated a negative control vector lacking the SV40 promoter, by subjecting pGL3 control vector to a MluI/HindIII digest (Fermentas). After dephosphorylation and Klenow (NEB) end-filling of 5' overhangs the blunt ended vector was ligated with T4 ligase (Fermentas) and transformed into bacterial host cells as described above.

About 10^4 MDA-MB468 cells were seeded in triplicates. By using 0.3 μ l of the FuGENE-HD Transfection Reagent (Promega), all wells were co-transfected with 240 ng of pGL3-based firefly luciferase plasmids and 26 ng of pGL4.74 renilla luciferase vectors (Promega) as internal controls. After 44 h, DLRA (Promega) assessed the influence of the uORF on the translation efficacies of the CDS and the transcriptional activity of different promoter constructs. Light signal intensity was quantified in a Glomax 96 well luminometer (Promega), choosing a sample volume of 40 μ l for both luciferin reagents. Normalization of transfection efficacies was done by dividing the total amount of luciferase light units in each pGL3 sample by the co-transfected pGL4.74 renilla light units, yielding RLU.

The RLU quantities of the mutated TTG (c.-3067A>T) uORF constructs were normalized on the RLU values of the ATG-uORF-constructs in each (n=5) trial. RLU quantities of promoter constructs were normalized on RLUs of endfilled control vectors for enabling signal-to noise ratio estimation. For both, the uORF and the promoter experiments, standard deviations were calculated via Excel 2003 (Microsoft). A two-sided Student's T-test with an error rate of 1% ($\alpha=0.01$) was performed to test for significance.

ACKNOWLEDGEMENTS

We thank Carsten Golletz (UK Bonn) for expert technical assistance. This work was supported by the Center for Computational Sciences (SRFN) and a further intramural grant (Stufe 1) of Johannes Gutenberg University Mainz [to TH] and by the University Hospital Bonn [to GK].

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Kölliker, A. (1850) Mikroskopische Anatomie oder Gewebelehre des Menschen., *Verlag Wilhelm Engelmann Leipzig Germany*, **1**, 248.
2. Günther, H. (1921) Über den Muskelfarbstoff. *Virchows Archiv*, **230**, 146-178.
3. MacMunn, C.A. (1886) Researches on Myohämatin and Histohämatins. *Journ. of Physiol.*, **5**, 24.
4. Wittenberg, B.A. and Wittenberg, J.B. (1989) Transport of oxygen in muscle. *Annu. Rev. Physiol.*, **51**, 857-878.
5. Wittenberg, J.B. and Wittenberg, B.A. (2003) Myoglobin function reassessed. *J. Exp. Biol.*, **206**, 2011-2020.
6. Wittenberg, J.B. (1970) Myoglobin-facilitated oxygen diffusion: role of myoglobin in oxygen entry into muscle. *Physiol. Rev.*, **50**, 559-636.
7. Flögel, U., Merx, M.W., Godecke, A., Decking, U.K. and Schrader, J. (2001) Myoglobin: A scavenger of bioactive NO. *Proc. Natl. Acad. Sci. USA*, **98**, 735-740.
8. Flögel, U., Godecke, A., Klotz, L.O. and Schrader, J. (2004) Role of myoglobin in the antioxidant defense of the heart. *FASEB J.*, **18**, 1156-1158.
9. Shiva, S., Huang, Z., Grubina, R., Sun, J., Ringwood, L.A., MacArthur, P.H., Xu, X., Murphy, E., Darley-Usmar, V.M. and Gladwin, M.T. (2007) Deoxymyoglobin is a nitrite reductase that generates nitric oxide and regulates mitochondrial respiration. *Circ. Res.*, **100**, 654-661.
10. Helbo, S., Dewilde, S., Williams, D.R., Berghmans, H., Berenbrink, M., Cossins, A.R. and Fago, A. (2012) Functional differentiation of myoglobin isoforms in hypoxia-tolerant carp indicates tissue-specific protective roles. *Am. J. Physiol. Regul. Integr. Comp. Physiol.*, **302**, 693-701.

11. Ruck, P., Horny, H.P., Greschniok, A., Wehrmann, M. and Kaiserling, E. (1995) Nonspecific immunostaining of blast cells of acute leukemia by antibodies against nonhemopoietic antigens. *Hematol. Pathol.*, **9**, 49-56.
12. Zhang, P.J., Goldblum, J.R., Pawel, B.R., Fisher, C., Pasha, T.L. and Barr, F.G. (2003) Immunophenotype of desmoplastic small round cell tumors as detected in cases with EWS-WT1 gene fusion product. *Mod. Pathol.*, **16**, 229-235.
13. Flonta, S.E., Arena, S., Pisacane, A., Michieli, P. and Bardelli, A. (2009) Expression and functional regulation of myoglobin in epithelial cancers. *Am. J. Pathol.*, **175**, 201-206.
14. Kristiansen, G., Rose, M., Geisler, C., Fritzsche, F.R., Gerhardt, J., Lüke, C., Ladhoff, A.M., Knüchel, R., Dietel, M., Moch, H. *et al.* (2010) Endogenous myoglobin in human breast cancer is a hallmark of luminal cancer phenotype. *Br. J. Cancer*, **102**, 1736-1745.
15. Oleksiewicz, U., Daskoulidou, N., Liloglou, T., Tasopoulou, K., Bryan, J., Gosney, J.R., Field, J.K. and Xinarianos, G. (2011) Neuroglobin and myoglobin in non-small cell lung cancer: expression, regulation and prognosis. *Lung Cancer*, **74**, 411-418.
16. Galluzzo, M., Pennacchietti, S., Rosano, S., Comoglio, P.M. and Michieli, P. (2009) Prevention of hypoxia by myoglobin expression in human tumor cells promotes differentiation and inhibits metastasis. *J. Clin. Invest.*, **119**, 865-875.
17. Kristiansen, G., Hu, J., Wichmann, D., Stiehl, D.P., Rose, M., Gerhardt, J., Bohnert, A., ten Haaf, A., Moch, H., Raleigh, J. *et al.* (2011) Endogenous myoglobin in breast cancer is hypoxia-inducible by alternative transcription and functions to impair mitochondrial activity: a role in tumor suppression? *J. Biol. Chem.*, **286**, 43417-43428.
18. Kanatous, S.B., Mammen, P.P., Rosenberg, P.B., Martin, C.M., White, M.D., Dimaio, J.M., Huang, G., Muallem, S. and Garry, D.J. (2009) Hypoxia reprograms calcium signaling and regulates myoglobin expression. *Am. J. Physiol. Cell Physiol.*, **296**, 393-402.

19. Wystub, S., Ebner, B., Fuchs, C., Weich, B., Burmester, T. and Hankeln, T. (2004) Interspecies comparison of neuroglobin, cytoglobin and myoglobin: sequence evolution and candidate regulatory elements. *Cytogenet. Genome Res.*, **105**, 65-78.
20. Weller, P., Jeffreys, A.J., Wilson, V. and Blanchetot, A. (1984) Organization of the human myoglobin gene. *EMBO J.*, **3**, 439-446.
21. Kozak, M. (1987) An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucleic Acids Res.*, **15**, 8125-8148.
22. Meijer, H.A. and Thomas, A.A. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem. J.*, **367**, 1-11.
23. Hendgen-Cotta, U.B., Merx, M.W., Shiva, S., Schmitz, J., Becher, S., Klare, J.P., Steinhoff, H.J., Goedecke, A., Schrader, J., Gladwin, M.T. *et al.* (2008) Nitrite reductase activity of myoglobin regulates respiration and cellular viability in myocardial ischemiareperfusion injury. *Proc. Natl. Acad. Sci. USA*, **105**, 10256-10261.
24. Qiu, Y., Sutton, L. and Riggs, A.F. (1998) Identification of myoglobin in human smooth muscle. *J. Biol. Chem.*, **273**, 23426-23432.
25. Fraser, J., de Mello, L.V., Ward, D., Rees, H.H., Williams, D.R., Fang, Y., Brass, A., Gracey, A.Y. and Cossins, A.R. (2006) Hypoxia-inducible myoglobin expression in nonmuscle tissues. *Proc. Natl. Acad. Sci. USA*, **103**, 2977-2981.
26. Maquat, L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.*, **5**, 89-99.
27. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F. *et al.* (2012) Landscape of transcription in human cells. *Nature*, **489**, 101-108.

28. Raser, J.M. and O'Shea, E.K. (2005) Noise in gene expression: origins, consequences, and control. *Science*, **309**, 2010-2013.
29. Kozak, M. (1999) Initiation of translation in prokaryotes and eukaryotes. *Gene*, **234**, 187-208.
30. Davuluri, R.V., Suzuki, Y., Sugano, S. and Zhang, M.Q. (2000) CART classification of human 5' UTR sequences. *Genome Res.*, **10**, 1807-1816.
31. Cordaux, R. and Batzer, M.A. (2009) The impact of retrotransposons on human genome evolution. *Nat. Rev. Genet.*, **10**, 691-703.
32. Shen, S., Lin, L., Cai, J.J., Jiang, P., Kenkel, E.J., Stroik, M.R., Sato, S., Davidson, B.L. and Xing, Y. (2011) Widespread establishment and regulatory impact of Alu exons in human genes. *Proc. Natl. Acad. Sci. USA*, **108**, 2837-2842.
33. Cowley, M. and Oakey, R.J. (2013) Transposable elements re-wire and fine-tune the transcriptome. *PLoS Genet.*, **9**, e1003234.
34. Kanatous, S.B. and Mammen, P.P. (2010) Regulation of myoglobin expression. *J. Exp. Biol.*, **213**, 2741-2747.
35. Dolfini, D. and Mantovani, R. (2013) Targeting the Y/CCAAT box in cancer: YB-1 (YBX1) or NF-Y? *Cell Death Differ.*, **20**, 676-685.
36. Graeber, T.G., Osmanian, C., Jacks, T., Housman, D.E., Koch, C.J., Lowe, S.W. and Giaccia, A.J. (1996) Hypoxia-mediated selection of cells with diminished apoptotic potential in solid tumours. *Nature*, **379**, 88-91.
37. Michieli, P. (2009) Hypoxia, angiogenesis and cancer therapy: to breathe or not to breathe? *Cell Cycle*, **8**, 3291-3296.
38. Reynafarje, B. (1962) Myoglobin content and enzymatic activity of muscle and altitude adaptation. *J. Appl. Physiol.*, **17**, 301-305.

39. Hoppeler, H. and Vogt, M. (2001) Muscle tissue adaptations to hypoxia. *J. Exp. Biol.*, **204**, 3133-3139.
40. Masuda, K., Okazaki, K., Kuno, S., Asano, K., Shimojo, H. and Katsuta, S. (2001) Endurance training under 2500-m hypoxia does not increase myoglobin content in human skeletal muscle. *Eur. J. Appl. Physiol.*, **85**, 486-490.
41. Fordel, E., Geuens, E., Dewilde, S., De Coen, W. and Moens, L. (2004) Hypoxia/ischemia and the regulation of neuroglobin and cytoglobin expression. *IUBMB Life*, **56**, 681-687.
42. Gelfi, C., De Palma, S., Ripamonti, M., Eberini, I., Wait, R., Bajracharya, A., Marconi, C., Schneider, A., Hoppeler, H. and Cerretelli, P. (2004) New aspects of altitude adaptation in Tibetans: a proteomic approach. *FASEB J.*, **18**, 612-614.
43. Wittenberg, B.A. (2009) Both hypoxia and work are required to enhance expression of myoglobin in skeletal muscle. Focus on Hypoxia reprograms calcium signaling and regulates myoglobin expression. *Am. J. Physiol. Cell Physiol.*, **296**, 390-392.
44. Gorr, T.A., Wichmann, D., Pilarsky, C., Theurillat, J.P., Fabrizius, A., Laufs, T., Bauer, T., Koslowski, M., Horn, S., Burmester, T. *et al.* (2011) Old proteins - new locations: myoglobin, haemoglobin, neuroglobin and cytoglobin in solid tumours and cancer cells. *Acta Physiol. (Oxf)*, **202**, 563-581.
45. Wheelan, S.J., Church, D.M. and Ostell, J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952-1957.
46. Rebeiz, M. and Posakony, J.W. (2004) GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev. Biol.*, **271**, 431-438.
47. Scherf, M., Klingenhoff, A. and Werner, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.*, **297**, 599-606.

48. Saha, S., Bridges, S., Magbanua, Z.V. and Peterson, D.G. (2008) Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res.*, **36**, 2284-2294.
49. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403-410.
50. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621-628.
51. Hemmerlein, B., Galuschka, L., Putzer, N., Zischkau, S. and Heuser, M. (2004) Comparative analysis of COX-2, vascular endothelial growth factor and microvessel density in human renal cell carcinomas. *Histopathology*, **45**, 603-611.

LEGENDS TO FIGURES

Fig. 1: Revised exon-intron structure of the human *MB* gene and its former version (top).

Exons are shown as dark boxes with their sizes in bp written on top. Untranslated and CDS-encoding exons are tagged with a “u” or “c”, respectively. The most prominent protein encoding splice variants from muscle (variants 2 and 5) and from cancer-related samples (variants 9-11, 13) are shown, including the start and stop codons of the *MB*-CDS.

Fig. 2: Quantification of *MB* start exon transcription in different human tissues by RNA-Seq read counting. Normalized RPKM values (reads per exon size (in kb) per total reads of the dataset (in Mio)) indicate the transcriptional representation of preferential start exon and CDS containing exons.

Fig. 3: Expression analysis of *MB* splice variants *in vitro* in normoxic and hypoxic MDA-MB468 breast cancer and DLD-1 colon cancer cells. Numbers in the left half of the table show cDNA copy numbers determined by qRT-PCR, fold-changes in the right half of the table indicate hypoxic upregulation (* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; $n=3$). Primers which are able to distinguish between transcript variants 2, 9-11 and 13, are indicated by arrows above (forward) and underneath (reverse) their respective target exons. The solid, dashed and dotted lines that frame fields of the table indicate promoters common between the splice forms (also shown left to each transcript variant’s graph).

Fig. 4 (a) Expression analysis of *MB* splice variants in breast cancer biopsies (named A to U). cDNA copy numbers (per 10 ng cDNA) inferred by qRT-PCR are plotted on the y axis (bars). Values for variants 9-11 have been summed up and are shown relative to expression of the minor cancer-associated transcript 13 and the muscle-typic variant 2. (b) Correlation between

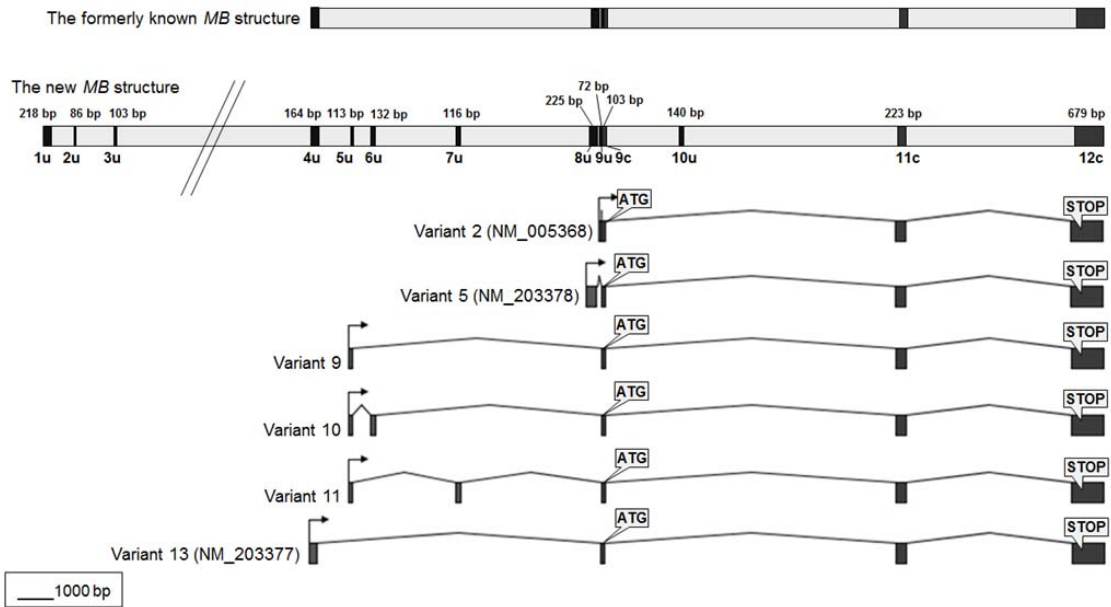
MB expression and hypoxia markers in breast tumor specimen. The fold-increase of *MB* variants and *VEGFA* mRNA expression (bars) in tumors versus normal tissue, inferred by qRT-PCR, is shown in relation to the intensity of CAIX immunostaining, as applied to the same tumor sections.

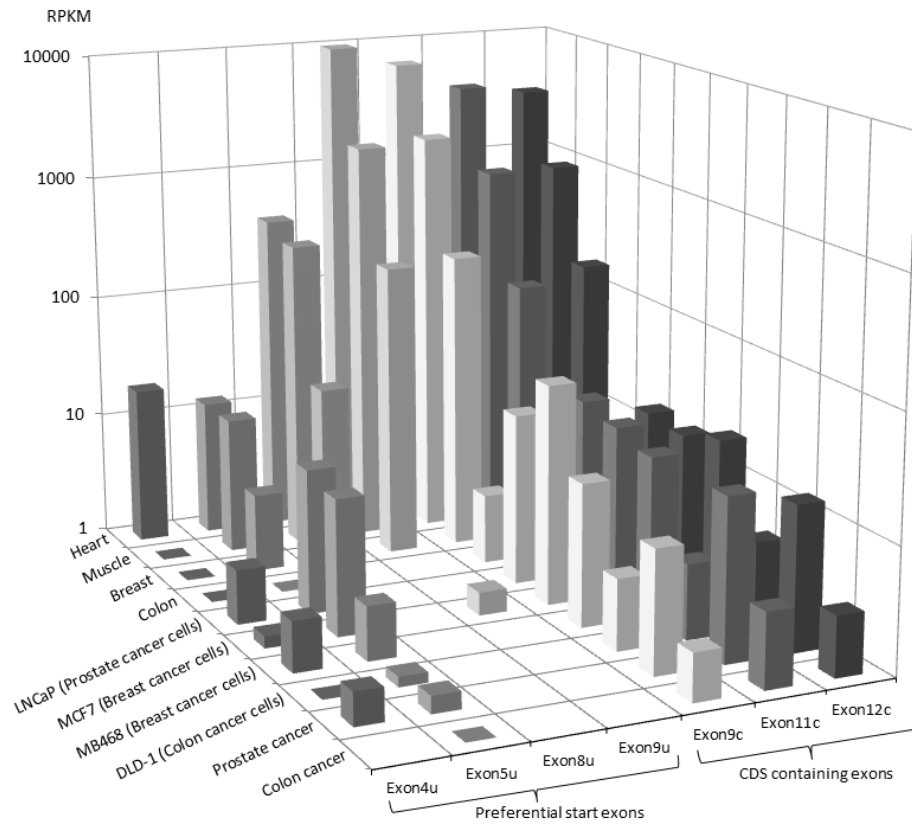
Fig. 5: Functional analysis of the promoter section upstream of exon 5u. DLRA were conducted on promoter sections that were PCR-amplified from primer sites represented by black arrows. The numbers on the arrows indicate the bp distance to the transcription start site of the *MB* splice variants 9, 10 and 11. An *in silico*-predicted promoter region and a commonly observed cis-acting promoter element (NF-Y box) are indicated. The average RLUs of each construct were plotted in the bar chart; standard deviations are indicated by error bars (* $p < 0.05$; *** $p < 0.001$; $n=3$).

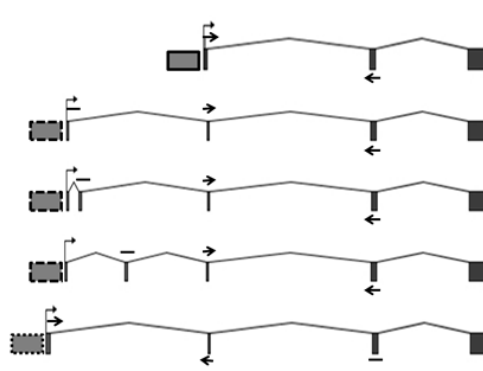
Fig. 6: Functional analysis of a uORF in *MB* cassette exon 7u. DLRA results, plotted as a bar chart, show translational activity of the uORF-containing construct (ATG) and its point mutated version TTG (c.-3067A>T) for comparison. Standard deviations are indicated by error bars (***) $p < 0.001$; $n=5$).

ABBREVIATIONS

CAIX	Carbonic anhydrase IX
CDS	Coding sequence
DCIS	Ductal carcinoma in situ
DLRA	Dual luciferase reporter assay
ER α	Estrogen receptor alpha
EST	Expressed sequence tag
FFPET	Formalin-fixed Paraffin-embedded tissue
HIF1 α	Hypoxia inducible factor 1 alpha
HRE	hypoxia response element
IRES	Internal ribosome entry site
MB	Myoglobin
NO	Nitric oxide
ORF	Open reading frame
RLU	Relative light units
RPKM	Reads per kilobase exon size and per Million reads of each dataset
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
SRA	Sequence read archive
TSS	Translational start site
UTR	Untranslated region
uORF	Upstream open reading frame
VEGFA	Vascular endothelial growth factor A







	Total amount at 21% O ₂		Fold increase upon 72h, 1% O ₂	
	MDA-MB468	DLD-1	MDA-MB468	DLD-1
Var 2	52	71	2.4**	1.5
Var 9	36 665	2 167	3.5***	1.9*
Var 10	19 393	629	3.2*	2.1***
Var 11	27 435	1 151	3.0***	1.5***
Var 13	1 708	98	2.2* (see Ref.(17))	0.9

