Intragenic Enhancers Act as Alternative Promoters.

UNIVERSITÀ DEGLI STUDI DI TORINO

Intragenic enhancers act as alternative promoters

Kowalczyk et al.

Revised manuscript 01.10.2011

Introduction

Transcriptional start sites (TSSs) of all currently defined classes of polyadenylated RNA are thought to be defined by one type of *cis*-element (generically referred to as promoters) which initiate cell-specific mRNA transcripts. However, it is now clear that the transcriptome is far more complex than originally proposed (Carninci et al., 2005; Kapranov et al., 2007; Kapranov et al., 2005). A large proportion of the mammalian genome is transcribed in a developmental stage- and cell type- specific manner to produce many different classes of RNA (Mattick et al., 2010). Understanding how cell- and developmental stage-specific expression is determined will be greatly facilitated by identifying the origins of these different classes of RNA. This should allow us to evaluate their potential significance in more detail before embarking on appropriate functional assays.

Recently it was shown that active promoters (whose chromatin is marked by high levels of H3K4me3 and low levels of H3K4me1 (Heintzman et al., 2009; Heintzman et al., 2007)) are transcribed to generate short bi-directional RNAs centred around transcriptional start sites (TSSs) (Core et al., 2008; He et al., 2008; Preker et al., 2008; Seila et al., 2008), but transcriptional elongation only occurs in the direction of the gene to produce spliced, poly(A)$^+$ RNAs (mRNA). Such *cis*-acting elements are now referred to as divergent promoters (Core et al., 2008; Seila et al., 2009). Of interest, it was recently shown that some intergenic, and possibly some intragenic enhancers, (whose chromatin is marked by high levels of H3K4me1 and low levels of H3K4me3) may also be transcribed to produce short bi-directional transcripts, called eRNAs,

(which may (Kim et al., 2010) or may not be polyadenylated (De Santa et al., 2010)). However, no elongated poly(A)$^+$ transcripts originating from these enhancers were detected (De Santa et al., 2010; Kim et al., 2010). Therefore, it was concluded that, although similar in some respects, promoters and enhancers produce different classes of RNA transcripts: enhancers are not promoters (Kim et al., 2010).

It is estimated that mammalian genomes contain more enhancers than promoters (Heintzman et al., 2009) and where analyzed, approximately 40% of enhancers lie within the body of a gene (Heintzman et al., 2009; Heintzman et al., 2007; Visel et al., 2009). The gene hosting the enhancers may lie 10s-1000s kb away from the genes regulated by the elements and often appears to be unrelated to the target gene. In view of the recent observations on transcription of intergenic enhancers it was of interest to investigate in detail how activation of intragenic enhancers might affect transcription and expression of their host genes.

To address this we initially analysed a well characterised set of five erythroid-specific enhancers, four of which lie within the body of the mouse *Nprl3* gene and one lies upstream from the promoter (R1 to R4, Figure 1A and B). In erythroid cells, these enhancers coincide with *DnaseI* hypersensitive sites (DHSs), bind erythroid-specific and widely expressed transcription factors and recruit low levels of RNA polymerase II (RNAP2). The chromatin associated with these enhancers is characteristically modified by high levels of H3K4me1 and low levels of H3K4me3 (De Gobbi et al., 2007). Moreover, they have

been shown to physically interact with the α-globin promoters (Lower et al., 2009; Vernimmen et al., 2007) in a tissue-specific manner. Finally, it has been shown in transgenic experiments and by natural deletions that these elements are essential for high levels of α-globin expression (Anguita et al., 2002; Higgs and Wood, 2008).

Analysing the transcription of such intragenic enhancers is obscured by the transcription of the host gene (in this case *Nprl3*) and consequently transcription of these elements has been largely ignored. To address this problem we analysed transcription of the *Nprl3* gene after deleting its constitutive promoter. This revealed a previously unknown feature of these intragenic enhancers: they act as highly active, alternative tissue-specific promoters for the gene containing them. Using a genome-wide approach we have shown that a large proportion of activated intragenic enhancers behave as alternative promoters producing an abundant new class of long poly(A)$^+$ RNAs (referred here as to as messenger RNAs derived from enhancers or meRNAs). The expression of meRNAs explains a substantial proportion of the complexity of the transcriptome and how this changes from one cell type to another.

Results

Erythroid-specific expression of the *Nprl3* gene continues in the absence of its constitutive promoter

Nprl3 is a well characterised, widely expressed gene lying adjacent to the α-globin cluster (Figure 1A). It contains four intragenic enhancers and a fifth intergenic enhancer lying 1.4 kb upstream from its constitutive promoter (Figure 1B); four of the five elements are associated with multiple conserved sequences (Hughes et al., 2005). These elements (R1 to R4) act as erythroid-specific enhancers of the α-globin genes. This locus thus provides a paradigm for examining in detail the effects of enhancers on transcription.

Since transcription of intragenic enhancers may be masked by transcription of the gene that contains them we used homologous recombination to delete the *Nprl3* promoter (ΔP6) allowing us to examine any independent transcription from the enhancers (Figure 1B; Figure S1A). Homozygosity for this promoter deletion caused late embryonic lethality with a variety of cardiac outflow abnormalities although the haematological profile was normal (Kowalczyk in preparation). Using RNA-FISH we showed that nascent transcription of *Nprl3* was abolished in non-erythroid cells but surprisingly not in erythroid cells (Figure S1B and C). Similarly, stable mRNA was detected by qPCR in erythroid cells but not in a variety of non-erythroid cells. Even in the absence of the *Nprl3* promoter, abundant expression of the *Nprl3* gene (~50% of wild type level) persisted in erythroid cells (fetal liver in Figure 1C).

We have shown that expression of the *Nprl3* gene is upregulated in human (Lower et al., 2009) (Figure S1D) and mouse erythroid cells (Figure S1E) and the new observations presented here suggest that a substantial proportion of the expression of *Nprl3* mRNA seen in erythroid cells is derived from an erythroid-specific alternative promoter(s) distinct from the canonical *Nprl3* promoter. An EST in the mouse genome (AK036633), annotated as a gene isoform by Ensembl (ENSMUSG00000020289) has a start site lying downstream from the main promoter of the *Nprl3* gene. Annotation of this alternative transcript shows a unique alternative first exon (AFE) (Figure 1B), which is formed from within intron 2 of the *Nprl3* gene. Beyond this AFE the isoform structure appears identical to the full length transcript. Interestingly, the AFE coincides with one of the known intragenic enhancers (R3).

First, we tested wild type mouse tissues by RT-PCR to verify the existence of this alternative transcript. Primers spanning the unique AFE-exon3 junction generated two specific PCR products only in bone marrow cDNA (Figure 1D). Sequence analysis confirmed both to be specific products capturing AFE-exon3 splice junctions. Although it seemed likely that these are erythroid-specific transcripts, since bone marrow contains a mixture of cells from all haematopoietic lineages, we specifically tested purified populations of erythroid (Ter119+) and non-erythroid (Ter119-) cells. Several erythroid-specific PCR products were amplified and sequence analysis showed a series of related exonic junctions, where different donor splice sites within the AFE are used (Figure 1E). These observations show that although R3 has the

hallmarks of an enhancer, when activated in erythroid cells, it can also behave as an alternative, independent, internal promoter of the *Nprl3* gene.

Transcription of *Nprl3* persists in the absence of both the constitutive promoter and the erythroid-specific alternative promoter

In a further attempt to abolish transcription of the *Nprl3* gene in erythroid cells, we deleted both the canonical *Nprl3* promoter and the R3 enhancer (ΔP6-R3) (Figure S2A). The *ΔP6-R3*$^{-/-}$ embryos had similar phenotypes to the promoter knockout (Kowalczyk in preparation). Analysing expression of *Nprl3* mRNA in this mouse model again showed no expression in the brain (as expected), but surprisingly still showed about 20% (with respect to wild type) of *Nprl3* mRNA in erythroid cells (Figure 1F).

Since deletion of both the constitutive promoter and the alternative erythroid-specific promoter still failed to abolish *Nprl3* transcription in erythroid cells, we next set out to determine the origin(s) of the remaining transcription in the *Nprl3* locus. We analysed total RNA from erythroid and non-erythroid cells of wild type, *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ mice using custom tiled arrays. Again, expression of the *Nprl3* gene in both *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ was undetectable in brain (Figure S2B). By contrast, in erythroid cells of both knockouts the only non-transcribing region of the *Nprl3* gene coincided with the extent of each deletion.

Of interest we noted that in both of these knockout mice there was a clear peak of erythroid-specific transcription upstream of the *Nprl3* promoter, which coincides with the conserved intergenic enhancer R4. Transcription associated with such intergenic enhancers (eRNAs) was previously described as short and bi-directional rather than divergent, as occurs from promoters (De Santa et al., 2010; Kim et al., 2010).

Enhancers in and around the *Nprl3* locus produce poly(A)$^-$ eRNAs

Although it seemed possible that transcription from the enhancers contributes to the erythroid-specific expression of mRNA from the *Nprl3* gene, short eRNAs could not account for expression extending throughout the gene (Figure S2B). To reveal the mechanism by which transcription throughout the *Nprl3* gene continues despite deleting two of its known transcription start sites (TSSs) we analysed transcription using high-throughput sequencing (RNA-Seq). Since previous attempts failed to resolve whether eRNAs are polyadenylated (De Santa et al., 2010; Kim et al., 2010), we analysed both poly(A)$^+$ and poly(A)$^-$ RNA from wild type erythroid cells.

First we analysed the poly(A)$^-$ RNA of the WT erythroid cells and compared this with high-resolution chromatin maps (Figure 2). Strand-specific analysis showed that there is abundant primary transcription in the direction of *Nprl3* transcription (bottom strand), but also prominent and discrete antisense RNA transcripts (~1kb in length, top strand) associated with each of the enhancers including the intergenic element (R1 to R4, Figure 2). The intergenic enhancer

R4 was associated with a separate peak of antisense transcription which is distinct from that associated with the *Nprl3* constitutive promoter (see red column, Figure 2). The intergenic location of this element allowed us to see bi-directional poly(A)$^-$ eRNAs originating from this element, as previously noted for other intergenic enhancers (Kim et al., 2010). However, the remainder of the enhancers are embedded in the body of the *Nprl3* gene and therefore any sense transcription originating from these elements was masked by transcription of the host gene.

Activated enhancers direct expression of abundant poly(A)$^+$ isoforms extending to the poly(A) addition site of the *Nprl3* gene

We next analysed the poly(A)$^+$ RNA. To account for any new and/or alternatively spliced transcripts we used algorithms (Tophat and Cufflinks) which do not depend on gene annotation (Trapnell et al., 2009; Trapnell et al., 2010) to reconstruct gene isoforms from the spliced poly(A)$^+$ RNA-Seq data (Figure 3A). Unexpectedly, this predicted four new isoforms extending from the intragenic enhancers (R2 and R3, DHS-12) and, surprisingly, also from the nearby intergenic enhancer (R4) to the polyA site of the *Nprl3* gene. These long transcripts encode unique, alternative first exons (AFEs) spliced onto an adjacent annotated exon and the remainder of the transcript is spliced and polyadenylated in the same manner as the host gene.

Each of the predicted, alternative isoforms were verified by RT-PCR between the unique 5'exons (AFEs) and distal *Nprl3* exons confirming their spliced

structure and erythroid-specificity in wild type cells (Figure S3A) and in the two knockout models (*ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$) (Figure S3B and C).

To further characterise the nature of the enhancer derived polyadenylated transcripts and their abundance with respect to the levels of RNA from the intact gene, we performed RNA-Seq in erythroid cells derived from *ΔP6-R3*$^{-/-}$ mice (Figure 3B). This shows that even in the absence of the consitutive *Nprl3* promoter and two intragenic enhancers (R3 and DHS-12) abundant transcripts extending from the remaining enhancers (R2 and 4) to the poly(A) addition site of the *Nprl3* gene are still detected (Figure S3B and C). Compared to the adjacent, constitutively expressed *Mpg* gene, normalised RNA-seq data show these enhancer isoforms (in total) represent ~20% of the poly(A)$^{+}$ RNA from the intact *Nprl3* gene (Figure 3B); consistent with previously established qPCR data (Figure 1F). None of these transcripts produced proteins that could be detected by western blot analysis (Figure S3D and E).

Together these data show that the intragenic enhancers (R2, R3 and DHS-12) and an upstream intergenic enhancer (R4) act as alternative erythroid-specific promoters. Significant levels of divergent transcription from these enhancers occurs independently of the *Nprl3* promoter and resembles that seen from canonical promoters.

Genome-wide identification of intragenic enhancers

Having identified new, alternative poly(A)$^+$ RNA isoforms with start sites coinciding with the α-globin enhancers, we next determined whether this phenomenon occurs at other intragenic enhancers throughout the genome. To identify enhancers and to facilitate their correlation with RNA-Seq data, we used previously defined chromatin signatures that distinguish enhancers from promoters. Enhancers are marked by *DnaseI* hypersensitive sites (DHSs) with a high level of H3K4me1 and low levels of H3K4me3 (Heintzman et al., 2009; Heintzman et al., 2007) whereas promoters are associated with DHSs marked by high levels of H3K4me3 and low levels of H3K4me1. By considering only those elements that unequivocally display the H3K4me1 high H3K4me3 low signature (Figure 4A) we identified 3358 erythroid enhancers (enhancer set) of which 1794 lie within gene bodies (~54%). Transcription start sites (TSSs) based on current gene annotation were used to define promoters (promoter set) independently of their chromatin marks. The composite profiles (including RNAP2, H3K4me3, H3K4me1) comparing predicted enhancers and annotated promoters (Figure 4B and C) showed that these two classes of elements have different chromatin and transcription factor signatures.

To ensure that these selection criteria had identified *bona fide* enhancers we performed further analysis to confirm that these elements were clearly distinguished from canonical promoters. This showed that sequences in the enhancer set are predominantly bound by tissue-specific transcription factors (in this case Gata1, Scl, Klf1, Ldb1), and the co-factor p300; the associated chromatin is also modified by H3K27ac (Figure 4A). All of these observations are consistent with previous criteria used to identify active enhancers

(Creyghton et al., 2010; Heintzman et al., 2009; Heintzman et al., 2007; Rada-Iglesias et al., 2011). Further bioinformatic analysis clearly showed characteristic differences in DNA sequence and coding potential between the enhancer and promoter sets (Figure S4A to E).

To extend the analysis to another species and a well characterised non-erythroid cell type we used the same approach to identify enhancers using previously published data from primary fetal human lung fibroblasts (Bernstein et al., 2010) and identified 21374 intragenic and 17664 intergenic enhancers in these cells (Figure S5A).

Many intragenic enhancers produce poly(A)$^-$ eRNAs

Genome wide analysis of the poly(A)$^-$ RNAs in erythroid cells showed variable levels of antisense transcription originating from intragenic enhancers (Figure 4D) and similar results were obtained from human primary lung fibroblasts using nascent transcription (global run-on, GRO) (Core et al., 2008) (Figure S5E). In erythroid cells 876 (49%) intragenic enhancers were transcribed at detectable levels (Figure 4D and E) and similarly in human fibroblasts transcription could be detected at 8,775 (41%) intragenic enhancers (Figure S5E and F). Since intergenic enhancers are transcribed in both directions (Kim et al., 2010), we also analysed sense transcription for intragenic enhancers. Sense poly(A)$^-$ transcription from the intragenic enhancers, normally masked by transcription of the host gene, was weakly seen in erythroid (mouse) (Figure 4F) and more clearly in non-erythroid (human) cells

(Figure S5G). These findings show that many intragenic enhancers are transcribed in both sense and antisense directions into short poly(A)$^-$ eRNA. Of note the differences in the levels of eRNA expression are reflected by different degrees of activating histone modification marks (Figure 4G and S5H).

Many intragenic enhancers contribute to a new class of full length poly(A)$^+$ RNAs (meRNAs)

Based on the observations made at the *Nprl3* locus we next determined whether intragenic enhancers might frequently generate alternative poly(A)$^+$ isoforms. Within the body of active genes it would only be possible to detect new enhancer driven isoforms which produced AFEs. Hence we developed a stringent pipeline to identify previously unannotated AFEs (see Extended Experimental Procedures) and this was validated since it detected all the confirmed AFEs within the *Nprl3* gene. In erythroid cells we detected 176 enhancers producing AFE transcripts, analysis of which showed they use conventional splice signals. We verified 13 new junctions between the identified AFEs and the appropriate exon of the host gene by RT-PCR (Figure S3A and S6).

We next analysed the distribution of all newly identified erythroid AFEs around the intragenic enhancers (Figure 5A). This showed a highly significant enrichment of AFEs within 1000 bp downstream (but not upstream) of the midpoint of enhancers. These findings show that although many intragenic enhancers are transcribed from both sense and antisense strands of DNA,

RNA from the sense strand (with respect to the host gene) is transcribed and spliced to produce long poly(A)$^+$ mRNA transcripts. We refer to this new class of RNA transcripts as mRNA associated enhancer RNAs (meRNAs). Therefore as in the *Nprl3* gene a large proportion of enhancers throughout the genome behave as alternative intragenic promoters although they retain the chromatin signature of an enhancer rather than a promoter.

Enhancer driven poly(A)$^+$ RNAs are abundant full length transcripts

By deleting the promoter of the *Nprl3* gene we showed that the intragenic enhancers act as promoters independently of the canonical *Nprl3* promoter. This principle is also clearly exemplified by other genes whose canonical promoters are inactive in erythroid cells, unmasking transcription from intragenic erythroid-enhancers (e.g. see *D18Ertd653e* (Figure 5B), *Acmsd* (Figure 5C), *Abat* (Figure 5D), *Tg* (Figure 6A)). These examples also allowed us to assess the levels of expression from these enhancers by comparing the expression of enhancer driven RNA to expression of the closest neighbouring gene. This revealed that the amount of meRNA produced is as variable as that produced from canonical promoters. For example, at the mouse *D18Ertd653e* and *Acmsd* loci, meRNA isoforms are expressed at levels comparable to their neighbouring genes, *4933403F05Rik* and *Ccnt2* respectively (Figure 5B and C). A similar meRNA isoform in the *Abat* locus is expressed at 40% of that of *Tmem186* gene (Figure 5D). By deleting the canonical promoter of the *Nprl3* gene we have shown that meRNAs may

account for up to 50% of the poly(A)$^+$ transcripts derived from this gene in erythroid cells (Figure 1C).

Having identified these meRNAs by RNA-Seq, we have also showed intact meRNAs by northern blot analysis (e.g. *Tg* Figure 6A and B; *Znfx1* Figure 6C and D) confirming that they are abundant full length transcripts extending from the intragenic enhancers to the poly(A) addition site of the gene that contains them. The levels meRNAs determined by northern blot analysis correspond to those determined by RNAseq. We have therefore shown that intragenic enhancers may act as transcriptional start sites, producing abundant, full length, polyadenylated transcripts. meRNA transcripts arising from enhancers (as opposed to canonical promoters) therefore represent a new, previously unrecognised class of RNA.

meRNAs add to the complexity of the transcriptome

With this new understanding that abundant, full length, polyadenylated transcripts may originate both from promoters and enhancers, we reviewed the current Refseq annotation. We noted that some enhancer driven transcripts have already been annotated as isoforms of their associated protein-coding genes. For example *D18Ertd653e* has two isoforms one from its canonical promoter and one from an erythroid specific enhancer (Figure 7A). Surprisingly, in erythroid cells, we found 139 “active TSSs” with the chromatin signature of an enhancer rather than a promoter (Figure 7B). Similar analysis of primary human lung fibroblasts (Figure 7C), K562

(erythroleukemia), GM12878 (B cell) showed that between 1 and 7.5% of active TSSs correspond to enhancers rather than promoters. Although the number of TSSs with an enhancer signature varied from tissue to tissue, a relatively small fraction overlap between tissues; most are specific to each cell type (Figure 7D).

The complexity of meRNAs in the transcriptome will be much greater than we have detected in this study. We could only identified enhancer driven RNAs via their association with AFEs. However, many enhancers may initiate meRNAs by transcribing an existing exon, rather than an AFE, to produce an extended polyadenylated transcript. This principle is illustrated by the enhancer-driven RNAs derived from the *Abat* gene in erythroid cells (in which the canonical promoter is inactive) (Figure S7). Given that there may be many more enhancers than protein coding genes determining cell identity (Bulger and Groudine, 2010; Heintzman et al., 2009) in hundreds of different cell types, enhancer driven meRNAs will account for a substantial degree of the complexity and abundance of poly(A)$^+$ RNA in the transcriptome.

Discussion

Enhancers and promoters are currently considered as two distinct classes of functional *cis*-elements. Promoters are defined as regions which initiate the transcription of a gene (associated with transcriptional start sites, TSS) whereas enhancers are distally positioned elements which regulate transcription from canonical promoters in time and space. Although these functional definitions remain correct, data presented here show that *bona fide* intragenic enhancers frequently behave as alternative promoters (alternative TSS) producing a new class of abundant, full length poly(A)$^+$ mRNA which we refer to as meRNAs (messenger RNA derived from enhancers).

Transcription from intragenic enhancers closely resembles that from promoters since engaged RNAP2 produces short bidirectional poly(A)$^-$ RNA transcripts, but also produces abundant, full length, spliced poly(A)$^+$ transcripts extending in the direction of the host gene (meRNAs). Although they initiate at intragenic enhancers, meRNAs still reflect the host gene's structure using the same splicing and polyadenylation signals, although a proportion use cryptic splice signals within the intron containing the enhancer to produce non-coding, alternative first exons (AFEs). As for conventional promoters, the mechanism underlying the directionality of meRNAs remains to be determined (Seila et al., 2009).

Until now the existence of this new class of meRNAs has been obscured by transcripts driven from the canonical promoters, which overlap meRNAs. Because of this confounding problem, researchers have exclusively focussed

on RNAs originating from intergenic regions (e.g. long intergenic non-coding RNAs or lincRNAs (Guttman et al., 2009; Khalil et al., 2009)). However once aware of transcription originating from intragenic enhancers, meRNAs that produce an AFE can readily be identified even when the host gene is being transcribed (e.g. *Nprl3* Figure 3A, *Znfx1* Figure 6D). However, the expression and processing of meRNAs can be seen most clearly when the canonical promoter has been deleted (e.g. as shown for the *Nprl3* gene Figure 3B) or when the canonical promoter is naturally inactive (e.g. *Abat* Figure 6, *Acmsd* Figure S7 and *D18Ertd653e* gene, Figure 7). The levels of both eRNAs and meRNAs expression from activated enhancers is just as variable as that seen from canonical promoters. In some cases expression of meRNAs may be as great as that directed from canonical promoters of neighbouring genes (Figure 5B, C, D) and readily detectable on northern blots (Figure 6B, C).

An important question is why intragenic enhancers act as alternative promoters? Physical interaction via looping has now been demonstrated for many enhancers and promoters in different cell types and is thought to be a fundamental feature of their action (Lieberman-Aiden et al., 2009). It is possible that the juxtaposition between these two classes of *cis*-elements, which is associated with high levels of transcription at the canonical promoter, may also promote transcription from the interacting enhancer. When the enhancer is located in intergenic regions, this could produce short poly(A)$^-$ transcripts or perhaps in some cases lincRNAs (Cabili et al., 2011). By contrast, in the case of an intragenic enhancer, the splicing and

polyadenylation machinery of the host gene could cause induced transcription to produce the spliced and poly(A)$^+$ isoforms (meRNAs) described here.

What is the biological significance of this new class of RNA? There are currently at least XX subclasses of RNA. Although the roles of some RNAs (e.g. mRNA, rRNA, tRNA, snoRNA, miRNA) are fully or partially established, in most cases (e.g. lincRNAs) their role is unclear (ref). meRNAs resemble isoforms of the host gene but in the case of *Nprl3*, although abundant, the meRNAs are not translated into any detectable protein. Global analysis of meRNAs also suggests that, in general, meRNAs have relatively low protein encoding potential. meRNAs are abundant, complex and show tissue and developmental stage specificity. Therefore, rather than simply producing transcriptional noise, it would be surprising if evolution has not used some meRNAs, or their processed RNA products, for important biological functions which can now be evaluated.

A major task in understanding the transcriptome is to identify the different classes of RNA so that the functional role of each subclass can be determined. Whatever their functional role, this study shows that meRNAs constitute a complex and abundant new class of RNA. Using stringent criteria we identified 1794 intragenic enhancers in erythroid cells of which at least 876 express eRNA and 179 produce meRNAs identified via their associated AFEs. However due the high stringency with which we identified enhancers and the unknown frequency with which enhancer driven transcripts use the donor splice sites of the host gene rather than produce AFEs (Figure S7) we

have greatly underestimated the full contribution meRNAs to the transcriptome of erythroid cells. Nevertheless, even using this very stringent analysis, in just one tissue (erythroid cells) we have already demonstrated that 139 of annotated, active TSSs correspond to an enhancer rather than a promoter (Figure 5B).

Taking all cell types into account, it has been estimated that mammalian genomes contain many more enhancers than promoters (Bulger and Groudine, 2010; Heintzman et al., 2009) and up to 45% of these enhancers are intragenic. Analysis of just three cell types (IMR90, K562, GM12878) in this study showed very little overlap in enhancer driven (meRNAs) between cell types. Therefore meRNAs may account for a substantial proportion of transcriptome complexity and how this changes from one cell type to another. This new insight will require the field to re-annotate the transcriptome.

Clearly, distinguishing enhancer driven meRNAs from the protein coding mRNAs of host genes containing the enhancers will discriminate between genes whose expression is increased in a specific cell type from those genes whose expression may simply increase as a result of containing enhancers for other genes (bystander effect) (Cajiao et al., 2004). Identifying meRNAs together with eRNAs may solve a long standing problem of analysing when and where enhancers are activated during commitment, differentiation and development at a single cell level.

Experimental procedures

Primary cells and cell lines

Mouse primary erythroid cells were sorted from the spleens of acetylphenylhydrazine-treated (to induce acute hemolytic anemia) mice based on the expression of Ter119 antigen (Spivak et al., 1973; Vernimmen et al., 2009). Mouse erythroid cells were grown from fetal livers following published protocol (Dolznig et al., 2001). The human erythroid progenitors were obtained from mononuclear cells in the two-phase liquid cell culture (Brown et al., 2006; Fibach et al., 1989).

Deletional constructs

The targeting constructs (pP6, pR3) were assembled in pNTflox vector. Homology arms were cloned onto each side of a *loxP* flanked PGK-*neo* cassette. The ΔP6 deleted segment spans 2315bp between coordinates chr11:32,166,133-32,168,448; the ΔP6-R3 deleted segment spans 12403bp between coordinates chr11:32,156,045-32,168,448. All coordinates were obtained with the mouse Build 37 (NCBI37/mm9) as reference. Characteristics of deleted elements, ES cells targeting, screening and generation of mouse models are described in the Extended Experimental Procedures.

RNA blot

RNA was polyA selected using the PolyATract mRNA isolation system (Promega). Northern blots were performed using NorthernMax-Gly kit (Ambion) following the manufacturer's protocol.

RNA-sequencing (RNA-Seq)

For RNA-Seq library preparation total RNA was split into poly(A)$^+$ RNA and poly(A)$^-$ RNA using the PolyATract mRNA isolation system (Promega). Poly(A)$^+$ RNA libraries were produced using the Illumina mRNA-Seq pair-end kit after depletion of globin transcripts using GlobinClear (Ambion). Poly(A)$^-$ RNA libraries were produced using the Illumina DGE Small RNA Sample Prep kit with minor modifications, after depleted of ribosomal transcripts with RiboMinus Eukaryote Kit for RNA-sequencing (Invitrogen).

Chromatin immunoprecipitation (ChIP) and ChIP-sequencing (ChIP-Seq)

ChIPs were performed as described previously (De Gobbi et al., 2007). ChIP-seq libraries were prepared and sequenced using the standard Illumina protocol.

***DNaseI* assay and *DNaseI*-Sequencing (*DNaseI*-Seq)**

Nuclei from primary erythroid cells (Ter119+) were digested with 8 increasing concentrations of *DnaseI* (Roche) (Higgs et al., 1990). 1.5 mg of DNA from the mid-phase digestions was blunt-ended with T4 DNA Polymerase (NEB), and prepared for Illumina GAII sequencing. The *DNaseI*-digested material was amplified using PCR primer PE1.0 and PE2.0 (Illumina).

Antibodies

The following antibodies were used for ChIP: anti-H3K4me1 (Upstate 07-436), anti-H3K4me3 (Upstate 07-473), anti-H3K27me3 (Upstate 07-449), anti-H3K27ac (Abcam ab4729), anti-RNA-PolII (Santa Cruz Biotechnology H224).

Protein blots were performed using anti-Nprl3 (C16B) (Lunardi et al., 2009), anti-GST (Santa Cruz Biotechnology sc-138) and anti-Gapdh (Cell Signaling, 3683).

Accession numbers

Sequencing datasets described in this study have been deposited at the NCBI GEO (GSE27921).

References

Anguita, E., Hughes, J., Heyworth, C., Blobel, G.A., Wood, W.G., and Higgs, D.R. (2004). Globin gene activation during haemopoiesis is driven by protein complexes nucleated by GATA-1 and GATA-2. EMBO J *23*, 2841-2852.

Anguita, E., Sharpe, J.A., Sloane-Stanley, J.A., Tufarelli, C., Higgs, D.R., and Wood, W.G. (2002). Deletion of the mouse alpha-globin regulatory element (HS -26) has an unexpectedly mild phenotype. Blood *100*, 3450-3456.

Bernstein, B.E., Stamatoyannopoulos, J.A., Costello, J.F., Ren, B., Milosavljevic, A., Meissner, A., Kellis, M., Marra, M.A., Beaudet, A.L., Ecker, J.R.*, et al.* (2010). The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol *28*, 1045-1048.

Brown, J.M., Leach, J., Reittie, J.E., Atzberger, A., Lee-Prudhoe, J., Wood, W.G., Higgs, D.R., Iborra, F.J., and Buckle, V.J. (2006). Coregulated human globin genes are frequently in spatial proximity when active. J Cell Biol *172*, 177-187.

Bulger, M., and Groudine, M. (2010). Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev Biol *339*, 250-257.

Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., and Rinn, J.L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev *25*, 1915-1927.

Cajiao, I., Zhang, A., Yoo, E.J., Cooke, N.E., and Liebhaber, S.A. (2004). Bystander gene activation by a locus control region. EMBO J *23*, 3854-3863.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M.C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C.*, et al.* (2005). The transcriptional landscape of the mammalian genome. Science *309*, 1559-1563.

Cheng, Y., Wu, W., Kumar, S.A., Yu, D., Deng, W., Tripic, T., King, D.C., Chen, K.B., Zhang, Y., Drautz, D.*, et al.* (2009). Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res *19*, 2172-2184.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M.A., Frampton, G.M., Sharp, P.A.*, et al.* (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc Natl Acad Sci U S A.

De Gobbi, M., Anguita, E., Hughes, J., Sloane-Stanley, J.A., Sharpe, J.A., Koch, C.M., Dunham, I., Gibbons, R.J., Wood, W.G., and Higgs, D.R. (2007). Tissue-specific histone modification and transcription factor binding in alpha globin gene expression. Blood *110*, 4503-4510.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol *8*, e1000384.

Dolznig, H., Boulme, F., Stangl, K., Deiner, E.M., Mikulits, W., Beug, H., and Mullner, E.W. (2001). Establishment of normal, terminally differentiating

mouse erythroid progenitors: molecular characterization by cDNA arrays. FASEB J *15*, 1442-1444.
Fibach, E., Manor, D., Oppenheim, A., and Rachmilewitz, E.A. (1989). Proliferation and maturation of human erythroid progenitors in liquid culture. Blood *73*, 100-103.
Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P.*, et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223-227.
He, Y., Vogelstein, B., Velculescu, V.E., Papadopoulos, N., and Kinzler, K.W. (2008). The antisense transcriptomes of human cells. Science *322*, 1855-1857.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.
Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.
Higgs, D.R., and Wood, W.G. (2008). Long-range regulation of alpha globin gene expression during erythropoiesis. Curr Opin Hematol *15*, 176-183.
Higgs, D.R., Wood, W.G., Jarman, A.P., Sharpe, J., Lida, J., Pretorius, I.M., and Ayyub, H. (1990). A major positive regulatory region located far upstream of the human alpha-globin gene locus. Genes Dev *4*, 1588-1601.
Hughes, J.R., Cheng, J.F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E., and Higgs, D.R. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc Natl Acad Sci U S A *102*, 9830-9835.
Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A.T., Stadler, P.F., Hertel, J., Hackermuller, J., Hofacker, I.L.*, et al.* (2007). RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science *316*, 1484-1488.
Kapranov, P., Drenkow, J., Cheng, J., Long, J., Helt, G., Dike, S., and Gingeras, T.R. (2005). Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. Genome Res *15*, 987-997.
Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.
Khalil, A.M., Guttman, M., Huarte, M., Garber, M., Raj, A., Rivea Morales, D., Thomas, K., Presser, A., Bernstein, B.E., van Oudenaarden, A.*, et al.* (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proc Natl Acad Sci U S A *106*, 11667-11672.
Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S.*, et al.* (2010).

Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O.*, et al.* (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289-293.
Lower, K.M., Hughes, J.R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., Goriely, A., Ayyub, H., Sloane-Stanley, J., Vernimmen, D.*, et al.* (2009). Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proc Natl Acad Sci U S A *106*, 21771-21776.
Lunardi, A., Chiacchiera, F., D'Este, E., Carotti, M., Dal Ferro, M., Di Minin, G., Del Sal, G., and Collavin, L. (2009). The evolutionary conserved gene C16orf35 encodes a nucleo-cytoplasmic protein that interacts with p73. Biochem Biophys Res Commun *388*, 428-433.
Mattick, J.S., Taft, R.J., and Faulkner, G.J. (2010). A global view of genomic information--moving beyond the gene and the master regulator. Trends Genet *26*, 21-28.
Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279-283.
Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.
Seila, A.C., Core, L.J., Lis, J.T., and Sharp, P.A. (2009). Divergent transcription: a new feature of active promoters. Cell Cycle *8*, 2557-2564.
Spivak, J.L., Toretti, D., and Dickerman, H.W. (1973). Effect of phenylhydrazine-induced hemolytic anemia on nuclear RNA polymerase activity of the mouse spleen. Blood *42*, 257-266.
Tallack, M.R., Whitington, T., Yuen, W.S., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey, T.L.*, et al.* (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res *20*, 1052-1063.
Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.
Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.
Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO J *26*, 2041-2051.
Vernimmen, D., Marques-Kranc, F., Sharpe, J.A., Sloane-Stanley, J.A., Wood, W.G., Wallace, H.A., Smith, A.J., and Higgs, D.R. (2009).

Chromosome looping at the human {alpha} globin locus is mediated via the major upstream regulatory element (HS -40). Blood.
Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F.*, et al.* (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854-858.

Figure legends

Figure 1

Nprl3 promoter knockout failed to abolish *Nprl3* transcription in erythroid cells

(A) The mouse α-globin cluster (11q). The globin genes are shown as red boxes. The *Nprl3* gene containing enhancers is in green. Other genes are shown as black boxes. Genes above the line are transcribed from left to right and those below the line from right to left

(B) The genomic structure of the *Nprl3* gene with associated mRNA isoforms. *Cis* elements are shown above the *Nprl3* with R and P representing regulatory and promoter elements respectively. DHS-12 is a mouse specific *DnaseI* hypersensitive site, which shows similar transcription factors binding profile to other enhancers (R1 to R4), but has not been conserved (Anguita et al., 2004). The direction of the *Nprl3* gene transcription is shown as a black arrow. A cartoon forms of annotated *Nprl3* mRNA isoforms are shown below the genomic structure. The P6 deletion (ΔP6) is indicated as a grey rectangle (see also Figure S1A). The deletion between P6 and R3 is show as a dashed box. The alternative first exon (AFE) of the alternative transcript (AK036633; Ensembl ENSMUSG00000020289) coincides with enhancer R3 and is highlighted within the red rectangle. Quantitative RT-PCR (qPCR) primers (light blue) used in **(C)** and **(F)** span the junction between exon 7 and 8. RT-PCR primers (dark blue) used in **(D), (E)** span the junction between AFE and

exon 3. Exons (black boxes); introns (white boxes); splicing of mature transcripts (dashed lines)

(C) mRNA expression of *Nprl3* across mouse fetal tissues at embryonic day E14.5. Fetal liver at this stage of development is a source of erythroid cells.

(D) Specific products corresponding to the alternative transcript were detectable only in cDNA sample from mouse bone marrow (red arrowheads pointing to the PCR products). Conventional sequencing revealed that the ~300bp band corresponds to the predicted product size (311bp) and the ~200bp band was a splice variant of the alternative transcript, which used an alternative splice site within the AFE. All tissues showed a band (~800bp) corresponding to the unspliced *Nprl3* transcript

(E) The alternative transcript is present in purified mouse erythroid cells (Ter119+). Similarly to **(D)** (~800bp) corresponds to the unspliced *Nprl3* transcript. A ladder of bands specific to erythroid cells corresponds to the PCR products containing splice variants of AFE

For **(D)** and **(E)**, RT-PCR was performed with primers shown in **(B)** using cDNA from a panel of mouse WT tissues **(D)** and primary mouse non-erythroid (Ter119-) and erythroid (Ter119+) cells **(E)**. Mouse genomic DNA was used as a positive control and negative control excluded polymerase. RT(-) reactions excluded reverse transcriptase

(F) mRNA expression of *Nprl3* in mouse *ΔP6-R3* fetal tissues at embryonic day E14.5. Fetal liver and brain were used as erythroid and non-erythroid tissue respectively. For **(C)** and **(F)**, the qPCR primers used are shown on **(B)** and span exon 7-8 junction. *Nprl3* expression was calculated relative to mouse *Gapdh*. For each tissue, the mean of expression in WT is set to 1, and expression in linked samples is expressed relative to this mean. All error bars represent ± SD; n≥3.

Figure 2

The epigenetic and transcriptional landscape of the *Nprl3* locus in mouse erythroid cells.

The mouse *Nprl3* locus with high-resolution maps of poly(A)$^-$ transcription, chromatin states and erythroid-specific transcription factor binding. Poly(A)$^-$ RNA-Seq data was split into top (light blue) and bottom strand (dark blue). Arrows indicate bi-directional transcription from the centre of enhancer R4. The y-axis represents reads density. The enhancers and the *Nprl3* promoter are highlighted as grey and red columns respectively. UCSC Genes annotation is shown at the bottom.

Figure 3

New long poly(A)$^+$ RNAs within the *Nprl3* locus in wild type and mutant (ΔP6-R3) erythroid cells

(A) Overview of the poly(A)$^+$ transcription within the *Nprl3* locus in WT erythroid cells. Spliced reads from RNA-Seq data are displayed split into two classes: reads associated with AFE splice junctions and with annotated splice junctions. Cufflinks transcripts reconstruction compared to the UCSC Genes annotation. The Cufflinks isoform associated with enhancer R3 was found within Ensembl genome annotation. The enhancers (grey columns) and the *Nprl3* CpG island are shown.

(B) The mouse *Nprl3* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data for wild type and ΔP6-R3 erythroid cells. The y-axis represents fragments per base pair per million reads aligned. The enhancers and the *Nprl3* promoter are highlighted as grey and red columns respectively. UCSC Genes annotation is shown in purple and UCSC CpG Island annotation is shown as orange boxes. High-resolution maps for the chromatin markers *DnaseI* hypersensitivity, H3K4me1 and H3K4me3 are shown below. The y-axes represent reads density.

Figure 4

Genome-wide identification of enhancers in erythroid cells.

(A) All detected mouse erythroid DHSs were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. The same sort order was used for all panels displayed here and its direction is shown as red (H3K4me3) and blue triangle (H3K4me1). Analysis of the *DnaseI* hypersensitivity, H3K4me1, H3K4me3, H3K27ac, p300, Gata1, Scl, Klf1 and Ldb1 data in the same sort order clearly shows the segregation of the DHS sites into H3K4me1 (enhancers) and H3K4me3 (promoters) enriched populations of which most erythroid specific transcription factors are bound to enhancers. The red rectangle indicated the cut-off used for identification of enhancers. Each panel shows the distribution of signal in a 4-kb window centred in the middle of each DHS.

Chromatin profiles normalized for number of peaks for the mouse erythroid intragenic enhancers population is shown in **(B)** and all annotated mouse TSSs (UCSC Genes) is shown in **(C)**. Color-coding for each chromatin mark is shown below.

(D) Mouse erythroid enhancers were sorted based on the level of antisense poly(A)$^-$ transcription. Red and blue rectangles contain high and low transcribing populations of enhancers respectively.

(E) and **(F)** The cumulative poly(A)$^-$ transcription associated with intragenic enhancers in the antisense **(E)** and sense **(F)** direction (relative to the transcription of the host gene) is seen to originate close to the midpoint of the enhancers and extend ~1kb in the antisense direction in mouse erythroid cells. The poly(A)$^-$ transcription in the sense direction is masked by the host gene transcription (see relatively higher background in **(F)** in comparison to **(E)**)

(G) The comparison between high (red) and low (blue) transcribing enhancers is displayed as enrichment of various factors. The high and low transcribing enhancer populations are as indicated in **(D)**. The fold difference for each factor is indicated above the graphs.

Figure 5

AFEs are associated with enhancers and produced meRNAs are expressed at similar levels to neighbouring protein-coding genes

(A) The idealized structure of a gene containing an enhancer is shown. The TSS is shown as a black arrow. The midpoint of the enhancer is represented as a red arrow and aligned with the other data. mRNA forms are represented below the gene; the AFE is shown in red. Exons (black boxes); introns (white boxes); splicing of mature transcripts (dashed lines).

The frequency of AFE is shown in 200 bp bins relative to the midpoint of intragenic enhancers over a 5kb window.

(B), **(C)** and **(D)** give examples of meRNAs which are expressed from within genes with inactive canonical promoters (*D18Ertd653e*, *Acmsd*, *Abat*). On each panel normalised poly(A)$^+$ RNA-Seq data in wild type erythroid cells is displayed at the top and cartoon form of UCSC Genes annotation at the bottom. Genes are as grey rectangles containing black exons. The transcription from canonical promoters and enhancers is indicated as black and red arrows respectively. The extent of the enhancer transcripts (*meD18Ertd653e* in **B**, *meAcmsd* in **C** and *meAbat* in **D**) is highlighted within beige rectangles. On each panel a green dashed line connects the alternative first exon of each meRNA to the first exon of neighbouring gene to compare meRNA expression levels to mRNA from protein-coding genes. In **C** the expression of *meAcmsd* is compared to the second exon of *Ccnt2* gene because the overlapping signal from first exons of *Ccnt2* and *AK013506* genes.

Figure 6

Erythroid enhancers give rise to intact and full length poly(A)$^+$ RNA (meRNA)

(A) The mouse *Tg* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data in wild type erythroid cells. The y-axes represent the normalised expression value, fragments per base pair per million reads aligned. Splice reads detected in erythroid cells are displayed below the RNA-Seq track. UCSC Genes annotation is shown.

The canonical promoter of the *Tg* gene is inactive in erythroid cells, but black arrow indicates the TSS from this promoter (encoding for protein-coding mRNA). A beige rectangle highlights the extent of the enhancer transcript (me*Tg*). A red arrow indicates the TSS from this enhancer (encoding for meRNA).

(B) and **(C)** Northern blots show intact and full length *meTg* RNA **(B)** and *meZnfx1* RNA **(C)** of expected sizes. Both meRNAs are present in mouse erythroid cells (two biological replicates), but are absent in non-erythroid cells (brain, ES, L929 cells). The *Tg* mRNA from the canonical promoter is not expressed in any of cells tested **(B)**. The intact and full length *Znfx1* mRNA from the canonical promoter is present in all cells tested **(C)**. Beta actin RNA is shown as a loading control.

(D) The mouse *Znfx1* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data in wild type erythroid cells. The y-axes represent the normalised expression value, fragments per base pair per million reads aligned. Reads spanning splice junctions containing unannotated exon (new

spliced reads) is displayed separately from spliced reads containing annotated reads and both are shown below the RNA-Seq track. UCSC Genes annotation is shown. Ensembl annotation shows the annotation of alternative *Znfx1* transcript which we found here to be an erythroid-specific enhancer driven poly(A)$^+$ RNA (*meZnfx1*).

The canonical promoter of the *Znfx1* gene (black arrow) is active in erythroid cells and therefore to some degree masks the transcription from the erythroid enhancer. A beige rectangle highlights the extent of the enhancer transcript (me*Znfx1*). A red arrow indicates the start and the direction of transcription from this enhancer (encoding for meRNA).

Figure 7

Tissue specific enhancer transcripts within current transcriptome annotation.

(A) The mouse *D18Ertd653e* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data for wild type erythroid and brain cells. The y-axes represent the normalised expression value, fragments per base pair per million reads aligned.

The canonical promoter of the *D18Ertd653e* gene is associated with a CpG island and a black arrow indicates the start and the direction of transcription from this promoter (encoding for protein-coding mRNA). Transcription from this promoter is present in the brain but absent in the erythroid cells.

A beige column highlights the extent of the enhancer transcript (me*D18Ertd653e*). A red arrow indicates the start and the direction of transcription from this enhancer (encoding for meRNA). Transcription from this enhancer is present in the erythroid but absent in the brain cells. The spliced reads detected in erythroid cells are displayed below the RNA-Seq tracks.

UCSC Genes annotation including the annotated enhancer transcript (the annotated AFE is shown with an arrow) is shown. UCSC CpG Island annotation is shown as orange box. High-resolution maps for the chromatin markers H3K27me3, *DnaseI* hypersensitivity (DHS), H3K4me1 and H3K4me3 are shown below. The y-axes represent reads density.

(B) The mouse transcription start sites (from UCSC Genes and “Refseq” gene annotations) which overlap with a single DHS site in Ter119+ cells (13,506

sites), were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the chromatin marks, as in **Figure 4A**, shows 139 of TSSs active in erythroid cells resembles an enhancer chromatin signature (indicated by the red rectangle).

(C) The human transcription start sites (from UCSC Genes and "Refseq" gene annotations) which overlap with a single DHS site in IMR90 cells (16,508 sites), were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the chromatin marks, as in **Figure S5A**), shows 693 of TSSs active in IMR90 cells resembles an enhancer chromatin signature (indicated by the red rectangle).

In both **(B)** and **(C)** the upper panels show the 2000 most H3K4me1 enriched TSSs and the lower panels show the 2000 most H3K4me3 enriched TSSs.

(D) A venn diagram of the active TSSs in human IMR90, K562 and GM12878 cells with an enhancer chromatin signature shows the highly tissue specific nature of meRNAs (enhancer derived poly(A)$^+$ transcripts).

Supplementary Figure Legends

Figure S1

***Nprl3* expression and generated deletions, related to Figure 1**

(A) ΔP6 model.

(i) A schematic diagram of the *Nprl3* gene and deleted segment (grey rectangle).

(ii) A diagram showing WT, targeted and recombined alleles (from top to bottom). In the targeted allele P6 segment is substituted by *neo*. In the recombined allele *neo* is replaced by a single loxP site. The extent of the homology arms is shown in black and loxP sites as red arrowheads. *KpnI* sites are indicated as black arrows and the probe (PV) as green rectangle. Primers used in **(iv)** are shown as blue arrows.

(iii) Representative Southern blot shows restriction fragments of genomic DNA digested with *KpnI* and hybridised with the PV probe for WT, $\Delta P6^{+/flox}$ and $\Delta P6^{+/-}$ mutant mice. The sizes of expected restriction fragments for each type of allele hybridised with the PV probe are given to the right of the autoradiograph.

(iv) PCR analysis of $\Delta P6^{+/-}$ *x* $\Delta P6^{+/-}$ offspring using 14RecF/14RecR/14WTR primers.

(B) Representative images showing the *Nprl3* transcript (red) in DAPI-stained nuclei (blue). *Nprl3* transcription is seen in homozygous mutant embryo erythroid cells from the liver but not in brain. Bar is 5μm.

(C) Nascent transcription from the *Nprl3* gene in WT, heterozygous ($\Delta P6^{+/-}$) and homozygous ($\Delta P6^{-/-}$) mutant E14.5 brain and liver, detected by RNA FISH scored as % of active genes in E14.5 brain and liver. The reduced signal found in mutant brain cells is not observed in mutant erythroid cells.

(D) mRNA expression of *Nprl3* in differentiating human erythroid cells. Two-phase *in vitro* cultures (Fibach et al., 1989) were used for erythroid differentiation. The three stages shown here (early, intermediate, late) were enriched in cells CD36+/GPA-/CD71+, GPA+/CD71+, GPA-/CD71+ and correspond to human proerythroblasts, polychromatic and orthochromatic erythroblasts respectively. CD36 – glycoprotein IV, GPA – glycophorin A, CD71 – transferrin receptor. Epstein Barr Virus (EBV) transformed B lymphocyte lines were used as non-erythroid cells.

(E) mRNA expression of *Nprl3* in mouse non-erythroid (Ter119-) erythroid cells (Ter119+). For primers see table in Extended Experimental Procedures.

Figure S2

***ΔP6-R3*$^{-/-}$ model and analysis of transcription in the *Nprl3* locus of the *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ model, related to Figure 3**

(A) ΔP6-R3 model.

(i) A schematic diagram of the *Nprl3* gene and deleted segment (grey rectangle).

(ii) A diagram showing WT, targeted and recombined alleles (from top to bottom). A clone already deleted for the P6 segment (single loxP site remaining) was retargeted. In the targeted allele R3 segment is substituted by *neo*. In the recombined allele *neo* is replaced by a single loxP site. The extent of the homology arms for R3 and P6 regions is shown in black and grey respectively. loxP sites are shown as red arrowheads. *BstEII* sites are indicated as black arrows and the probe (LEFT2) as yellow rectangle. Primers used in **(iv)** are shown as blue arrows.

(iii) Representative Southern blot shows restriction fragments of genomic DNA digested with *BstEII* and hybridised with the LEFT2 probe for WT, *ΔP6-R3*$^{+/flox}$ and *ΔP6-R3*$^{+/-}$ mutant mice. The sizes of expected restriction fragments for each type of allele hybridised with the LEFT2 probe are given to the right of the autoradiograph.

(iv) PCR analysis of *ΔP6-R3*$^{+/-}$ *x* *ΔP6-R3*$^{+/-}$ offspring using 1421RecF/1421RecR/1421WTR primers.

(B) Total RNA expression profile within *Nprl3* locus in *ΔP6*$^{-/-}$, *ΔP6-R3*$^{-/-}$ and WT fetal tissues on a genomic tiled array. RNA signal from erythroid (fetal liver) and non-erythroid (fetal brain) tissues is displayed in red and black

respectively. Red/Black and grey bars represent logarithmic level of the RNA enrichment relative to genomic input above and below zero respectively. The deletions are indicated as black boxes which extend as grey transparent boxes over the *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ tracks. *Cis* elements are shown as black boxes above the tracks. The annotated genes are in purple and erythroid-specific transcript (AK036633) in red.

Figure S3

New long poly(A)$^+$ RNAs within the *Nprl3* locus, related to Figure 3

In **(A)-(C)** a schematic diagram of *Nprl3* gene is shown. Exons (black) are numbered to indicate the direction of gene transcription. The positions of key *cis*-regulatory elements of α-globin genes are indicated. Each Cufflinks reconstructed RNA isoform is displayed as a mature transcript (dashed lines). Splicing between the AFE and downstream exons was validated by PCR with reverse transcription (RT–PCR). The primers are shown as blue arrows and exons within the expected PCR product are in red. Bottom panel shows RT–PCR products in erythroid and non-erythroid tissues on 2% agarose gels. These products are erythroid specific, of the expected sizes and were validated by Sanger sequencing.

(A) Validation of the new RNA isoforms in WT

(B) Validation of the new RNA isoforms in ΔMCS-P6 mouse model

(C) Validation of the new RNA isoforms in ΔMCS-P6-R3 mouse model

(D) Western blot analysis of Nprl3 protein in ΔMCS-P6 model. Gapdh was used as a loading control. The protein products are of the expected sizes. The non-specific bands are common in non-erythroid and erythroid cells.

(E) Nprl3 antibody titration. The putative mouse Nprl3 transcript initiating in exon2 was expressed and purified as a GST tagged protein (left panel,

coomassie stain). A dilution series of quantified GST-Nprl3 protein was blotted and probed with 1/500 α-Nprl3 antibody. Middle panel: 100, 50 and 10 ng of GST-Nprl3, short 2 minute exposure. Right panel: 1, 0.5, 0.1 and 0.05 ng of GST-Nprl3, long 30 minute exposure.

Figure S4

Comparison of the DNA sequence between of enhancers and promoters, related to Figure 4

All plots are shown in the sense direction of the associated gene. Each graph shows the distribution of signal in a 8-kb window centred on the middle of each enhancer and promoter respectively.

(A) The cumulative occurrence of the CG dinucleotide in either strand is shown for promoters and intragenic enhancers are shown in red and blue respectively. Promoters show a strong enrichment for the CG dinucleotide because they are associated with CpG islands, intragenic enhancers do not.

(B) The cumulative conservation scores (Phastcons UCSC) for promoters and intragenic enhancers are shown in blue and red respectively. The cumulative association with annotated coding regions (UCSC Known Gene) for promoters and intragenic enhancers are shown in purple and green respectively. The cumulative coding potential of the underlying DNA sequence, independent of gene annotation (UCSC Exoniphy), for promoters and intragenic enhancers are shown in grey and gold respectively.

The association of promoters with downstream protein coding exons produces a strong skewed signal of conservation, whereas no such conservation is seen at canonical promoters

(D) and **(E)** The cumulative occurrence of promoter associated motifs (Kozak in blue, B recognition element (BRE) in red, CAATT boxes in purple and splice donor sites in green) are shown for promoter and intragenic enhancers

are shown in **(D)** and **(E)** respectively. Enhancers lack the CAATT box motif, the Kozak consensus sequence and other highly positioned features present at canonical promoters.

(F) The cumulative occurrence of the two published binding motifs for the erythroid transcription factor Gata1 in either strand are shown for promoters (T/GATAAA in blue and CAG(N9)GATA in red) and intragenic enhancers (T/GATAAA in purple and CAG(N9)GATA in green). Enhancers are highly enriched with DNA motifs, which direct the binding of erythroid transcription factors.

Figure S5

Genome-wide identification of enhancers and associated transcription in fetal human primary lung fibroblasts, related to Figure 4

(A) All human lung fibroblast DHSs were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the *DNAseI* hypersensitivity, H3K4me1 and H3K4me3 data in the same sort order clearly shows the segregation of the DHS sites into H3K4me1 and H3K4me3 enriched populations. The red rectangle indicated the cut-off used for identification of all enhancers. Each panel shows the distribution of signal in a 4-kb window centered in the middle of each DHS.

(B) The transcription start sites (from UCSC Genes and "Refseq" gene annotations) which overlap with a single DHS site in human fetal fibroblasts (16,508 sites), were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the chromatin marks, as in **(A)**, shows a similar, although smaller (~4%), H3K4me1 enriched population, showing that a proportion of annotated TSSs, at least within lung fibroblast cells, resembles an enhancer chromatin signature (indicated by the red rectangle). Each panel shows the distribution of signal in a 4-kb window centred in the middle of each TSS.

Chromatin profiles normalized for number of peaks for the human lung fibroblasts enhancers population is shown in **(C)** and all annotated human TSSs ("Known Gene") is shown in **(D)**. Color-coding for each chromatin mark is shown below.

(E) Human lung fibroblasts enhancers were sorted based on the level of antisense poly(A)$^-$ transcription. Red and blue rectangles contain high and low transcribing populations of enhancers respectively.

(F) The cumulative poly(A)$^-$ transcription (using global run-on method (GRO)) associated with intragenic enhancers in the antisense direction (relative to the transcriptional direction of the host gene) is seen to originate close to the midpoint of the enhancers and extends ~1kb in the sense direction in primary human lung fibroblasts.

(G) The cumulative poly(A)$^-$ transcription associated with intragenic enhancers in the sense direction (relative to the transcriptional direction of the host gene) is masked by the gene transcription (see relatively higher background in comparison to **(F)**) and is seen to originate close to the midpoint of the enhancers in primary human lung fibroblasts.

(H) The comparison between high (red) and low (blue) transcribing enhancers is displayed as enrichment of various factors. The high and low transcribing enhancer populations are as indicated in **(D)**. The fold difference for each factor is indicated above the graphs.

Figure S6

Examples of loci producing meRNAs (new long poly(A)$^+$ spliced isoforms from enhancers), related to Figure x

Loci in **(A,** *Zfpm1* or *Fog1* locus**)** and **(B,** *4922503N01Rik* locus**)** characterised by ChIP-Seq profiles (H3K4me3, H3K4me1, H3K4me3, RNAP2, Gata1 (Cheng et al., 2009), Klf1 (Tallack et al., 2010) and Scl (Kassouf et al., 2010)), *DNAseI* hypersensitivity and RNA-Seq. Each labeled AFE splice junction (1, 2, 3, 4, 5) was verified by junction-specific RT-PCR and is shown in **(C).**

(C) Junction-specific RT-PCRs (1-5) of RNA-Seq reads highlighted in **(A)** and **(B)**. A band representing unspliced nascent RNA is also detected in both erythroid and non-erythroid (**C**-5) due to its relatively small unspliced size. Corresponding reaction products of expected sizes are displayed.

Figure S7

***meAbat* is expressed from an inactive gene in erythroid cells, but only a proportion of the transcript has AFE, related to Figure 7**

The mouse *Abat* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data for wild type Ter119+ and brain cells. The y-axis represents the normalised expression value, fragments per base pair per million reads aligned. A red column highlights the *Abat* promoter region. A grey column highlights the extent of the enhancer transcript. UCSC Genes annotation is shown in purple. High-resolution maps for the chromatin markers *DnaseI* hypersensitivity (DHS), H3K4me1 and H3K4me3 are shown below. The y-axes represent reads density.

(B) Shows a zoomed view of the region of the *Abat* locus containing the enhancer transcript. The amount of signal from only spliced reads is shown in black (spliced signal). First exons show spliced signal only at the donor junction (gray column and black arrow) an exon which is both spliced to and transcribed through using only the donor signal shows a decreased acceptor relative to donor signal (red *). Below this in red are the predicted structures of the two isoforms produced by the alternate usage of the AFE and second exon acceptor sites. Normalised expression signal for brain poly(A)$^+$ full length canonical transcript and erythroid enhancer transcript are shown in black. The region corresponding to the unspliced AFE is shown in brown. UCSC Genes annotation is shown in purple. High-resolution maps for the

chromatin markers *DnaseI* hypersensitivity (DHS), H3K4me1 and H3K4me3 are shown below. The y-axes represent reads density.

Molecular Cell, Volume *45*

Supplemental Information

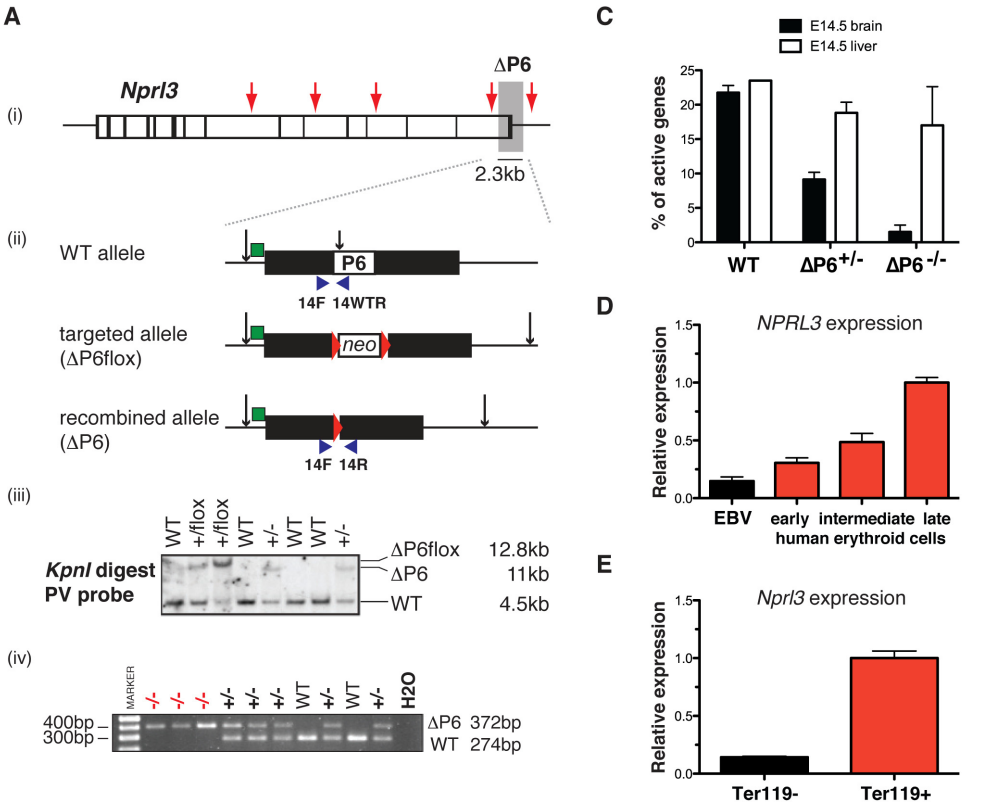
Intragenic Enhancers Act as Alternative Promoters

Monika S. Kowalczyk, Jim R. Hughes, David Garrick, Magnus D. Lynch, Jacqueline A. Sharpe, Jacqueline A. Sloane-Stanley, Simon J. McGowan, Marco De Gobbi, Mona Hosseini, Douglas Vernimmen, Jill M. Brown, Nicola E. Gray, Licio Collavin, Richard J. Gibbons, Jonathan Flint, Stephen Taylor, Veronica J. Buckle, Thomas A. Milne, William G. Wood, and Douglas R. Higgs

Inventory of Supplemental Information

The Supplemental Information file contains seven Supplemental Figures, Supplemental Experimental Procedures and Supplemental References.

Supplemental Figures



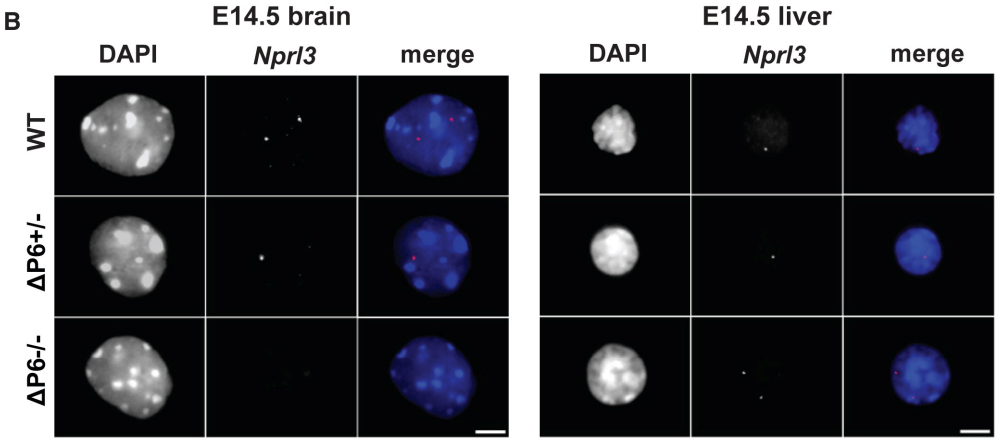


Figure S1

Figure S1

Nprl3 expression and generated deletions, related to Figure 1

(A) ΔP6 model.
(i) A schematic diagram of the *Nprl3* gene and deleted segment (grey rectangle).
(ii) A diagram showing WT, targeted and recombined alleles (from top to bottom). In the targeted allele the P6 segment is substituted by *neo*. In the recombined allele *neo* is replaced by a single loxP site. The extent of the homology arms is shown in black and loxP sites as red arrowheads. *KpnI* sites are indicated as black arrows and the probe (PV) as a green rectangle. Primers used in **(iv)** are shown as blue arrows.
(iii) Representative Southern blot shows restriction fragments of genomic DNA digested with *KpnI* and hybridised with the PV probe for WT, $\Delta P6^{+/flox}$ and $\Delta P6^{+/-}$ mutant mice. The sizes of expected restriction fragments for each type of allele hybridised with the PV probe are given to the right of the autoradiograph.
(iv) PCR analysis of $\Delta P6^{+/-}$ *x* $\Delta P6^{+/-}$ offspring using 14F/14R/14WTR primers. Primers used here are depicted in (***ii***). (+/-) and (-/-) denotes heterozygous ($\Delta P6^{+/-}$) and homozygous mutant ($\Delta P6^{-/-}$) offspring respectively.
(B) Representative images showing the *Nprl3* transcript (red) in DAPI-stained nuclei (blue). *Nprl3* transcription is seen in homozygous mutant embryo erythroid cells from the liver but not in brain. Bar is 5μm.
(C) Nascent transcription from the *Nprl3* gene in WT, heterozygous ($\Delta P6^{+/-}$) and homozygous ($\Delta P6^{-/-}$) mutant E14.5 brain and liver, detected by RNA FISH scored as % of active genes in E14.5 brain and liver cells. The reduced signal found in mutant brain cells is not observed in mutant erythroid cells.
(D) mRNA expression of *Nprl3* in differentiating human erythroid cells. Two-phase *in vitro* cultures (Fibach et al., 1989) were used for erythroid differentiation. The three stages shown here (early, intermediate, late) were enriched in cells CD36+/GPA-/CD71+, GPA+/CD71+, GPA-/CD71+ and correspond to human proerythroblasts, polychromatic and orthochromatic erythroblasts respectively. Epstein Barr Virus (EBV) transformed B lymphocyte lines were used as non-erythroid cells. CD36 – glycoprotein IV, GPA – glycophorin A, CD71 – transferrin receptor.
(E) mRNA expression of *Nprl3* in mouse non-erythroid (Ter119-) erythroid cells (Ter119+).

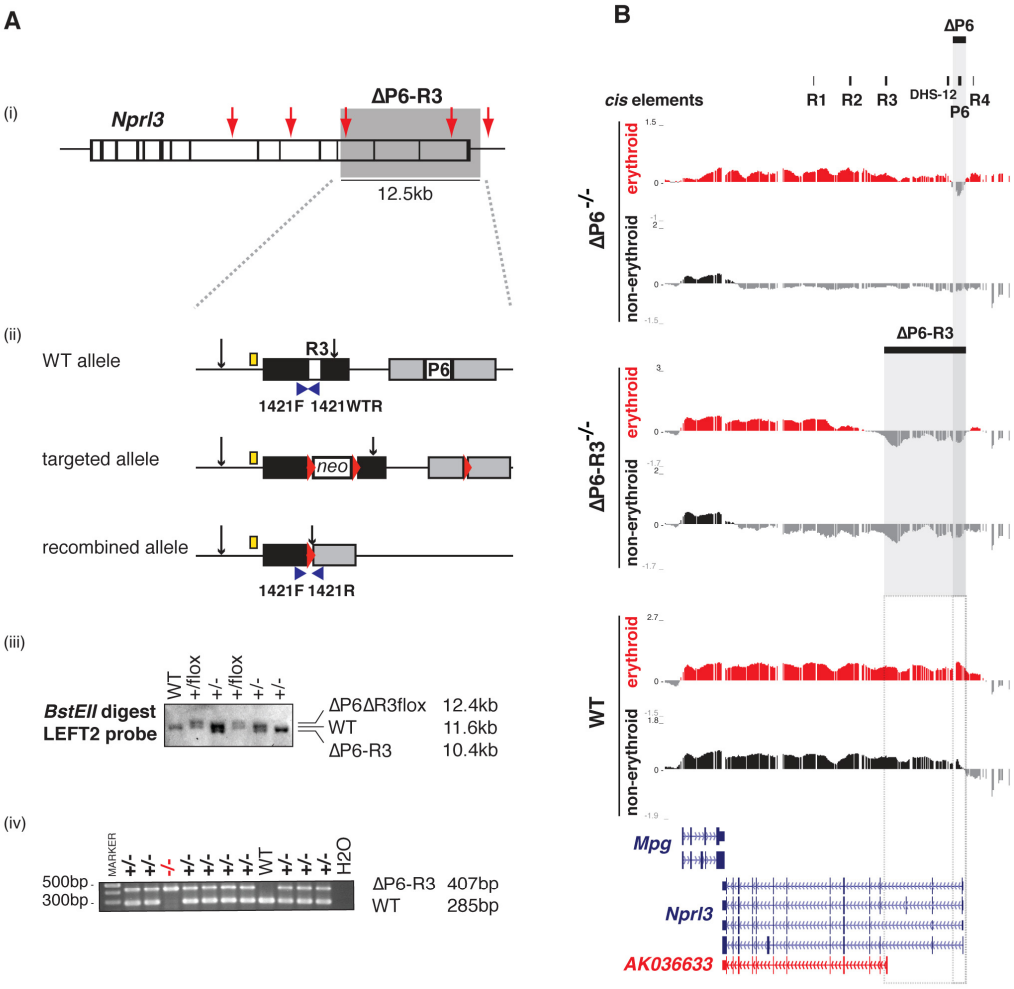

A
(i)
Nprl3
ΔP6-R3
12.5kb
(ii)
WT allele
R3
P6
1421F 1421WTR
targeted allele
neo
recombined allele
1421F 1421R
(iii)
BstEII digest
LEFT2 probe
WT
+/flox
+/-
+/flox
+/-
+/-
ΔP6ΔR3flox 12.4kb
WT 11.6kb
ΔP6-R3 10.4kb
(iv)
MARKER
+/+
+/+
-/-
+/+
+/+
+/+
+/+
WT
+/+
+/+
+/-
+/-
H2O
500bp
300bp
ΔP6-R3 407bp
WT 285bp
B
ΔP6
cis elements
R1
R2
R3
DHS-12
P6
R4
ΔP6 -/-
erythroid
non-erythroid
ΔP6-R3
ΔP6-R3 -/-
erythroid
non-erythroid
WT
erythroid
non-erythroid
Mpg
Nprl3
AK036633

Figure S2

Figure S2

ΔP6-R3$^{-/-}$ model and analysis of transcription in the *Nprl3* locus of the *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ model, related to Figure 3

(A) ΔP6-R3 model.
(i) A schematic diagram of the *Nprl3* gene and deleted segment (grey rectangle).
(ii) A diagram showing WT, targeted and recombined alleles (from top to bottom). A clone already deleted for the P6 segment (single loxP site remaining) was retargeted. In the targeted allele the R3 segment is substituted by *neo*. In the recombined allele *neo* is replaced by a single loxP site. The extent of the homology arms for R3 and P6 regions is shown in black and grey respectively. loxP sites are shown as red arrowheads. *BstEII* sites are indicated as black arrows and the probe (LEFT2) as a yellow rectangle. Primers used in **(iv)** are shown as blue arrows.
(iii) Representative Southern blot shows restriction fragments of genomic DNA digested with *BstEII* and hybridised with the LEFT2 probe for WT, *ΔP6-R3*$^{+/flox}$ and *ΔP6-R3*$^{+/-}$ mutant mice. The sizes of expected restriction fragments for each type of allele hybridised with the LEFT2 probe are given to the right of the autoradiograph.
(iv) PCR analysis of *ΔP6-R3*$^{+/-}$ *x* *ΔP6-R3*$^{+/-}$ offspring using 1421F/1421R/1421WTR primers. Primers used here are depicted in (***ii***). (+/-) and (-/-) denotes heterozygous (*ΔP6-R3*$^{+/-}$) and homozygous mutant (*ΔP6-R3*$^{-/-}$) offspring respectively.
(B) Total RNA expression profile within *Nprl3* locus in *ΔP6*$^{-/-}$, *ΔP6-R3*$^{-/-}$ and WT fetal tissues on a genomic tiled array. RNA signal from erythroid (fetal liver) and non-erythroid (fetal brain) tissues is displayed in red and black respectively. Red/Black and grey bars represent logarithmic level of the RNA enrichment relative to genomic input above and below zero respectively. The deletions are indicated as black boxes which extend as grey transparent boxes over the *ΔP6*$^{-/-}$ and *ΔP6-R3*$^{-/-}$ tracks. *Cis* elements are shown as black boxes above the tracks. The annotated genes are in purple and erythroid-specific transcript (AK036633) in red.

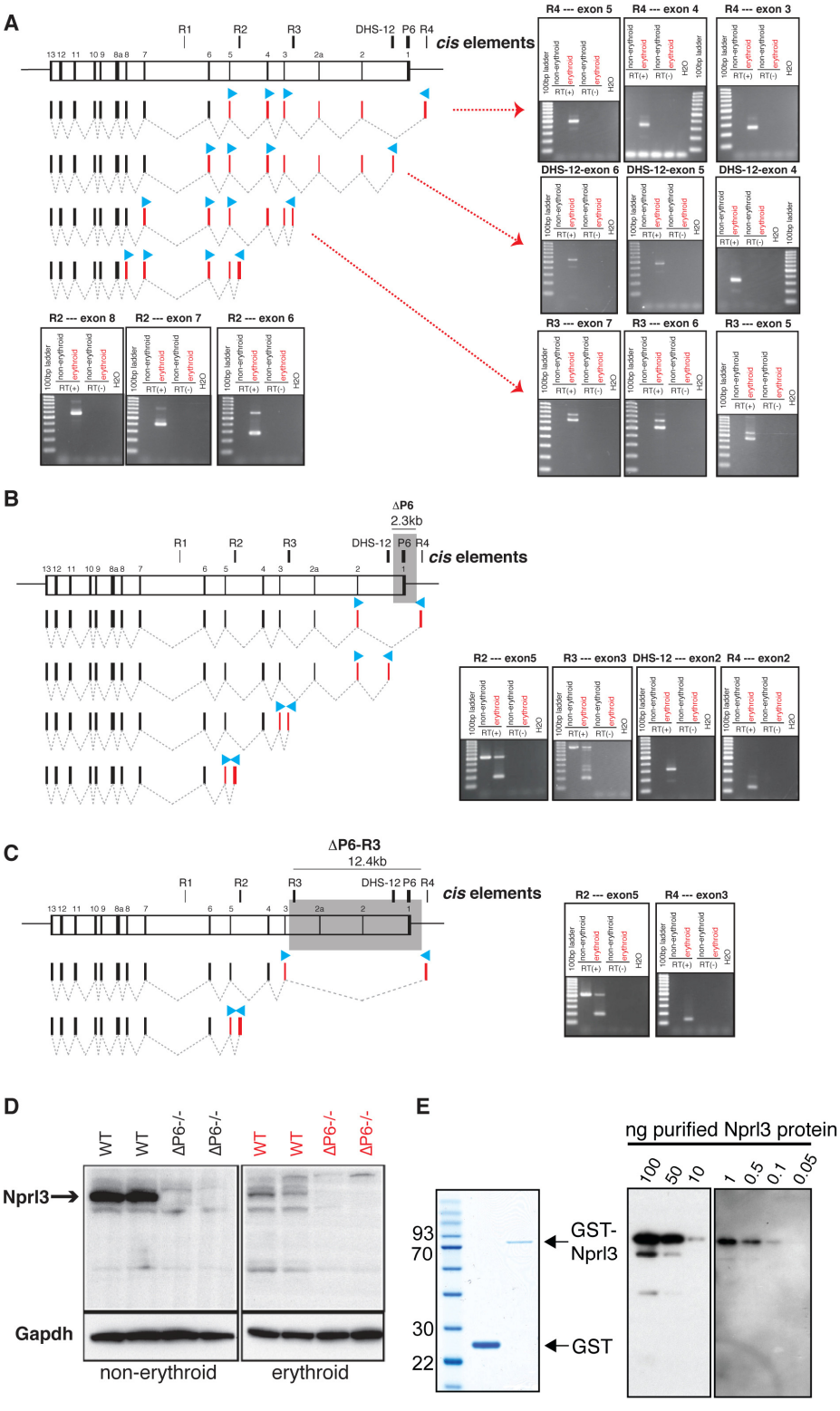

A
R1
R2
R3
DHS-12
P6
R4
cis elements
R4 --- exon 5
R4 --- exon 4
R4 --- exon 3
DHS-12-exon 6
DHS-12-exon 5
DHS-12-exon 4
R3 --- exon 7
R3 --- exon 6
R3 --- exon 5
R2 --- exon 8
R2 --- exon 7
R2 --- exon 6
B
ΔP6
2.3kb
cis elements
R2 --- exon5
R3 --- exon3
DHS-12 --- exon2
R4 --- exon2
C
ΔP6-R3
12.4kb
cis elements
R2 --- exon5
R4 --- exon3
D
WT
WT
ΔP6-/-
ΔP6-/-
WT
WT
ΔP6-/-
ΔP6-/-
Nprl3
Gapdh
non-erythroid
erythroid
E
ng purified Nprl3 protein
100
50
10
1
0.5
0.1
0.05
93
70
30
22
GST-Nprl3
GST

Figure S3

Figure S3

Long poly(A)$^+$ RNAs within the *Nprl3* locus, related to Figure 3

In **(A)-(C)** a schematic diagram of the *Nprl3* gene is shown. Exons (black) are numbered to indicate the direction of gene transcription. The positions of key *cis*-regulatory elements of the α-globin genes are indicated. Each Cufflinks reconstructed RNA isoform is displayed as a mature transcript (dashed lines). Splicing between the AFE and downstream exons was validated by PCR with reverse transcription (RT–PCR). The primers are shown as blue arrows and exons within the expected PCR product are shown in red. RT–PCR products in erythroid and non-erythroid tissues were visualised on 2% agarose gels. These products are erythroid specific, of the expected sizes and were validated by Sanger sequencing.

(A) Validation of the RNA isoforms in WT

(B) Validation of the RNA isoforms in ΔP6 mouse model

(C) Validation of the RNA isoforms in ΔP6-R3 mouse model

(D) Western blot analysis of Nprl3 protein in ΔP6 model. The anti-Nprl3 antibody was raised against C-terminal 390 amino acids of Nprl3 (Lunardi et al., 2009). Gapdh was used as a loading control. The protein products are of the expected sizes. The non-specific bands are common in non-erythroid and erythroid cells.

(E) Nprl3 antibody titration. The putative mouse *Nprl3* transcript initiating in exon2 was expressed and purified as a GST tagged protein (left panel, coomassie stain). A dilution series of quantified GST-Nprl3 protein was blotted and probed with 1/500 α-Nprl3 antibody. Middle panel: 100, 50 and 10 ng of GST-Nprl3, short 2 minute exposure. Right panel: 1, 0.5, 0.1 and 0.05 ng of GST-Nprl3, long 30 minute exposure.

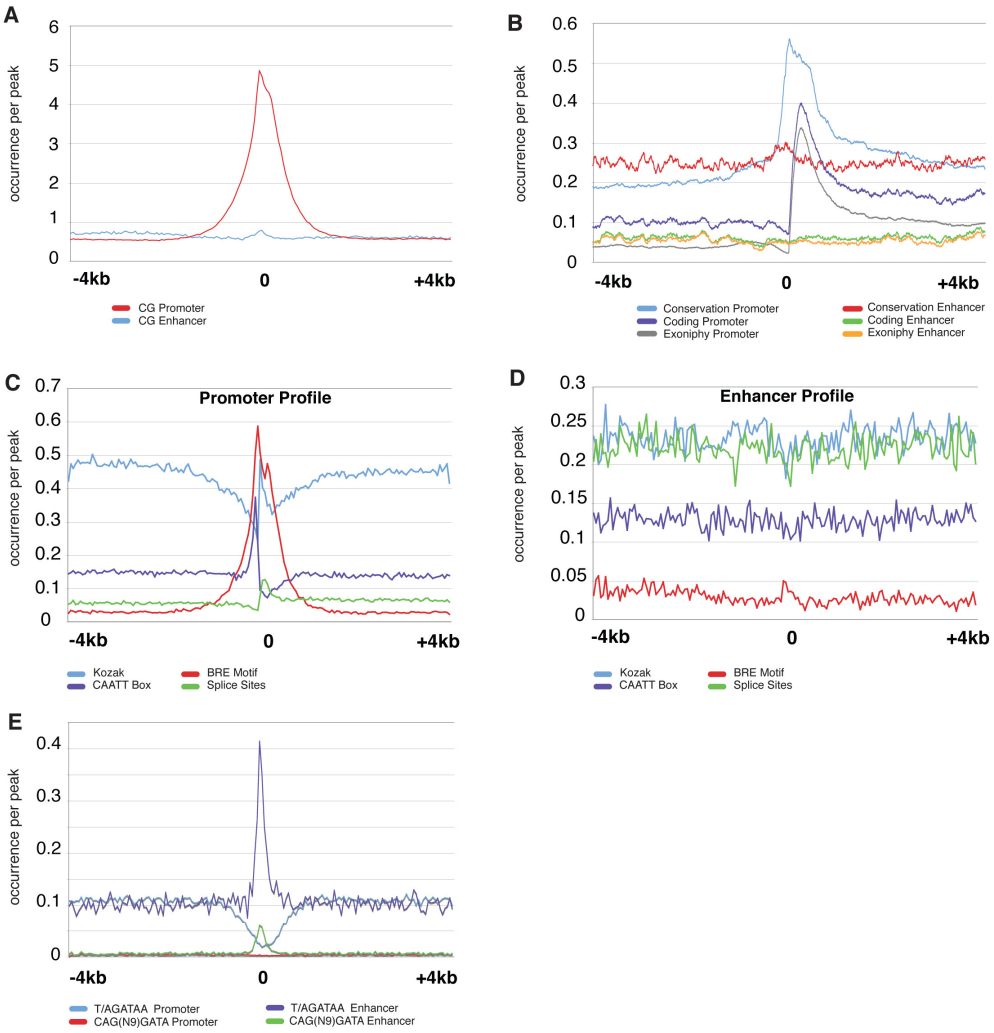

A
occurrence per peak
6
5
4
3
2
1
0
-4kb
0
+4kb
CG Promoter
CG Enhancer
B
occurrence per peak
0.6
0.5
0.4
0.3
0.2
0.1
0
-4kb
0
+4kb
Conservation Promoter
Coding Promoter
Exoniphy Promoter
Conservation Enhancer
Coding Enhancer
Exoniphy Enhancer
C
Promoter Profile
occurrence per peak
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4kb
0
+4kb
Kozak
CAATT Box
BRE Motif
Splice Sites
D
Enhancer Profile
occurrence per peak
0.3
0.25
0.2
0.15
0.1
0.05
0
-4kb
0
+4kb
Kozak
CAATT Box
BRE Motif
Splice Sites
E
occurrence per peak
0.4
0.3
0.2
0.1
0
-4kb
0
+4kb
T/AGATAA Promoter
CAG(N9)GATA Promoter
T/AGATAA Enhancer
CAG(N9)GATA Enhancer

Figure S4

Figure S4

Comparison of the DNA sequence composition of enhancers and promoters, related to Figure 4

All plots are shown in the sense direction of the associated gene. Each graph shows the distribution of signal in a 8-kb window centred on the middle of each enhancer and promoter respectively.

(A) The cumulative occurrence of the CG dinucleotide in either strand is shown for promoters and intragenic enhancers in red and blue respectively. Promoters show a strong enrichment for the CG dinucleotide because they are associated with CpG islands, intragenic enhancers do not.

(B) The cumulative conservation scores (Phastcons UCSC) for promoters and intragenic enhancers are shown in blue and red respectively. The cumulative association with annotated coding regions (UCSC Known Gene) for promoters and intragenic enhancers are shown in purple and green respectively. The cumulative coding potential of the underlying DNA sequence, independent of gene annotation (UCSC Exoniphy), for promoters and intragenic enhancers are shown in grey and gold respectively.

The association of promoters with downstream protein coding exons produces a strong skewed signal of conservation, whereas no such conservation is seen at canonical promoters

(C) and **(D)** The cumulative occurrence of promoter associated motifs (Kozak in blue, B recognition element (BRE) in red, CAATT boxes in purple and splice donor sites in green) are shown for promoter and intragenic enhancers are shown in **(C)** and **(D)** respectively. Enhancers lack the CAATT box motif, the Kozak consensus sequence and other highly positioned features present at canonical promoters.

(E) The cumulative occurrence of the two published binding motifs for the erythroid transcription factor Gata1 in either strand are shown for promoters (T/GATAAA in blue and CAG(N9)GATA in red) and intragenic enhancers (T/GATAAA in purple and CAG(N9)GATA in green). Enhancers are highly enriched with DNA motifs, which direct the binding of erythroid transcription factors.

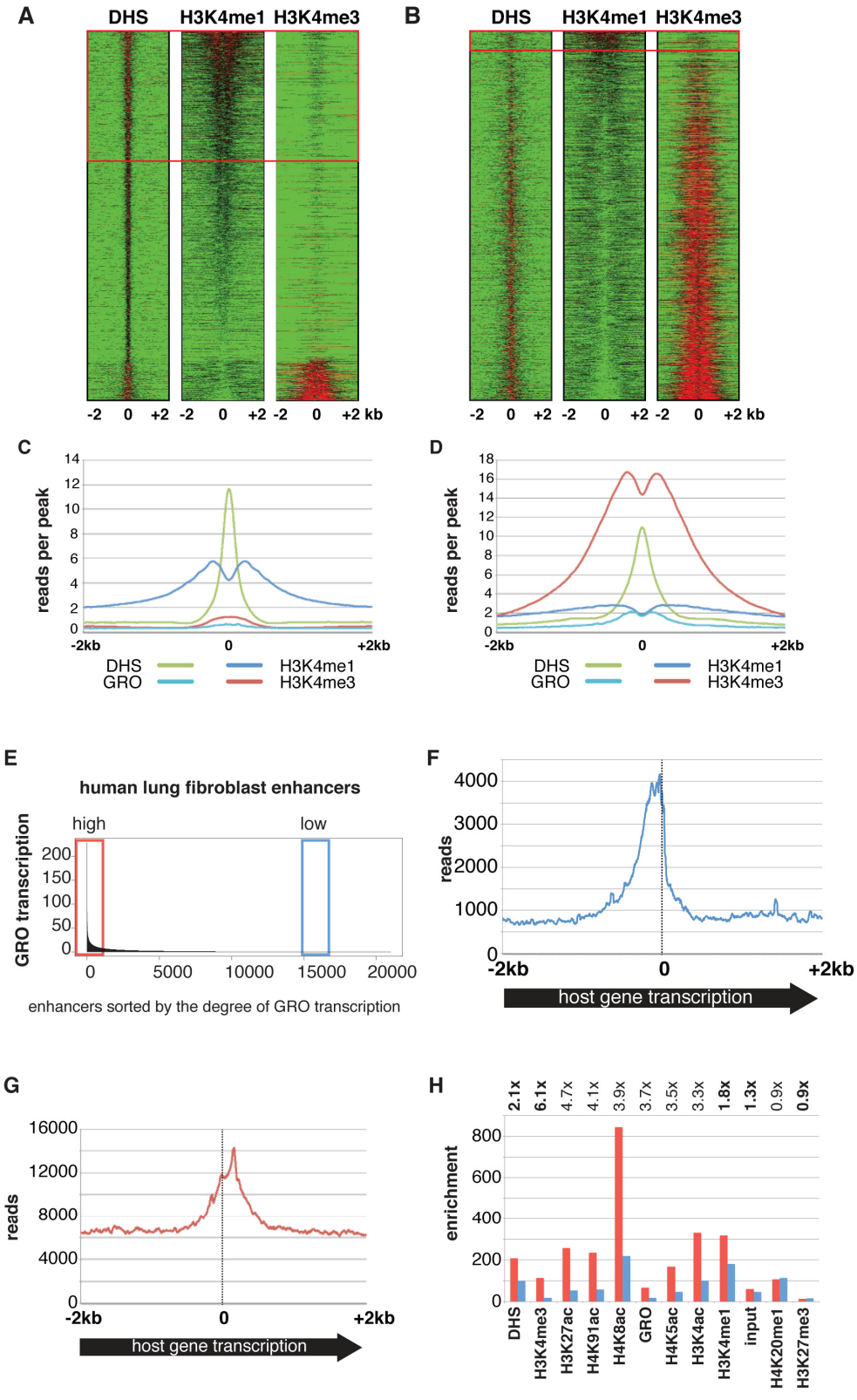

A
DHS
H3K4me1
H3K4me3
-2 0 +2 -2 0 +2 -2 0 +2 kb
B
DHS
H3K4me1
H3K4me3
-2 0 +2 -2 0 +2 -2 0 +2 kb
C
reads per peak
-2kb
0
+2kb
DHS
GRO
H3K4me1
H3K4me3
D
reads per peak
-2kb
0
+2kb
DHS
GRO
H3K4me1
H3K4me3
E
human lung fibroblast enhancers
high
low
GRO transcription
enhancers sorted by the degree of GRO transcription
F
reads
-2kb
0
+2kb
host gene transcription
G
reads
-2kb
0
+2kb
host gene transcription
H
enrichment
2.1x
6.1x
4.7x
4.1x
3.9x
3.7x
3.5x
3.3x
1.8x
1.3x
0.9x
0.9x
DHS
H3K4me3
H3K27ac
H4K91ac
H4K8ac
GRO
H4K5ac
H3K4ac
H3K4me1
input
H4K20me1
H3K27me3

Figure S5

Figure S5

Genome-wide identification of enhancers and associated transcription in fetal human primary lung fibroblasts, related to Figure 4

(A) All human lung fibroblast DHS sites were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the *DnaseI* hypersensitivity, H3K4me1 and H3K4me3 data in the same sort order clearly shows the segregation of the DHS sites into H3K4me1 and H3K4me3 enriched populations. The red rectangle indicated the cut-off used for identification of all enhancers. Each panel shows the distribution of signal in a 4-kb window centered in the middle of each DHS.

(B) The transcription start sites (from UCSC Genes and "Refseq" gene annotations) which overlap with a single DHS site in human fetal fibroblasts (16,508 sites), were sorted based on the difference in enrichment of H3K4me1 and H3K4me3. Analysis of the chromatin marks, as in **(A)**, shows a similar, albeit smaller (~4%), H3K4me1 enriched population, showing that a proportion of annotated TSSs, at least within lung fibroblast cells, resembles an enhancer chromatin signature (indicated by the red rectangle). Each panel shows the distribution of signal in a 4-kb window centred in the middle of each TSS.

Chromatin profiles normalized for number of peaks for the human lung fibroblasts enhancers population is shown in **(C)** and all annotated human TSSs (UCSC Genes) is shown in **(D)**. Color-coding for each chromatin mark is shown below.

(E) Human lung fibroblasts enhancers were sorted based on the level of antisense poly(A)$^-$ transcription. Red and blue rectangles contain high and low transcribing populations of enhancers respectively.

(F) The cumulative poly(A)$^-$ transcription (using global run-on method (GRO)) associated with intragenic enhancers in the antisense direction (relative to the transcriptional direction of the host gene) is seen to originate close to the midpoint of the enhancers and extends ~1kb in the sense direction in primary human lung fibroblasts.

(G) The cumulative poly(A)$^-$ transcription associated with intragenic enhancers in the sense direction (relative to the transcriptional direction of the host gene) is masked by the gene transcription (see relatively higher background in comparison to **(F)**) and is seen to originate close to the midpoint of the enhancers in primary human lung fibroblasts.

(H) The comparison between high (red) and low (blue) transcribing enhancers is displayed as enrichment of various factors. The high and low transcribing enhancer populations are as indicated in **(D)**. The fold difference for each factor is indicated above the graphs.

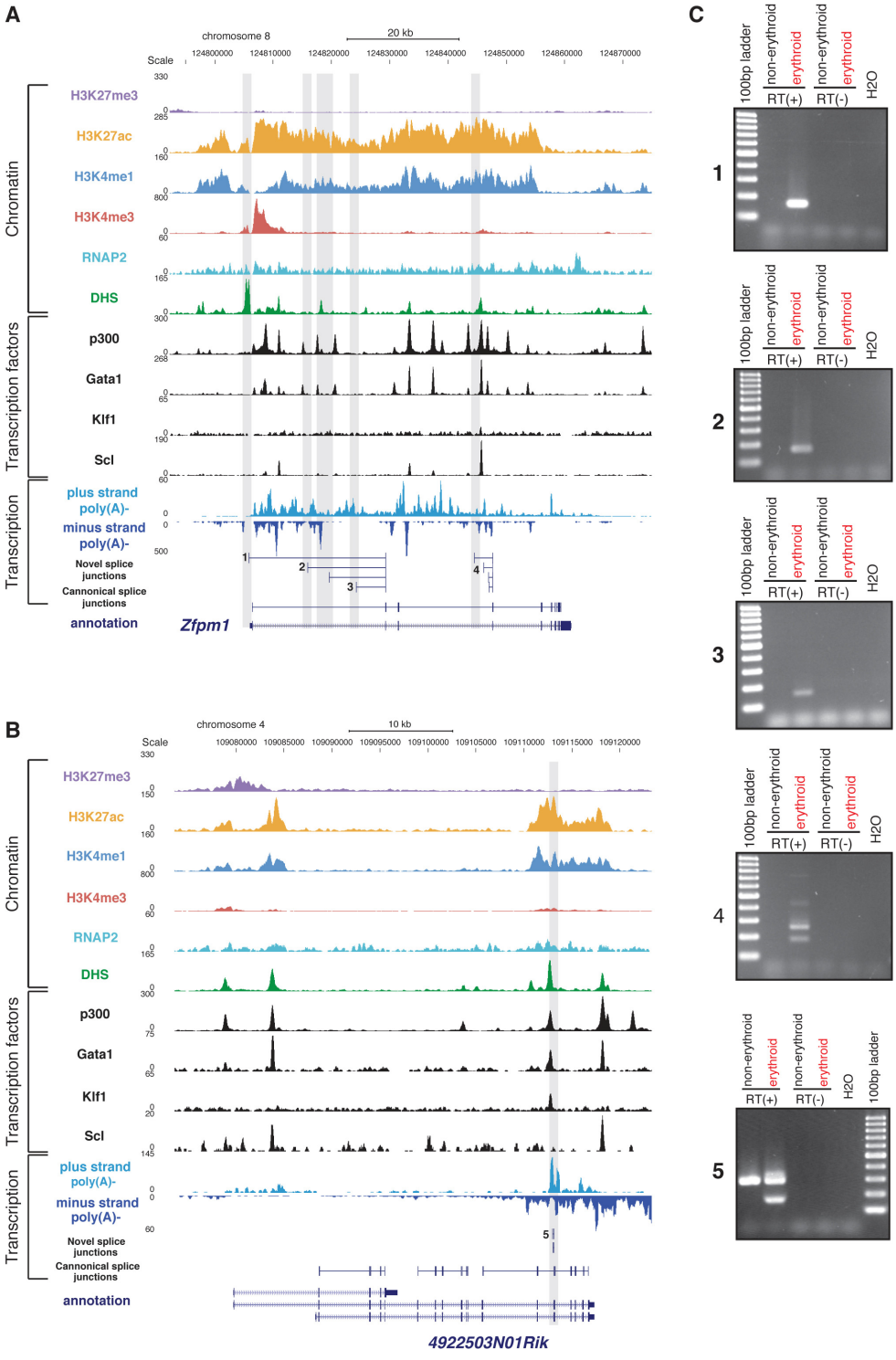

A
chromosome 8
20 kb
Scale
124800000
124810000
124820000
124830000
124840000
124850000
124860000
124870000
Chromatin
H3K27me3
H3K27ac
H3K4me1
H3K4me3
RNAP2
DHS
Transcription factors
p300
Gata1
Klf1
Scl
Transcription
plus strand poly(A)-
minus strand poly(A)-
Novel splice junctions
Cannonical splice junctions
annotation
Zfpm1
B
chromosome 4
10 kb
Scale
109080000
109085000
109090000
109095000
109100000
109105000
109110000
109115000
109120000
Chromatin
H3K27me3
H3K27ac
H3K4me1
H3K4me3
RNAP2
DHS
Transcription factors
p300
Gata1
Klf1
Scl
Transcription
plus strand poly(A)-
minus strand poly(A)-
Novel splice junctions
Cannonical splice junctions
annotation
4922503N01Rik
C
100bp ladder
non-erythroid
erythroid
RT(+)
RT(-)
H2O
1
2
3
4
5

Figure S6

Figure S6

Examples of loci producing meRNAs (long poly(A)$^+$ spliced isoforms from enhancers), related to Figure 5

Loci in **(A,** *Zfpm1* or *Fog1* locus**)** and **(B,** *4922503N01Rik* locus**)** characterised by ChIP-Seq profiles (H3K27me3, H3K27ac, H3K4me1, H3K4me3, RNAP2, p300 (Birney et al., 2007), Gata1 (Cheng et al., 2009), Klf1 (Tallack et al., 2010) and Scl (Kassouf et al., 2010)), *DnaseI* hypersensitivity and RNA-Seq. Each labelled AFE splice junction (1, 2, 3, 4, 5) was verified by junction-specific RT-PCR and is shown in **(C).**

(C) Junction-specific RT-PCRs (1-5) of RNA-Seq reads highlighted in **(A)** and **(B)**. A band representing unspliced nascent RNA is also detected in both erythroid and non-erythroid (**C**-5) due to its relatively small unspliced size. Corresponding reaction products of expected sizes are displayed.

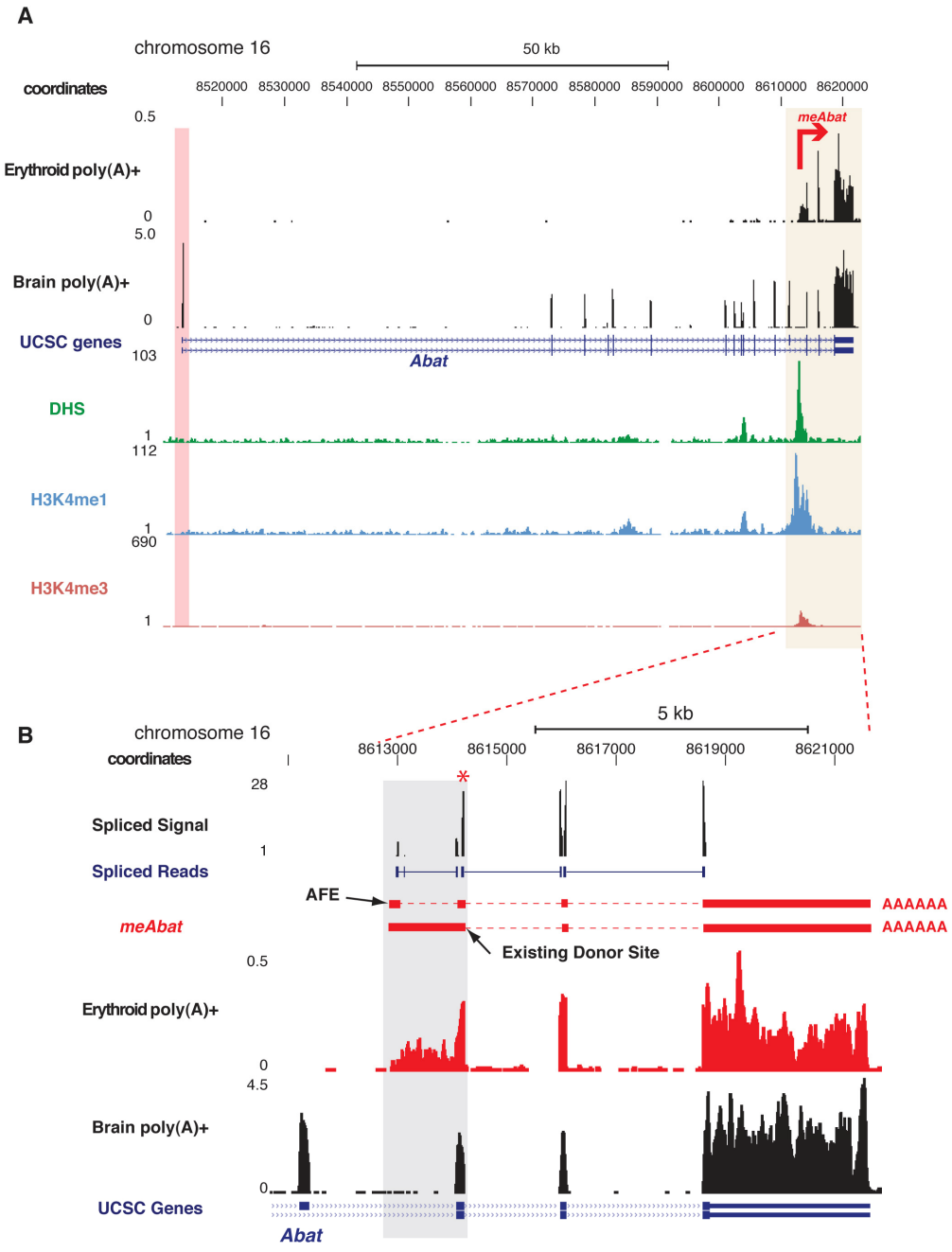

A
chromosome 16
50 kb
coordinates
8520000 8530000 8540000 8550000 8560000 8570000 8580000 8590000 8600000 8610000 8620000
meAbat
0.5
Erythroid poly(A)+
0
5.0
Brain poly(A)+
0
UCSC genes
103
Abat
DHS
1
112
H3K4me1
1
690
H3K4me3
1
B
chromosome 16
5 kb
coordinates
8613000 8615000 8617000 8619000 8621000
28
Spliced Signal
1
Spliced Reads
AFE
meAbat
AAAAAA
AAAAAA
Existing Donor Site
0.5
Erythroid poly(A)+
0
4.5
Brain poly(A)+
0
UCSC Genes
Abat

Figure S7

Figure S7

meAbat is expressed from an inactive gene in erythroid cells, but only a proportion of the transcript has AFE, related to Figure 7

The mouse *Abat* locus with high-resolution maps of normalised poly(A)$^+$ RNA-Seq data for wild type Ter119+ and brain cells. The y-axis represents the normalised expression value, fragments per base pair per million reads aligned. A red column highlights the *Abat* promoter region. A grey column highlights the extent of the enhancer transcript. UCSC Genes annotation is shown in purple. High-resolution maps for the chromatin markers *DnaseI* hypersensitivity (DHS), H3K4me1 and H3K4me3 are shown below. The y-axes represent read density.

(B) Shows a zoomed view of the region of the *Abat* locus containing the enhancer transcript. The signal from reads which are involved in a splicing event (Spliced Signal in black) shows the quantitative usage of the acceptor and donor sites of each expressed exon (the intron/exon usage of the expressed isoforms is shown in the Spliced Reads track in blue). It can be seen that the first and last exons only show splicing signals at the acceptor and donor sites respectively as expected. The spliced signal at the donor site of the downstream exon (marked with a red asterisk) is seen to be greater than that of the same exon's acceptor site signal (see downstream exon in the transcript for comparison).

This shows that the donor site of this exon is used more frequently than the acceptor site and is used at the same frequency as all of the downstream exons. Conversely the acceptor site of this exon (red asterisk) is used at the same frequency as the first exon (if the minor splice variant represented by the thin second bar in Spliced Reads is ignored).

Taken together with the erythroid-specific high levels of poly(A)$^+$ transcription in the first "intron" (highlighted by grey bar normalised expression signal for Erythroid poly(A)$^+$ in red and Brain poly(A)$^+$ in black) demonstrates the existence of two isoforms of the *meAbat* transcript (*meAbat* in red shows the two predicted isoforms). One isoform initiates at the enhancer element and uses a cryptic splice site within the intron and splices to the acceptor site of the downstream canonical exon to produce an alternative first exon (shown as black arrow and labeled "AFE"). A second isoform initiates at the enhancer element but uses the existing donor site of a canonical exon to splice to the next downstream canonical exon (shown as black arrow and labeled "Existing Donor site") to produce a large first exon which includes the whole canonical exon.

The second isoform would not be evident in the presence of transcription from the canonical promoter whereas the AFE containing isoform would be detectable. Estimation from the usage of the splice sites specific to the AFE isoform compared to the common splice junctions show this isoform to represent half the occurrence of the isoform lacking the AFE. UCSC Genes annotation is shown in purple.

Supplemental Experimental Procedures

Control cells and cell lines

Epstein Barr Virus (EBV) transformed B lymphocytes were derived from normal individuals. Primary mouse embryonic stem cells (E14-TG2aIV 129/Ola). Mouse fibroblasts (L929 cells) were a gift from Dr Karl Morten (The Nuffield Department of Obstetrics and Gynaecology, University of Oxford).

Characterisation And Deletion Of Annotated *Nprl3* Transcription Start Sites

Promoter element P6

The first targeted region contains the previously described promoter region of the *Nprl3* gene, which is associated with a CpG island (Flint et al., 1997; Vyas et al., 1992). The evolutionary conservation study of that region showed that the promoter is conserved in at least 16 species (Hughes et al., 2005). A number of ESTs map to the P6 segment, all originate within the CpG island of the *Nprl3* gene and have conserved counterparts in other species (e.g. human). The putative protein encoded by these mRNAs has been characterised elsewhere (Kowalczyk in preparation).

Enhancer R3

It was shown that the multi-species conserved element R3, which corresponds to an erythroid-specific *DnaseI* hypersensitive site (HS-21 in mouse and HS-33 in human) and includes general and tissue-specific transcription factor binding sites, acts as an alternative transcription start site for the *Nprl3* gene in erythroid cells in mouse. One mRNA transcript is currently described (AK036633) and annotated as a gene isoform by Ensembl (ENSMUSG00000020289). This isoform has its putative start site within the R3 segment and derives from a *Mus musculus* adult male bone cDNA (RIKEN full-length enriched library, clone: 9830144A18) and maps to chr11: 32132670-32156413 (mm9).

Deletional constructs

The targeting constructs (pP6, pR3) were assembled in pNTflox vector. Homology arms were cloned onto each side of a *loxP* flanked PGK-*neo* cassette. The ∆P6 deleted segment spans 2315bp between coordinates chr11:32,166,133-32,168,448; the ∆P6-R3 deleted segment spans 12403bp between coordinates chr11:32,156,045-32,168,448. All coordinates were obtained with the mouse Build 37 (NCBI37/mm9) as a reference.

ES Cell Gene Targeting

The linearized constructs (pP6, pR3) were electroporated into E14TG2a mouse embryonic stem cells. Correctly targeted clones were identified by Southern blot analysis. For *ES∆P6*$^{+/flox}$, DNA was digested with: *HindIII* and hybridized with NEO (G418 cassette with primers NEO-F: ATGGGATCGGCCATTGAACAAG NEO-R: CAGAAGAACTCGTCAAGAAG) and small BAL probes (mm9, chr11:32158914-32159468), *KpnI* and hybridized with PV probe (mm9, chr11:32162468-32163063). For *ES∆R3*$^{+/flox}$, DNA was digested with: *KpnI* and hybridized with HVR probe (mm9 chr11:32203961-32204848), *BstEII* and hybridized with LEFT2 probe (mm9 chr11: 32151189-32151823).

Additionally, the P6-targeted cell line (*ES∆-P6*$^{+/flox}$) underwent *in vitro Cre*-recombination using the previously described *Cre*-expressing plasmid (Araki et al., 1995) to remove the *neo* cassette and obtain the *ES∆P6*$^{+/-}$ cell line. *ES∆P6*$^{+/-}$ was then used in a second round of electroporation to target the R3 region according to the procedure above. Targeted clones lacking P6 segment and targeted for R3 on the same chromosome (*ES∆P6*$^{+/-}$*∆R3*$^{+/flox}$) were identified by Southern blot analysis. DNA was initially digested with: *KpnI* and hybridized with HVR, R3-probe (mm9 chr11: 32156257-32157046) and NEO probes, *BstEII* and hybridized with LEFT2 probe. To discriminate between *cis* and *trans* configuration DNA was digested with *SfiI* and *SdaI* and hybridized with HVR probe.

Mouse Models

All animal work was carried out according to UK Home Office regulations, under appropriate project licenses. Chimeric mice were generated by ES cells injections into C57BL/6 blastocysts. Male agouti chimeras were crossed with C57BL/6 mice. Agouti F1 were genotyped by Southern blotting, PCR and sequencing. For *∆P6*$^{+/flox}$, DNA was digested with *KpnI* and hybridized with PV probe. For *∆R3*$^{+/flox}$, DNA was digested with *ScaI* and hybridized with HVR probe. For *∆P6*$^{+/-}$*∆R3*$^{+/flox}$, DNA was digested with *BstEII* and hybridized with

LEFT2 probe. The germline transmissions of all targeted alleles were obtained ($\Delta P6^{+/flox}$ and $\Delta P6^{+/-}\Delta R3^{+/flox}$). The *neo* cassette was subsequently removed by *in vivo Cre*-recombination (Mao et al., 1999) which resulted in two mouse models: $\Delta P6^{+/-}$ and $\Delta P6\text{-}R3^{+/-}$. For schematic diagrams and final mapping see also Figure S1A and S2A.

Conventional Sequencing

Sequencing of double stranded DNA was performed using Big Dye chemistry. DNA electrophoresis was performed using ABI-3730 DNA Analyser (Applied Biosystems). All sequences were analysed using Sequencher 4.6 (Gene Codes Corp.) and Macvector (Symantec) software.

Tilling array design and RNA hybridization

Total RNA was extracted with Tri Reagent (Sigma) and *DnaseI* treated with TURBO DNA-*free* (Ambion). Total RNA was converted into double stranded complementary DNA (ds cDNA) using SuperScript Double-Stranded cDNA Synthesis Kit (Invitrogen). Test ds cDNA and input DNA (sonicated wild type mouse DNA) were labeled with Cy5-dCTP and with Cy3-dCTP (GE Healthcare) respectively using the Bioprime DNA Labeling System (Invitrogen). ds cDNA was analyzed as previously described with minor modifications (De Gobbi et al., 2006; Wallace et al., 2007) on custom tiled Agilent arrays (platform 4x44K) covering a ~1Mb region of mouse α-globin cluster (chr11:31,644,970-32,644,588; mm9).

RT-PCR and qPCR

For all cell types, total RNA was extracted with Tri Reagent (Sigma) and *DnaseI* treated with TURBO DNA-*free* (Ambion).

For reverse transcription reactions, 1 μg of total RNA was converted into cDNA using Superscript II (Invitrogen). Templates with (+RT) and without (−RT) reverse transcriptase were prepared to detect genomic contamination.

For quantitative real-time expression analysis, qPCR reactions were performed using the TaqMan universal PCR mastermix (Applied Biosystems) and TaqMan gene expression assays (Applied Biosystems) or custom primers. Expression in all cell types was calculated relative to a control sequence in the *Gapdh* gene or 18S ribosomal RNA gene (Eurogentec RT-CKFT-18S). All reactions were performed in duplicate on each template using Sequence Detection System 7000 thermocycler (Applied Biosystems). Details of primers and probes: human *NPRL3* (Applied Biosystems;

Hs00429220_m1), mouse *Nprl3* (Eurogentec; exon 7-8 junction; F-GCTATTGAACGGAGCCTGAAA, R-AGCAGAGACTTCTCGTCACTGAGA, probe – CCATCCGCCCGTACCATGCC), *18S* (Eurogentec; 18S Genomic Control Kit FAM-TAMRA RT-CKFT-18S), mouse *Gapdh* (Eurogentec; F-CCTGGCCAAGGTCATCCATGACAACTTT, R-CTTCACCACCATGGAGAAGGC, probe – AGGCCGAGAATGGGAAGCTTGTCATC).

RT-PCR validation of enhancers associated poly(A)$^+$ RNAs

For validating poly(A)$^+$ transcripts, we chose 13 transcripts. The RNA was isolated using Trizol (Sigma). Each RNA sample was *DnaseI*-treated (TURBO DNA-free, Ambion) before reverse transcription using the High Capacity cDNA synthesis kit (Applied Biosystems) with random priming. RT-PCR reactions were performed using Advantage 2 Polymerase Mix (Clontech) with GC-melt for amplification of GC-rich templates on a Biometra TRIO-Thermoblock. Optimum annealing temperatures were determined for each set of primers in the range of 56-60°C, and extension time at 68°C was in the range of 30-180 seconds depending on the size of the product.

Northern blot probes

Probes were PCR amplified from mouse erythroid cDNA, cloned into pGEM-T Easy vector (Promega) and finally radioactively labeled using asymmetric PCR (single stranded DNA probes). Probes were generated using the following primers sets: *Znfx1* (Znfx1-4F GACTGCAGCCACATCTTTGA and Znfx1-4R CAATCTGCACTCGTTCCTCA) and *Tg* (Tg2F AACTTCCATCCAGACGGTTG and Tg2R GTTGAAAACTGGCCCTGGTA).

RNA-FISH

RNA-FISH was performed as described previously (Brown et al., 2006). Probes used for detection of mouse *Nprl3* gene by RNA-FISH were pools of plasmids covering the 3' end of the mouse *Nprl3* gene (mm9, chr11:32136151-32148096). Probes were labelled by nick translation with biotin-16-dUTP and detected with one layer of Avidin Cy3.5 (GE Healthcare). Slides were scored and imaged with a BioRad Radiance 2000 confocal system mounted on a BX51 Olympus microscope using Lasersharp software.

Recombinant Nprl3 protein expression

Nprl3 cDNA fragments were subcloned into a pGEX6p1 vector and the constructs were transduced into Rosetta 2 cells (Merck 4 Biosciences, 71402-

4). 1 liter of Rosetta 2 cultures grown at 30°C were induced for 3 hours with 0.4mM IPTG. Cell pellets were resuspended in a 20mM Tris-Hcl pH 7.5, 1M KCl, 20% Glycerol, 5mM EDTA solution; and sonicated on a Branson sonicator at 74% amplitude for 1 min. in total, in 20 second bursts with a 59.9 sec pause. Supernatants were incubated with a GST bead slurry (GE catalog #17-0756-01) at 4°C for 4 hours, washed multiple times with the resuspension buffer followed by washed with a series of decreasing KCl concentrations (ie. 0.5M, 0.3M, 0.1M) and eluted with 300μl 10mM Glutathione/50mM Tris.HCl (pH 8.0). Purified samples were quantified using a Coomassie Plus protein assay kit (Fisher PN23236) and used for western blotting experiments.

Protein blots

Protein extracts were prepared by RIPA extraction. Total denatured protein extracts were separated by SDS-PAGE on a Bis-Tris gradient gel (Invitrogen) and transferred to Immobilon-P polyvinylidene difluoride membrane (Millipore). For Nprl3 protein detection, the membrane was incubated overnight at 4°C with primary antibody (1:500) (C16) (Lunardi et al., 2009) and then for 1 hour with the HRP-conjugated secondary anti–rabbit IgG antibody (1:5000) (BD). Gapdh (Cell Signalling Technology) was used as a loading control. The HRP signal was detected with ECL detection reagent (GE Healthcare).
The recombinant protein used to immunize the rabbits was the "short" isoform of human C16orf35/NPRL3 (genbank NP_001034565) fused to MBP. It corresponds to the c-terminal 390 aminoacids of the “longest” isoform. (Lunardi et al., 2009).

High-Throughput Whole Genome Methods

RNA-Sequencing (RNA-Seq)

For RNA-Seq library preparation, the total RNA quality was assessed using Agilent Bioanalyser. RNA with overall RIN score >9 was used. Poly(A)$^-$ fraction was separated from total RNA using PolyATract mRNA isolation system (Promega) retaining the poly(A)$^-$ fraction. Poly(A)$^+$ mRNA was depleted of globin transcripts using GlobinClear (Ambion). Poly(A)$^-$ fraction was depleted of ribosomal transcripts by using RiboMinus Eukaryote Kit for RNA-sequencing (Invitrogen) followed by RNA purification on RiboMinus Concentration Module (Invitrogen). The quality of obtained RNA samples was assessed using PicoChip (Agilent). The poly(A)$^+$ libraries were prepared from *Nprl3* extended knock-out (*ΔP6-R3*$^{-/-}$) cultured fetal liver cells and wild-type Ter119+ cells using the mRNA-Seq pair-end kit (Illumina).

For the poly(A)$^-$ library, poly(A)$^-$ RNA from wild type Ter119+ cells was heat fragmented and the accuracy of the fragmentation reaction was assessed on Agilent Bioanalyser. The RNA was purified using RiboMinus Concentration Module (Invitrogen). Finally, the poly(A)$^-$ library was prepared according to DGE Small RNA Sample Prep kit with minor modifications (Illumina). Briefly, the RNA fragments underwent end repair with TAP and PNK and cleaned up using RiboMinus Concentration Module (Invitrogen). The 5' adapter was ligated using T4 RNA ligase for 6 hours at 20°C and the excess of the adaptor was removed on NucAway column (Ambion). Similarly, the 3' adaptor was ligated and the RNA cleaned on NucAway column (Ambion). Next, the RNA underwent RT reaction using SuperScript II (Invitrogen) and cDNA was amplified with Phusion DNA Polymerase (10-15 cycles). The amplified cDNA underwent size selection at 100-350bp. All RNA libraries were sequenced using massively parallel sequencing (Illumina, GAII) with 50 base single or pair-end reads for poly(A)$^-$ and poly(A)$^+$ respectively.

Chromatin Immunoprecipitation (ChIP) And Chip-Sequencing (ChIP-Seq)

For ChIP-Seq experiments, Ter119+ cells were fixed with 1% formaldehyde for 10 minutes at RT and chromatin was sonicated to a size <500 bp. Immunoprecipitations were performed, after an overnight incubation with the appropriate antibody, with protein A agarose (Millipore). A sample containing no antibody was used as a negative control and both immunoprecipitated DNA and input control were purified by phenol and chloroform extraction followed by ethanol precipitation.

Subsequently the material was analysed by real time PCR (ABI Prism 7000 Sequence Detection System, Applied Biosystems) using a series of PCR amplicons and 5'FAM-3'TAMRA probes across the α-globin locus. ChIP-seq libraries were prepared and sequenced using the standard Illumina protocol, with the modification that samples where amplified prior to size selection (150-200 bp).

Genome Alignment

Single-end reads for ChIP-Seq, the *DnaseI*-Seq, the poly(A)$^-$ and global Run-on (GRO) samples were aligned to the appropriate genome build (mm9 for mouse data and hg18 for human) using bowtie (version 0.12.3, http://bowtie-bio.sourceforge.net/index.shtml) (Langmead et al., 2009). To prevent the exclusion of the duplicated globin genes bowtie was run with the –m reporting option set to 2 to allow reads to map twice to the genome. To exclude over-amplified products from these data sets, reads that map to the exact same genomic position were collapsed into a single representative read.

The paired-end mRNA samples were aligned to mm9 using Tophat (version 1.1.4b) (Trapnell et al., 2009) with the inner mate difference – r set to 200 bp. *de novo* RNA transcripts reconstruction was generated using Cufflinks (Trapnell et al., 2010).
The number of reads aligned in each sequencing assay was 23 million (H3K27me3), 23 million (H3K4me3), 16 million (H3K4me1), 32 million (H3K27ac) and 20 million (DHS). A total of 67 and 37 million pair-end reads were generated from *Nprl3* extended knock-out and wild type respectively. Additionally, a total of 25 million single-end reads were generated from wild type.
Bowtie alignments were converted to genome wide density tracks (BigWig) and the output of TopHat was separated into spliced and unspliced reads and viewed in UCSC Genome Browser (Kent et al., 2002) (BAM file).

Production of Genome-Wide Tracks

Genome-wide tracks of the ChIP-Seq, *DnaseI*-Seq and nascent transcription were produced using the in-house perl tool sam2bigwig.pl, which produces a track of read density over a set window size and increment of movement across the genome. For the more diffuse nascent transcription data (poly(A)$^-$ and GRO) a window of 600 bp and an increment of 60 bps were used. For the ChIP-Seq and *DnaseI*-Seq data a window of 300 bp and an increment of 30 bps was used. The tracks were displayed in the UCSC genome database in bigwig format.
Prior to analysis the stranded poly(A)$^-$ and GRO RNA datasets were split into their forward and reverse strands using the bitwise alignment score from the SAM file and each strand was analysed individually. Genome-wide plots of poly(A)$^+$ transcription were normalized for library size and quality using an adaptation of the method employed by the Cufflinks algorithm (Trapnell et al., 2010). The amount of transcription associated with each base pair of the MM9 genome build was expressed as the number of reads aligned to the base position divided by the total number of millions of reads aligned in the experiment to give the value fragments per base pair per million aligned.

Peak Finding From *DnaseI*-Seq

Peak detection for the *DnaseI*-Seq in erythroid (Ter119+), human fetal lung fibroblasts (IMR90), erythroleukemia (K562) and B cells (GM12878) was performed with the PeakRanger algorithm (Feng et al., 2011), with appropriate input controls. Peaks for the human cell lines (IMR90, K562, GM12878) were overlapped and annotated with the Encode generated file (ftp://hgdownload.cse.ucsc.edu/apache/htdocs/goldenPath/hg18/encodeDCC/wgEncodeMapability/wgEncodeDukeRegionsExcluded.bed6.gz), which

represents regions in the genome which strongly overreact in high-throughput sequencing experiments due to large copy number differences between the real genome and the genome build and normalize poorly. A similar set of regions was generated in-house for the mouse by stringently peak calling our input data and overlapping our peaks with the resultant set of regions. These regions in both human and mouse were excluded from all downstream analysis.

The genomic context of *DnaseI*-Seq peaks was determined by comparison to Refseq and UCSC Genes annotation (UCSC Genome Browser) and split into three categories, TSS associated (within 1 kb of an annotated transcription start site), intragenic (lies within the body of an annotated gene, but not within 1 kb of an annotated transcription start site) and intergenic (does not lie within an annotated gene or within 1 kb of a TSS). Intragenic and TSS peaks were annotated with the transcriptional strand of the associated gene.

Peak Quantification

The Ter119+ *DnaseI* peaks were analysed for their enrichment for each of the ChIP-seq datasets produced here. IMR90 *DnaseI* peaks were analysed for their enrichment in all of the chromatin modifications publically available for this cell line. K562 and GM12878 *DnaseI* peaks were analysed for their enrichment in H3K4me1 and H3K4me3. Using the in-house perl script normwindow.pl, the density of reads within a specified window around the *DnaseI* peak was quantified and the density of reads in the input was subtracted from this value. To analyse the relative amount of H3K4me1 and H3K4me3 enrichment for each *DnaseI* peak normwindow.pl was used to quantify the density of reads, for H3K4me1 and H3K4me3, within a 1kb window around each peak. The density of H3K4me3 was subtracted from the H3K4me1 density and the peak list was sorted from the most relatively H3K4me1 enriched to the most relatively H3K4me3 enriched (Figure 4A).

Defining enhancers

Sorting the *DnaseI* peaks by the difference between H3K4me1 and H3K4me3 shows there to be three discernable populations within the peaks. One population which is predominantly H3K4me3 marked, a population predominantly H3K4me1 marked and a population which is evenly and/or poorly marked.

By current understanding enhancers are represented by the population more enriched by H3K4me1 than H3K4me3 (Birney et al., 2007; Heintzman et al., 2009; Heintzman et al., 2007), hence we arrived at an empirical set of cut offs to capture this population of *DnaseI* peaks in our dataset, based on the *DnaseI* hypersensitivity of the peaks and the difference in read density

between H3K4me1 and H3K4me3 (Figure 4A). The cumulative distribution of high-throughput sequencing data over the peaks (Figure 4B and C) was generated using the in-house perl script Quantpile.pl and displayed in Microsoft Excel. The sorted distribution of high-throughput sequencing data over individual peaks or TSSs (Figure 4A, Figure 7B and C, Figure S5A and B) was generated using the in-house perl script Hotpile.pl and visualized in R using the gplots library.

Quantitation of poly(A)$^-$ transcription from intragenic enhancers

Of the 1794 intragenic enhancers 280 were removed from this analysis due to their proximity to strong sources of antisense transcription. For the remaining 1514 analyzable intrangenic enhancers the amount of poly(A)$^-$ transcription associated with intragenic peaks was determined antisense to the transcriptional direction of the associated gene.

To normalise for transcriptional signals transcribing through the peak, rather than originating from the peak the difference between the number of antisense reads aligned in a 1kb window downstream and a 1kb window upstream of the midpoint of the peak, relative to gene transcription, was quantified.

AFE pipeline

Alternative first exons (AFEs) were isolated using a two-step process. Firstly exons where identified within the spliced poly(A)$^+$ RNA-Seq which spliced to the acceptor site of an annotated exon within RefSeq or UCSC Genes annotation but were not themselves annotated. Secondly, to remove unannotated internal exons, any exons which showed evidence within the poly(A)$^+$ RNA-Seq of providing a donor site for an upstream exon were removed.

Previously Published Genome-Wide Datasets Used in This Study

Previously published datasets are as follows:
Gata1 (Cheng et al., 2009), Scl (Kassouf et al., 2010), and Klf1 (Tallack et al., 2010) occupancy in mouse erythroid cells; Ldb1 in MEL cells (Soler et al., 2010); p300 in MEL cells (ENCODE Consortium (Birney et al., 2007); Micheal Snyder, Stanford and Weissman, Yale); RNA-Seq from mouse brain (Mortazavi et al., 2008).

IMR90 cell line chromatin modifications were from the UCSD Human Reference Epigenome Mapping Project (GSE16256). IMR90 cell line chromatin accessibility data were from the University of Washington Human Reference Epigenome Mapping Project (SRA010036). Nascent transcription in the IMR90 cell line generated using the global run on method (GRO) from (Core et al., 2008) (SRX003135 and SRX003136).
K562 and GM12878 cell line chromatin modifications and chromatin accessibility data are from the ENCODE Consortium (Birney et al., 2007) via the UCSC table browser, the files used in the analysis are detailed in the table below. The ChIP data was produced by the Bernstein laboratory in the Broad Institute of MIT and Harvard and the chromatin accessibility data was produced by the Crawford laboratory in Duke University.

Cell Line	**Data**	**File Name**
K562	H3K4me1 (Bernstein, Broad)	wgEncodeBroadHistoneK562H3k4me1StdRawDataRep2.fastq wgEncodeBroadHistoneK562H3k4me1StdRawDataRep1.fastq
K562	H3K4me3 (Bernstein, Broad)	wgEncodeBroadHistoneK562H3k4me3StdRawDataRep1.fastq, wgEncodeBroadHistoneK562H3k4me3StdRawDataRep2.fastq
K562	Chromatin accessibility (Crawford Duke)	wgEncodeOpenChromDnaseK562RawDataRep1.fastq wgEncodeOpenChromDnaseK562RawDataRep2.fastq
GM12878	H3K4me1 (Bernstein, Broad)	wgEncodeBroadHistoneGm12878H3k4me1StdRawDataRep1.fastq wgEncodeBroadHistoneGm12878H3k4me1StdRawDataRep2.fastq
GM12878	H3K4me3 (Bernstein, Broad)	wgEncodeBroadHistoneGm12878H3k4me3StdRawDataRep2.fastq wgEncodeBroadHistoneGm12878H3k4me3StdRawDataRep1.fastq
GM12878	Chromatin accessibility (Crawford Duke)	wgEncodeOpenChromDnaseGm12878RawDataRep1.fastq wgEncodeOpenChromDnaseGm12878RawDataRep2.fastq wgEncodeOpenChromDnaseGm12878RawDataRep3.fastq
MEL	p300 (Snyder Standford; Weissman Yale)	wgEncodeSydhTfbsMelP300sc584IggrabRawDataRep1.fastq wgEncodeSydhTfbsMelP300sc584IggrabRawDataRep2.fastq

Glossary

meRNAs

Multi-exonic, spliced and polyadenylated RNA that originates from intragenic enhancer elements (described here). The name refers to the type of RNA (**m**ulti-exonic) and its origin (**e**nhancer)

eRNAs

Short (~1 kb), bi-directional RNA transcripts first described to originate from intergenic enhancers (Kim et al., 2010). The polyadenylation status of these RNAs was unclear (polyadenylated (Kim et al., 2010), not polyadenylated (De Santa et al., 2010)). No elongated poly(A)$^+$ transcripts were detected together with eRNAs originating from intergenic enhancers (De Santa et al., 2010; Kim et al., 2010).
Of note, similar RNAs were seen around active promoters (Core et al., 2008; Preker et al., 2008; Seila et al., 2008).

host gene

Gene containing enhancers (hosting enhancers). The gene hosting the enhancers may lie 10s-1000s kb away from the genes regulated by the elements and often are unrelated to the target gene.

Supplemental References

Araki, K., Araki, M., Miyazaki, J., and Vassalli, P. (1995). Site-specific recombination of a transgene in fertilized eggs by transient expression of Cre recombinase. Proc Natl Acad Sci U S A *92*, 160-164.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E.*, et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Brown, J.M., Leach, J., Reittie, J.E., Atzberger, A., Lee-Prudhoe, J., Wood, W.G., Higgs, D.R., Iborra, F.J., and Buckle, V.J. (2006). Coregulated human globin genes are frequently in spatial proximity when active. J Cell Biol *172*, 177-187.

Cheng, Y., Wu, W., Kumar, S.A., Yu, D., Deng, W., Tripic, T., King, D.C., Chen, K.B., Zhang, Y., Drautz, D.*, et al.* (2009). Erythroid GATA1 function revealed by genome-wide analysis of transcription factor occupancy, histone modifications, and mRNA expression. Genome Res *19*, 2172-2184.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845-1848.

De Gobbi, M., Viprakasit, V., Hughes, J.R., Fisher, C., Buckle, V.J., Ayyub, H., Gibbons, R.J., Vernimmen, D., Yoshinaga, Y., de Jong, P.*, et al.* (2006). A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. Science *312*, 1215-1217.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol *8*, e1000384.

Feng, X., Grossman, R., and Stein, L. (2011). PeakRanger: a cloud-enabled peak caller for ChIP-seq data. BMC Bioinformatics *12*, 139.

Fibach, E., Manor, D., Oppenheim, A., and Rachmilewitz, E.A. (1989). Proliferation and maturation of human erythroid progenitors in liquid culture. Blood *73*, 100-103.

Flint, J., Thomas, K., Micklem, G., Raynham, H., Clark, K., Doggett, N.A., King, A., and Higgs, D.R. (1997). The relationship between chromosome structure and function at a human telomeric region. Nat Genet *15*, 252-257.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A*., et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet *39*, 311-318.
Hughes, J.R., Cheng, J.F., Ventress, N., Prabhakar, S., Clark, K., Anguita, E., De Gobbi, M., de Jong, P., Rubin, E., and Higgs, D.R. (2005). Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences. Proc Natl Acad Sci U S A *102*, 9830-9835.
Kassouf, M.T., Hughes, J.R., Taylor, S., McGowan, S.J., Soneji, S., Green, A.L., Vyas, P., and Porcher, C. (2010). Genome-wide identification of TAL1's functional targets: insights into its mechanisms of action in primary erythroid cells. Genome Res *20*, 1064-1083.
Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.
Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S*., et al.* (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182-187.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol *10*, R25.
Lunardi, A., Chiacchiera, F., D'Este, E., Carotti, M., Dal Ferro, M., Di Minin, G., Del Sal, G., and Collavin, L. (2009). The evolutionary conserved gene C16orf35 encodes a nucleo-cytoplasmic protein that interacts with p73. Biochem Biophys Res Commun *388*, 428-433.
Mao, X., Fujiwara, Y., and Orkin, S.H. (1999). Improved reporter strain for monitoring Cre recombinase-mediated DNA excisions in mice. Proc Natl Acad Sci U S A *96*, 5037-5042.
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.
Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851-1854.
Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849-1851.
Soler, E., Andrieu-Soler, C., de Boer, E., Bryne, J.C., Thongjuea, S., Stadhouders, R., Palstra, R.J., Stevens, M., Kockx, C., van Ijcken, W*., et al.* (2010). The genome-wide dynamics of the binding of Ldb1 complexes during erythroid differentiation. Genes Dev *24*, 277-289.
Tallack, M.R., Whitington, T., Yuen, W.S., Wainwright, E.N., Keys, J.R., Gardiner, B.B., Nourbakhsh, E., Cloonan, N., Grimmond, S.M., Bailey,

T.L.*, et al.* (2010). A global role for KLF1 in erythropoiesis revealed by ChIP-seq in primary erythroid cells. Genome Res *20*, 1052-1063.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Vyas, P., Vickers, M.A., Simmons, D.L., Ayyub, H., Craddock, C.F., and Higgs, D.R. (1992). Cis-acting sequences regulating expression of the human alpha-globin cluster lie within constitutively open chromatin. Cell *69*, 781-793.

Wallace, H.A., Marques-Kranc, F., Richardson, M., Luna-Crespo, F., Sharpe, J.A., Hughes, J., Wood, W.G., Higgs, D.R., and Smith, A.J. (2007). Manipulating the mouse genome to engineer precise functional syntenic replacements with human sequence. Cell *128*, 197-209.