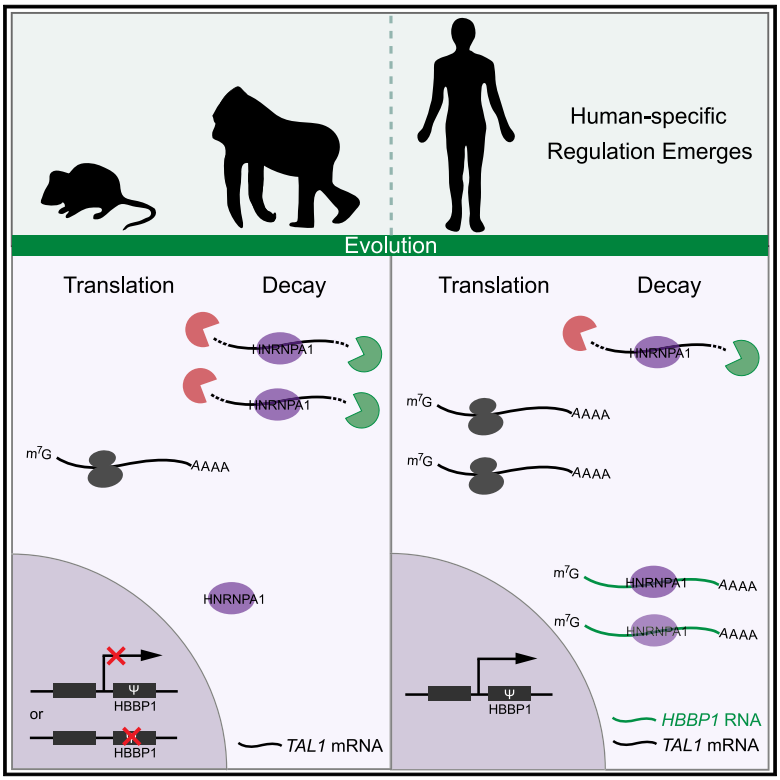
Article

Developmental Cell

Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia

Graphical Abstract



Highlights

- Pseudogenic *HBBP1* confers human-specific essentiality in erythropoiesis
- HBBP1 competitively binds the RNA-binding protein HNRNPA1 to stabilize TAL1 mRNA
- HBBP1/TAL1 interaction contributes to milder symptoms of β-thalassemia patients
- Compared with retropseudogene, duplicated pseudogene often binds RBP but not miRNA

Authors

Yanni Ma, Siqi Liu, Jie Gao, ..., Lihong Shi, Yong E. Zhang, Jia Yu

Correspondence

yanni_ma@126.com (Y.M.), shilihongxys@ihcams.ac.cn (L.S.), zhangyong@ioz.ac.cn (Y.E.Z.), j-yu@ibms.pumc.edu.cn (J.Y.)

In Brief

Ma et al. generate an expressional landscape of pseudogenes and unmask the human-specific significance of *HBBP1* in normal and pathogenic erythropoiesis, which is mediated by competitive HNRNPA1 binding between *HBBP1* and *TAL1* RNAs. This study has far-reaching implications for pseudogene functions, their underlying mechanisms, and species-specific regulation.

Ma et al., 2021, Developmental Cell *56*, 1–16
February 22, 2021
https://doi.org/10.1016/j.devcel.2020.12.019

CellPress

Developmental Cell

CellPress

Article

Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia

Yanni Ma,[1,2,21,*] Siqi Liu,[1,2,21] Jie Gao,[3,21] Chunyan Chen,[4,5,21,22] Xin Zhang,[6,21] Hao Yuan,[4,5] Zhongyang Chen,[1,2] Xiaolin Yin,[7] Chenguang Sun,[1,2] Yanan Mao,[4,5] Fanqi Zhou,[1,2] Yi Shao,[4,5] Qian Liu,[8] Jiayue Xu,[1,2] Li Cheng,[1] Daqi Yu,[4,5] Pingping Li,[7] Ping Yi,[9] Jiahuan He,[1,2] Guangfeng Geng,[3] Qing Guo,[3] Yanmin Si,[1,2] Hualu Zhao,[1,2] Haipeng Li,[10,19] Graham L. Banes,[10,11] He Liu,[12] Yukio Nakamura,[13] Ryo Kurita,[14] Yue Huang,[1] Xiaoshuang Wang,[1,2] Fang Wang,[1,2] Gang Fang,[15,16,17] James Douglas Engel,[18] Lihong Shi,[3,*] Yong E. Zhang,[4,5,19,20,*] and Jia Yu[1,2,3,23,*]

[1]State Key Laboratory of Medical Molecular Biology, Institute of Basic Medical Science, Chinese Academy of Medical Sciences (CAMS) & School of Basic Medicine, Peking Union Medical College (PUMC), Beijing 100005, China

[2]Key Laboratory of RNA and Hematopoietic Regulation, Chinese Academy of Medical Sciences, Beijing 100005, China

[3]State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Institute of Hematology & Blood Diseases Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Tianjin 300020, China

[4]Key Laboratory of Zoological Systematics and Evolution & State Key Laboratory of Integrated Management of Pest Insects and Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China

[5]University of Chinese Academy of Sciences, Beijing 100049, China

[6]Laboratory of Molecular Cardiology & Medical Molecular Imaging, First Affiliated Hospital of Shantou University Medical College, Shantou 515041, China

[7]923rd Hospital of the Joint Logistics Support Force of the Chinese People's Liberation Army, Guangxi 530021, China

[8]Shantou University Medical College, Shantou 515041, China

[9]Department of Obstetrics and Gynecology, the Third Affiliated Hospital of Chongqing Medical University (General Hospital), Chongqing 401120, China

[10]Chinese Academy of Sciences Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, Shanghai 200031, China

[11]Wisconsin National Primate Research Center, University of Wisconsin Madison, 1220 Capitol Court, Madison, WI 53715, USA

[12]Beijing Key Laboratory of Captive Wildlife Technology, Beijing Zoo, Beijing 100044, China

[13]Cell Engineering Division, RIKEN BioResource Research Center, Ibaraki 305-0074, Japan

[14]Department of Research and Development, Central Blood Institute, Japanese Red Cross Society, Tokyo 105-8521, Japan

[15]NYU Shanghai, 1555 Century Avenue, Shanghai 20012, China

[16]Department of Biology, 1009 Silver Center, New York University, New York, NY 10003, USA

[17]School of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China

[18]Department of Cell and Developmental Biology, University of Michigan Medical School, Ann Arbor, MI 48109, USA

[19]CAS Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming 650223, China

[20]Chinese Institute for Brain Research, Beijing 102206, China

[21]These authors contributed equally

[22]Present address: School of Ecology and Environment, Northwestern Polytechnical University, Xi'an 710072, China

[23]Lead contact

*Correspondence: yanni_ma@126.com (Y.M.), shilihongxys@ihcams.ac.cn (L.S.), zhangyong@ioz.ac.cn (Y.E.Z.), j-yu@ibms.pumc.edu.cn (J.Y.)

https://doi.org/10.1016/j.devcel.2020.12.019

SUMMARY

The human genome harbors 14,000 duplicated or retroposed pseudogenes. Given their functionality as regulatory RNAs and low conservation, we hypothesized that pseudogenes could shape human-specific phenotypes. To test this, we performed co-expression analyses and found that pseudogene exhibited tissue-specific expression, especially in the bone marrow. By incorporating genetic data, we identified a bone-marrow-specific duplicated pseudogene, *HBBP1* (η-globin), which has been implicated in β-thalassemia. Extensive functional assays demonstrated that *HBBP1* is essential for erythropoiesis by binding the RNA-binding protein (RBP), HNRNPA1, to upregulate *TAL1*, a key regulator of erythropoiesis. The HBBP1/TAL1 interaction contributes to a milder symptom in β-thalassemia patients. Comparative studies further indicated that the HBBP1/TAL1 interaction is human-specific. Genome-wide analyses showed that duplicated pseudogenes are often bound by RBPs and less commonly bound by microRNAs compared with retropseudogenes. Taken together, we not only demonstrate that pseudogenes can drive human evolution but also provide insights on their functional landscapes.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

INTRODUCTION

It has long been believed that gene duplication drives evolution by generating new genes (Chen et al., 2013; Ohno, 1970; Zhang, 2003). However, most duplicates are deleted or become pseudogenes with disabling mutations compared with functional paralogs (Innan and Kondrashov, 2010; Lynch and Conery, 2000). Mammalian genomes harbor thousands of pseudogenes and the majority (>95%) are generated by RNA- or DNA-level duplications (Kovalenko and Patrushev, 2018; Pei et al., 2012; Zhang and Zheng, 2008), which are retropseudogenes (processed) and duplicated (unprocessed) pseudogenes, respectively. Since their initial discovery, pseudogenes have been generally considered non-functional (Graur and Li, 2000; Jacq et al., 1977; Podlaha and Zhang, 2010). However, technical progress has started to reveal the functionality of pseudogenes in humans. Besides pseudogenes that encode a truncated protein and represent annotation errors (Cheetham et al., 2020; Li et al., 2013), they can be transcribed as long noncoding RNAs (lncRNAs) regulating paralogs. Among various functional mechanisms of pseudogenes, antisense transcription-related regulation, competition with miRNA and RNA-binding proteins (RBPs), appear to be common (Kovalenko and Patrushev, 2018; Poliseno, 2012). However, most of these studies were conducted in pathogenesis, especially tumorigenesis (Karreth et al., 2015; Poliseno et al., 2010). Studies of pseudogenes' functions under normal conditions are scarce, with the exception of a few cases, such as *XIST* in X chromosome inactivation (Duret et al., 2006) and *HMGA1-p* in insulin resistance (Chiefari et al., 2010). Duplicated coding genes could shape human-specific biology (Dennis and Eichler, 2016; Kaessmann, 2010; Zhang and Long, 2014); therefore, it can be speculated that functional pseudogenes play similar roles, especially considering their low levels of conservation across species (Glenfield and McLysaght, 2018). However, evolutionary studies are rare, with *BRAFP1* as an outlier. Specifically, human *BRAFP1* and mouse *Braf-rs1* emerged by independent retropositions, both of which upregulate BRAF by binding miRNAs and functioning as competitive endogenous RNAs (Karreth et al., 2015). Thus, the function of *BRAFP1* is not human-specific.

In this study, we aimed to generate an overview of the function of pseudogenes under normal conditions by implementing an integrative strategy in humans (Figure 1A), thereby prioritizing candidates for in-depth studies of their function, mechanism, clinical relevance, and evolution. We found that one duplicated pseudogene, hemoglobin subunit beta pseudogene 1 (*HBBP1*), is essential for human erythropoiesis by competitively binding RNA-heterogeneous nuclear ribonucleoprotein A1 (HNRNPA1) and upregulating T cell acute lymphocytic leukemia 1 (TAL1), a key regulator of erythropoiesis (Hattangadi et al., 2011; Porcher et al., 1996; Robb et al., 1995). Moreover, the essentiality of *HBBP1* and its interaction with TAL1 is human-specific. More broadly, we found that duplicated pseudogenes like *HBBP1* are more often bound by RBPs compared with retropseudogenes. Together with their poor conservation, these RBP-bound duplicated pseudogenes may also play human-specific functions.

RESULTS

Pseudogenes tend to be tissue-specific especially in bone marrow

To get an overview of pseudogenes' functionality and to prioritize functional candidates, we examined their transcriptional profiles and genetic associations after accounting for sequence similarities (STAR methods; Figure S1). Specifically, based on human proteome atlas (HPA) transcriptome dataset, we performed weighted co-expression network analyses (WGCNA), which clustered genes with similar expression profiles as co-expression modules. We identified eight tissue-specific modules and eight broadly transcribed modules (Figures 1B and S2A; Table S1). We corroborated two known patterns of pseudogene expression (Guo et al., 2014; Kalyana-Sundaram et al., 2012; Pei et al., 2012): (1) pseudogenes were generally unexpressed and contributed only an average of 8% of the genes in each module (Figures 1C and S2B), although they constituted 25.2% gene models (Figure 1A), and (2) pseudogenes were enriched in tissue-specific modules rather than broadly transcribed modules relative to protein-coding genes (Figure 1C). In addition, we found that the proportions of pseudogenes within eight tissue-specific modules were uneven, with 29.5% in the bone-marrow-specific module (M10) (Figures 1D and S2C), which was even higher than the overall proportion (25.2%). The second top module was testis-specific module (M1), in which 11.3% genes were pseudogenes. Testis expression of pseudogenes has been observed previously (Pei et al., 2012), which echoes the emergence of new testis-biased duplicate coding genes driven by sexual selection and the permissive transcription environment (Kaessmann, 2010). In contrast, the brain-specific module (M6) showed one of the lowest proportions (0.9%) of pseudogenes. This depletion could be due to the functional constraint imposed by the adult brain network, which also contains few newly emerged protein-coding genes (Shao et al., 2019).

We speculated on the significance of 109 pseudogenes dominantly transcribed in bone marrow. Since hematopoietic stem cells (HSCs) differentiate into distinct lineages in the bone marrow, we examined the expression of 109 pseudogenes during hematopoietic lineage differentiation and found that more than 50% genes show lineage-specific expression (Figure 1E). The lineage specificity is further corroborated by 50 pseudogenes, which show significant expressional correlation with lineage-specific transcription factors in six cell types (Figure 1F).

To enhance the functional link between 109 pseudogenes and hematopoiesis, we examined genome-wide association study (GWAS) data and identified 13 pseudogenes harboring single nucleotide polymorphisms (SNPs) associated with disease susceptibility (Figures 1G and S2D). Four pseudogenes are associated with diseases related to the function of blood cells. These included O_2 uptake response (*AC084871.1*), systemic lupus erythematosus (*ODCP*, *RGS17P1*), and β-thalassemia (*HBBP1*), which is the only case implicated in hematopathy supported by three independent studies (Figures 1G and S2D). Based on these findings and the significant correlation of *HBBP1* with the erythroid transcriptional factors *TAL1* and *KLF1* ($r > 0.80$, Figure 1F) (Nandakumar et al., 2016), as well as its erythroid-specific expression (Figure S2E), we focused on this pseudogene in subsequent studies.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

CellPress

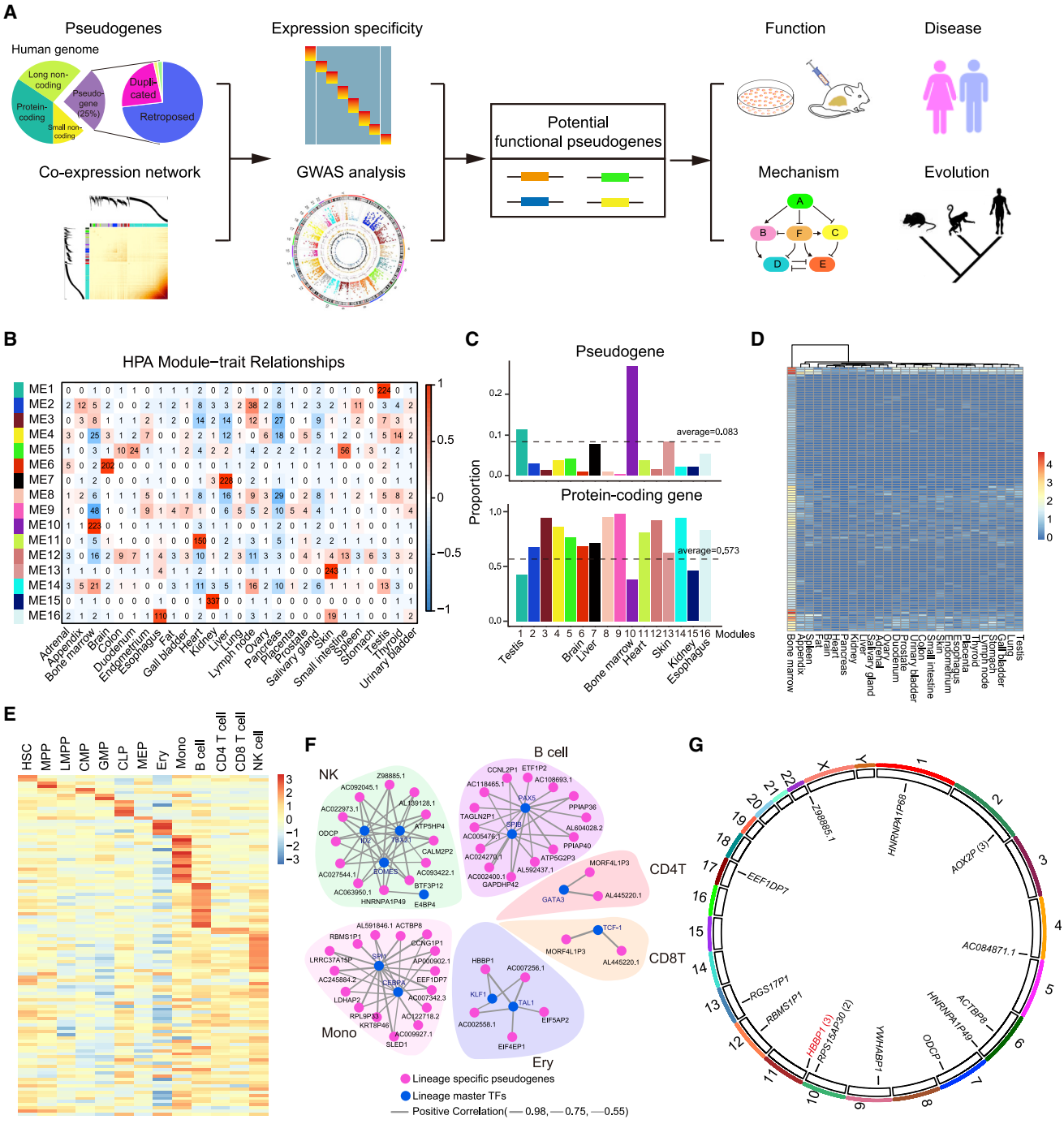


Figure 1. Pseudogenes show tissue-specific expression, especially in bone marrow

(A) The overall design for genome-wide analyses of pseudogenes and characterization of functional candidates.

(B) Heatmap showing the correspondence between tissues and co-expression modules. Color-coding indicates the correlation coefficient (−1 to 1).

(C) The proportion of pseudogenes and protein-coding genes among all genes within each module.

(D) Transcription profiling of pseudogenes in M10 across tissues.

(E) Transcription profiling of pseudogenes in M10 across hematopoietic lineages. HSC, hematopoietic stem cells; MPP, multipotent progenitors; LMPP, lymphoid-primed multipotential progenitors; CMP, common myeloid progenitors; GMP, granulocyte-monocyte progenitor; CLP, common lymphoid progenitors; MEP, megakaryocyte-erythrocyte progenitor; Ery, erythroblasts; Mono, monocytes; NK, natural killer cell.

(F) Pearson correlation between lineage-specific pseudogenes and transcription factors (TFs).

(G) Chromosomal distribution of disease-associated pseudogenes in M10. For pseudogenes with multiple disease-associated SNPs, the number of SNPs is shown in parentheses.

See also Figures S1 and S2; Table S1.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

Overall, our analyses demonstrate pseudogenes' tissue-specific expression and suggest their functions in both normal and pathological bone marrow.

HBBP1 is highly expressed in human erythroid cells as a noncoding RNA

We quantified the expression of *HBBP1* in erythropoiesis. First, we confirmed that *HBBP1* was expressed exclusively in bone marrow and particularly in bone marrow erythroblasts (Figures 2A and 2B). We also confirmed the annotated gene model (Figure 2A) by amplifying the termini (Figure S2F) and detecting the expected band (660 bp) in northern blot of blood (Figure 2C). We then reanalyzed RNA-sequencing (RNA-seq) datasets of erythropoiesis (Table S1) induced in hematopoietic stem/progenitor $CD34^+$ cells (HSPCs) derived from umbilical cord blood (UCB) and adult peripheral blood (Liu et al., 2018; Shi et al., 2014). For both systems, we detected high *HBBP1* expression, especially in the later stage of differentiation (Figure 2D). Taking day 8 as an example, the *HBBP1* expression level was higher than that of 99.9% of long intergenic noncoding RNAs (lincRNAs) and pseudogenes and 98.9% of protein-coding genes (Figure 2E). Such temporally regulated high expression indicated *HBBP1*'s potential importance during erythropoiesis.

Since pseudogenes tend to be unexpressed (Figure 1C), the high expression of *HBBP1* suggests that it is a misannotated, truncated protein-coding gene (Xu and Zhang, 2016). Although *HBBP1* harbors nonsense mutations or frame-disruptive indels relative to the closely related coding paralog, *HBG* (Koop et al., 1986; Storz, 2016), the middle sequence could be translated as an open reading frame (ORF) of 77 amino acids (Figure S2G). We, therefore, mapped the mass spectra data from the Human Proteome Map (HPM) (Kim et al., 2014) and identified uniquely mapping peptides for all coding paralogs (Figure 2F). However, no *HBBP1* peptides were detected, which was consistent with the absence of a signal in the protein expression assay (Figure 2G). Thus, *HBBP1* ORF is not translated.

Taken together, expression profiling demonstrated *HBBP1* as a highly expressed erythroid-specific lncRNA.

HBBP1 is essential for erythroid development and differentiation

Given *HBBP1*'s loss in the mouse (Hardison, 2012), we could not utilize a mouse model to investigate its function and mechanism. We, therefore, implemented five biologically or technically complementary human-centric systems, resembling at least one stage of erythroid development or differentiation (Figure S3A). We first induced the differentiation of human embryonic stem cells (hESCs) toward the erythroid lineage (Figure 3A). We generated two deletion mutants of the whole locus (2–3 Kb, $\Delta HBBP1^{-/-}$) and one deletion mutant of the transcription start site (TSS, 260 bp, $\Delta TSS^{-/-}$) (Figure 3B). For all mutants, we confirmed the presence of deletions and the absence of *HBBP1* expression (Figures S3B–S3D).

We then investigated the effects of *HBBP1* knockout on hematopoiesis. After 8 days of hESC differentiation into hematopoietic precursors, we detected no significant difference in the generation of HSPCs between three knockout and wild-type (WT) hESCs (5.6%–8.9% versus 7.0%) (Figures 3C and S3E). Thus, *HBBP1* is dispensable for hematopoiesis initiated by hESC differentiation. In the erythroid differentiation stages, whereas erythroblasts from WT hESCs increased from 9.5% on day 11 to 85.6% on day 20, almost no erythroblasts (0.2%–1.5%) were detected from either $\Delta HBBP1^{-/-}$ or $\Delta TSS^{-/-}$ cells (Figures 3C and 3D). The impact on early erythroid differentiation (day 11) suggests that *HBBP1* mediates early erythroid-lineage development. Consistently, we also detected fewer early erythroid-committed progenitors, i.e., burst-forming unit-erythroid colonies (BFU-E) from either $\Delta HBBP1^{-/-}$ or $\Delta TSS^{-/-}$ cells in comparison with WT cells (approximately 2 versus 58) (Figure 3E). Hemoglobin induction was also impaired (Figure S3F). Collectively, deletion of the *HBBP1* locus or impairment of its transcription leads to defective erythroid-lineage commitment.

We wondered whether *HBBP1* regulates erythropoiesis at later developmental stages given its even higher expression at this point (Figure 2D). We thus implemented a second system by inducing erythroid differentiation from UCB-derived HSPCs, followed by incorporation of two short hairpin RNAs (shRNAs) (shHBBP1-1 and shHBBP1-2), enabling downregulation of HBBP1 transcripts. The rationale was that a partial knockdown would yield a weaker phenotype compared with a complete knockout, and thus erythropoiesis may progress to the later stages. We confirmed that both shRNAs specifically targeted HBBP1 transcripts but not paralogous β-globin family members (Figure S4A) and that HBBP1 expression was reduced to 20%–30% (Figure S4B). In accordance with the hESC-based system, there were fewer mature erythroid cells at two time points (Figures 3F and 3G), which was also consistent with the inhibition of hemoglobin synthesis (Figure S4B).

These analyses were based on HSPCs of UCB and these cells approximate erythropoiesis at birth (Figure S3A). Given *HBBP1*'s high expression at birth and in adulthood, we predicted that *HBBP1* functions similarly during adult erythropoiesis. Thus, we introduced a third system by extracting HSPCs from adult bone marrow (BM) and conducting loss-of-function studies through knockdown of HBBP1. Analogously, there were fewer erythroblasts and lower expression of hemoglobin (Figure S4C–S4E). Note that HBBP1 depletion appears to be correlated with globin switching from fetal to adult β-globin (Sankaran et al., 2008; Xu et al., 2011) where the proportion of γ-globin [γ/(γ+β)] becomes lower (Figures S4B and S4E).

Having demonstrated the importance of *HBBP1* based on three *in vitro* systems, we aimed to confirm its *in vivo* function. We implemented a fourth strategy by establishing a xenograft model where human HSPCs from UCB were transplanted into the BM of immune-deficient mice. After ensuring a similar transplantation efficiency for shRNA-infected cells and WT cells (Figure S4F), we observed fewer erythroid progenitors and precursors after transplantation of the shRNA-infected cells (Liu et al., 2020) (Figures 3H and 3I).

Although the knockdown data indicated the HBBP1 transcripts' functionality, the knockout phenotype could be explained by *HBBP1*'s endogenous transcription especially considering that this locus contains regulatory elements that mediate long-range chromatin interactions (Huang et al., 2017; Ivaldi et al., 2018; Liu et al., 2017, 2020) and that *HBBP1*'s transcription may affect the interactions. To corroborate the function of transcripts, we performed two complementary experiments in the fifth system, i.e., the HUDEP-2 cell line (Figure S3A), which

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

CellPress

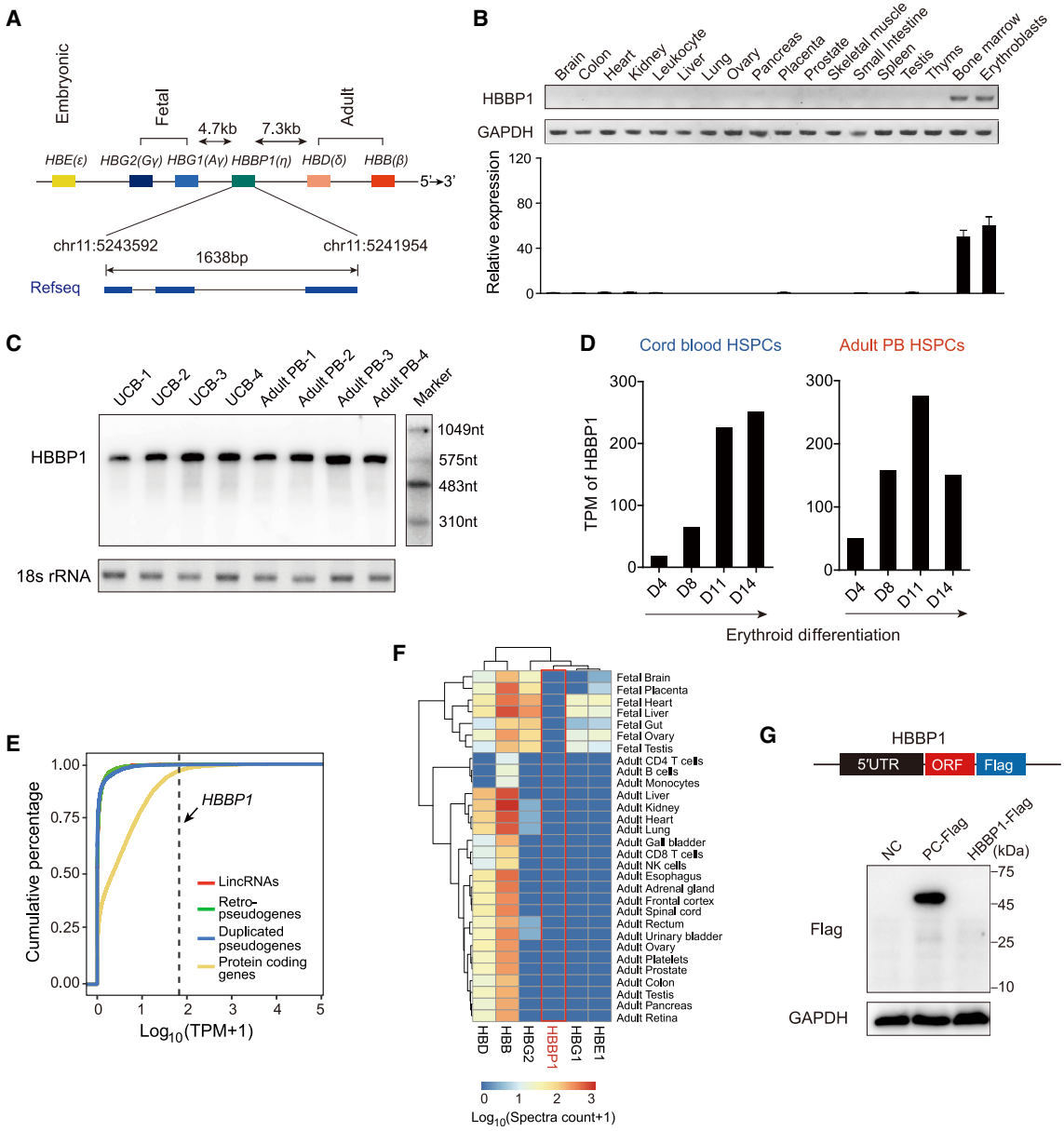


Figure 2. *HBBP1* is expressed abundantly in human erythroid cells as a lncRNA

(A) Schematic diagram of the human β-globin gene cluster with expression stages labeled.
(B) Quantification (q-PCR) of *HBBP1* expression in adult tissues. Data represent mean ± SD.
(C) Northern blot of *HBBP1* expression in umbilical cord blood (UCB) and adult peripheral blood (PB).
(D) Expression intensity (transcripts per million, TPM) of *HBBP1* expression during erythroid differentiation of HSPCs.
(E) Expression intensity distribution of coding and noncoding genes in UCB-derived erythroblasts at day 8 of differentiation.
(F) Proteomic profiling of the β-globin cluster based on the Human Proteome Map.
(G) Protein expression assay in 293 T cells. Both the *HBBP1* open reading frame (ORF) and a positive control (*PABPC1*) were fused with Flag.
See also Figure S2; Table S1.

represents immortalized erythroblasts derived from UCB (Kurita et al., 2013). First, we introduced mutations around one splice junction and generated two HUDEP-2 cell lines harboring junction mutations (Figure S4G). Since sequences around junctions are critical for splice-site recognition (Fairbrother et al., 2002), these mutations completely eliminated mature HBBP1 transcripts but up- and downregulated the primary transcripts moderately (Figure S4H). Both cell lines showed decreased expression of globin genes as well as decreased hemoglobin levels (Figures S4I and S4J). Second, we constructed *HBBP1* knockout HUDEP-2 cells and overexpressed exogenous HBBP1. *HBBP1* knockout leads to inhibition of erythroid differentiation, and its overexpression partially rescued the phenotype (Figures S4K–S4N). Thus, both lines of data consistently support

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

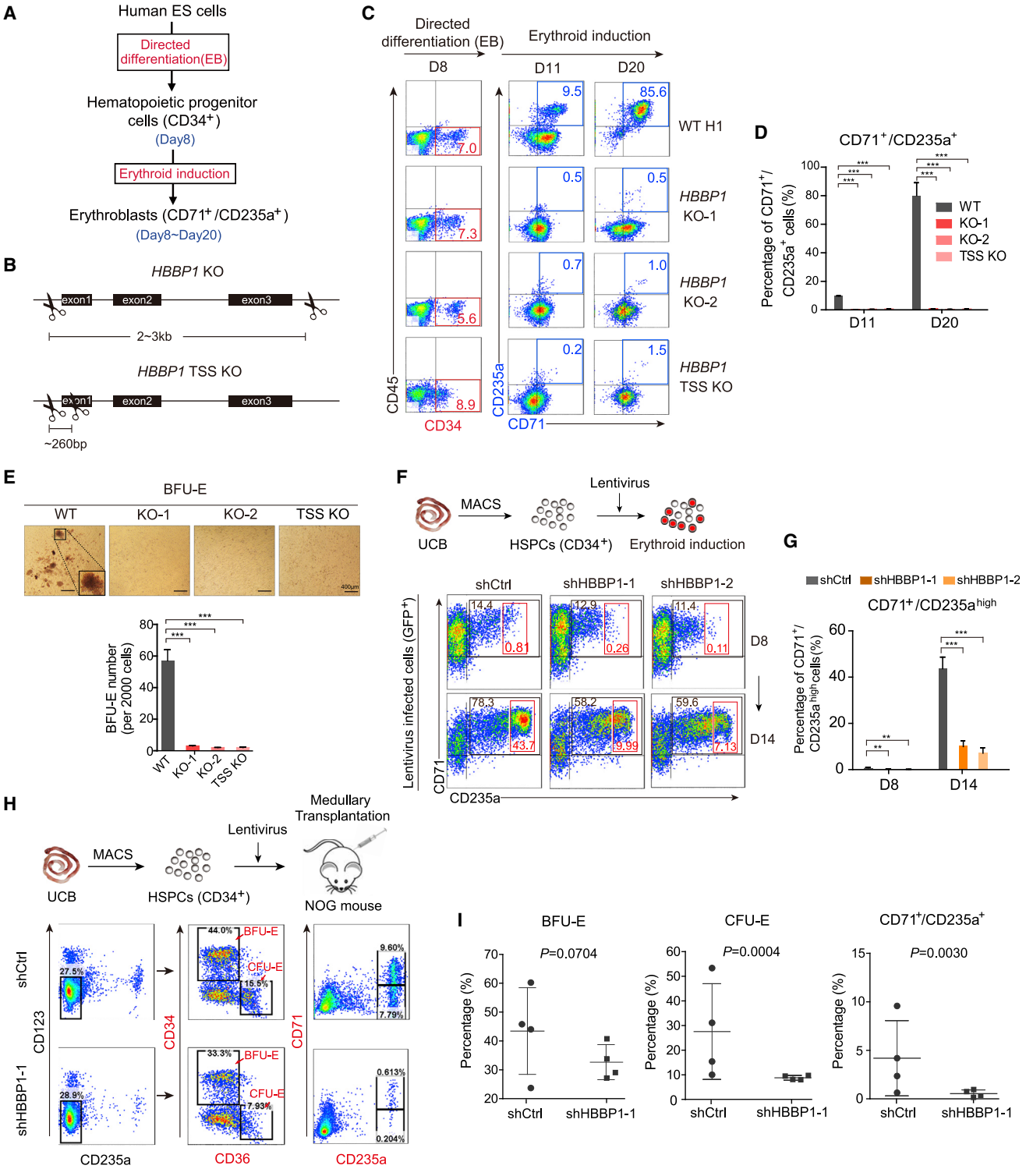


Figure 3. *HBBP1* is essential for human erythropoiesis

(A) Schematic diagram illustrating erythroid differentiation from hESCs.

(B) The strategy to knock out the *HBBP1* locus and its transcription start site in hESCs.

(C–E) Representative flow cytometry plots (C) and percentage of erythroblasts (D) show the induction of HSPCs and erythroblasts from hESCs at indicated days; (E) shows morphology and quantification of BFU-E colonies.

(legend continued on next page)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

the functionality of *HBBP1* transcripts for erythroid differentiation.

In summary, our loss-of-function studies supported the importance of *HBBP1* transcripts for erythroid-lineage commitment and maturation.

HBBP1 RNA binds HNRNPA1 and stabilizes *TAL1* mRNA

We next investigated the mechanism by which *HBBP1* transcripts drive erythropoiesis based on UCB-derived HSPCs. By performing differential expression analysis based on RNA-seq data in HBBP1 knockdown cells, we observed the inhibition of genes promoting erythroid differentiation (Figures 4A and S5A; Table S3). In particular, key erythropoiesis regulatory genes (e.g., *TAL1*) were downregulated (Figure 4B). Decreased expression of TAL1 was consistently observed in the *HBBP1* splice junction mutated and *HBBP1* KO HUDEP-2 cells, while TAL1 expression was partially rescued by HBBP1 overexpression (Figure S5B).

Since subcellular localization provides clues about mechanism, we determined the cytoplasmic location of HBBP1 (Figures 4C and S5C). Since no miRNAs were predicted to bind HBBP1 (Figure S5D), we hypothesized that HBBP1 might interact with RBP(s) to execute its function. By conducting RNA pull-down assays, mass spectrometry (MS), and immunoblotting in HUDEP-2 cells, we reasoned HNRNPA1 and HNRNPA2B1 were candidate HBBP1-interacting RBPs for the following reasons: (1) they were ranked to 3rd and 7th in terms of protein abundance in MS data, respectively (Figure 4D; Table S4), and (2) they bind HBBP1 specifically while the other ten abundant proteins bind a random control RNA too (Figure 4E). However, the reverse immunoprecipitation with antibodies against HNRNPA1 and HNRNPA2B1 revealed the enrichment of HBBP1 only in HNRNPA1 immunoprecipitates (Figure 4F). We predicted that *HBBP1* RNA harbors two candidate HNRNPA1 binding motifs, i.e., hexamers beginning at nucleotides (nt) 371 and nt 580, respectively. The electrophoretic mobility shift assay showed that HNRNPA1 bound only the motif at nt 371 in a dose-dependent manner, which was consistent with the relatively higher prediction score for this motif (Figure 4G). In parallel, enhanced cross-linking and immunoprecipitation (eCLIP) experiments showed that HBBP1 RNA fragments harboring the nt 371 motif interact with HNRNPA1 in HUDEP-2 cells, while the other fragments or a random control could not (Figure S5E). Collectively, these results support the notion that HNRNPA1 interacts with HBBP1 at the nt 371 motif (TAGGCA).

Since cytoplasmic HNRNPA1 mediates RNA decay (Geissler et al., 2016; Zhao et al., 2009), we hypothesized that HBBP1 affects the RNA stability of erythropoiesis regulators by binding HNRNPA1. To test this possibility, we identified 40 genes exhibiting the most significant differential expression between knockdown and WT cells during erythroid differentiation (Figure S5F). We further identified eight with significantly different half-lives (Figure 4H). Among them, *TAL1* was the only known essential master regulator driving erythropoiesis (Hattangadi et al., 2011; Porcher et al., 1996; Robb et al., 1995), and its expression profile is highly correlated with HBBP1 (Figure 1F); thus, we subsequently focused on this gene. The half-life of *TAL1* mRNA was decreased by ~50% following HBBP1 knockdown but increased by ~50% following HBBP1 overexpression (Figure 4I). Similar patterns were observed for TAL1 protein following HBBP1 knockdown or overexpression (Figure 4J).

Thus, HBBP1 seemingly regulates TAL1 by binding HNRNPA1.

HBBP1 RNA stabilizes *TAL1* mRNA as a RBP decoy

We further dissected how HNRNPA1 underlies HBBP1-mediated *TAL1* mRNA stabilization. We found that *TAL1* mRNAs were also enriched in immunoprecipitates of HNRNPA1 (Figure 5A). The half-life of *TAL1* mRNA was increased after HNRNPA1 knockdown, suggesting that HNRNPA1 mediates *TAL1* mRNA decay (Figures 5B and S5G). Moreover, the enrichment of *TAL1* mRNA in HNRNPA1 immunoprecipitates was enhanced after HBBP1 knockdown and reduced after HBBP1 overexpression (Figure 5C).

These data suggest a model where the depletion of HBBP1 decreases the stability of TAL1 by disturbing the RNA interaction network mediated by HNRNPA1 (Figure 5D). The feasibility of this model is supported by the relative abundance of three genes and their subcellular location: (1) as aforementioned, *HBBP1*'s expression is higher than 99% of coding genes (Figure 2E) and its expression level is even higher than *TAL1* in the later stage of differentiation (~200 versus ~150, Figure S5H); (2) HBBP1 is different from TAL1 and HNRNPA1 binding does not cause HBBP1's shorter half-life (Figure 5B); (3) although HNRNPA1 transcript is even more abundant (~1,500, Figure S5H), only a small proportion of protein is cytoplasmic (Figure S5I); and (4) fluorescence signal showed that the diffused low level of cytoplasmic HNRNPA1 is co-localized with HBBP1 and TAL1 RNA, both of which tend to be spatially close too (Figure S5J).

Therefore, once *HBBP1* is disrupted, HNRNPA1 is released to degrade *TAL1* mRNA. Since *TAL1* orchestrates the erythropoiesis process, both early fate commitment and late maturation are affected. To prove this causality, we disrupted the HNRNPA1 binding motif in HUDEP-2 cells. That is, we generated one heterozygous mutant where the motif and neighboring nucleotides were changed and HBBP1 showed a moderate downregulation (Figures S6A and S6B). We detected decreased *TAL1* mRNA stability and lower globin expression (Figure S6B), although the downregulation of HBBP1 may also contribute. Therefore, to further enhance this causal relationship, we performed two rescue experiments. First, we conducted rescue assays by simultaneously delivering HBBP1 and HNRNPA1 shRNAs (Figure S6C). Although HNRNPA1 knockdown did not fully rescue the impairment of erythroid differentiation possibly due to its pleiotropy as an RBP, but the defect was significantly moderate, as shown by increasing erythroblasts (Figures 5E and 5F) as well as

(F and G) Representative flow cytometry plots (F) of erythroblasts and the percentage (G) of CD71^{+}/CD235a^{high} erythroblasts with HBBP1 knockdown. Compared with CD235a^{+}, CD235a^{high} cells are more mature.

(H and I) *In vivo* transplantation assay. Representative flow cytometry plots (H) and quantification (I) of erythroid progenitors after transplantation for 15 weeks. If applicable, data are represented as the mean ± SD. **: t test $p < 0.01$, ***$p < 0.001$. See also Figures S3 and S4; Table S3.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

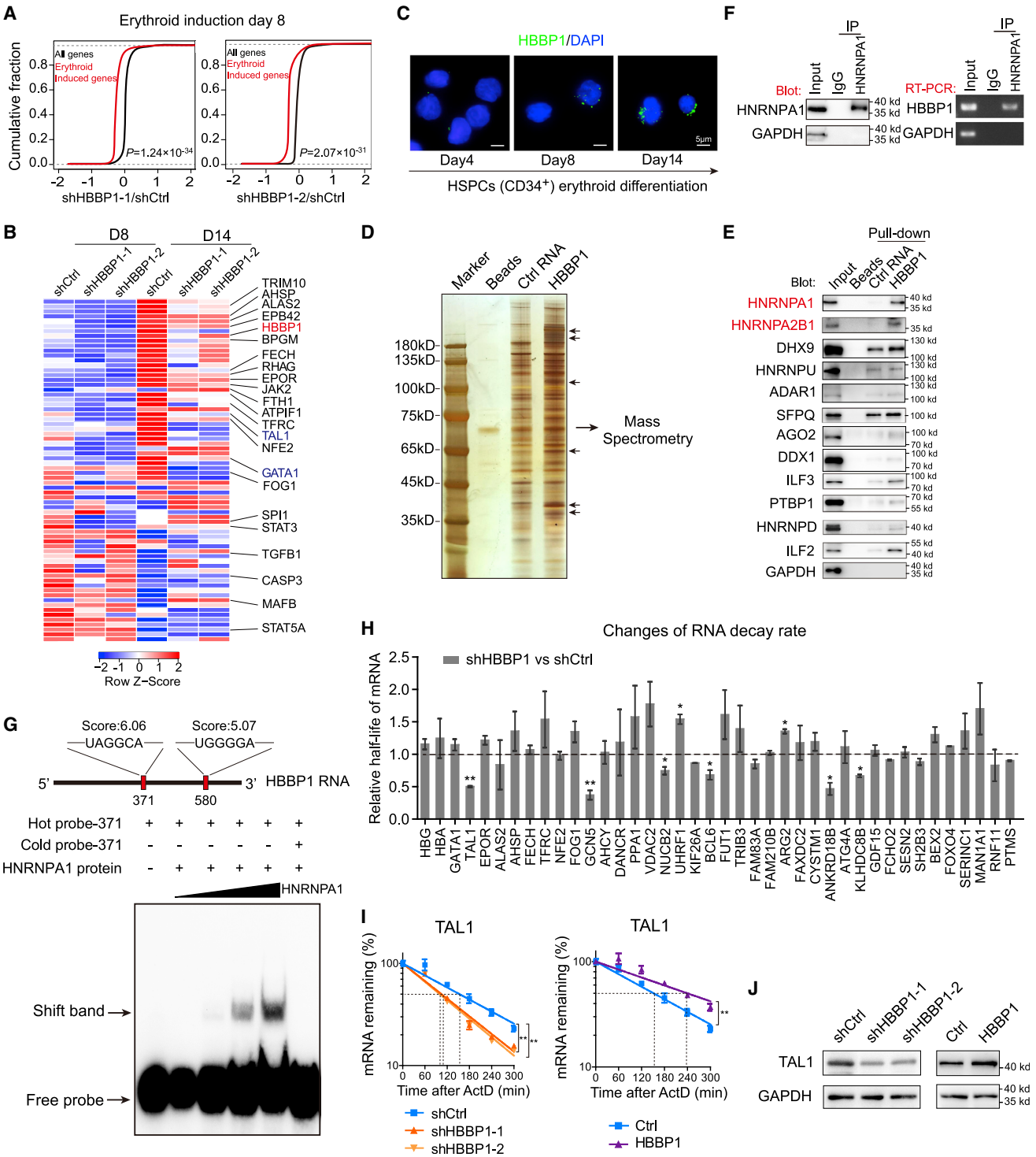


Figure 4. *HBBP1* RNA binds to HNRNPA1, which is associated with stabilization of TAL1

(A) Cumulative distribution of "all genes" and "erythroid induced genes" under HBBP1 reduction on day 8.

(B) Heatmap depicting significantly differentially expressed genes associated with erythroid differentiation after HBBP1 knockdown.

(C) Subcellular localization of HBBP1 during erythroid differentiation induced from UCB-derived HSPCs.

(D) Silver staining of HBBP1 interacting proteins in HUDEP-2 cell lysates.

(E) Western blot validation of proteins interacting with biotinylated HBBP1 RNA in HUDEP-2 cell lysates.

(F) Western blot and RT-PCR analyses showing the specificity of HNRNPA1 antibody (left) and the presence of HBBP1 in the HNRNPA1 immunoprecipitations (right).

(legend continued on next page)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

upregulation of TAL1 and globin expression (Figure S6D). Second, we overexpressed TAL1 after HBBP1 knockdown and overexpressed HBBP1 after TAL1 knockdown. TAL1 overexpression partially rescued the defect of erythroid differentiation induced by HBBP1 repression (Figures S6E and S6F), while the promotion of HBBP1 overexpression on erythropoiesis fully disappeared when TAL1 was depleted (Figure S6G). Thus, it seems that HBBP1's effect is largely dependent on TAL1.

If HBBP1 stabilizes *TAL1* mRNA, we expected that *HBBP1* and *TAL1* exhibit correlated expression across human individuals. The peripheral blood datasets of the Genotype-Tissue Expression project (GTEx Consortium et al., 2015) showed that both *HBBP1* and *TAL1* were indeed correlated (Spearman ρ = 0.63; Figure 5G) despite that erythroid cells in peripheral blood are generally terminally differentiated and contain few RNAs (Ji et al., 2011).

Thus, we concluded that HBBP1 drives erythropoiesis by competitively binding HNRNPA1 and stabilizing TAL1.

The stronger HBBP1/TAL1 interaction may contribute to a milder β-thalassemia symptom

We next hypothesized that the HBBP1/TAL1 interaction is involved in the pathogenesis of β-thalassemia for two reasons: (1) *HBBP1* harbors variants associated with the severity of β-thalassemia (Giannopoulou et al., 2012; Nuinoon et al., 2010) (Figures 1G and S2D) and (2) TAL1 reactivates γ-globin transcription in adults and relieves the symptoms of β-thalassemia (Wienert et al., 2015). Thus, TAL1 may underlie the aforementioned HBBP1 depletion associated lower fetal globin percentages (Figure S4E). By analyzing blood samples from 303 β-thalassemia patients, we found consistent results (Figures 6A, 6B, and S6H): (1) higher levels of *HBBP1* and *TAL1* transcripts were associated with milder β-thalassemia symptoms, and (2) both *HBBP1* and *TAL1* were significantly ($p < 0.01$) correlated with fetal hemoglobin (HbF, $\alpha 2\gamma 2$) although the correlation for *TAL1* is moderate (0.19). Note that, consistent with the fact that the intronic SNP of *HBBP1* (rs2071348) is associated with the severity of β-thalassemia (Figure S2D), the alternative C allele is associated with higher HBBP1 and HbF levels ($p < 0.05$, Figure S6I). This correlation could be caused by the regulatory role of this SNP.

To obtain direct evidence that HBBP1/TAL1 interaction underlies the upregulation of γ-globin, we modified our previous third system (HSPCs extracted from adult BM) by adding hydroxyurea (HU), which induces γ-globin expression (Mabaera et al., 2008; Platt et al., 1984). After induction, γ-globin and $\gamma/(\gamma+\beta)$ increased in association with elevated *HBBP1* and *TAL1* expression (Figure 6C). Following HBBP1 knockdown, TAL1 was decreased, while the increase in γ-globin and $\gamma/(\gamma+\beta)$ was strongly attenuated (Figure 6D).

In summary, these preliminary data suggest that the HBBP1/TAL1 interaction facilitates the reactivation of γ-globin and contributes to milder β-thalassemia symptoms.

The essentiality of *HBBP1* transcripts in erythropoiesis is human specific

The low conservation of *HBBP1* across mammals (Hardison, 2012; Opazo et al., 2008) led us to investigate the hypothesis that its essentiality is human- or primate-specific. First, we confirmed that *HBBP1* was not conserved across mammals. After its origination in the ancestor of placental mammals, *HBBP1* was subject to three independent pseudogenizations in primate, pig, and armadillo, and two independent losses in rodents and elephant, while it appeared to be maintained as a protein-coding gene in the dog and horse lineages (Figures 7A and S7A; Table S5). In addition, *HBBP1* was poorly conserved across primates (Figure S7B). Phylogenetic analyses further revealed a human-specific slowdown of exonic evolution compared to other primates or hominoids (Figure S7C). Thus, these findings indicated that *HBBP1* has evolved neutrally in primates, except in humans, where it is subject to functional constraint.

Second, since the regulatory role of HBBP1 in humans is reflected by its correlation with TAL1 (Figure 5G), we investigated whether this correlation is shared by other mammals. Reanalysis of public blood RNA-seq data showed that *HBBP1* was not widely expressed in macaques, horses, and pigs (Figure 7B; Table S1). In dogs, it was expressed but uncorrelated with *TAL1* (Figure 7C). More interestingly, *HBBP1* was also not or poorly expressed in chimpanzee RNA-seq samples. To confirm this, we examined five additional blood samples from chimpanzees by qRT-PCR and observed the consistent absence of HBBP1 (Figure 7B).

Third, if HNRNPA1-mediated regulation is human-specific, the motif (TAGGCA) should not be conserved in non-human mammals. Consistently, the motif emerged by a G->A substitution in the placental mammal ancestor and then became independently lost three times: deletion of the third exon harboring this motif in bushbaby (Figure 7A), a point mutation in rhesus macaque (TAAGCA), and an indel in pig (TAGG–, Figure 7D). Similarly, HNRNPA1 did not bind either the ancestral or rhesus macaque version (Figure S7D). Furthermore, although exogenous HBBP1 expression of the extant version could partially rescue the phenotype in *HBBP1* knockout cells, expression of the ancestral version does not show any rescue effect (Figures 7E, 7F, and S7E). The lack of evolutionary constraint on the gene sequence, expression, and motif suggests that *HBBP1* has been functionless for the majority of its evolution but has recently acquired its function as a human-specific essential regulator of erythropoiesis. Since HNRNPA1 binds and degrades TAL1 in mouse (Figures S7F and S7G), the following parsimonious process seems most likely (Figure 7G): (1) HNRNPA1-TAL1 interaction was ancestrally shared across mammals and TAL1 was upregulated via an unknown common mechanism, (2) after the emergence of *HBBP1* in the mammalian ancestor, its pseudogenizations or deletions occurred after the mammalian split, and (3) *HBBP1* transcription was reactivated and co-opted into the HNRNPA1/TAL1 network during human evolution as a replacement for the pre-existing mechanism of TAL1 upregulation.

(G) Motif analysis. Two putative HNRNPA1 binding motifs were predicted in the *HBBP1* transcript where the nt 371 motif was validated.
(H) Changes in the relative half-lives of the top 40 significantly differentially expressed genes after HBBP1 knockdown.
(I and J) Half-life of *TAL1* mRNA (I) and TAL1 protein (J) in HUDEP-2 cells with HBBP1 repression or overexpression. Dashed lines indicate 50% decrease of half-lives.
If applicable, data are represented as the mean ± SD. *t test $p < 0.05$, **$p < 0.01$. See also Figure S5; Table S4.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

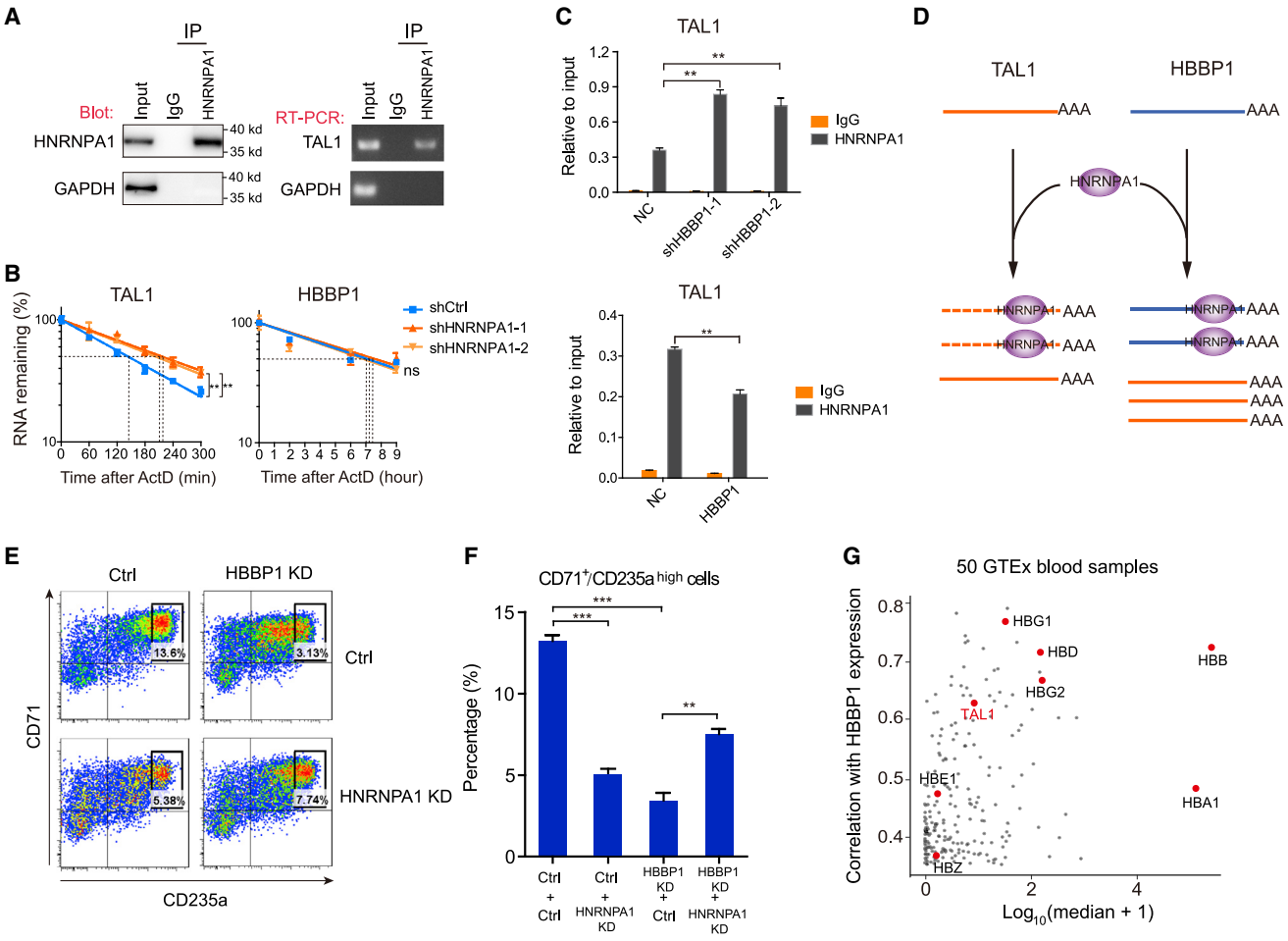


Figure 5. *HBBP1* RNA stabilizes *TAL1* mRNA by competitively binding HNRNPA1

(A) Western blot and RT-PCR analyses showing the specificity of HNRNPA1 antibody (left) and the presence of TAL1 mRNA in HNRNPA1 immunoprecipitations (right).

(B) Half-life of *TAL1* and *HBBP1* RNA after HNRNPA1 depletion.

(C) The relative amount of *TAL1* mRNA interacting with HNRNPA1 when HBBP1 was depleted or overexpressed.

(D) Working model of the mechanism by which HBBP1 functions as an HNRNPA1 decoy to increase *TAL1* mRNA stability. Dashed line indicates degradation of *TAL1* mRNA.

(E and F) Representative flow cytometry plots (E) and percentage of $CD71^{+}/CD235a^{high}$ erythroblasts (F) with simultaneous HBBP1 and HNRNPA1 knockdown during HSPCs erythroid differentiation.

(G) Protein-coding genes positively correlated with *HBBP1* in 50 reprocessed GTEx blood samples.

If applicable, data are represented as the mean ± SD. ns, no statistical significance. **t test $p < 0.01$; ***$p < 0.001$. See also Figure S6; Table S1.

Overall, our analyses suggested that *HBBP1* is co-opted into a critical developmental program during human evolution.

An appreciable proportion of duplicated pseudogenes could function as RBP decoys

Finally, we ask whether *HBBP1* represents an outlier or indicates a pattern in duplicated pseudogenes acting as RBP decoys. To approach this question, we analyzed the transcriptome and interactome features of duplicated pseudogenes relative to retropseudogenes.

First, in accordance with previous reports (Glenfield and McLysaght, 2018; Pei et al., 2012), we found that duplicated pseudogenes/paralog pairs showed stronger correlations in expression compared to retropseudogenes/paralog pairs (median correlation, 0.17 versus 0.10) (Figure S7H; Table S6). Thus, like *HBBP1*, duplicated pseudogenes are more likely to be transcribed under similar regulatory contexts as their paralogs. In contrast, retropseudogenes showed weaker correlations with their paralogs, likely due to loss of promoters during the retroposition process. Next, we attempted to analyze interaction features for these two groups of pseudogenes, including sense/antisense pairing, miRNA binding, and RBP binding (Cheetham et al., 2020; Poliseno, 2012). Duplicated pseudogenes are processed as siRNAs less frequently than retropseudogenes (10% versus 20%) (Guo et al., 2014), suggesting less sense/antisense pairing for the former group (Tam et al., 2008). Therefore, we compared the two groups only with respect to their preferential binding by miRNAs and RBPs. Using public

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

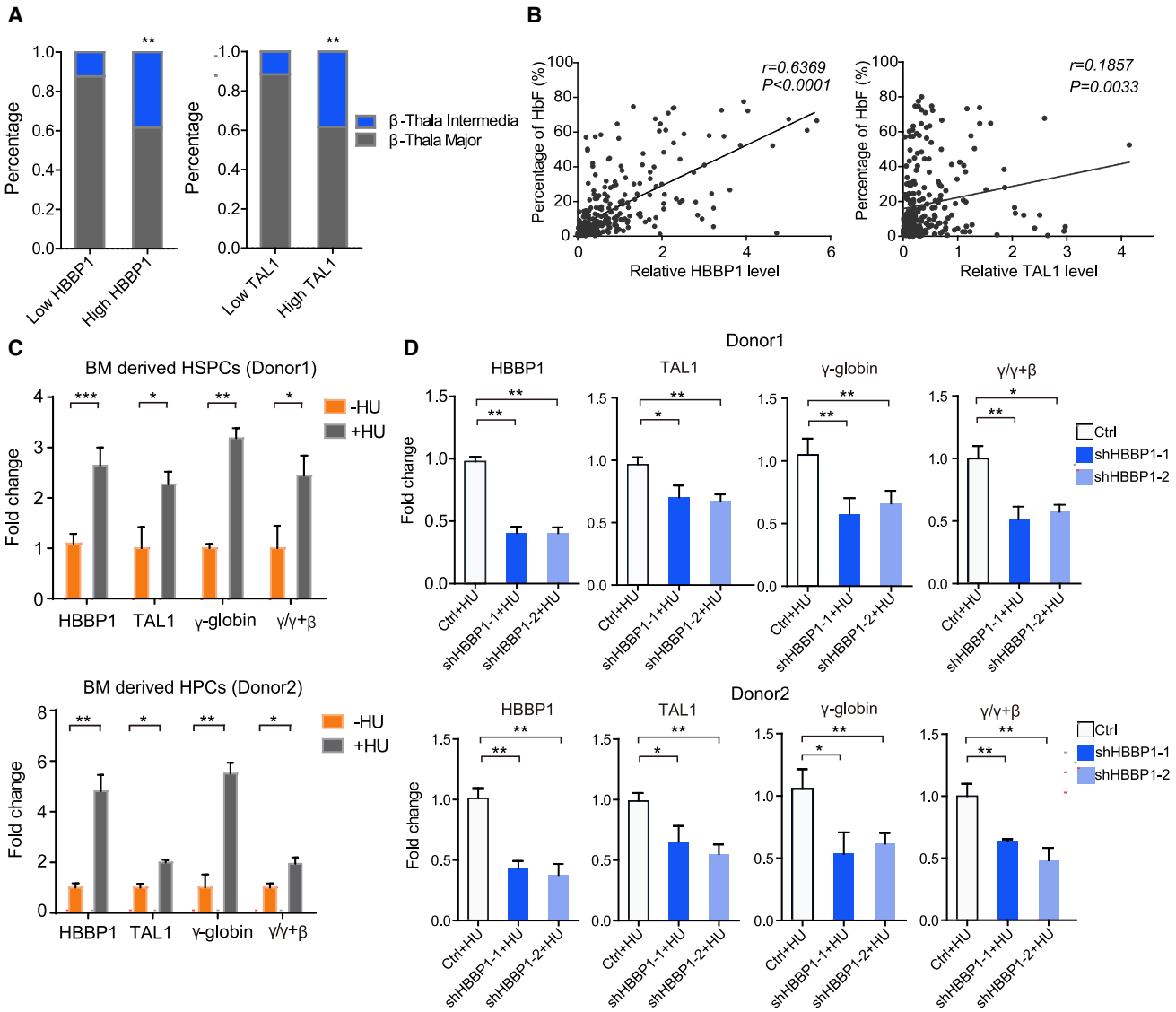


Figure 6. The stronger HBBP1-TAL1 interaction contributes to a benign symptom in β-thalassemia patients

(A) Major (severe) β-thalassemia patients are overrepresented among patients with low *HBBP1* or *TAL1* expression.

(B) Positive correlation of fetal hemoglobin (HbF) level with HBBP1 and TAL1 expression in the blood of β-thalassemia patients.

(C) Relative expression of HBBP1, TAL1, and γ-globin as well as γ/γ+β ratio in erythroid cells after hydroxyurea (HU) treatment.

(D) Relative expression of TAL1, γ-globin, and γ/γ+β ratio with HU treatment followed by HBBP1 knockdown.

If applicable, data are represented as mean ± SD. *t test $p < 0.05$; **$p < 0.01$; ***$p < 0.0001$. See also Figure S6.

data, we found that duplicated pseudogenes were less targeted by miRNAs (6% versus 16%), while they were more often bound by RBPs relative to retropseudogenes (6% versus 1%, Figure 7H; Table S7). The same preference was also observed for pseudogenes in the bone-marrow-specific module (Figure S7I).

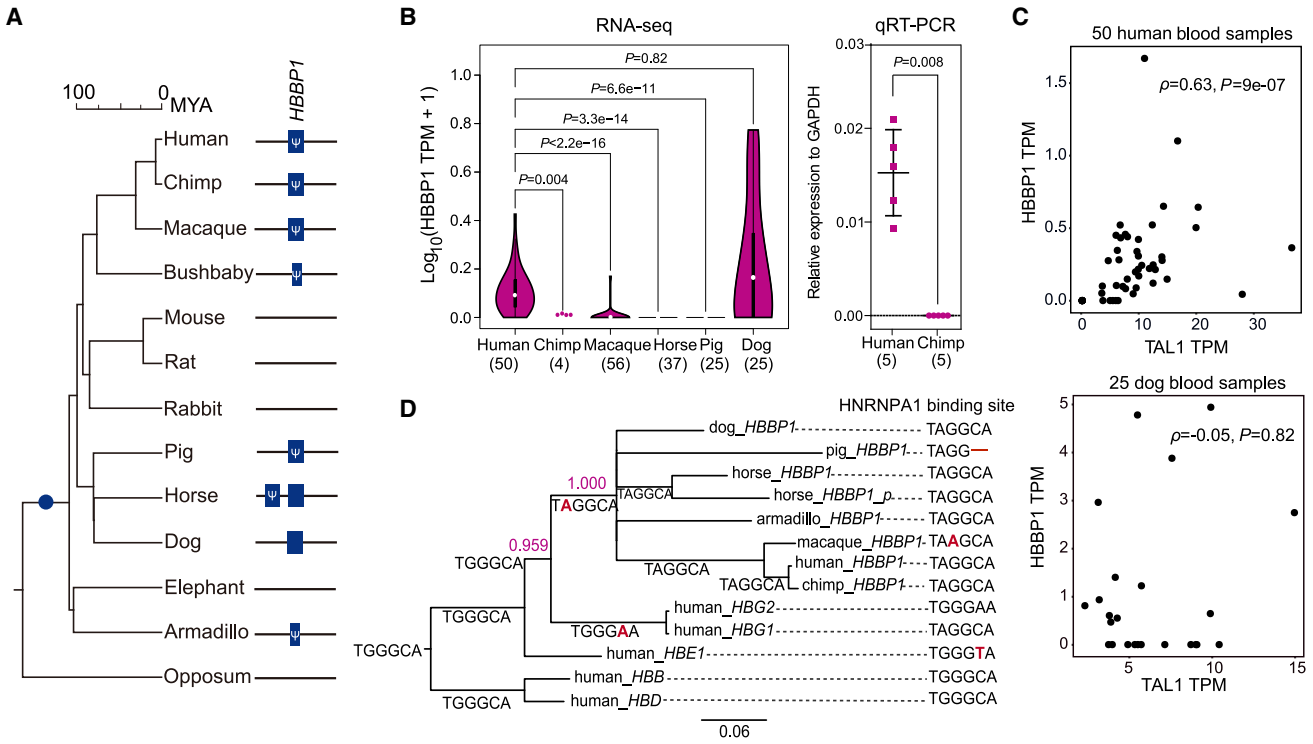
The 104 RBP-binding duplicated pseudogenes share two features similar to *HBBP1*: (1) stronger co-regulation with their coding paralogs as revealed by higher correlation compared with pseudogenes without RBP-binding (0.31 versus 0.15, Figure 7I), (2) pseudogene-specific binding for the majority (62%) of cases (Figure 7J). Together with pseudogenes' low conservation (Glenfield and McLysaght, 2018), a proportion of duplicated pseudogenes may shape human evolution by interacting with RBPs.

DISCUSSION

In this study, we performed a genome-wide analysis of human pseudogenes and an in-depth study of one pseudogene, i.e., *HBBP1*. This study has far-reaching implications regarding the functions of pseudogenes, their underlying mechanisms, and species-specific phenotypes.

Possibly due to the traditional view that pseudogenes are functionless evolutionary relics, only dozens of human pseudogenes have been functionally characterized (Cheetham et al., 2020; Poliseno, 2012; Poliseno et al., 2015). Herein, using co-expression network analyses, we cataloged the tissue-specific pseudogenes and found a high abundance of pseudogenes in

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019



(legend on next page)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

BM and testis. The unexpected enrichment of pseudogenes in BM is consistent with the overrepresentation of primate-specific duplicate coding genes expressed in BM (Shao et al., 2019) and the presence of human-specific coding genes increasing the hematopoietic stem cell reservoir (Costantini et al., 2019). Subsequent analyses highlighted the importance of *HBBP1* during physiological and pathological erythropoiesis. Furthermore, some hematopoietic lineage-specific pseudogenes showed strong correlation with lineage-specific master transcription factors and displayed disease associations, suggesting their general functional relevance in human hematopoiesis. Therefore, these analyses emphasize the significance of functional pseudogenes and echo the emerging opinion that pseudogenes should not be excluded from functional analyses (Cheetham et al., 2020).

Our meta-analyses further demonstrated that duplicated pseudogenes and retropseudogenes preferentially bind RBP and miRNAs, respectively. This pattern, together with the known pattern that duplicated pseudogenes are less likely to function as antisense regulators (Guo et al., 2014; Tam et al., 2008), can be used to guide future studies of pseudogenes. These patterns are illuminating. On one hand, knowledge of the functional mechanisms of duplicated pseudogenes is limited, due to the prior focus on retropseudogenes (Chiefari et al., 2010; Karreth et al., 2015; Poliseno et al., 2010). On the other hand, the difference between the two groups of pseudogenes suggests that the underlying mutational mechanism could predetermine the evolutionary trajectory of pseudogenes. That is, DNA-level duplications could contain the 5′ termini of genes, while the 3′ termini are lost (Dougherty et al., 2018). In contrast, human retropseudogenes emerge mainly via long interspersed nuclear element 1(L1)-mediated retroposition, in which L1 uses the polyA tails of mRNAs as primers (Kaessmann et al., 2009; Tan et al., 2016). Thus, compared to retropseudogenes, the duplicated pseudogenes more often carry promoters and lose 3′ UTRs, which limits the opportunity to evolve as antisense regulators or miRNA decoys.

Finally, *HBBP1* represents the co-opted pseudogene shaping human-specific regulation. Actually, the *HBBP1* locus regulates γ-globin expression by affecting DNA folding (Huang et al., 2017), indicating that both its DNA region and transcripts are functional. Regardless of this complexity, the human-specific indispensability of the *HBBP1* transcript is significant in two aspects. It is conventionally believed that essential genes are highly conserved (Goldstein et al., 2009); however, recent evidence suggests that newly originated protein-coding genes could be essential (Chen et al., 2010; Kasinathan et al., 2020; Ross et al., 2013). *HBBP1* expands this emerging concept by demonstrating that a pseudogene can be revived and rapidly acquire essentiality. Although the fitness consequences remain enigmatic, the reprogrammed erythropoiesis may provide better energetic support for human-specific endurance running (O'Bleness et al., 2012). More than that, this layer of human erythropoiesis represents an underexplored direction. That is, while previous studies focuses on visible traits (O'Bleness et al., 2012), more cryptic uniqueness of human biology, including development and pathogenesis, might be hereon revealed.

In conclusion, pseudogenes represent a new layer in the flow of genetic information. The highly integrative framework implemented in this study provides a prototype for determining the function of pseudogenes under normal and pathological conditions. Exploration of species-specific regulatory functions of pseudogenes or even studies of population-specific pseudogenes (Ewing et al., 2013) are expected to blossom in future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - hESC Cells
  - Human hematopoietic stem/progenitor $CD34^+$ cells (HSPCs)
  - HUDEP-2 cells
  - Human and chimpanzee samples
- METHOD DETAILS
  - Gene annotation and reannotation
  - Transcriptome (RNA-seq) analyses
  - GWAS data analyses
  - hESC based erythroid differentiation
  - CRISPR/Cas9 mediated knockout and mutation

Figure 7. *HBBP1* shapes human-specific regulation and duplicated pseudogenes are more often bound by RBPs relative to retropseudogenes

(A) Distribution of *HBBP1* orthologs along the mammalian phylogenetic tree. Blue dot indicates the origin of *HBBP1* via duplication. Blue boxes indicate a protein-coding gene, while "Ψ" indicates a pseudogene.

(B) Violin plot of *HBBP1* expression (RNA-seq) of adult mammals (left) and qRT-PCR data in chimpanzees and humans (right). Sample sizes are shown in brackets.

(C) Spearman correlation of *HBBP1* and *TAL1* expression across individual blood samples.

(D) Evolution of the *HNRNPA1* binding motif. The Bayesian posterior probability is marked. Nucleotide changes are marked in red.

(E and F) The half-life of *TAL1* mRNA (E) and relative expression of globin genes (F) in HUDEP-2 cells with overexpression of wild-type HBBP1 and HNRNPA1 motif-mutated HBBP1.

(G) Evolutionary trajectories of *HBBP1* and HNRNPA1 mediating HBBP1-TAL1 interaction.

(H) Proportions of retroposed or duplicated pseudogenes bound by miRNAs or RBPs, respectively.

(I) Distribution of the correlation between pseudogenes and their coding paralogs.

(J) Distribution of duplicated pseudogenes in terms of binding specificity. ns, no statistical significance.

If applicable, data are represented as mean ± SD. **t test $p < 0.01$, ***$p < 0.001$. See also Figure S7 and Tables S5, S6, and S7.

- Mouse transplantation
- Plasmid construction, lentivirus production and infection
- RNA extraction and Quantitative real-time PCR (qRT-PCR)
- Rapid amplification of cDNA ends (RACE)
- RNA *in situ* hybridization and immunofluorescence
- Cell fractionation
- RNA pull-down assays
- RNA immunoprecipitation (IP)
- Western blotting, silver staining and mass spectrometry
- RNA electrophoretic mobility shift assay (EMSA)
- Enhanced UV crosslinking, immunoprecipitation and qPCR (eCLIP-qPCR)
- Interactome analyses
- Evolutionary analyses

● QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.devcel.2020.12.019.

ACKNOWLEDGMENTS

We thank Drs. Weiwei Zhai, Bin He, Qi Zhou, Manyuan Long, Xionglei He, Maria D. Vibranovski, Siqin Ge, and Zhijin Liu for helpful comments. This work was supported by the National Key Research and Development Program of China (2019YFA0801800, 2019YFA0802600, 2016YFA0100601, 2016YFA0102300, 2017YFA0103102, and 2018YFC1406902), CAMS Innovation Fund for Medical Sciences (2017-I2M-3-009, 2019-I2M-2-001, 2016-I2M-3-002, 2017-I2M-1-015, and 2016-I2M-1-018), the National Natural Science Foundation of China (81530007, 31725013, 81970101, 31771444, 81870092, 81870089, 81890990, 917313002, 31970565, and 31771410), the CAMS (2017-I2M-B&R-04; 2018RC310015, 2017PT31035), CAS (ZDBS-LY-SM005, no. XBZG-ZDSYS-201913), the Open Research Program of Chinese Institute for Brain Research (Beijing), and funds for public research and capacity building in Guangdong (2014A02021102). Computing was supported by the HPC Platform of BIG.

AUTHOR CONTRIBUTIONS

J.Y., Y.E.Z., L.S., and Y.M. conceived the study; Y. M., S.L. C.S., F.Z., Y.S., and H.Z. performed experimental analyses; J.G., G.G., and Q.G performed the HSPC functional study; C.C., H.Y.,Y.M., Y.S., and Z.C. performed the bioinformatics analysis; X.Z., Q.L., L.C., and Y.H performed erythroid differentiation from hESCs; X.Y., P. L., P.Y., H.L., G.L.B.,Y.N., and R.K. provided materials and reagents; X.W. and F.W. provided technical assistance; Y.E.Z., J.Y., Y.M., L.S., and X.Z. wrote the manuscript; G.L.B., J.D.E., and G.F. edited the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 28, 2020
Revised: November 16, 2020
Accepted: December 28, 2020
Published: January 20, 2021

REFERENCES

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. *34*, 525–527.

Cheetham, S.W., Faulkner, G.J., and Dinger, M.E. (2020). Overcoming challenges and dogmas to understand the functions of pseudogenes. Nat. Rev. Genet. *21*, 191–201.

Chen, S., Krinsky, B.H., and Long, M. (2013). New genes as drivers of phenotypic evolution. Nat. Rev. Genet. *14*, 645–660.

Chen, S., Zhang, Y.E., and Long, M. (2010). New genes in Drosophila quickly become essential. Science *330*, 1682–1685.

Chiefari, E., Iiritano, S., Paonessa, F., Le Pera, I., Arcidiacono, B., Filocamo, M., Foti, D., Liebhaber, S.A., and Brunetti, A. (2010). Pseudogene-mediated posttranscriptional silencing of HMGA1 can result in insulin resistance and type 2 diabetes. Nat. Commun. *1*, 40.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Costantini, T.W., Chan, T.W., Cohen, O., Langness, S., Treadwell, S., Williams, E., Eliceiri, B.P., and Baird, A. (2019). Uniquely human CHRFAM7A gene increases the hematopoietic stem cell reservoir in mice and amplifies their inflammatory response. Proc. Natl. Acad. Sci. USA *116*, 7932–7940.

Darriba, D., Taboada, G.L., Doallo, R., and Posada, D. (2012). JModelTest 2: more models, new heuristics and parallel computing. Nat. Methods *9*, 772.

Dennis, M.Y., and Eichler, E.E. (2016). Human adaptation and evolution by segmental duplication. Curr. Opin. Genet. Dev. *41*, 44–52.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dougherty, M.L., Underwood, J.G., Nelson, B.J., Tseng, E., Munson, K.M., Penn, O., Nowakowski, T.J., Pollen, A.A., and Eichler, E.E. (2018). Transcriptional fates of human-specific segmental duplications in brain. Genome Res. *28*, 1566–1576.

Duret, L., Chureau, C., Samain, S., Weissenbach, J., and Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. Science *312*, 1653–1655.

Ewing, A.D., Ballinger, T.J., Earl, D., Broad Institute, Harris, C.C., Ding, L., Wilson, R.K., and Haussler, D. (2013). Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. Genome Biol *14*, R22.

Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. (2002). Predictive identification of exonic splicing enhancers in human genes. Science *297*, 1007–1013.

Faulkner, G.J., Forrest, A.R., Chalk, A.M., Schroder, K., Hayashizaki, Y., Carninci, P., Hume, D.A., and Grimmond, S.M. (2008). A rescue strategy for multimapping short sequence tags refines surveys of transcriptional activity by CAGE. Genomics *91*, 281–288.

Flicek, P., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., et al. (2012). Ensembl 2012. Nucleic Acids Res. *40*, D84–D90.

Geissler, R., Simkin, A., Floss, D., Patel, R., Fogarty, E.A., Scheller, J., and Grimson, A. (2016). A widespread sequence-specific mRNA decay pathway mediated by hnRNPs A1 and A2/B1. Genes Dev. *30*, 1070–1085.

Giannopoulou, E., Bartsakoulia, M., Tafrali, C., Kourakli, A., Poulas, K., Stavrou, E.F., Papachatzopoulou, A., Georgitsi, M., and Patrinos, G.P. (2012). A single nucleotide polymorphism in the HBBP1 gene in the human beta-globin locus is associated with a mild beta-thalassemia disease phenotype. Hemoglobin *36*, 433–445.

Glenfield, C., and McLysaght, A. (2018). Pseudogenes provide evolutionary evidence for the competitive endogenous RNA hypothesis. Mol. Biol. Evol. *35*, 2886–2899.

Graur, D., and Li, W.H. (2000). Fundamentals of Molecular Evolution (Sinauer).

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigó, R., and Sammeth, M. (2012). Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. *40*, 10073–10083.

GTEx Consortium (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.

Guo, X., Lin, M., Rockowitz, S., Lachman, H.M., and Zheng, D. (2014). Characterization of human pseudogene-derived non-coding RNAs for functional potential. PLoS One *9*, e93972.

Haeussler, M., Zweig, A.S., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Hinrichs, A.S., Gonzalez, J.N., et al. (2019). The UCSC Genome Browser database: 2019 update. Nucleic Acids Res. *47*, D853–D858.

Hardison, R.C. (2012). Evolution of hemoglobin and its genes. Cold Spring Harbor Perspect. Med. *2*, a011627.

Hart, T., Komori, H.K., LaMere, S., Podshivalova, K., and Salomon, D.R. (2013). Finding the active genes in deep RNA-seq gene expression studies. BMC Genomics *14*, 778.

Hattangadi, S.M., Wong, P., Zhang, L., Flygare, J., and Lodish, H.F. (2011). From stem cell to red cell: regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. Blood *118*, 6258–6268.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Huang, P., Keller, C.A., Giardine, B., Grevet, J.D., Davies, J.O.J., Hughes, J.R., Kurita, R., Nakamura, Y., Hardison, R.C., and Blobel, G.A. (2017). Comparative analysis of three-dimensional chromosomal architecture identifies a novel fetal hemoglobin regulatory element. Genes Dev. *31*, 1704–1713.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Ivaldi, M.S., Diaz, L.F., Chakalova, L., Lee, J., Krivega, I., and Dean, A. (2018). Fetal gamma-globin genes are regulated by the BGLT3 long noncoding RNA locus. Blood *132*, 1963–1973.

Jacq, C., Miller, J.R., and Brownlee, G.G. (1977). A pseudogene structure in 5S DNA of Xenopus laevis. Cell *12*, 109–120.

Ji, P., Murata-Hori, M., and Lodish, H.F. (2011). Formation of mammalian erythrocytes: chromatin condensation and enucleation. Trends Cell Biol. *21*, 409–415.

Kadota, K., Ye, J., Nakai, Y., Terada, T., and Shimizu, K. (2006). ROKU: a novel method for identification of tissue-specific genes. BMC Bioinformatics *7*, 294.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kaessmann, H., Vinckenbosch, N., and Long, M. (2009). RNA-based gene duplication: mechanistic and evolutionary insights. Nat. Rev. Genet. *10*, 19–31.

Kalyana-Sundaram, S., Kumar-Sinha, C., Shankar, S., Robinson, D.R., Wu, Y.M., Cao, X., Asangani, I.A., Kothari, V., Prensner, J.R., Lonigro, R.J., et al. (2012). Expressed pseudogenes in the transcriptional landscape of human cancers. Cell *149*, 1622–1634.

Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. Nature *478*, 483–489.

Karreth, F.A., Reschke, M., Ruocco, A., Ng, C., Chapuy, B., Léopold, V., Sjoberg, M., Keane, T.M., Verma, A., Ala, U., et al. (2015). The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. Cell *161*, 319–332.

Kasinathan, B., Colmenares, S.U., 3rd, McConnell, H., Young, J.M., Karpen, G.H., and Malik, H.S. (2020). Innovation of heterochromatin functions drives rapid evolution of essential ZAD-ZNF genes in Drosophila. eLife *9*, e63368.

Katoh, K., and Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. *9*, 286–298.

Kennedy, M., D'Souza, S.L., Lynch-Kattman, M., Schwantz, S., and Keller, G. (2007). Development of the hemangioblast defines the onset of hematopoiesis in human ES cell differentiation cultures. Blood *109*, 2679–2687.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Koop, B.F., Goodman, M., Xu, P., Chan, K., and Slightom, J.L. (1986). Primate η-globin DNA sequences and man's place among the great apes. Nature *319*, 234–238.

Kovalenko, T.F., and Patrushev, L.I. (2018). Pseudogenes as functionally significant elements of the genome. Biochemistry (Mosc) *83*, 1332–1349.

Goldstein, E.S., Krebs, J.E., and Kilpatrick, S.T. (2009). Lewin's Essential Genes (Jones and Bartlett Publishers).

Kumar, S., Stecher, G., Suleski, M., and Hedges, S.B. (2017). TimeTree: a resource for timelines, timetrees, and divergence times. Mol. Biol. Evol. *34*, 1812–1819.

Kurita, R., Suda, N., Sudo, K., Miharada, K., Hiroyama, T., Miyoshi, H., Tani, K., and Nakamura, Y. (2013). Establishment of immortalized human erythroid progenitor cell lines able to produce enucleated red blood cells. PLoS One *8*, e59890.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lasham, A., Herbert, M., Coppieters 't Wallant, N., Patel, R., Feng, S., Eszes, M., Cao, H., and Reid, G. (2010). A rapid and sensitive method to detect siRNA-mediated mRNA cleavage in vivo using 5′ RACE and a molecular beacon probe. Nucleic Acids Res. *38*, e19.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, J.H., Liu, S., Zhou, H., Qu, L.H., and Yang, J.H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. *42*, D92–D97.

Li, M.J., Liu, Z., Wang, P., Wong, M.P., Nelson, M.R., Kocher, J.P., Yeager, M., Sham, P.C., Chanock, S.J., Xia, Z., et al. (2016). GWASdb v2: an update database for human genetic variants identified by genome-wide association studies. Nucleic Acids Res. *44*, D869–D876.

Li, W., Yang, W., and Wang, X.J. (2013). Pseudogenes: pseudo or real functional elements? J. Genet. Genomics *40*, 171–177.

Liu, J., Li, Y., Tong, J., Gao, J., Guo, Q., Zhang, L., Wang, B., Zhao, H., Wang, H., Jiang, E., et al. (2018). Long non-coding RNA-dependent mechanism to regulate heme biosynthesis and erythrocyte development. Nat. Commun. *9*, 4386.

Liu, X., Chen, Y., Zhang, Y., Liu, Y., Liu, N., Botten, G.A., Cao, H., Orkin, S.H., Zhang, M.Q., and Xu, J. (2020). Multiplexed capture of spatial configuration and temporal dynamics of locus-specific 3D chromatin by biotinylated dCas9. Genome Biol. *21*, 59.

Liu, X., Zhang, Y., Chen, Y., Li, M., Zhou, F., Li, K., Cao, H., Ni, M., Liu, Y., Gu, Z., et al. (2017). In situ capture of chromatin interactions by biotinylated dCas9. Cell *170*, 1028–1043.e19.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science *290*, 1151–1155.

Mabaera, R., West, R.J., Conine, S.J., Macari, E.R., Boyd, C.D., Engman, C.A., and Lowrey, C.H. (2008). A cell stress signaling model of fetal hemoglobin induction: what doesn't kill red blood cells may make them stronger. Exp. Hematol. *36*, 1057–1072.

Nandakumar, S.K., Ulirsch, J.C., and Sankaran, V.G. (2016). Advances in understanding erythropoiesis: evolving perspectives. Br. J. Haematol. *173*, 206–218.

Nicholas, K.B., Nicholas, H.B., and Deerfield, D.W. (1997). GeneDoc: analysis and visualization of genetic variation. EMBNEW News *4*, 14.

Nuinoon, M., Makarasara, W., Mushiroda, T., Setianingsih, I., Wahidiyat, P.A., Sripichai, O., Kumasaka, N., Takahashi, A., Svasti, S., Munkongdee, T., et al. (2010). A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. Hum. Genet. *127*, 303–314.

O'Bleness, M., Searles, V.B., Varki, A., Gagneux, P., and Sikela, J.M. (2012). Evolution of genetic and genomic features unique to the human lineage. Nat. Rev. Genet. *13*, 853–866.

Ohno, S. (1970). Evolution by Gene Duplication (George Alien & Unwin Limited Company, Springer-Verlag).

Opazo, J.C., Hoffmann, F.G., and Storz, J.F. (2008). Differential loss of embryonic globin genes during the radiation of placental mammals. Proc. Natl. Acad. Sci. USA *105*, 12950–12955.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods *14*, 417–419.

Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X.J., Harte, R., Balasubramanian, S., Tanzer, A., Diekhans, M., et al. (2012). The GENCODE pseudogene resource. Genome Biol. *13*, R51.

Platt, O.S., Orkin, S.H., Dover, G., Beardsley, G.P., Miller, B., and Nathan, D.G. (1984). Hydroxyurea enhances fetal hemoglobin production in sickle cell anemia. J. Clin. Invest. *74*, 652–656.

Podlaha, O., and Zhang, J. (2010). Pseudogenes and their evolution. In eLS (John Wiley & Sons) http://umich.edu/~zhanglab/publications/2010/Podlaha_2010_ELS.pdf.

Poliseno, L. (2012). Pseudogenes: newly discovered players in human cancer. Sci. Signal. *5*, re5.

Poliseno, L., Marranci, A., and Pandolfi, P.P. (2015). Pseudogenes in human cancer. Front. Med. (Lausanne) *2*, 68.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W.J., and Pandolfi, P.P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. Nature *465*, 1033–1038.

Porcher, C., Swat, W., Rockwell, K., Fujiwara, Y., Alt, F.W., and Orkin, S.H. (1996). The T cell leukemia oncoprotein SCL/tal-1 is essential for development of all hematopoietic lineages. Cell *86*, 47–57.

Raleigh, D.R., Marchiando, A.M., Zhang, Y., Shen, L., Sasaki, H., Wang, Y., Long, M., and Turner, J.R. (2010). Tight junction-associated MARVEL proteins marveld3, tricellulin, and occludin have distinct but overlapping functions. Mol. Biol. Cell *21*, 1200–1213.

Rhead, B., Karolchik, D., Kuhn, R.M., Hinrichs, A.S., Zweig, A.S., Fujita, P.A., Diekhans, M., Smith, K.E., Rosenbloom, K.R., Raney, B.J., et al. (2010). The UCSC Genome Browser database: update 2010. Nucleic Acids Res. *38*, D613–D619.

Robb, L., Lyons, I., Li, R., Hartley, L., Köntgen, F., Harvey, R.P., Metcalf, D., and Begley, C.G. (1995). Absence of yolk sac hematopoiesis from mice with a targeted disruption of the scl gene. Proc. Natl. Acad. Sci. USA *92*, 7075–7079.

Rodriguez, J.M., Carro, A., Valencia, A., and Tress, M.L. (2015). APPRIS WebServer and WebServices. Nucleic Acids Res. *43*, W455–W459.

Ross, B.D., Rosin, L., Thomae, A.W., Hiatt, M.A., Vermaak, D., de la Cruz, A.F.A., Imhof, A., Mellone, B.G., and Malik, H.S. (2013). Stepwise evolution of essential centromere function in a Drosophila Neogene. Science *340*, 1211–1214.

Sankaran, V.G., Menne, T.F., Xu, J., Akie, T.E., Lettre, G., Van Handel, B., Mikkola, H.K., Hirschhorn, J.N., Cantor, A.B., and Orkin, S.H. (2008). Human fetal hemoglobin expression is regulated by the developmental stage-specific repressor BCL11A. Science *322*, 1839–1842.

Schulz, H., Ruppert, A.K., Herms, S., Wolf, C., Mirza-Schreiber, N., Stegle, O., Czamara, D., Forstner, A.J., Sivalingam, S., Schoch, S., et al. (2017). Genome-wide mapping of genetic determinants influencing DNA methylation and gene expression in human hippocampus. Nat. Commun. *8*, 1511.

Shao, Y., Chen, C., Shen, H., He, B.Z., Yu, D., Jiang, S., Zhao, S., Gao, Z., Zhu, Z., Chen, X., et al. (2019). GenTree, an integrated resource for analyzing the evolution and function of primate-specific coding genes. Genome Res. *29*, 682–696.

Shi, L., Cui, S., Engel, J.D., and Tanabe, O. (2013). Lysine-specific demethylase 1 is a therapeutic target for fetal hemoglobin induction. Nat. Med. *19*, 291–294.

Shi, L., Lin, Y.H., Sierant, M.C., Zhu, F., Cui, S., Guan, Y., Sartor, M.A., Tanabe, O., Lim, K.C., and Engel, J.D. (2014). Developmental transcriptome analysis of human erythropoiesis. Hum. Mol. Genet. *23*, 4528–4542.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

Storz, J.F. (2016). Gene duplication and evolutionary innovations in hemoglobin-oxygen transport. Physiology *31*, 223–232.

Tam, O.H., Aravin, A.A., Stein, P., Girard, A., Murchison, E.P., Cheloufi, S., Hodges, E., Anger, M., Sachidanandam, R., Schultz, R.M., et al. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. Nature *453*, 534–538.

Tan, S.J., Cardoso-Moreira, M., Shi, W.W., Zhang, D., Huang, J.W., Mao, Y.A., Jia, H.X., Zhang, Y.Q., Chen, C.Y., Shao, Y., et al. (2016). LTR-mediated retroposition as a mechanism of RNA-based duplication in metazoans. Genome Res. *26*, 1663–1675.

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

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

Wang, G.D., Zhai, W., Yang, H.C., Wang, L., Zhong, L., Liu, Y.H., Fan, R.X., Yin, T.T., Zhu, C.L., Poyarkov, A.D., et al. (2016). Out of southern east Asia: the natural history of domestic dogs across the world. Cell Res. *26*, 21–33.

Wang, L., Wang, S., and Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics *28*, 2184–2185.

Wienert, B., Funnell, A.P., Norton, L.J., Pearson, R.C., Wilkinson-White, L.E., Lester, K., Vadolas, J., Porteus, M.H., Matthews, J.M., Quinlan, K.G., et al. (2015). Editing the genome to introduce a beneficial naturally occurring mutation associated with increased fetal globin. Nat. Commun. *6*, 7085.

Xu, J., Peng, C., Sankaran, V.G., Shao, Z., Esrick, E.B., Chong, B.G., Ippolito, G.C., Fujiwara, Y., Ebert, B.L., Tucker, P.W., et al. (2011). Correction of sickle cell disease in adult mice by interference with fetal hemoglobin silencing. Science *334*, 993–996.

Xu, J., and Zhang, J. (2016). Are human translated pseudogenes functional? Mol. Biol. Evol. *33*, 755–760.

Yang, H., He, B.Z., Ma, H., Tsaur, S.C., Ma, C., Wu, Y., Ting, C.T., and Zhang, Y.E. (2015). Expression profile and gene age jointly shaped the genome-wide distribution of premature termination codons in a Drosophila melanogaster population. Mol. Biol. Evol. *32*, 216–228.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Zhang, H., Meltzer, P., and Davis, S. (2013). RCircos: an R package for Circos 2D track plots. BMC Bioinformatics *14*, 244.

Zhang, J.Z. (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. *18*, 292–298.

Zhang, Y.E., Landback, P., Vibranovski, M.D., and Long, M. (2011). Accelerated recruitment of new brain development genes into the human genome. PLoS Biol *9*, e1001179.

Zhang, Y.E., and Long, M. (2014). New genes contribute to genetic and phenotypic novelties in human evolution. Curr. Opin. Genet. Dev. *29*, 90–96.

Zhang, Z., and Zheng, D. (2008). Pseudogene evolution in the human genome. In eLS (John Wiley & Sons). https://doi.org/10.1002/9780470015902.a0020836.

Zhao, T.T., Graber, T.E., Jordan, L.E., Cloutier, M., Lewis, S.M., Goulet, I., Côté, J., and Holcik, M. (2009). hnRNP A1 regulates UV-induced NF-kappaB signalling through destabilization of cIAP1 mRNA. Cell Death Differ. *16*, 244–252.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD71	eBioscience	Cat#17-0719-42; RRID: AB_10671393
CD235a	eBioscience	Cat#12-9987-82; RRID: AB_466300
CD34	eBioscience	Cat#25-0349-42; RRID: AB_1963576
CD123	BD	Cat#562517; RRID: AB_11153668
CD36	BD	Cat#561536; RRID: AB_10893348
mCD45	BioLegned	Cat#103116; RRID: AB_312981
hCD45	BD	Cat#555483; RRID: AB_2649445
HNRNPA2B1	Abcam	Cat#ab6102; RRID: AB_305293
HNRNPA1	Abcam	Cat#ab5832; RRID: AB_305145
AGO2	Proteintech	Cat#10686-1-AP; RRID: AB_2096299
ILF2	Proteintech	Cat#14714-1-AP; RRID: AB_2280445
ADAR1	Proteintech	Cat#14330-1-AP; RRID: AB_2273600
ILF3	Proteintech	Cat#19887-1-AP; RRID: AB_10666431
DDX1	Proteintech	Cat#11357-1-AP; RRID: AB_2092222
HNRNPD	Proteintech	Cat#12770-1-AP; RRID: AB_2117318
SFPQ	Proteintech	Cat#15585-1-AP; RRID: AB_10697653
DHX9	Proteintech	Cat#17721-1-AP; RRID: AB_2092506
HNRNPU	Proteintech	Cat#16365-1-AP; RRID: AB_1059648
PTBP1	Proteintech	Cat#12582-1-AP; RRID: AB_2176402
TAL1	Abcam	Cat#ab155195 N/A
GAPDH	Abcam	Cat#60004-1-Ig; RRID: AB_2107436
Goat anti-mouse Alexa Fluor 594	Abcam	Cat#ab150116; RRID: AB_2650601
Chemicals, peptides, and recombinant proteins		
DMEM	Gibco	Cat#C11995500BT
F12	Gibco	Cat#C11330500BT
nonessential amino acids (NEAA)	Gibco	Cat#11140-050

(*Continued on next page*)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
trypsin-EDTA	Gibco	Cat#25200072
β-mercaptoethanol	Sigma	Cat#M7522
hbFGF	R&D Systems	Cat#100-18B
Matrigel	Corning	Cat#356234
collagenase B	Worthington	Cat#LS004176
StemPro-34	Invitrogen	Cat#10639011
BMP-4	R&D Systems	Cat#120-05ET
Glutamax	Sigma	Cat#35050061
monothioglycerol (MTG)	Sigma	Cat#M1753
ascorbic acid	Sigma	Cat#A8960
IMDM	Gibco	Cat#I3390
knockout serum replacement (KSR)	Gibco	Cat#10828028
human holo transferrin	Sigma	Cat#T4132
recombinant human insulin	Sigma	Cat#I6634
Inositol	Sigma	Cat#I5125
Folic acid	Sigma	Cat#F7876
Ferric nitrate	Sigma	Cat#F8508
Ferrous sulfate	Sigma	Cat#F8633
hSCF	PeproTech	Cat#300-07
hIL-3	PeproTech	Cat#200-03
hIL-6	PeproTech	Cat#200-06
hIL-11	PeproTech	Cat#200-11
hTPO	PeproTech	Cat#300-18
hVEGF	PeproTech	Cat#100-44
hEPO	PeproTech	Cat#100-64
semisolid medium	Stem Cell Technologies	Cat#04435
HISTOPAQUE®	Sigma	Cat#10771
BIT	Stem Cell Technologies	Cat#09500
penicillin and streptomycin	Gibco	Cat#15140-122
Hydrocortisone	Sigma	Cat#H2270
StemSpan SFEM	Stem Cell Technologies	Cat#09650
FLT3 ligand	PeproTech	Cat#300-19
doxycycline	Sigma	Cat#9891
dexamethasone	Sigma	Cat#9891D2915
TRIzol	Invitrogen	Cat#15596026
Triton X-100	Sigma	Cat#A110694
DAPI	Sigma	Cat#DUO82040
benzidine	Sigma	Cat#D9143
Protease Inhibitor cocktail	Roche	Cat#S8830
HNRNPA1	Abcam	Cat#ab123212
ECL substrate	Millipore	Cat#WBKLS0500
SuperScript III	Thermo Fisher	Cat#18080044
TIANamp Genomic DNA Kit	TIANGEN	Cat# DP304
Critical commercial assays		
Anti-PE Multisort Kit	MACS	Cat#130-090-757
CD235a (Glycophorin A) MicroBeads	MACS	Cat#130-050-501
EasySep™ Human CD34 Positive Selection Kit	Stem Cell Technologies	Cat#17856
Human MTC Panel I	TaKaRa	Cat#636742
Human MTC Panel II	TaKaRa	Cat#636743

(*Continued on next page*)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
CD34 microbeads	Miltenyi Biotec	Cat#130-046-702
HNRNPA1 human shRNA plasmid kit	Origene	Cat#TL312381
QuikChange Site-Directed Mutagenesis Kit	Stratagene,	Cat#200518
ViraPower Lentiviral Packaging System	Invitrogen	Cat#K497500
SYBR Green PCR kit	Applied Biosystems	Cat#A25742
SMARTer RACE 5'/3' Kit	TaKaRa	Cat#634858
RNAscope Fluorescent Multiplex Kit	Advanced Cell Diagnostics	Cat#320293
NE-PER Nuclear and Cytoplasmic extraction Reagents Kit	Thermo Fisher Scientific	Cat#78835
RiboMAX large scale RNA production systems-T7	Promega	Cat#P1300
Pierce RNA 3' End Desthiobiotinylation Kit	Thermo Scientific	Cat#20163
Pierce Magnetic RNA-protein Pull-Down Kit	Thermo Scientific	Cat#20164
Streptavidin agarose beads	Invitrogen	Cat#65601
Light Shift Chemiluminescent RNA EMSA Kit	Pierce	Cat#20158
protein A beads	Invitrogen	Cat#10002D
Deposited data		
Raw and analyzed data	This paper	GEO: GSE136235
Experimental models: cell lines		
H1 hESC	NIH	WA01
HUDEP-2	RIKEN BioResource Research Center	N/A
Experimental models: organisms/strains		
NOG mice	NOD.Cg-PrkdcscidIl2rgtm1Sug/JicCrl	N/A
Oligonucleotides		
FISH Probe Hs-HBBP1	Advanced Cell Diagnostics	REF: 496481
FISH Probe Hs-TAL1	Advanced Cell Diagnostics	REF: 510091
FISH Probe 3-Plex Negative Control	Advanced Cell Diagnostics	REF: 320871
FISH Probe 3-Plex Positive Control	Advanced Cell Diagnostics	REF: 320861
Recombinant DNA		
pX330-sgRNA	Addgene	Cat#42230
lentiCRISPR v2	Addgene	Cat#52961
pLKO.1	Addgene	Cat#10878
Software and algorithms		
GraphPad Software	GraphPad Software, Inc	https://www.graphpad.com
FlowJo	BD Biosciences, San Jose, CA, USAFlowJo	https://www.flowjo.com
Chromas	Technelysium, Inc	http://technelysium.com.au/wp/chromas/
STAR	Dobin et al., 2013	https://www.elsevier.com/authors/author-resources/research-data/data-base-linking
RSEM	Li and Dewey, 2011	https://github.com/deweylab/RSEM
Kallisto	Bray et al., 2016	https://pachterlab.github.io/kallisto/download.html
RSeQC	Wang et al., 2012	http://rseqc.sourceforge.net/#download-rseqc
WGCNA	Langfelder and Horvath, 2008	http://bioconductor.org/biocLite.R
DAVID	Huang da et al., 2009	https://david.ncifcrf.gov/
DEGseq	Raleigh et al., 2010	http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html
Salmon	Patro et al., 2017	https://github.com/COMBINE-lab/salmon/

(*Continued on next page*)

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Cufflinks	Trapnell et al., 2012	http://cole-trapnell-lab.github.io/cufflinks/install/
TopHat	Trapnell et al., 2009	http://ccb.jhu.edu/software/tophat/index.shtml
Rcircos	Zhang et al., 2013	https://cran.r-project.org/web/packages/RCircos/index.html
MAFFT	Katoh and Toh, 2008	http://mafft.cbrc.jp/alignment/software/mafft-7.245-gcc_fc6.x86_64.rpm
GeneDoc	Nicholas et al., 1997	http://nrbsc.org/gfx/genedoc/index.html
jModelTest	Darriba et al., 2012	https://github.com/ddarriba/jmodeltest2
PAML	Yang, 2007	http://abacus.gene.ucl.ac.uk/software/#downloads-and-installation
RAxML	Stamatakis, 2014	https://codeload.github.com/stamatak/standard-RAxML/zip/master
GenTree	Shao et al., 2019	http://gentree.ioz.ac.cn
sgRNA	Cong et al., 2013	http://crispr.mit.edu

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jia Yu (j-yu@ibms.pumc.edu.cn).

Materials availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

Data and code availability

The datasets generated during this study are available at GEO database (GSE136235). Since most codes used in this study are quite straightforward, we only upload relatively complex codes for coexpression network analyses and evaluation of RNA-seq quantification software. The URL is https://github.com/Zhanglab-IOZ/HBBP1-project.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

hESC Cells

The H1 hESC line (NIH code WA01), obtained from the WiCell Research Institute, was maintained on irradiated MEF (mouse embryonic fibroblasts) feeder cells in the hESC media consisting of DMEM/F12 (1/1; Gibco, Cat#C11995500BT, Cat#C11330500BT), 20% knockout serum replacement (KSR, Gibco, Cat#10828028), 100 μM nonessential amino acids (NEAA, Gibco, Cat#11140-050), 2 mM glutamax (Sigma, Cat#35050061), 100 μM β-mercaptoethanol (Sigma, Cat#M7522), and 40 ng/mL hbFGF (R&D Systems, Cat#100-18B) in 6-well tissue culture plates. Cells were fed with fresh hESC medium daily and passaged to new feeders, which was followed with dissociation via 0.25% trypsin-EDTA (Gibco, Cat#25200072).

Human hematopoietic stem/progenitor CD34$^+$ cells (HSPCs)

Human umbilical cord blood and bone marrow samples were isolated by HISTOPAQUE® (Sigma, Cat#10771) density gradient centrifugation. Mononuclear cells (MNC) were passed through CD34 microbeads (Miltenyi Biotec, Cat#130-046-702) for further purification.

The purified CD34$^+$ cells were differentiated *ex vivo* toward the erythroid lineage as described previously with minor modifications (Shi et al., 2013). The CD34$^+$ cells were cultured in the basic Iscove's modified Dulbecco's medium (IMDM) supplemented with 20% BIT (v/v) (Stem Cell Technologies, Cat#09500), 40 μg/ml inositol (Sigma), 10 μg/ml folic acid (Sigma), 160 μM monothioglycerol (Sigma), 90 ng/ml ferrous nitrate (Sigma), 900 ng/ml ferrous sulfate (Sigma), 2 mM L-glutamine and 1% penicillin and streptomycin (v/v) (Gibco, Cat#15140-122). From day 0 to 8, 5 ng/ml hIL-3, 100 ng/ml hSCF, 1 μM hydrocortisone (Sigma) and 3 IU/ml hEPO was added into the culture medium. From day 8 to 14, hSCF and hEPO were included in the culture medium and only hEPO was added to the culture during day 14 to 18. Cells were maintained at a density between 5×10^4 and 2×10^6 cells/ml, with medium change every 3 to 4 days. Cells were incubated at 37 °C with 5% CO_2.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

HUDEP-2 cells

Human umbilical cord blood CD34$^+$-derived erythroid progenitor cell (HUDEP-2) was obtained from RIKEN BioResource Research Center. HUDEP-2 cells were cultured in StemSpan SFEM® medium (Stem Cell Technologies, Cat#09650) in the presence of hSCF (50 ng/ml), hEPO (3 IU/ml), doxycycline (1 μg/ml) (Sigma, Cat# 9891), dexamethasone (10^{-6} M) (Sigma, Cat# 9891D2915) and 1% penicillin and streptomycin.

Human and chimpanzee samples

Human umbilical cord blood and bone marrow were obtained from The Third Affiliated Hospital of Chongqing Medical University (Chongqing, China) and The 303 Hospital (Guangxi, China). Peripheral blood samples of 5 healthy human donors and 303 β-thalassemia patients were collected from The 303 Hospital (Guangxi, China). Five peripheral blood samples of chimpanzees were collected from Beijing Zoo and Nanjing Hongshan Forest Zoo. Human informed consent was obtained via the Ethical Committee at each hospital; inclusion of chimpanzee samples was approved by each institution's Animal Care and Use Committee or equivalent.

Mononuclear cells (MNCs) were isolated from umbilical cord blood and bone marrow by density gradient [d=1.077 g/ml] (Sigma). CD71$^+$ cells were enriched from MNCs through positive immune-magnetic selection (PE-conjugated CD71 primary antibody, BD, 555537; Anti-PE Multisort Kit, MACS, 130-090-757). CD71$^+$/CD235$^+$ cells were enriched from CD71$^+$ cells using CD235a (Glycophorin A) MicroBeads (MACS, 130-050-501). Human tissues cDNA were derived from Human MTC Panel I and II (TaKaRa, 636742, 636743).

METHOD DETAILS

Gene annotation and reannotation

For mammals covered in this study, we downloaded Ensembl v91 annotation files (released in Dec. 2017) (Flicek et al., 2012). Unless otherwise noted, we took longest principle transcript (Rodriguez et al., 2015) to represent each coding gene.

In human, we classified genes into lincRNAs, duplicated pseudogenes (unprocessed pseudogenes), retropseudogenes (processed pseudogenes) and protein coding genes. We further downloaded 13,598 pseudogenes listed in the often-used database (http://www.pseudogene.org/Human/Human90.txt), whose corresponding paralogous protein-coding genes were annotated (Pei et al., 2012). To be conservative, we extracted 12,503 entries that were also annotated as retropseudogenes or duplicated pseudogenes by Ensembl 91 and used these pseudogene/coding-gene pairs in the downstream analyses.

We noticed that the *HBBP1* locus was poorly annotated for most mammals. To enable cross-species analyses, we reannotated these loci. Specifically, we re-examined the phylogenetic distribution of *HBBP1* along the mammalian tree based on two previous work (Hardison, 2012; Opazo et al., 2008) and the UCSC genome browser. With the UCSC syntenic genomic alignment, we confirmed a truncated version of *HBBP1* in armadillo (Opazo et al., 2008) and rejected the gene loss scenario reported in (Hardison, 2012). Analogously, we also corrected that bushbaby lost the third exon (Hardison, 2012). After that, we downloaded the RefSeq sequence of *HBBP1* in human from the UCSC genome browser since the RACE experiments better supported RefSeq gene model compared to Ensembl gene models. Ensembl annotates *HBBP1*'s orthologs in dog (*ENSCAFG00000024181*). For all other species (e.g. house), *HBBP1* is unannotated. In these cases, we extracted orthologous regions guided by the UCSC (Haeussler et al., 2019) syntenic genomic alignment (Net track) of mammals and predicted the gene structure based on human *HBBP1* (Table S5).

Transcriptome (RNA-seq) analyses

To perform a genome-wide transcriptional profiling of pseudogenes, investigate the expression level of HBBP1 across mammals, and analyze pseudogene/coding-gene pair correlation, we took advantage of multiple publicly available RNA-seq datasets including our own previously generated datasets (Table S1). If there are multiple similar datasets available, we chose the dataset with higher quality in terms of tissues of interest, sample number, strandness and read length. For HPA data, we used the previous quantification results stored in the GenTree database (Shao et al., 2019), which is based on alignments generated by STAR (v2.4.0k) and subsequent RSEM (v1.2.19) quantification. For the other datasets, we downloaded the raw sequencing data and performed our own quantification via Kallisto, which is able to differentiate similar paralogs (Bray et al., 2016). We took the default parameters except that the read length and strandness were specified according to description of each individual dataset. Herein, if strandness is unknown (Table S1), we estimated the strandness with RSeQC (Wang et al., 2012). We generated gene-level quantification results by summing up transcript-level results calculated by Kallisto.

Specifically, for human, we used multiple distinct RNA-seq datasets for different purposes (Table S1). First, by following (Shao et al., 2019), we implemented WGCNA v1.68 (Langfelder and Horvath, 2008) and performed the coexpression network analysis based on the HPA transcriptome data. WGCNA emphasize the stronger correlations by raising the pairwise between-gene expressional correlations to a power β (≥ 1). With these transformed correlations, WGCNA builds coexpression modules by performing hierarchical clustering (cutting branches off the dendrogram, Figure S2A) and genes in the same module often share similar functionality. The parameters are largely similar to (Shao et al., 2019) except that we increased the coefficient of variation cutoff to 0.12 to remove genes less variable across tissues (Kang et al., 2011) and chose the soft threshold as 7 to better fit the data (the scale-free R^2>0.85). Besides the CV cutoff, we also specified the expression cutoff, i.e., FPKM>0.5 (Hart et al., 2013), with FPKM as fragments per kilobase of transcript sequence per millions base pairs mapped. 29,122 coding and noncoding genes were included in the subsequent coexpression analyses. We identified 16 coexpression modules. For 8 tissue-specific modules, we ran GO enrichment

analysis via DAVID (Huang da et al., 2009) to estimate whether genes included in these modules play functions related with the tissue of interest and generally found consistent patterns. Second, RNA-seq datasets of 13 human hematopoietic populations were download from GEO database (GSE74912), which were reprocessed by Kallisto. Both Shannon entropy and Akaike's Information Criterion (AIC) were used to identify lineage specifically expressed pseudogenes: shannon entropy after data processing was used to rank various lineage-specific patterns, and the minimum AIC estimate (MAICE) was used to identify outliers (Kadota et al., 2006). For each lineage, up-regulated outliers were deemed as lineage-specific pseudogenes. Pearson Correlation Coefficient was used to quantify the correlation between lineage-specific pseudogenes and lineage master transcription factors. Third, to examine the temporal regulation of *HBBP1* during human erythropoiesis, we re-analyzed our two previous published datasets (Liu et al., 2018; Shi et al., 2014) with Kallisto. For SRP124244, we only used two subsets, SRX3367085 and SRX3367086 for which no specific genes are disturbed and thus the cells approximate normal erythropoiesis. Forth, we also used these two datasets to examine the transcriptional intensity of *HBBP1* relative to other genes. Finally, to examine the correlation between HBBP1 and TAL1 in human blood and the correlation between pseudogenes and their corresponding coding genes in various tissues, we downloaded GTEx v7 gene-level quantification dataset (transcripts per million mapped reads, TPM). It represents the largest human transcriptome dataset so far (GTEx Consortium et al., 2015). To control the heterogeneity, we focused on 1,862 high-quality (RNA integrity value not lower than 7) GTEx samples, which were derived from Europeans with an age between 20 and 50 years old. For 27 tissues with at least 10 samples, we followed (Glenfield and McLysaght, 2018) and calculated the correlation between pseudogenes and coding genes. In brief, we focused on 6,245 pseudogene and coding gene pairs for both of which median TPM are greater than 0.1 in at least one tissue, and calculated Spearman ρ for each tissue.

For non-human mammals, we downloaded blood transcriptomes (Table S1) covering chimpanzee, rhesus monkey, wolf (dog), horse, and pig. Bushbaby and armadillo were not included due to the truncation of *HBBP1*. For chimpanzee, only blood related datasets (blood, tissue admixture including blood, bone marrow) were re-analyzed. For rhesus monkey and pig, the experiments were designed to test the response of blood transcriptome after infection. So, we only used the normal control samples before the infection. The reason that we took a dataset (SRP073500) of wolf is that we didn't find any blood RNA-seq dataset of dog. However, since they are essentially one species sharing genetic flow (Wang et al., 2016), we used dog annotation subsequently. Since SRP073500 consists of both paired-end and single-end RNA-seq data, we only used the former type of data given its accuracy in quantifying the expression of paralogs. Moreover, for this wolf dataset, three samples (SRR3402521, SRR3402522, SRR3402527) are strongly diverged from all the other samples in terms of expression correlation of all genes and thus we discarded these samples in subsequent analyses.

In this reanalysis process, we noticed another two technical issues. First, GTEx has its own quantification pipeline, which is not based on Kallisto. To make the results of GTEx comparable with other mammals, we randomly picked 50 blood samples out of 1,862 samples and performed quantification with Kallisto. Second, for blood transcriptomes, globin-depletion was sometimes performed since the globin transcripts are extraordinarily abundant and limit the detection power of other transcripts. Thus, for blood transcriptomes in GTEx or other datasets, we examined whether globin-depletion was performed and whether this procedure affected the quantification of *HBBP1* (Table S1). We only found that dedicated experiments were taken to deplete globin transcripts in pig and wolf. In pig, antisense oligonucleotides were designed, which specifically targeted *HBA* and *HBB* transcripts. In wolf, Globin-Zero^{TM} kit was added, which was actually designed for human (*HBA*, *HBB*, *HBG*) and mouse (*HBA*, *HBB*). Thus, given the divergence (21.2%) between *HBG1* (the most closely related paralog of *HBBP1*) and *HBBP1*, we conclude that the quantification of *HBBP1* is less likely to be affected for both species.

For transcriptome analysis with *HBBP1* repression, differentially expressed genes were identified with DEGseq (Raleigh et al., 2010). Specifically, for the cumulative distribution plot, ''induced genes'' represents up-regulated genes in EPO induction day 14 compared with day 8. Gene expression is quantified as fragments per kilobase of transcript sequence per millions base pairs mapped (FPKM, Table S3).

For half-life analyses, a more conservative dataset of genes were selected given the following criterion: 1) genes up-regulated (fold change $\geq$ 1.5) in control group at erythroid differentiation day 14 compared to day 8, but down-regulated (fold change $\geq$ 1.5) when HBBP1 was knocked down (in both shHBBP1-1 and shHBBP1-2) compared to shCtrl for either day 8 or day 14; 2) genes down-regulated (fold change $\geq$ 1.5) in control group at erythroid differentiation day 14 compared to day 8, but up-regulated (fold change $\geq$ 1.5) when HBBP1 was knocked down (in both shHBBP1-1 and shHBBP1-2) compared to shCtrl for either day 8 or day 14; 3) the maximum expression of the genes in the above samples was greater than 20. Finally, 40 genes were selected to examine their half-lives with HBBP1 knockdown.

Note, mapping uncertainty is a notorious issue when studying pseudogenes or other types of duplicated genes and numerous strategies have been developed (Bray et al., 2016; Cheetham et al., 2020; Faulkner et al., 2008; Zhang et al., 2011). Thus, one key reason that we relied on STAR/RSEM and Kallisto is their reported accuracy in quantifying expression of similar paralogs (Bray et al., 2016). To strengthen this conclusion, we performed an independent evaluation including Kallisto, the STAR (Dobin et al., 2013) /RSEM (Li and Dewey, 2011) pipeline together with another two popular methods including Salmon (Patro et al., 2017) and STAR/Cufflinks (Trapnell et al., 2012) pipeline. Specifically, we extracted 882 Ensembl v75 (Flicek et al., 2012) annotated coding paralogs with a synonymous substitution rate (K_S) ranging between 0 to 0.04. Since the non-synonymous substitution rate (K_A) is generally much smaller than K_S (Graur and Li, 2000), this cutoff ensures that the mRNA-level identity is higher than 96%. We simulated RNA-seq reads derived from these coding paralogs with flux-simulator (v 1.2.1) (Griebel et al., 2012), which incorporated known sequencing bias. Four simulation runs with different sequencing depth (20 or 30 million), read property (paired-end or single-end) or read length

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

(100 or 150 bp) were implemented. We ran all software with default parameters excepting for Cufflinks, wherein a parameter '-u' is added to achieve a "more accurate" result as described in the manual. Herein, RSEM and Cufflinks were used for quantification based on mapping data generated by STAR, while both Salmon and Kallisto take mapping-free strategies. The reason that we used STAR rather than the conventional TopHat for Cufflinks is simply due to its high speed and accuracy relative to TopHat. After quantification, we then examined whether the differential expression (fold changes) between paralogs could be accurately quantified with four methods. Overall, all four methods reached a decent performance with the Pearson correlation coefficient between actual values and inferred values reaching 0.71∼0.96 and their performance generally increase with high-depth longer paired-end data (Figures S1A–S1D). Kallisto, Salmon and STAR/RSEM show similar performance (0.83∼0.96), while TopHat/Cufflinks has a relatively inferior performance (0.71∼0.86). We expect that the actual performance of Kallisto and STAR/RSEM is even better since majority (97%) of 6,245 pseudogene pairs analyzed in this work have an identity lower than 96% (Figure S1E).

For our focal gene, i.e. *HBBP1*, we further examined its identity relative to its paralogous β-globin family members (Table S2). Both mRNA-level and protein-level identity is limited (72-77%, 65-73%, respectively). For a short peptide of seven amino acids, two amino acid differences are expected even for its most closely related paralogs, *HBG1* or *HBG2*. Thus, transcriptional level (RNA-seq, >=100 bp read length) and proteomic level analyses would not be confounded. We also confirmed that all primers, shRNAs or sgRNAs specifically target *HBBP1* with at least 6 nucleotides differences relative to other paralogs.

GWAS data analyses

We began with gene-SNP association data stored in our previously developed GenTree database (Shao et al., 2019). In brief, we downloaded the GWAS records from GWASdb2 (Li et al., 2016) and excluded outlier studies with too many (>130) SNPs. To account for similarity between pseudogenes and coding paralogs, we further removed SNPs with the upstream and downstream 50 bp window (101 bp in total) equally mapped to multiple positions in the genome.

We further retained pseudogenes with SNPs in the flanking 10 Kb region meeting one of the following two criterions: 1) more than one GWAS SNPs link the gene of interest to the same disease ontology term (Li et al., 2016); 2) the gene has a high-confidence SNP-trait association ($P < 10^{-5}$) (Schulz et al., 2017).

We counted pseudogenes and the nearby SNPs in M10, and plotted their chromosomal distribution with Rcircos (Zhang et al., 2013).

hESC based erythroid differentiation

Hematopoietic differentiation of hESCs were carried out as previously described (Kennedy et al., 2007). In brief, hESCs maintained on MEF were transferred to Matrigel (Corning, Cat#356234) treated plate and cultured in hESC media for 24 hours before hematopoietic and erythroid differentiation. Then hESCs were digested with collagenase B (1 mg/mL; Worthington, Cat#LS004176) at 37 °C for 20 minutes followed by trypsin-EDTA (0.05%) treatment for 2 minutes to get small clusters (10-20 cells) for EB (Embryoid Body) formation. The clusters were re-plated in 6-well ultralow cluster plates (Corning) in 2 mL aggregation media per well, using StemPro-34 (Invitrogen, Cat#10639011) as basic media supplemented with 10 ng/mL BMP-4 (R&D Systems, Cat#120-05ET), 2 mM glutamax, 0.4 mM monothioglycerol (MTG, Sigma, Cat#M1753), and 50 μg/mL ascorbic acid (Sigma, Cat#A8960). The cells were maintained for 24 hours at 37 °C in 5% CO_2, 5% O_2, and 90% N_2. Then, 1 mL of the media was carefully removed and 2 mL fresh aggregation media supplemented with 7.5 ng/mL hbFGF was added to reach a final concentration of 5 ng/mL hbFGF (phase I media). The EBs were incubated for 72 hours and changed to phase II media, which consisted of phase I media supplemented with growth factors including hSCF (100 ng/mL, PeproTech, Cat#300-07), hIL-3 (20 ng/mL, PeproTech, Cat#200-03), hIL-6 (10 ng/mL, PeproTech, Cat#200-06), hIL-11 (5 ng/mL, PeproTech, Cat#200-11), hVEGF (5 ng/mL, PeproTech, Cat#100-44), and hEPO (3 IU/ml, PeproTech, Cat#100-64). The aggregates were then cultured for additional 4 days and the differentiation efficiency was determined on day 8 by flow cytometry.

The EBs were harvested on hematopoietic differentiation day 8 and treated with 0.25% trypsin-EDTA at 37 °C for 7 minutes. Equal volume of serum-containing media was added to stop trypsinization. Then the EBs were passaged 6 times through a 20-gauge needle followed by passing through cell strainer to dissociated to single cells. $CD34^+$ cells were purified using EasySep™ Human CD34 Positive Selection Kit (Stem Cell Technologies, Cat#17856) according to the manufacturer's manual. $CD34^+$ cells (93%±4% purity) were induced to erythroid differentiation using erythroid differentiation medium on the basis of IMDM (Gibco, Cat#I3390) supplemented with 20% knockout serum replacement (KSR, Gibco), 200 μg/mL human holo transferrin (Sigma, Cat#T4132), 10 μg/mL recombinant human insulin (Sigma, Cat#I6634), 40 μg/mL Inositol (Sigma, Cat#I5125), 10 μg/mL Folic acid (Sigma, Cat#F7876), 90 ng/mL Ferric nitrate (Sigma, Cat#F8508), 900 ng/mL Ferrous sulfate (Sigma, Cat#F8633), 0.4 mM monothioglycerol (MTG, Sigma), and 50 μg/mL ascorbic acid (Sigma), and 2 mM glutamax (Gibco).

In Phase I erythroid differentiation (day 8 to 12), $CD34^+$ cells were cultured in erythroid differentiation medium at 2×10^5 cells/mL in the presence of 1mM Hydrocortisone (Sigma, Cat#H2270) and growth factors (all purchased from PeproTech) including hSCF (100 ng/mL), hIL-3 (10 ng/mL), hIL-6 (10 ng/mL), hIL-11 (5 ng/mL), hTPO (5 ng/mL, PeproTech, Cat#300-18), and hEPO (3 IU/mL). Four days later, cells were cultured in Phase II erythroid differentiation medium containing hSCF (100 ng/mL), hIL-3 (20 ng/mL), hEPO (3 IU/mL) for 6 days (day 12 to day18). On day 18, the cells were re-suspended in Phase III medium supplemented with hSCF (100 ng/mL) and hEPO (3 IU/mL) and cultured for 6 days (day 18 to 24). For the last stage of differentiation, cells were maintained in Phase IV medium containing only EPO (3 IU/mL) for 7 days. The surface antigens including CD71 and CD235a were used to evaluate the erythroid differentiation by flow cytometry and samples were harvested at different time points for qRT-PCR.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

For the formation of BFU-E, $CD34^+$ cells (2000 cells/mL) were plated in triplicate in semisolid medium containing 1% methylcellulose (Stem Cell Technologies, Cat#04435) in IMDM supplemented with 10 μg/mL recombinant human insulin (Sigma), 200 μg/mL human holo-transferrin, 1×10^{-4} M 2-mercaptoethanol (β-ME, Sigma), glutamax (2 mM), hSCF (100 ng/mL), hIL-3 (10 ng/mL), hIL-6 (10 ng/mL), and hEPO (3 IU/mL). Cells were maintained in humidified air with 5% CO_2 and the BFU-E colonies were counted after 14 days under the light microscope.

CRISPR/Cas9 mediated knockout and mutation

The single guide RNA (sgRNA) sequences corresponding to *HBBP1* or transcription start site (TSS) of *HBBP1* were chosen to minimize the likelihood of off-target cleavage based on the publicly available online tool (http://crispr.mit.edu/). For *HBBP1* knockout, the annealed oligos were cloned into pX330. A pair of pX330-sgRNA plasmids and a puromycin-selection plasmid were co-electroporated into H1 hESCs. After 48 hours of puromycin-selection, electroporated cells were trypsinized and re-plated, hESC colonies were picked up in the 7th day. For *HBBP1* TSS knockout, the annealed oligos were cloned into lentiCRISPR v2. A pair of lentivirus were co-infected into H1 hESCs and then selected by puromycin for 7 days and were trypsinized and re-plated. hESC colonies were also picked up in the 7th day. *HBBP1* KO HUDEP-2 cells were constructed in the same way. To validate, genomic DNA was extracted from the cells and the knockout colonies were defined as having PCR amplification of smaller band as the specific primers spanned the deletions. Furthermore, the PCR products were directly sequenced to confirm the knockouts.

In a similar way, the single guide RNAs (sgRNAs) targeting HNRNPA1 binding motif of *HBBP1* were used to construct the splice junction mutations and motif deletions in HUDEP-2 cells. The PCR products were sequenced to confirm these mutations. The oligos and primers used in CRISPR/Cas9 were listed in Table S2.

Mouse transplantation

For transplantation experiments, isolated human cord blood hematopoietic progenitor $CD34^+$ cells were expanded in StemSpan SFEM (Stem Cell Technologies) supplemented with 100 ng/ml SCF, TPO, FLT3 ligand and IL-6 for 2 days. 1×10^7 $CD34^+$ cells were infected with the *HBBP1* control vector and short hairpin RNA (shRNA) for 24 h and re-suspended in a density of 5×10^5/10 μl. After the NOG mice were subject to sub-lethal irradiation (2.5 Gy) for six hours, 1×10^6 cells (in 20 μl medium) were injected directly into the hip bone marrow of these mice. The mice were sacrificed after about 3 months of transplantation.

Plasmid construction, lentivirus production and infection

To knockdown *HBBP1*, the shRNA sequences were cloned into lentiviral shRNA expression control vector pLKO.1 with puromycin resistance. The HNRNPA1 human shRNA plasmid kit was purchased from Origene (TL312381). To overexpress *HBBP1* and *TAL1*, the respective cDNAs were amplified from K562 cells and cloned into pSIN lentiviral expression vector. Mutations on the HNRNPA1 binding motif harbored by the HBBP1 cDNA were induced by the QuikChange Site-Directed Mutagenesis Kit (Stratagene, Cat#200518). Lentiviruses were packaged using ViraPower Lentiviral Packaging System (Invitrogen, Cat#K497500). For infection, HUDEP-2 cells or primary erythroid cells at day 4 of expansion or differentiation were incubated with lentiviruses for 12 h before washing the excess virus. After 48 h of infection, the cells were selected with 1 μg/ml puromycin until the end of culture. For RNA-sequencing (RNA-seq), differentiated erythroid cells at day 8 and 14 were collected, $PolyA^+$ RNA was enriched and sequenced at Novogene (Tianjin, China). The primers used for shRNA sequences, overexpression of *HBBP1* and *TAL1* or the mutant sequences are listed in Table S2.

RNA extraction and Quantitative real-time PCR (qRT-PCR)

Total RNA was extracted using TRIzol (Invitrogen, Cat#15596026), and cDNA was synthesized using a reverse transcription system from Promega. For whole blood RNA extraction, the peripheral blood samples of human and chimpanzee were centrifuged to collect blood cells and then extracted by TRIzol as above. We performed qRT-PCR with the SYBR Green PCR kit (Applied Biosystems, Cat#A25742). For assessment of mRNA half-life, 1 μM actinomycin D was added to block the transcription and cells were collected at different time points for RNA extraction. rs2071348 of HBBP1 was genotyped by PCR combined with Sanger sequencing. Primer and probe sequences used for qRT-PCR are shown in Table S2.

Rapid amplification of cDNA ends (RACE)

To examine the transcriptional product of *HBBP1*, RACE was performed with SMARTer RACE 5'/3' Kit (TaKaRa, Cat#634858) according to the manufacturer's instruction. In brief, 10 ng-1 μg of total RNA was converted into 5'- and 3'- RACE-Ready first-strand cDNA respectively. Two rounds of nested PCR were carried out with the universal primer (provided in the kit) and gene-specific primer. The final PCR product was subjected to Sanger sequencing.

To directly confirm the specificity of the two *HBBP1* shRNAs, we employed modified 5′-RNA-linker-mediated RACE (Lasham et al., 2010). siRNAs usually cleave their target following a canonical pattern: 10 bp downstream from the 5'-end of the antisense strand. RISC-associated Argonaute 2 at this position cleaves the target RNA into two distinct fragments: a 5' fragment with a 3' hydroxyl group, and a 3' fragment with a 5' phosphate. The cleaved 3' fragment with a 5' phosphate can be ligated to a RNA linker and amplified through a 5' RACE primer located in RNA linker and the downstream gene-specific primer. PCR products with the expected sizes would be used to judge the specificity of the shRNAs. A brief experimental protocol is described here: 1) 10ng-1μg total RNA was directly ligated to the RNA linker without prior treatment before first-strand cDNA was synthesized using random primer; 2) two

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

rounds of nested PCR were carried out with the 5' RACE primer and gene-specific primer; 3) the PCR product was subjected to Sanger sequencing to ascertain the junction between the RNA linker sequence and the target RNA sequence. Primer sequences are listed in Table S2.

RNA *in situ* hybridization and immunofluorescence

RNA in situ hybridization was performed using the RNAscope Fluorescent Multiplex Kit (Advanced Cell Diagnostics, Cat#320293) according to manufacturer's instruction. In brief, cells differentiated from $CD34^+$ HSPCs with EPO induction or HUDEP-2 cells were fixed in PBMC-Fix for 1 h at room temperature, dehydrated in increasing concentrations of ethanol (50%, 70%, and 100%) and permeabilized with Protease III. A target probe (Hs-HBBP1, REF: 496481; Hs-TAL1, REF: 510091) or control probes (3-Plex Negative Control, REF: 320871; 3-Plex Positive Control, REF: 320861) were then hybridized for 2 h at 40 °C, followed by incubation with signal amplification reagents. The hybridized probes were visualized with fluorescence microscopy.

For HNRNPA1 co-localization experiments, the cells stained with the RNA probes were further blocked with 10% serum and incubated with mouse anti-HNRNPA1 antibody (Abcam, ab5832), which was followed by goat anti-mouse Alexa Fluor 594 (Abcam, ab150116) secondary antibody incubation. The nuclear was stained with DAPI (Sigma, Cat#DUO82040). The signals were visualized with the fluorescence microscopy.

Cell fractionation

Nuclear-cytoplasmic fractionation was performed using the NE-PER Nuclear and Cytoplasmic extraction Reagents Kit (Thermo Fisher Scientific, Cat#78835) according to the manufacturer's instruction. Briefly, cells were suspended in CER I followed by adding CER II on ice for 30 minutes. After centrifugation, the supernatant was collected as cytoplasmic fraction and the remainder with NER washing was considered as nuclear pellets. RNAs from both fractions were extracted with Trizol reagent (Invitrogen).

RNA pull-down assays

To enable RNA Pull-down assays, HBBP1 and Ctrl (control) RNA were *in vitro* transcribed using RiboMAX large scale RNA production systems-T7 (Promega, Cat#P1300). A single biotinylated nucleotide was added to the 3' terminus of an RNA strand with the Pierce RNA 3' End Desthiobiotinylation Kit (Thermo Scientific, Cat#20163). RNA Pull-down assay was performed with the Pierce Magnetic RNA-protein Pull-Down Kit (Thermo Scientific, Cat#20164) according to the manufacturer's instruction. In brief, 50 pmol of biotinylated RNA was mixed with 50 μl washed Streptavidin agarose beads (Invitrogen, Cat#65601) and incubated at room temperature with agitation for 30 minutes. 2×10^7 HUDEP-2 cells were lysed in standard lysis buffers (25 mM Tris-HCl pH 7.4, 150 mM NaCl, 1% NP-40, 1 mM EDTA, 5% glycerol) supplemented with P.I. and 1 U/μl SUPERase•In on ice for 30 minutes followed by centrifugation. Then the RNA-binding beads were added to the cell lysate and rotated at 4 °C for 1 h. Beads were washed with 20 mM Tris-Cl three times and boiled in 1×SDS Loading buffer. The proteins were detected by Western blot or separated in gradient gel electrophoresis followed by mass spectrometry (MS) identification.

RNA immunoprecipitation (IP)

To enable RNA Immunoprecipitation, 50 μl of protein A Sepharose beads were incubated with 5 μg of mouse anti-HNRNPA2B1 antibody (Abcam, ab6102) and mouse anti-HNRNPA1 antibody (Abcam, ab5832) or 5 μg of mouse pre-immune IgG in 500 μl wash buffer (20 mM HEPES PH 7.9, 150 mM NaCl, 0.5 mM EDTA PH 8.0, 10 mM KCl, 1.5 M $MgCl_2$, 0.5% NP-40, 10% glycerine, 1.5 mM DTT, 1 mM PMSF) at 4 °C for 4-6 h. Then the beads were washed three times with wash buffer and kept on ice until use. Harvested HUDEP-2 cells were suspended in 400 μl lysis buffer (20 mM HEPES PH 7.9, 150 mM NaCl, 0.5 mM EDTA PH 8.0, 10 mM KCl, 1.5 M $MgCl_2$, 0.5% NP-40, 10% glycerine, 1.5 mM DTT, 1×Protease Inhibitor cocktail (Roche, Cat#S8830), 10 U/ml RNase Inhibitor and then left on ice for 30 minutes followed by centrifugation. The supernatant was transferred to tubes with antibody or pre-immune IgG-coated beads, and IP was performed by tube rotating at 4 °C overnight. Following IP, the beads were washed two times. At last, the RNA was extracted with Trizol reagent (Invitrogen).

Western blotting, silver staining and mass spectrometry

Protein samples were separated on 10% SDS-PAGE gels and then transferred to PVDF membranes. Membranes were blocked with 5% skimmed milk in TBS-T and incubated with primary antibodies at 4 °C overnight. Membranes were washed and incubated with HRP-conjugated secondary antibodies for 1 h at room temperature. Finally, membranes were washed and visualized with ECL substrate (Millipore, Cat#WBKLS0500).

Silver staining was performed according to the instruction. Briefly, gels were fixed with 40% ethanol:10% acetic acid solution for 30 minutes, and sensitized with 0.02% $Na_2S_2O_3$ for 30 minutes, stained with 0.2% $AgNO_3$: 0.02% formaldehyde for 20 minutes and developed with 3% Na_2CO_3: 0.01% formaldehyde. Reaction was stopped using 1.46% EDTA-2Na•$2H_2O$.

Mass spectrometry (MS) analyses were conducted based on the manufacturer's instructions at the Technology Center for Protein Sciences (Tsinghua University). Briefly, proteins fractions were generated by enzymatic digestion, which was followed by search against the UniProt database.

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

RNA electrophoretic mobility shift assay (EMSA)

To enable RNA EMSA, the biotin-labeled probe, cold probe (unlabeled probe) and mutant probe were synthesized by RiboBio Life Science Company. A total of 50 fmol of biotin-labeled RNA probes were incubated with various concentrations of HNRNPA1 (Abcam, ab123212) using the Light Shift Chemiluminescent RNA EMSA Kit (Pierce, Cat#20158) according to the manufacturer's protocol. Competition experiments were performed with 200-fold molar excess of the unlabeled probe (cold probe) pre-incubation. The reactions were incubated at room temperature for 30 minutes before adding RNA loading dye and separated by native 8% PAGE. The probes used for the RNA EMSA experiment are listed in Table S2.

Enhanced UV crosslinking, immunoprecipitation and qPCR (eCLIP-qPCR)

eCLIP experiments were performed as described previously with minor modifications (Van Nostrand et al., 2016). Briefly, cells were irradiated one time with 150 mJ/cm^2 at 254 nm in 10 cm plates with 3 ml cold PBS. Then cells were snap-frozen in the liquid nitrogen, and stored at -80 °C. The pellet was lysed with the buffer (50 mM Tris-HCl, pH 7.4; 100 mM NaCl; 1% NP-40; 0.1% SDS; 0.5% sodium deoxycholate; 1/200 Protease Inhibitor Cocktail III; 5 μl 0.1 M DTT) followed by further treatment with RNase I and Turbo DNase. The lysate was incubated with mouse anti-HNRNPA1 antibody (Abcam, ab5832) or 10 μg of mouse pre-immune IgG overnight at 4 °C for immunoprecipitation. 70 μl of protein A beads (Invitrogen, Cat#10002D) was added and incubated for 2 hours followed by washes. Following end repair and 3' adaptor ligation, size selection was conducted using Nupage 4–12% Bis-Tris protein gels followed by transfer to nitrocellulose membranes. RNAs on nitrocellulose were harvested and reverse transcribed using SuperScript III (Thermo Fisher, Cat#18080044). cDNA libraries were then prepared and qPCR was conducted using the sequence-specific primers (Table S2).

Interactome analyses

To infer the functionality of pseudogenes, we took advantage of two publicly available interactome datasets. First, to infer whether pseudogenes regulate coding genes by binding miRNA, we downloaded miRNA-pseudogene interaction data from starBase v3 (Li et al., 2014), which was predicted based on hundreds to thousands of miRNA-seq datasets, degradome-seq (miRNA-RNA pair sequencing) datasets and Argonaute CLIP-Seq (sequencing of RNA isolated by crosslinking and immunoprecipitation) datasets. To increase the reliability, we set multiple filters concurrently. To identify miRNAs inducing the cleavage of one pseudogene, we required that their interaction was supported by at least one Ago CLIP-experiment, one degradome-seq experiment and one significant negative (miRNA vs. pseudogene) transcriptional correlation in pan-cancer analyses. Moreover, both *P*-value and FDR (false discovery rate) were set as 0.001 to control for countless possible interactions between RNAs and miRNAs. We actually tried multiple different parameters and consistently observed that retropseudogenes are overrepresented compared to duplicated pseudogenes.

Second, to infer whether RNAs of pseudogenes and coding genes are bound by RBPs, we downloaded all 96 publicly available processed enhanced CLIP-seq (eCLIP) dataset (Table S7) assaying 73 diverse RBPs in hepG2 and K562 cells (Van Nostrand et al., 2016). Compared to the conventional noisy CLIP data, eCLIP appears to be more specific (Van Nostrand et al., 2016). To further increase the accuracy, we added the recommended filters: $P < 1\times10^{-5}$, fold change (eCLIP vs. size-matched input) > 8 (Van Nostrand et al., 2016). According to the chromosomal coordinates, we defined RBP-bound pseudogenes as those overlapping with binding peaks by at least one bp exonic nucleotide. Among 73 RBPs, 47 RBPs bound at least one pseudogene across two replicates.

Evolutionary analyses

To infer whether *HBBP1* evolves slowly in human branch compared to other primates, we aligned all its homologs with MAFFT (Katoh and Toh, 2008) and curated the alignment with GeneDoc (Nicholas et al., 1997). By implementing jModelTest (Darriba et al., 2012), we inferred the best substitution model as GTR. With this model, we estimated the number of exonic or intronic substitutions on each branch by PAML (baseml) (Yang, 2007) v4.9 package. Given the divergence time obtained from the TimeTree database (Kumar et al., 2017), we calculated the background evolutionary rate (μ_{exp}) of *HBBP1* as the sum of the number of substitutions on non-human branches divided by the sum of divergence times on non-human branches. To detect slowdown, μ_{exp} was compared with the observed evolutionary rate (μ_{obs}). The significance level was then determined by the Poisson probability: $P(x \le \xi_{obs}) = \sum_{x \le \xi_{obs}} e^{-\xi_{exp}} \xi_{exp}^{x} / x!$, where $\xi_{obs} = t_{human} L \mu_{obs}$ and $\xi_{exp} = 6.65 L \mu_{exp}$, $t_{human} = 6.65$ million years, and L (=660, 979 bp) is the length of exons and introns, respectively.

To reconstruct the evolutionary history of the nt 371 motif, we built the phylogenetic tree of *HBBP1* across mammals with respect to the other human β-globin cluster genes. Specifically, we aligned all three exons of homologs with MAFFT (Katoh and Toh, 2008), checked the alignment with GeneDoc (Nicholas et al., 1997) and built the tree with RAxML (Stamatakis, 2014). Analogously, the GTR substitution model was used. We performed 1000 bootstrapping and collapsed branches with scores lower than 60. The tree topology is the same if we used only the third exon harboring the motif. Given this phylogeny, we inferred how the binding motif of HNRNPA1 evolves by implementing the baseml tool. In this analysis, the bushbaby ortholog is not included since it lost the third exon. Both GTR and HKY85 models were used, which essentially made the same inference of ancestral nucleotides. So, we simply presented the results based on HKY85.

We analyzed whether *HBBP1* is under negative selection in primate evolution. First, we downloaded exonic phastCons scores based on whole genome alignments of primates from UCSC (Rhead et al., 2010). For each coding and noncoding gene model, we calculated the median value and examined how *HBBP1* contrasted with other genes. Exons of noncoding genes overlapping

Please cite this article in press as: Ma et al., Genome-wide analysis of pseudogenes reveals *HBBP1*'s human-specific essentiality in erythropoiesis and implication in β-thalassemia, Developmental Cell (2020), https://doi.org/10.1016/j.devcel.2020.12.019

with coding gene models were discarded. Since ancient genes are usually more constrained (Yang et al., 2015), we only compared *HBBP1* against genes with the same age, i.e., predating the split of placental mammals. Evolutionary age information was extracted from the GenTree database (Shao et al., 2019). In brief, we inferred when a pseudogene emerged in evolution based on gene synteny across the vertebrate phylogenetic tree. Since this age dataset was generated based on Ensembl v73, we mapped it to Ensembl v91 based on Ensembl accessions.

QUANTIFICATION AND STATISTICAL ANALYSIS

Experimental data are represented as mean ± SD. Student's t-test and Mann Whitney test were used to determine significance as indicated in legends. Sample sizes, including the number of cells or animals in each experiment and *P*-values were indicated in figures and figure legends. Statistical analysis was performed using GraphPad Prism 6.

Most computational data analyses were performed in the environment of R v3.0.0 except the coexpression network analyses enabled by R v3.6.1.