Transcription-Associated Mutation Promotes RNA Complexity in Highly Expressed Genes—A Major New Source of Selectable Variation

Shengkai Pan,[1,2,3] Michael W. Bruford,[2,4] Yusong Wang,[1] Zhenzhen Lin,[1,2] Zhongru Gu,[1,2,3] Xian Hou,[1] Xuemei Deng,[5] Andrew Dixon[2,6] Jennifer A. Marshall Graves,[7] and Xiangjiang Zhan*[,1,2,8]

[1]Key Laboratory of Animal Ecology and Conservation Biology, Institute of Zoology, Chinese Academy of Sciences, Beijing, China

[2]Cardiff University-Institute of Zoology Joint Laboratory for Biocomplexity Research, Beijing, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Organisms and Environment Division, School of Biosciences and Sustainable Place Institute, Cardiff University, Cardiff, United Kingdom

[5]National Engineering Laboratory for Animal Breeding and Key Laboratory of Animal Genetics, Breeding, and Reproduction of the Ministry of Agriculture, China Agricultural University, Beijing, China

[6]Emirates Falconers' Club, Abu Dhabi, UAE

[7]School of Life Science, La Trobe University, Melbourne, VIC, Australia

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

***Corresponding author:** E-mail: zhanxj@ioz.ac.cn.

Associate editor: Rebekah Rogers

The data generated from the saker falcon and chicken blood samples are available in the NCBI Short Read Archive database (PRJNA354494). The RNA-seq data of other chicken and mouse tissue samples are obtained from GenBank (for details, see supplementary table S4, Supplementary Material online).

Abstract

Alternatively spliced transcript isoforms are thought to play a critical role for functional diversity. However, the mechanism generating the enormous diversity of spliced transcript isoforms remains unknown, and its biological significance remains unclear. We analyzed transcriptomes in saker falcons, chickens, and mice to show that alternative splicing occurs more frequently, yielding more isoforms, in highly expressed genes. We focused on hemoglobin in the falcon, the most abundantly expressed genes in blood, finding that alternative splicing produces 10-fold more isoforms than expected from the number of splice junctions in the genome. These isoforms were produced mainly by alternative use of *de novo* splice sites generated by transcription-associated mutation (TAM), not by the RNA editing mechanism normally invoked. We found that high expression of globin genes increases mutation frequencies during transcription, especially on nontranscribed DNA strands. After DNA replication, transcribed strands inherit these somatic mutations, creating *de novo* splice sites, and generating multiple distinct isoforms in the cell clone. Bisulfate sequencing revealed that DNA methylation may counteract this process by suppressing TAM, suggesting DNA methylation can spatially regulate RNA complexity. RNA profiling showed that falcons living on the high Qinghai–Tibetan Plateau possess greater global gene expression levels and higher diversity of mean to high abundance isoforms (reads per kilobases per million mapped reads $\geq$18) than their low-altitude counterparts, and we speculate that this may enhance their oxygen transport capacity under low-oxygen environments. Thus, TAM-induced RNA diversity may be physiologically significant, providing an alternative strategy in lifestyle evolution.

Key words: **transcription-associated mutation, alternative splicing, *de novo* splice site, high expression, DNA methylation, hemoglobin.**

Article

Introduction

Understanding the mechanism generating RNA complexity is a core requirement for interpreting biological functional complexity. It is widely accepted that alternative splicing (AS), the inclusion of different exons in mRNA (Keren *et al.* 2010), is one of the major ways of generating RNA diversity (Jin *et al.* 2005; Kelemen *et al.* 2005; Chen *et al.* 2012; Nellore *et al.* 2016). Recent studies have indicated that 92–94% of human multi-exonic genes undergo AS (Pan *et al.* 2008; Wang *et al.* 2008).

Alternative splice sites can be authentic or *de novo* (supplementary fig. S1, Supplementary Material online). Authentic splice sites refer to those existing in the genome at exon boundaries, including canonical 5′GT-AG3′ and non-canonical splice sites (5′AT-AC3′; 5′AT-AG3′; 5′AT-AT3′; 5′AT-AA3′; 5′GT-AT3′; 5′GT-GG3′; 5′GT-AA3′; 5′GT-CA3′; 5′GC-AG3′; 5′GT-AC3′; Jackson 1991; Burset *et al.* 2000; Bernard *et al.* 2004). In contrast, *de novo* splice sites refer to splice sites newly created during transcription. Genes with more exons normally possess more authentic splice sites

than those with fewer, and are expected to yield more transcript isoforms through alternative combinations of exons (Grishkevich and Yanai 2014), for example, in the *Drosophila Dscam* gene (Graveley 2005). Genes containing *de novo* splice sites can also generate splice isoforms during transcription (e.g., *Drosophila 4f-rnp*) (Petschek *et al.* 1997).

RNA editing could play a major role in the generation of *de novo* splice sites because it can create the sites by modifying adenosine (*A*) to inosine (*I*), or cytosine (*C*) to uracil (*U*) (Rueter *et al.* 1999; Keegan *et al.* 2001). More than half of the splice isoforms found in the *Drosophila melanogaster* genome have been attributed to RNA editing and 3′-UTR extensions (Brown *et al.* 2014). However, RNA editing occurs rarely in mammals, birds, other insects, and nematodes (Frésard *et al.* 2015; Rosenthal *et al.* 2015), with only ~1,200 RNA editing sites found in human genome (Xu and Zhang *et al.* 2014), 152 in the diamondback moth (He *et al.* 2015), and 40 in chicken (Frésard *et al.* 2015). Hence, there may be another, more important source of *de novo* splice sites.

Previous work has suggested that some highly expressed genes with few exons may yet generate many isoforms (Grishkevich and Yanai 2014). Since such genes could generate few isoforms through AS via existing authentic splice sites, their high levels of isoform diversity must be derived from *de novo* splice sites created during expression. How these *de novo* splice sites form during transcription has not been determined to date. DNA methylation has recently been found to play an important role in regulating AS by impacting the use of authentic splice sites to promote exon exclusion or inclusion (Shukla *et al.* 2011; Ong and Corces 2014; Lev Maor *et al.* 2015; Yearim *et al.* 2015), but little is known about the role methylation plays in regulating *de novo* splice sites.

To answer these questions, we sequenced blood-derived transcriptomes and methylomes of wild saker falcons, aerial predators belonging to a group that includes the fastest land vertebrates. We compared the falcon transcriptomes with publicly available data from tissues in chicken ($N = 29$) and mouse ($N = 7$), and 12 additional chicken blood transcriptome profiles generated here. We focused on RNA diversity in hemoglobin genes because their expression is the most abundant in blood and they play a vital role in oxygen supply, and because of their relatively short length (only 2,000 bp with two or three exons corresponding to two or four authentic splice sites). Our comparisons reveal that transcription-associated mutation (TAM) is the predominant mode for the generation of novel splicing events, rather than previously proposed mechanisms such as RNA editing (Brown *et al.* 2016). We further showed that the expression levels of hemoglobin isoforms induced by TAM and the diversity of average to highly expressed isoforms (i.e., reads per kilobases per million mapped reads [*RPKM*] ≥18) were much greater in falcons on the high Qinghai–Tibetan Plateau than in their lowland relatives, suggesting an evolutionary strategy to survive their low-oxygen environment. We propose that generation of transcriptome complexity relates to biological function diversity.

Results

Data Generation and Processing

Blood samples for RNA extraction were collected from 30 wild saker falcon chicks from sites across Eurasia, and from 12 chickens originating from three breeds (Chahua, White Leghorn, and Tibetan), all at 3–5 weeks of age (supplementary tables S1 and S2 and fig. S2, Supplementary Material online). Next generation sequencing (NGS) generated ~3.6 gigabases (Gb) of clean RNA-seq data for each falcon and 5.0 Gb for each chicken (supplementary tables S1 and S2, Supplementary Material online). Reads from each individual were aligned to the reference genome (Zhan *et al.* 2013) using *SOAP* v2.21 (Li *et al.* 2009), followed by SNP calling using *SOAPsnp* v1.04 (Li *et al.* 2009), yielding 377,530 SNPs after filtering (supplementary table S3, Supplementary Material online). To obtain the highest quality data set, only those sites covered by reads in all individuals were analyzed. The final data set comprised 76,044 SNPs. To detect potential alternative splicing events, we mapped reads to the reference genomes using *Bowtie* 0.12.7 (Langmead *et al.* 2009) and identified potential splice junctions using *Tophat* 1.3.3 (Trapnell *et al.* 2009). The expression level of each gene was quantified using *RPKM* by mapping the transcriptome reads of each sample to the gene set. The same pipeline was also used for the analyses of chicken and mouse data using their reference genomes (Ensembl 83).

Population genomic analysis of the falcon samples identified three genetic clusters, namely, Qinghai–Tibetan Plateau (QH), Kazakhstan–Mongolia (KZ–MN), and West: (Moldova [MD] and Slovakia [SK]) (Pan *et al.* 2017). These clusters were used to analyze patterns of AS variation and expression.

Alternative Splicing in Highly Expressed Genes

We first explored the relationship between alternatively spliced transcripts of a gene and the level of its expression. Based on 11,000 expressed genes per falcon, we defined 14 expression levels and calculated the average transcript isoform number (AS number) caused by AS per gene for each level (fig. 1A). We observed that AS number is stable at low expression levels, but rapidly increases at higher levels Ln (*RPKM*) ≥10, corresponding to *RPKM* ≥20,000, fig. 1A).

To establish whether this relationship is general to other vertebrates, we carried out the same analysis on the RNA-seq data of the 12 newly sequenced chicken blood transcriptome profiles (supplementary table S2, Supplementary Material online), 29 chicken tissues, and 7 mouse tissues downloaded from GenBank (supplementary table S4, Supplementary Material online), and found similar correlations (supplementary fig. S3, Supplementary Material online). Genes with low expression (*RPKM* <3000) showed no correlation between AS and expression level, contradicting previous conclusions (Grishkevich and Yanai 2014). To investigate this relationship, we next plotted the average number of exons per gene in each falcon against its expression level and found that exon number was negatively correlated with gene expression in all birds (all $P < 0.0005$, fig. 1B and supplementary table S5, Supplementary Material online). Reanalysis of the data in

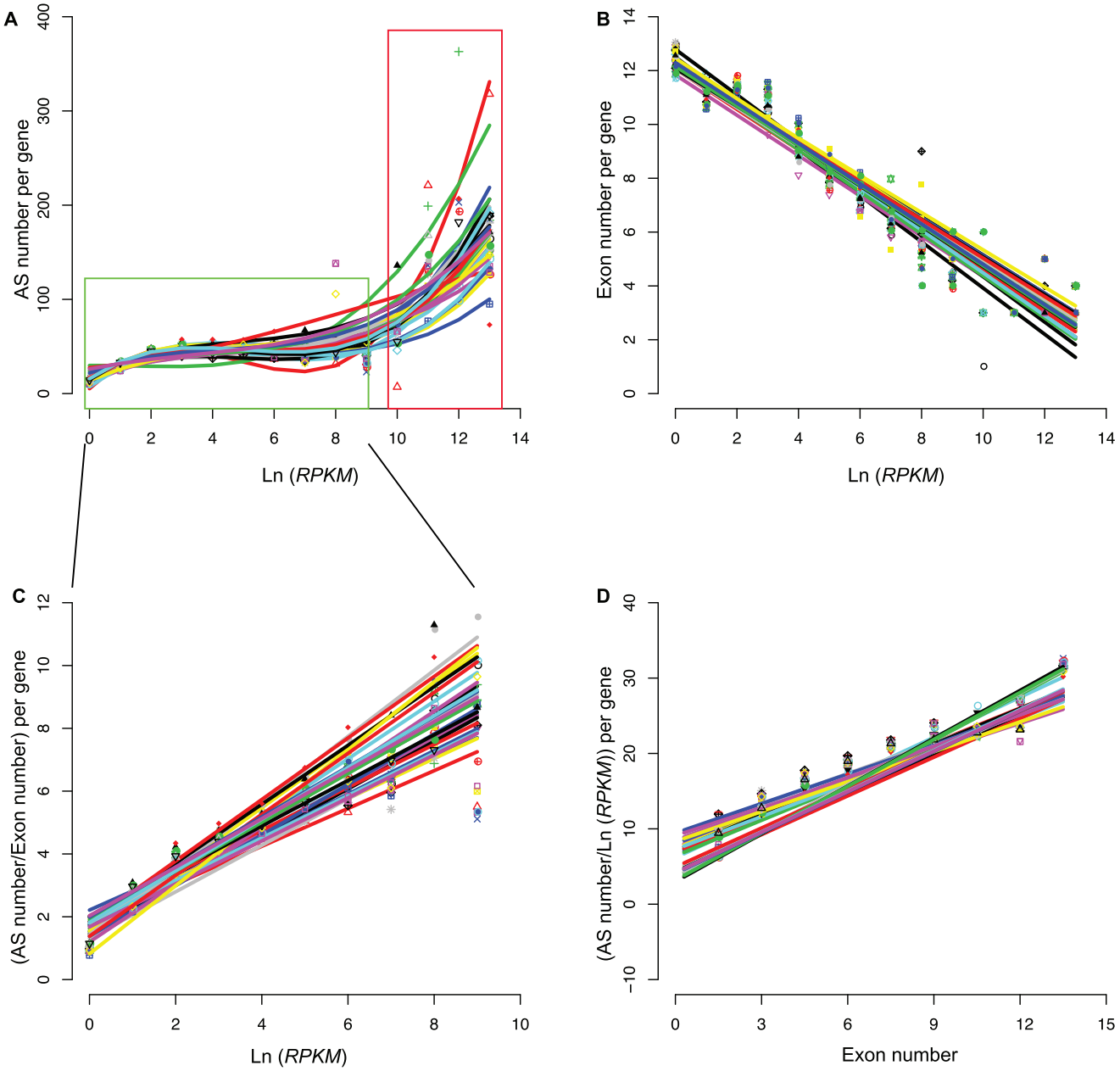

Fig. 1. Correlations among alternative splicing (AS) number, gene expression, and exon number. (*A*) Average AS number in each expression interval is plotted against expression (Ln (*RPKM*)) for all 30 saker falcons. Fourteen levels of expression are defined according to Ln-transformed gene expression levels (0–13). A curve for each individual was fitted and each falcon individual is indicated by a different color. Green rectangle denotes the observed stable stage and the red one denotes the sharp increase stage. Different individuals are denoted by different symbols. (*B*) Averaged exon number is plotted against expression (Ln (*RPKM*)), showing the negative correlation between them. (*C*) With the effect of exon number controlled, average AS number is positively correlated with expression level even at lower expression levels. (*D*) Average AS number after removing the effect of gene expression is positively correlated with exon number.

Grishkevich and Yanai (2014) reached the same conclusion (all $P<0.05$, supplementary fig. S4 and table S6, Supplementary Material online). Realizing that the relationship between AS and gene expression is contingent on the number of exons as well as gene expression levels, we calculated the AS number per exon by dividing by the mean exon number and observed a linear positive correlation with expression level over the range of low to moderately expressed genes (fig. 1*C*, all $P<0.01$, supplementary table S5, Supplementary Material online).

We conclude that the most highly expressed genes generate an extremely large number of alternative transcripts, even though the number of exons (and therefore authentic splice sites) is lower than average. In our research, the most highly represented genes in falcons were hemoglobins (*RPKM* >50,000). There were ~100 transcript isoforms per hemoglobin gene, although each of them has no more than three exons. Therefore, they represented an ideal model system in which to study how highly expressed genes can generate such a large number of isoforms.

Alternative Splicing in Falcon Hemoglobin Genes

Hemoglobin is a tetrameric protein comprising two α and two β chains in birds and other jawed vertebrates, and is responsible for dissolved oxygen transport in the blood via uptake of pulmonary oxygen, and for its release to tissue mitochondria. Similar to other birds, the falcon was found to possess three clustered α globin copies (5′-α^E-α^D-α^A-3′) and four β globin copies (5′-ρ-β^H-β^A-ϵ-3′) (Opazo *et al.* 2015), except that the first exon of the falcon α^A gene was truncated (supplementary fig. S5, Supplementary Material online). Expression analysis showed that α^D, α^A and β^H, β^A were the four most highly expressed genes in blood of these 3- to 5-week chicks, so we focused on the globin genes for detailed analysis.

We first examined AS for hemoglobin in falcons. Because the paralogous hemoglobin genes have high-sequence similarity, bioinformatic assembly of AS transcripts for hemoglobin genes using NGS RNA-seq reads was problematic, so we used spliced intron profiles to infer putative AS events among exons. Given that there are three exons in the falcon α^D, β^H, and β^A and two in α^A gene (supplementary fig. S5, Supplementary Material online), the number of introns was expected to be two and one, respectively, in the absence of alternative splicing. However, we observed 22 times this number of spliced introns for α^D, 33× for α^A, 41× for β^H, and 53× for β^A, on average (supplementary table S7, Supplementary Material online).

To verify the reality of these spliced introns, we extracted the splice junction sequences for the four genes with 200-bp flanking sequences from randomly selected spliced introns ($N = 393$), and aligned the RNA-seq reads to the fragments containing splice junctions (supplementary fig. S6, Supplementary Material online). The result showed that the vast majority of splice junctions (90%) were supported by at least one unique-hit read (supplementary table S8, Supplementary Material online), suggesting the reality of splice junctions.

We inferred potential isoforms per falcon using the information of spliced introns (Materials and Methods, supplementary fig. S7, Supplementary Material online), and then aligned the RNA-seq reads back to the candidate isoforms to test their reliability (Materials and Methods). Only isoforms with read coverage $\geq$90% were retrieved. This showed that falcons averaged 81 α and 71 β hemoglobin isoforms (supplementary table S9, Supplementary Material online). When applying a stricter threshold ($\geq$95% coverage), we still detected 66 and 45 isoforms (supplementary table S9, Supplementary Material online).

To independently confirm the hemoglobin isoform diversity inferred from NGS-based RNA-seq, we performed the whole transcript-length third-generation (Iso-Seq) sequencing on three new samples from different 3- to 5-week-old falcon chicks, this time identifying a total of 150 α and β hemoglobin transcript isoforms. We then compared these isoforms with those predicted from the RNA-seq data of three falcons randomly chosen from our NGS data set, and found that the Iso-Seq results with the new samples reproduced 80% of NGS transcript isoforms.

We examined whether the inferred transcripts were putatively functional. A transcript isoform is considered to be putatively functional if the AS event does not result in a frame-shift or premature stop codon in the predicted protein. Following this criterion, we calculated the proportion of intact isoforms with complete ORFs (open reading frames) for each individual, and found that on average 63% of NGS-based transcript isoforms were functional in each falcon (supplementary fig. S8, Supplementary Material online). A similar rate (65%) was also observed in the sequenced chicken transcriptome and both rates were significantly higher than those expected by chance (33%, $P < 0.01$, χ^2 test) given that the frequency of in-frame incorporation into intron is 1/3. Further analysis on the 3D structure of hemoglobin showed that ~70% of variants among these saker hemoglobin isoforms occurred in the heme-binding domain, implying that these variants might influence the heme-binding affinity, which warrants further experimental evidence (e.g., hemoglobin–oxygen affinity) in future. To verify the authenticity of these isoforms, we selected the top ten highly expressed isoforms (excluding canonical isoforms) for RT-qPCR in six falcons (three in QH and three in KZ–MN). The expression of these ten isoforms was found to account for 52% of the total hemoglobin produced (excluding canonical isoforms). All ten chosen isoforms were expressed in each individual (supplementary fig. S9, Supplementary Material online).

Position of AS and RNA Variations

Since each hemoglobin gene has only two or three exons (i.e., two or four authentic splice sites), it is theoretically impossible to generate more than three or seven isoforms by simply using the original authentic splice sites. Nevertheless, we identified on average 46, 41, 91, and 121 splice sites for α^D, α^A, β^H, and β^A, respectively, for each individual falcon, more than an order of magnitude greater than the number of authentic splice sites (supplementary table S10, Supplementary Material online).

We therefore reasoned that most of the splice sites must be newly created by somatic mutations. To test this hypothesis, we first estimated general levels of between-transcript variation (termed as RNA variation hereafter) in the four hemoglobin genes by calculating a polymorphism index π (Tajima 1989) in each population, and compared them across the whole gene set. We found that the four hemoglobin genes (α^D, α^A, β^H, and β^A) showed extreme π values in all populations (fig. 2A), two orders of magnitude higher than other genes with two or three exons. To assess whether the potentially high mutation loads were general or were associated with the high proliferative potential or altered replication fidelity during hematopoiesis in blood cells, we also compared mutation loads (RNA variation) between other genes with high, moderate, and low levels of gene expression in falcons. We found that mutation loads in highly expressed genes were indeed significantly higher than for genes with moderate or low expression (supplementary fig. S10, Supplementary Material online, $R^2 > 0.90$, and $P < 1E\text{-}7$).

We next compared the locations of identified *de novo* splice junctions with the positions of these identified RNA

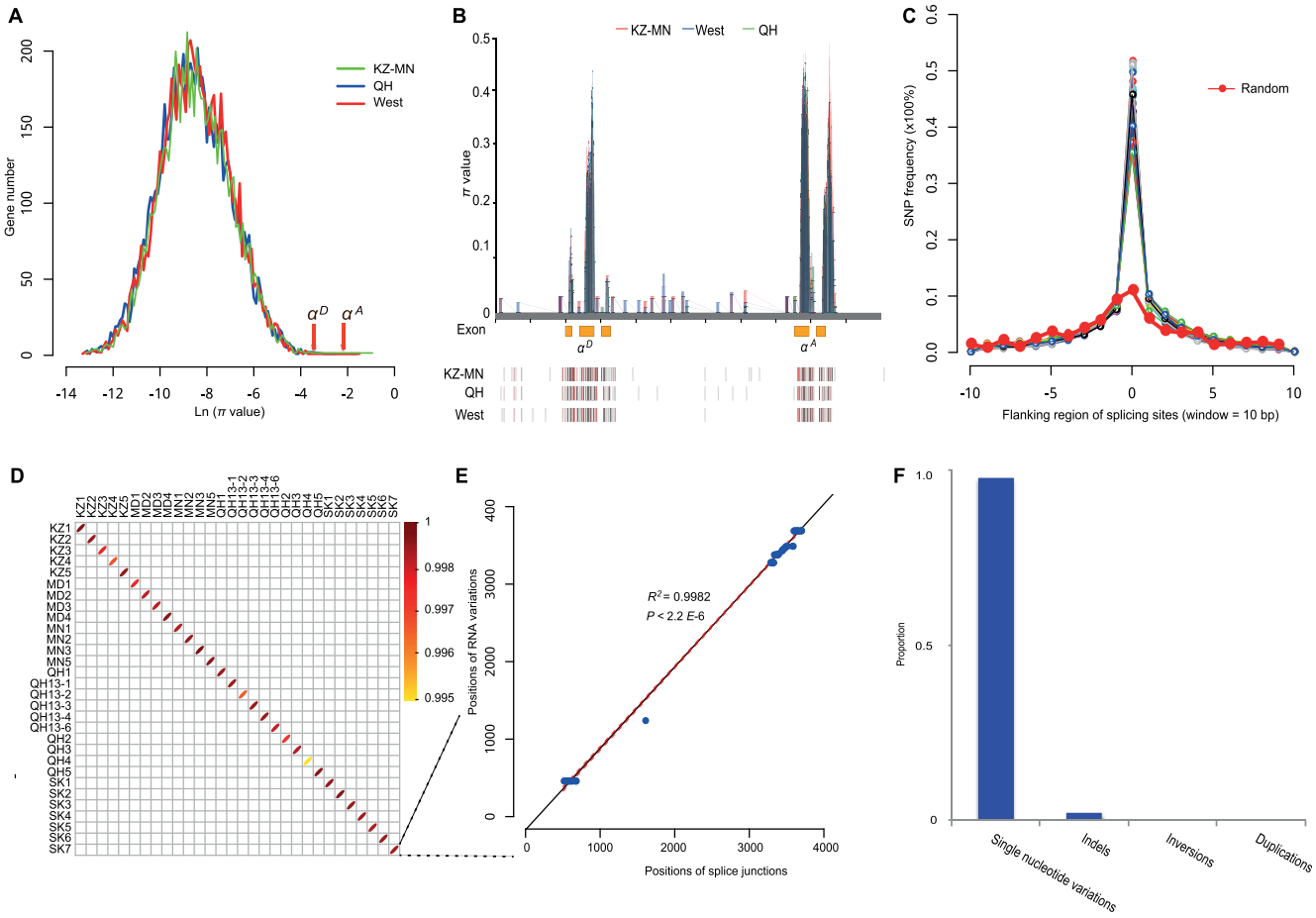

FIG. 2. Alternative splicing and variations for the α globin locus in the saker falcon. (*A*) π value distribution for the whole gene set in QH, KZ–MN, and West population clusters. (*B*) The distribution of π values using a 20-bp window at the α globin locus (α^E is not shown here since its π value is low), and the distribution of splice junctions on the transcribed strand in QH, KZ–MN, and West populations. The intensity of red color is in proportional to the utilization frequency of the splice junction. (*C*) The actual distribution of RNA variations near *de novo* splice junctions in each individual (represented by different colors). Simulation results are indicated in red. (*D*) The position of the splice junction is plotted against the position of RNA variant within the α globin locus in each falcon and the strength of correlation is denoted by red intensity. A representative individual is shown in *E*. (*F*) The proportions of single nucleotide variants, indels, inversions, and duplications identified in the hemoglobin locus.

variations for each of the four hemoglobin genes. We found that the positions of the splice junctions inferred from both RNA-seq and Iso-Seq overlapped (fig. 3*A*), and largely overlapped the locations of RNA variations (fig. 2*B* and supplementary fig. S11*A* and *B*, Supplementary Material online). There is a significant correlation between these RNA variations and splice junction positions, with coefficients (R^2) ranging from 0.945 to 0.999 for α globin genes (all $P < 2E\text{-}16$) and from 0.980 to 0.998 for β globin genes (all $P < 2E\text{-}16$) (fig. 2*D* and *E*; supplementary table S11, Supplementary Material online). In addition, the distribution of RNA variations showed that RNA variations were significantly enriched around splice sites (fig. 2*C*).

As well as single nucleotide variations, indels, inversions, and duplications are known to give rise to potential mutations in transcripts (Tümer 2013). However, aligning the RNA-seq reads to the four genes showed that single nucleotide variations accounted for almost all the RNA mutations we detected (98%; fig. 2*F*).

To establish the generality of our findings, we performed the equivalent analysis of chicken and mouse RNA-seq data, and identified the same patterns in highly expressed genes: high mutation load (supplementary figs. S12 and S13, Supplementary Material online), a high correlation ($R^2 > 0.95$) between RNA variations and splice junction positions (supplementary table S12, Supplementary Material online), and RNA variations significantly clustered with the splice sites (supplementary fig. S14, Supplementary Material online).

We conclude that *de novo* splice sites, as well as RNA variations, are nonrandomly distributed over the genome, and these sites coincide with the predictions of the hypothesis that RNA variations create *de novo* splice sites.

Source of RNA Variations in Hemoglobin Genes

To investigate the source of high RNA variation in hemoglobin genes, we first calculated the percentage of heterozygous and homozygous variations in transcripts for the four genes per falcon, on the assumption that *de novo* splice sites would be heterozygous. As expected, we found that heterozygous variations accounted for ~90% of the total (an average of 89.7% for α^D, 95.6% for α^A, 93.0% for β^H, 91.1% for β^A),

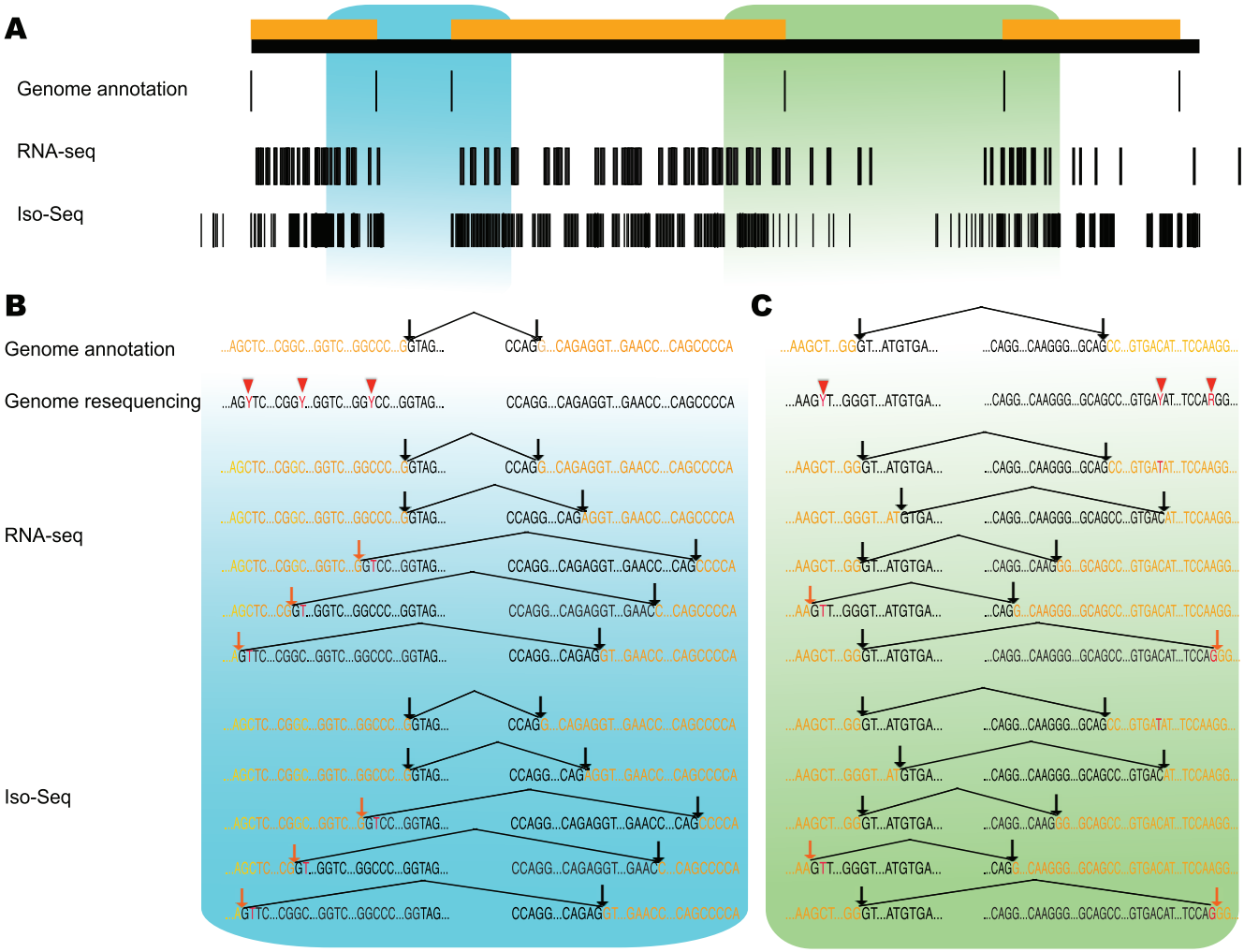

FIG. 3. Alternative splicing in hemoglobin genes based on the genome annotation, RNA-seq prediction and Iso-Seq sequencing of saker falcons, respectively. (*A*) Splice sites in the hemoglobin β^H gene inferred or identified using the three approaches. Orange rectangles represent exons and black lines splice sites. (*B*) and (*C*) Magnification of the selected regions in (*A*) (shadowed). Black bases mark authentic exon/intron boundaries. Red bases denote TAM-induced *de novo* mutations. Red and black arrows denote *de novo* and authentic splice sites, respectively. *Y* represents *C*/*T* and *R* *A*/*G*.

whereas homozygous variations account for only 10% ($P = 3.14E\text{-}9$, paired *t* test) (supplementary table S13, Supplementary Material online). Simulation results confirmed that at least 92% of these heterozygous variations were free from sequencing error (details see Materials and Methods; supplementary fig. S15, Supplementary Material online).

We then investigated whether this unusually high percentage of heterozygous RNA variations could be attributed to transcriptomic, or genomic germline variations. We analyzed the falcon reference genome, calling germline SNPs by mapping the NGS DNA sequencing reads from small insert libraries (170–800 bp: accession SRP018394) onto the genome (Zhan *et al.* 2013) (Materials and Methods). The germline variations of hemoglobin genes called from the genome sequences were found to be far fewer than those found among RNA transcripts (5 vs. 40 for α^D, 7 vs. 78 for α^A, 2 vs. 43 for β^H, 4 vs. 69 for β^A) ($P = 5.99E\text{-}3$, *t* test) (supplementary table S14, Supplementary Material online), accounting for only ~10% of total heterozygous RNA variations. To verify this finding from RNA-seq, we Sanger-sequenced the cloned genomic DNA for the four genes ($N = 3$), the results of which consistently confirmed that only 10% of the RNA variations could be linked to germline variations detected from the cloned sequences (Materials and Methods).

We further compared the RNA variation frequencies in highly expressed genes (i.e., $RPKM \geq 20{,}000$) with the frequency in intergenic regions across the genome. If heterozygous RNA variations originated mainly from germline variations, we would expect similar levels of variation for intergenic regions. However, our comparison showed a significantly greater RNA variation frequency associated with high gene expression ($1.2E\text{-}2$ vs. $8.0E\text{-}4$; $P < 0.01$, Poisson test). Therefore, we conclude that the large number of heterozygous RNA variations originate mainly from transcription. We made similar conclusions from the analyses of 11 different chicken tissues and 7 mouse tissues (all $P < 0.01$ Poisson test, Materials and Methods).

Mechanism for Generating RNA Variations during Transcription

We focused on heterozygous RNA variations, since they comprised the majority of total RNA variations. Previous studies have suggested that either or both of two main mechanisms, TAM (Kim and Robertson 2012; Park *et al.* 2012; Gaillard *et al.* 2013) and posttranscriptional modification (e.g., RNA editing; Brennicke *et al.* 1999), may underpin high RNA variations during transcription.

During TAM, DNA sequences in a gene are locally mutated during transcription. Among the total of six mutation types (summarized in Materials and Methods), four types, $C \rightarrow T$ ($G \rightarrow$ A on the other strand), $A \rightarrow G$ ($T \rightarrow C$), $G \rightarrow T$ ($C \rightarrow A$), and $A \rightarrow T$ ($T \rightarrow A$) have been found to be the most frequent (Mugal *et al.* 2009; Park *et al.* 2012). We calculated the proportion of these four mutation types and found that they accounted for 90% of the total heterozygous RNA variations identified at the two falcon hemoglobin loci (supplementary fig. S16, Supplementary Material online), suggesting a strong role for TAM in RNA variation production. Again, analysis of the expression profiles of hemoglobin genes for each chicken blood sample showed that the same four mutations comprised the majority of heterozygous RNA variations identified (on average 87%), suggesting a major role of TAM in the generation of transcriptome complexity.

Somatic variation of RNA sequence has been generally ascribed to RNA editing, by substitution and insertion/deletion (Brennicke *et al.* 1999; Gott and Emeson 2000). RNA editing requires specific RNA secondary structure conformations and leads to $A \rightarrow G$ or $C \rightarrow T$ substitutions (Gott and Emeson 2000; Tian *et al.* 2011), both of which, however, can also be caused by TAM. To test whether RNA editing contributes to the RNA variations observed, we used *RNAstructure* (Reuter and Mathews 2010) to check for secondary structure motifs associated with RNA editing, and their potential variations flanking the $A \rightarrow G$ and $C \rightarrow T$ sites in falcon α^D, α^A, β^H, and β^A genes. We found no evidence for secondary structure modification associated with the heterozygous RNA variations identified (supplementary fig. S17, Supplementary Material online), suggesting RNA editing may contribute little to the RNA diversity observed in hemoglobin genes.

It is reported that more TAMs occurred on nontranscribed than transcribed DNA strands (Francino and Ochman 1997; Gaillard *et al.* 2013). For the falcon hemoglobin genes, we estimated that 85% of TAM mutations came from nontranscribed DNA strands, and only 15% from transcribed DNA strands ($P = 1.30E\text{-}10$, χ^2 test), consistent with previous studies (Francino and Ochman 1997; Gaillard *et al.* 2013). Different from this bias, we found no strand preference of mutations in intergenic regions that are not transcribed (53% vs. 47%, $P = 0.95$, χ^2 test, supplementary notes, Supplementary Material online).

Because most TAM induced mutations occurred on nontranscribed DNA strands, it was important to establish whether the TAM mutations have a role in transcription. To test this, we used the relationship between hemoglobin synthesis and erythropoiesis as a model. Although erythrocytes are the end products of erythropoiesis, their precursor cells (basophilic normoblasts, polychromatic erythroblasts, orthochromatic normoblasts, or reticulocytes; supplementary fig. S18, Supplementary Material online) can also synthesize hemoglobin (Glass *et al.* 1975). Somatic proliferation of the first two cell types would result in the change in base sequences of transcribed strands to produce hemoglobin isoforms (supplementary fig. S18, Supplementary Material online). We therefore hypothesized that TAM induces somatic mutations mostly on nontranscribed DNA strands, but these are copied onto transcribed strands at cell division, generating *de novo* splice sites and distinct hemoglobin isoforms.

To test this hypothesis, we carried out deep NGS genome resequencing (>30-fold coverage) of blood of six falcons to confirm that most RNA variations are derived from somatic genomic mutations (supplementary notes, Supplementary Material online). We found that 80–90% of between-transcript RNA variations could be tracked back to the somatic mutations identified by resequencing the four hemoglobin genes.

We next determined whether TAM induced mutations could be passed down through cell division. If the four most frequent TAM mutations ($A \rightarrow T$, $C \rightarrow T$, $G \rightarrow T$, and $A \rightarrow G$) occur on nontranscribed DNA strands, the dominant mutation categories in the cDNA are expected to be $T \rightarrow A$, $G \rightarrow A$, $C \rightarrow A$, and $T \rightarrow C$ after cell division and transcription (Prediction 1; supplementary fig. S19, Supplementary Material online). Conversely, if TAM mutations occur on transcribed DNA strands, the dominant cDNA mutations would be $A \rightarrow T$, $C \rightarrow T$, $G \rightarrow T$, and $A \rightarrow G$ (Prediction 2; supplementary fig. S19, Supplementary Material online). To test these predictions, we carried out RT-PCR on three falcon RNA extracts and obtained cDNA sequences for β^H fragments (361 bp) using Sanger sequencing. We analyzed 95 clones for each individual. Our results support Prediction 1 and show that 80% of the mutation classes on cDNA are $T \rightarrow A$, $G \rightarrow A$, $C \rightarrow A$, and $T \rightarrow C$, (vs. 10% of $A \rightarrow T$, $C \rightarrow T$, $G \rightarrow T$, and $A \rightarrow G$; supplementary fig. S19, Supplementary Material online). We conclude that TAM mutations are passed down to somatic cell descendants.

To test whether TAM induced mutations could create *de novo* splice sites, we focused on a representative region of the β^H gene (fig. 3*B* and *C*). In this region, we identified four *de novo* 3′ splice sites (*GT*) caused by $C \rightarrow T$ mutations, which were consistently verified using different sequencing approaches including genome resequencing, RNA-seq, and Iso-Seq technologies (fig. 3*B* and *C*). We verified that the hemoglobin isoforms produced by these four *de novo* splice sites were distinct from those that could be produced by the authentic splice sites (fig. 3*B* and *C*). Specifically, the splicing events due to these *de novo* splice sites can increase/decrease the size of the excised intron by a few base pairs or create a new intron (fig. 3*C*).

In addition, to experimentally confirm that the *de novo* splice sites were created by somatic mutation, we analyzed 95 Sanger sequenced clones of β^H cDNA in one falcon (QH13-6) and identified five *de novo* splice sites ($C \rightarrow T$ mutations). Four

of these were identical to those found above (fig. 3B and C). We could confirm that these *T* alleles were caused by somatic as apposed to germline mutations because the genome resequencing reads of this individual (>30×) gave a ratio of reads supporting the "*T*" and "*C*" allele significantly deviated from 1:1 (2:26, 2:30, 3:28, 3:31, and 2:27 for the five splice sites, respectively; Fisher's exact test, all $P < 1.0E\text{-}6$; See Somatic Mutation Identification in Materials and Methods).

The Influence of DNA Methylation on TAM

What is the basis of the nonrandom distribution of *de novo* splice sites in hemoglobin genes (TAMs) that we observed (fig. 2B and C)? Previous research has suggested that DNA methylation affects authentic splice sites (Shukla *et al.* 2011; Ong and Corces 2014; Lev Maor *et al.* 2015; Yearim *et al.* 2015). However, there is no information on the influence of DNA methylation on the generation of *de novo* splice sites (i.e., TAMs). We therefore sought to understand the relationship between the TAM frequency and its DNA methylation level.

Blood-derived DNA from five falcons was therefore subjected to bilsulfate sequencing to obtain the methylation profile (Materials and Methods). Because the majority of TAMs (85%) occurred on the nontranscribed strands of falcon hemoglobin genes, we focused on the analysis of DNA methylation on these strands in the falcon.

We found that the promoter regions of highly expressed genes were hypomethylated, whereas promoters of genes with low expression were not (supplementary fig. S20, Supplementary Material online), consistent with previous findings that DNA methylation suppresses gene expression (Suzuki and Bird 2008). Interestingly, CpG methylation was higher in the gene bodies of the four hemoglobin genes than the promoters. There were significantly fewer CpG methylated sites in gene bodies in TAM regions than non-TAM regions (about half, $P = 5.16E\text{-}4$, paired *t*-test, supplementary table S15, Supplementary Material online) (fig. 4). We also observed similar patterns for the two main types of non-CpG methylation (CHH and CHG, where H represents any nucleotide) (fig. 4; $P = 8.27E\text{-}3$ and $7.99E\text{-}3$ for non-TAM vs. TAM regions for CHH and CHG, respectively; supplementary table S15, Supplementary Material online).

We conclude that DNA methylation may protect regions from transcription activated mutation. NMR spectroscopy has shown that secondary DNA structures are formed on single stranded DNA when CpG methylation occurs, stabilizing the DNA strand (Taqi *et al.* 2012). Therefore, it is likely that regions enriched with cytosine methylation on the nontranscribed DNA strand are more resistant to TAM during transcription because of the stabilization associated with methylated sequences.

Biological Significance of Transcript Diversity Caused by TAM

Given that falcons demonstrate extraordinary transcript diversity in the hemoglobin genes, we sought to understand whether this high diversity caused by TAM could be functionally significant. The falcons inhabiting the Qinghai–Tibetan Plateau provide an ideal model to test this hypothesis. Because this QH falcon population lives at the altitude >4,000 m (Pan *et al.* 2017), it is expected that they have evolved an optimized evolutionary strategy to cope with low-oxygen stress associated with high altitude. We therefore compared RNA profiles of QH falcons with those from the neighboring population (i.e., KZ–MN) at relatively low altitudes (500–1350 m).

RNA-seq analysis showed that overall expression levels of the four hemoglobin genes in QH were significantly higher than KZ–MN falcons ($RPKM = 690{,}347$ vs. 597,865 for α^A, 348,834 vs. 280,233 for α^D, 505,393 vs. 448,348 for β^H, and 106,075 vs. 95,927 for β^A, $P < 0.05$, Paired *t* test). We then plotted the number of *de novo* splice sites against their expression levels (i.e., supporting reads) in each population. Compared with KZ–MN, QH demonstrated a striking difference in its splice sites, with significantly more implicated at medium to high expression levels (≥4 supporting reads), and with significantly fewer implicated at low levels (fig. 5). Given that the number of splice sites correlates with that of transcript isoform diversity, this suggests that QH falcons possess a higher number of higher-than-average abundance isoforms. One explanation for this observation is adaptive gene expression (Zhang *et al.* 2009) where elevated gene expression promotes the fixation of favorable expression level-altering mutations.

To overcome difficulties in transcript assembly associated with NGS RNA-seq, we used the whole-transcript length Iso-Seq data and identified the total of 150 hemoglobin isoforms in the four hemoglobin genes (41 in α^A, 39 in α^D, 32 in β^A, and 38 in β^H) from the three QH and three KZ–MN falcons (fig. 6 and supplementary fig. S21, Supplementary Material online), the majority (40 in α^A, 38 in α^D, 31 in β^A, and 37 in β^H) of which were found to result from *de novo* splicing associated with TAMs. We next estimated the expression level of each isoform in the hemoglobin loci in each falcon by mapping its NGS RNA-seq reads to these isoforms. However, we found that Iso-Seq only detected medium to high abundance isoforms because its lowest detectable *RPKM* was 18, corresponding to four supporting RNA-seq reads. Our observation is consistent with previous reports (Li *et al.* 2014). Among the 150 isoforms obtained by Iso-Seq at the four loci, 68 (22 in α^A, 15 in α^D, 13 in β^A, and 18 in β^H) were expressed in both populations, 59 (18 in α^A, 17 in α^D, 11 in β^A, and 13 in β^H) only in the highland QH population, and 23 (1 in α^A, 7 in α^D, 8 in β^A, and 7 in β^H) only in the lowland KZ–MN population. Thus, the hemoglobin diversity induced by TAM was much higher in QH than KZ–MN falcons at medium to high gene expression levels ($P = 7.89E\text{-}3$, paired *t* test), supporting our RNA-seq results (fig. 5).

For the Iso-Seq isoforms coexpressed in both populations, the mean expression levels were higher in QH than that in KM in all the hemoglobin genes (all $P < 0.01$, paired *t* test). Moreover, 20 isoforms (7 in α^A, 4 in α^D, 4 in β^A, and 5 in β^H) had significantly higher expression in QH, but only ten (4 in α^A, 2 in α^D, 3 in β^A, and 1 in β^H) were significantly overexpressed in the lowland KZ–MN population (fig. 6). The same pattern was also observed when we compared QH

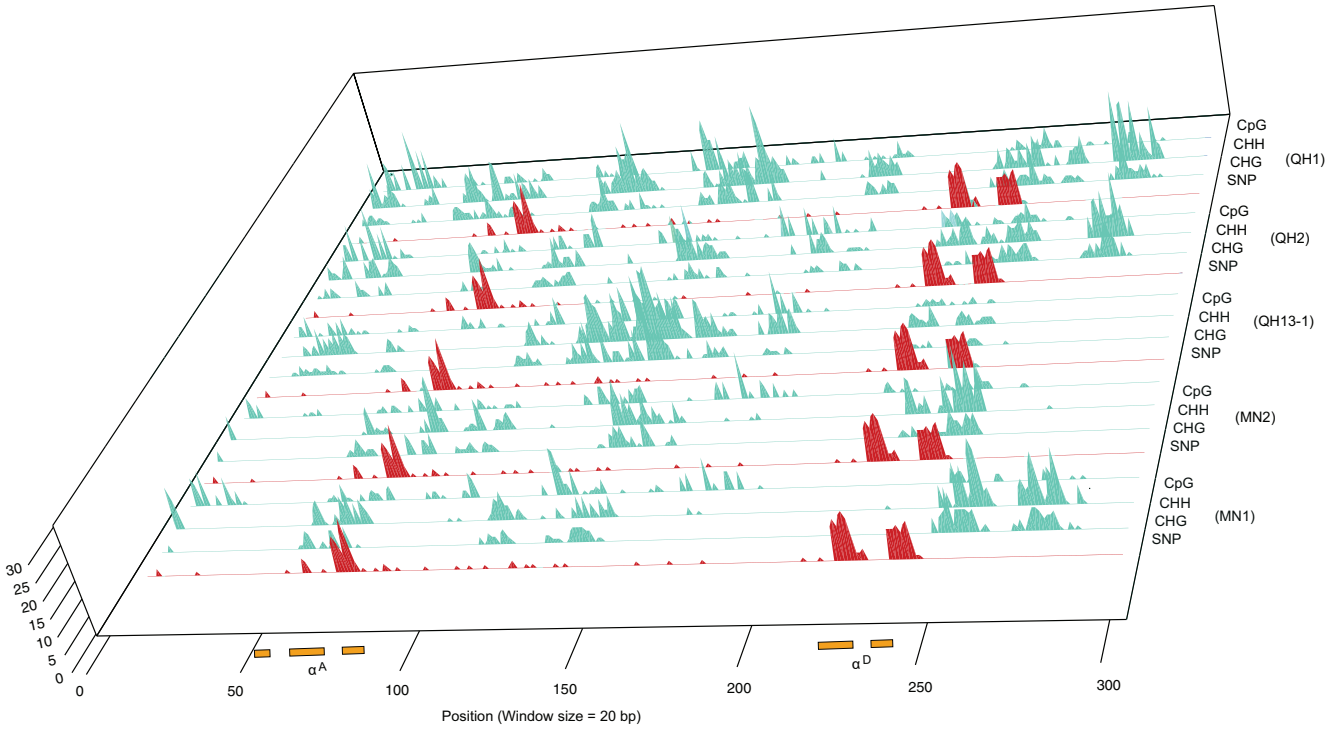

FIG. 4. Distribution of CpG, CHH, CHG, and RNA variants in α globin for five falcons (*Y* axis). In each sample, RNA variant is in red, and sequentially followed by CHG, CHH, and CpG in green. The window size is 20 bp. Exons of α^A and α^D are in orange.

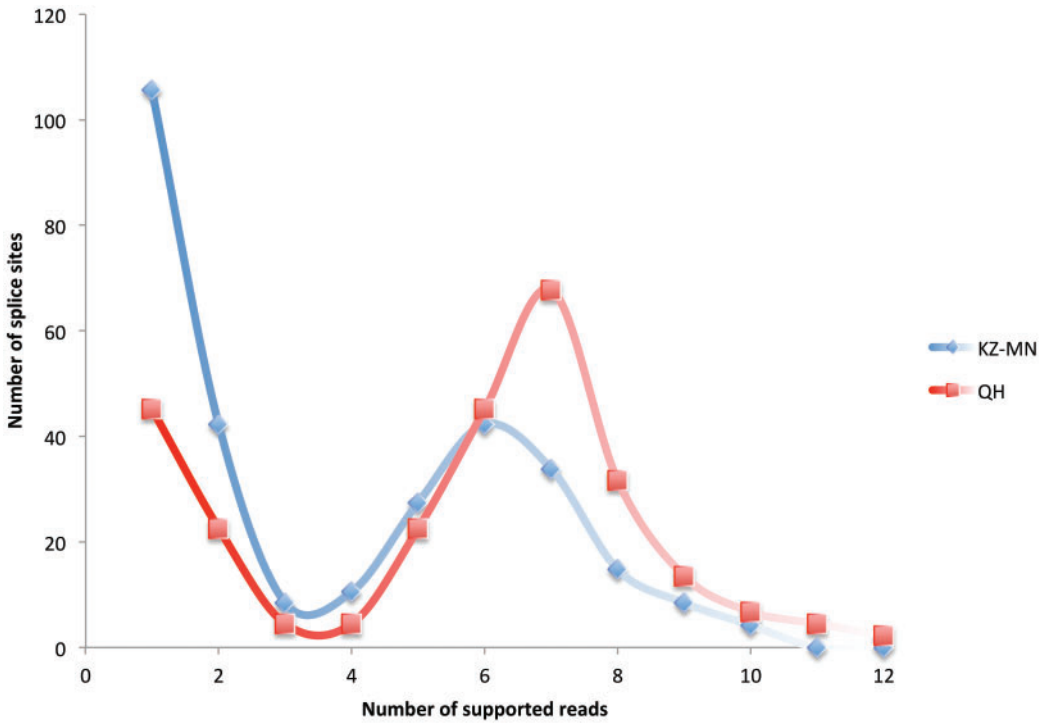

FIG. 5. Statistics of RNA-seq reads supporting splice sites in hemoglobin isoforms (excluding canonical ones) in QH (red) and KZ–MN (blue) falcon populations.

population with west (MD and SK) falcons (supplementary fig. S22, Supplementary Material online).

Because hemoglobin is the main oxygen carrier in the blood, the higher global expression of hemoglobin genes and higher diversity of higher than average expressed isoforms in QH falcons may enhance their capacity of transport oxygen (Storz and Moriyama 2008; DuBay and Witt 2014), enabling them to cope with hypoxia stress at high altitudes. A mechanistic understanding of effects of these changes in RNA expression and diversity on

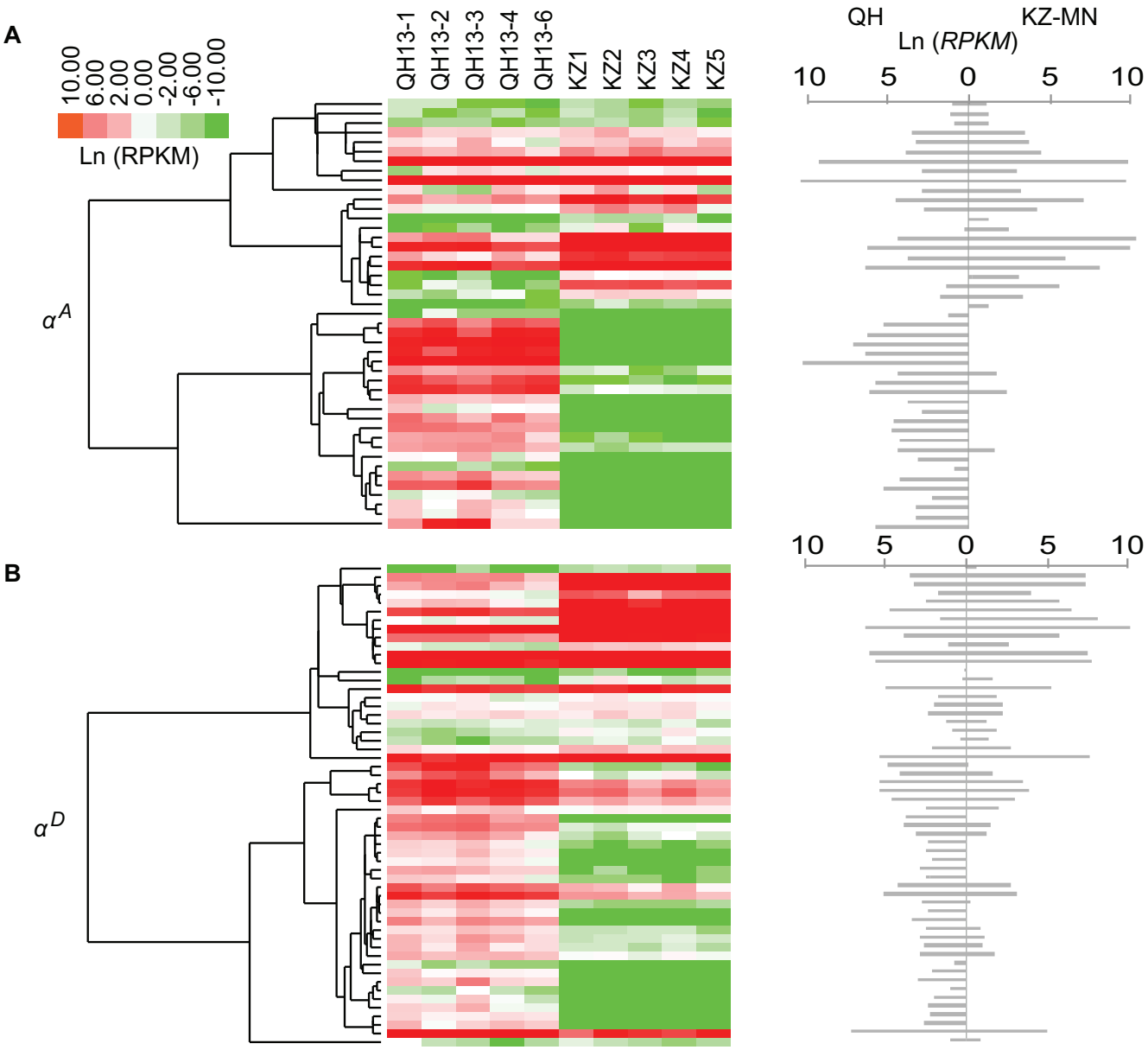

Fig. 6. Expression pattern of the α^A (*A*) and α^D (*B*) isoforms in the high altitude (QH) and low lowland (KZ–MN) populations of falcons. In each subfigure, the expression profiles of each isoform for each falcon in the two populations are shown on the left, and the mean expression values (*RPKM*) of each isoform in each population are shown on the right. The length of the gray bar is proportional to *RPKM*.

oxygen transport efficiency is a key requirement for further research.

Discussion

Our transcriptome analysis shows that *de novo* splice sites accounted for ~60% of the total splice sites identified in transcripts from the falcons. Remarkably, in the four most highly expressed falcon hemoglobin genes, *de novo* somatic splice sites amounted to over 90% of the total splice sites (supplementary fig. S23, Supplementary Material online), giving rise to a large number of alternative splicing isoforms (e.g., 81 in α globin). We observed extraordinarily high frequencies of TAM in these four genes, and proved that TAM, rather than RNA editing, produced most of *de novo* splice sites that we identified. Our results therefore reveal a new and highly active mode for the generation of transcriptome complexity (fig. 7).

Our results also provide some information on the mechanism of TAM, and the factors which counteract it. The abundant TAMs that we observed for highly expressed genes is likely to result from the lengthy exposure of single stranded DNA within the R-loop during the incessant transcription of these genes. We found that TAM occurs mainly on nontranscribed strands, in accordance with previous reports (Francino and Ochman 1997; Gaillard *et al.* 2013). TAM is known to be counteracted by DNA repair during transcription (Park *et al.* 2012), so this strand bias could be because the transcribed DNA strand is preferentially repaired compared with the nontranscribed strand in expressed genes (Bohr *et al.* 1985; Mellon *et al.* 1987; Mellon and Hanawalt 1989;

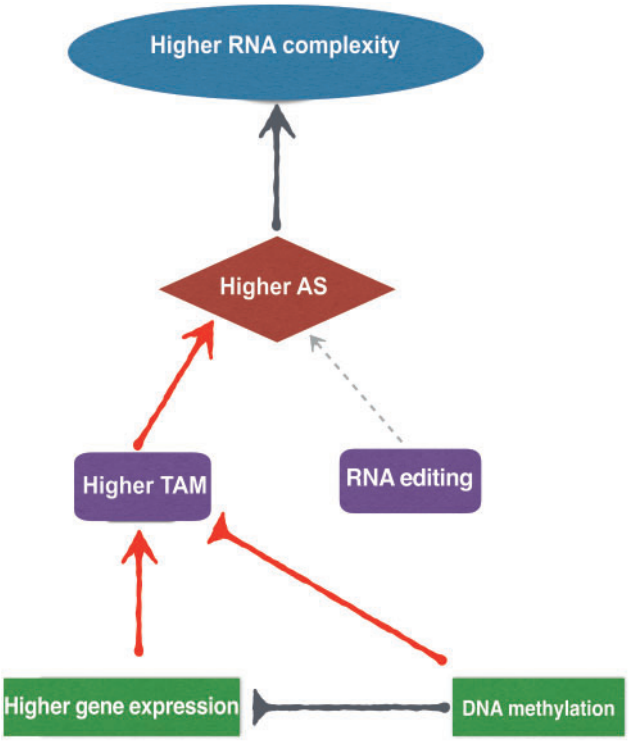

FIG. 7. Schematic showing how highly expressed genes generate and regulate RNA complexity.

Hanawalt *et al.* 1994). Alternatively, it might result from the protection of the transcribed strand by hybridization with the nascent RNA strand.

Taken together, our study proposes a new perspective on DNA methylation during transcription. Our analysis of falcon DNA methylomes showed that DNA methylation level in the promoter region is negatively correlated with gene expression, consistent with previous reports (Suzuki and Bird 2008). Notably, we found that there is much more DNA methylation in gene bodies than that in promoters, and this may regulate the generation of alternative splicing in these regions. Two mechanisms have been suggested by which DNA methylation could regulate splicing: modulation of the elongation rate of RNA polymerase II depending on whether the exon is excluded or included (Shukla *et al.* 2011; Ong and Corces 2014; Lev Maor *et al.* 2015), or an effect on recruiting splicing factors onto transcribed alternative exons (Lev Maor *et al.* 2015; Yearim *et al.* 2015). Both models considered only the authentic splice sites, which accounted for $<10\%$ of the total splice sites identified in the falcon hemoglobin genes studied here (supplementary fig. S23, Supplementary Material online). Our results suggest that these proposed mechanisms are inadequate to explain the regulation of the other 90% of splice sites. We found that gene regions with low methylation are prone to TAM induction in falcon hemoglobin genes, providing, to our knowledge, the first evidence that DNA methylation negatively regulates TAM (fig. 7). We propose that DNA methylation on the nontranscribed strand may spatially regulate the TAM events by promoting the formation of secondary DNA structures of single stranded DNA exposed during transcription, which represses TAM.

The discovery of this source of additional variation raises a question of whether somatic mutations caused by TAM have functional roles that could be selected for. We have shown that TAM mutations on the nontranscribed stands in the hemoglobin genes can be replicated onto transcribed DNA strands and become fixed in the daughter cells (supplementary fig. S18, Supplementary Material online), creating *de novo* splice sites and generating distinct isoforms in somatic cell clones. This has the potential to change or expand biological functions (e.g., increase in affinity of hemoglobin for oxygen). Previous research suggested that mutations caused by TAM are generally deleterious (Park *et al.* 2012), and in this study, we estimated that $\sim$40% of the TAM induced isoforms were nonfunctional, consistent with this expectation.

However, our comparative transcriptome analysis suggested that many TAM induced isoforms may be physiologically significant, because we discovered a generally elevated expression and higher transcript diversity of medium to highly expressed hemoglobin in QH falcons living at high altitudes. This expanded and fine-tuned repertoire of hemoglobin expression may underpin their response to the hypoxia environment on the plateau. Thus, genes with high expression, as well as having essential function in a cell (Zhang and He 2005; Gout 2010), may have their repertoire expanded by TAM, providing an evolutionary strategy that contributes to lifestyle adaptations.

Materials and Methods

Sample Preparation and Extraction

Blood samples were collected from 30 saker falcons and 12 chickens at 3–5 weeks in age. For saker falcons, four chicks were from Moldova, seven from Slovakia, ten from Qinghai, China, four from Mongolia, and five from Kazakhstan. Approximately 0.1 ml of blood was taken from each chick for RNA extraction and $\sim$25 μl for DNA extraction. For chickens, four individuals from each of three breeds (Chuahua, White Leghorn, and Tibetan breeds) were collected. About 150 μl of blood from each chicken was used for RNA and 25 μl for DNA extraction. Total RNA extraction was performed using the RNeasy Protect Animal Blood Kit (Qiagen), followed by mRNA purification, library construction, and sequencing as described previously (Zhan *et al.* 2013). Genomic DNA was extracted following the instruction of Blood & Cell Culture DNA Kit (Qiagen). Libraries with insert size of 300 bp were constructed, followed by NGS sequencing on an Illumina Hiseq2500 Genome Analyzer.

Data Generation

For each RNA sample, 200-bp fragments were collected by gel purification after ligation with an Illumina paired-end adapter oligo mix, followed by library construction (Illumina). The libraries were then sequenced using a Hiseq2000 Genome Analyzer (BGI-Shenzhen) and a Hiseq2500 for chickens (Novogene). Raw data were filtered as previously described (Pan *et al.* 2017). As a result, $\sim$3.5-Gb clean data were obtained for each saker and 5 Gb for each chicken, yielding the total of 114.75 Gb data from sakers and 61.8 Gb from

chickens (supplementary tables S1 and S2, Supplementary Material online).

Variant Detection

We sought to identify four types of variant: between-transcript RNA variations, indels (insertions and deletions), inversions, and duplications. RNA-seq reads were first mapped to the falcon (Zhan *et al.* 2013), chicken, and mouse reference genomes (Ensembl 83) using *SOAP* v2.21. SNP calling of each species was determined using *SOAPsnp* v1.04 with the criteria previously reported (Pan *et al.* 2017). Small insertions and deletions were identified using *SOAPindel* v2 (Li *et al.* 2013), and inversions and duplications using *SOAPsv* (Lin *et al.* 2015).

Splice Junction Prediction

To identify splice junctions, we first aligned the filtered RNA-seq reads for each falcon to the falcon reference genome using *Bowtie* 0.12.7. Splice junctions were then predicted using *Tophat* 1.3.3 with default parameters. Based on the splice junctions identified at each gene locus in a sample, we used the following approach to estimate the isoform number. Assuming that there are m donor and n acceptor sites in a given gene, it must possess at least $|(n-m)|+1$ isoforms. According to this rule, when $m = n$, there is a single isoform predicted. The same pipeline was adopted for splice junction identification in the chicken ($N = 29$ in supplementary table S4, Supplementary Material online and the 12 newly sequenced individuals) and mouse ($N = 7$). Mouse tissue data were downloaded from GenBank (supplementary table S4, Supplementary Material online).

Alternative Splicing and Variations in Falcon Hemoglobins

Spliced Intron Validation

To validate the presence of the predicted spliced introns, we need to check the accuracy of splice junctions. A total of 393 splice junctions at the α and β globin loci were randomly chosen, and 200-bp flanking each splice junction was extracted from the reference genome (Zhan *et al.* 2013). RNA-seq reads were mapped to the extracted sequences using *SOAP* v2.21 (supplementary fig. S6, Supplementary Material online). Finally, the sequencing depth of each splice junction was calculated based on the alignment results via an in-house *PERL* script.

Hemoglobin Isoform Estimation

To overcome difficulties in assembling hemoglobin isoforms using NGS RNA-seq reads, we used the pipeline illustrated in supplementary fig. S7, Supplementary Material online, to predict the hemoglobin isoforms for each falcon. Generally, candidate exons were determined by their two neighboring introns, and the three (α^A, β^H, and β^A) or two (α^D) ordered exons were considered as a candidate isoform. To obtain reliable isoforms, RNA-seq reads were mapped to these candidate isoforms, followed by read coverage estimation. To be conservative, only those isoforms with coverage $\geq$90% were accepted (supplementary table S9, Supplementary Material online). The same pipeline was used for the 12 chicken blood transcriptomes.

Third-Generation Iso-Seq Sequencing

Five micrograms of total RNA extracted from three QH saker falcons was reverse transcribed using the Clontech SMARTer cDNA synthesis kit in a PCR tube to generate full length cDNA. A single primer (primer IIA from the Clontech SMARTer kit 5′-AAGCAGTGGTATCAACGCAGAGTAC-3′) was used for the downstream PCR reactions after RT. The PCR products were then purified with AMPure PB beads and the quality control was performed based on the measurement with a 2100 BioAnalyzer (Agilent), followed by the size fractionation using the SageELF system to yield fragments with 2- to 3-kb insert-size, which was subjected to Iso-Seq SMRTBell library preparation (http://www.pacb.com/support/documentation/; last accessed May 10, 2016). The SMRTBell libraries were sequenced on a PacBio RS II platform using P6-C4 chemistry.

ToFU (Gordon *et al.* 2015) was used to determine the Iso-Seq raw reads (non-CCS subreads) as full length cDNAs if both the 5′ and 3′ cDNA primers were present and there was a polyA tail signal preceding the 3′ end. Next, an isoform-level clustering algorithm (Iterative Clustering for Error Correction) (Wang *et al.* 2016) was used to improve the consensus sequence accuracy. Consensus sequences were proofed using *Quiver* (Wang *et al.* 2016).

RT-qPCR Validation of Ten Hemoglobin Isoforms

The top ten highly expressed hemoglobin isoforms (excluding canonical isoforms) were used to conduct RT-qPCR in six individuals (QH1, QH3, QH5, MN1, KZ2, and KZ5). For each individual, RT-PCR was performed in a 20-μl reaction volume (supplementary table S17, Supplementary Material online), following the manufacturer's instructions (Promega).

Primers for the target isoforms were designed using *Primer* 3.0 (Koressaar and Remm 2007; Untergasser *et al.* 2012). For each isoform, qPCR was conducted in 20 μl using the cDNA obtained above as templates (supplementary table S18, Supplementary Material online) with *β-actin* as the reference gene (Forward primer: CACCGCAAATGCTTCTAAACC; Reverse: TTAATCCTGAGTCAAGCGCCA).

Organization and Composition of Hemoglobin Genes in Falcons

Since the organization and composition of hemoglobin genes are conserved among avian species (Opazo *et al.* 2015), we annotated each hemoglobin gene using the syntenic block analysis referring to the chicken genome.

Polymorphism Index (π) Calculation

π (Tajima 1989) was calculated for each gene based on its RNA variation using the formula:

$$\pi = \sum_{i<j}^{m} d_{ij}/(cL),$$

where d_{ij} is the nucleotide difference between individual i and individual j; m is the total number of individuals;

L is the length of the gene considered; and c is given as $m(m-1)/2$.

Distribution of RNA Variations Near De Novo Splice Sites

To investigate the distribution pattern of RNA variation near *de novo* splice sites, we randomly generated 80 SNP mutations within the DNA sequence of a given α^A gene, comparable to the average number of α^A between-transcript mutations (78) detected in the falcon population (supplementary table S14, Supplementary Material online). A *de novo* splice site is defined as in supplementary figure S1, Supplementary Material online. We examined the 50-bp flanking each *de novo* splice site, divided it (50 bp × 2) into ten windows (size = 10 bp), and counted the number of the mutations in each window. To calculate the mutation frequencies for each window (e.g., 10 bp around the splice site), we divided the total number of mutations in the 10-bp flanking regions for all the new splice sites by 10 × (the total number of new splice sites). We plotted the estimated mutation frequency along the window to obtain the simulated (random) distribution (red line in fig. 2C). This distribution was then compared with the observed RNA variation distributions in α^A gene for each individual (fig. 2C). The same pipeline was also used for the analysis of highly expressed genes in chickens and mice.

Quality Check of Heterozygous RNA Variations in α and β Globin Loci

Two factors could cause false heterozygous variation calling: misalignment with paralogs' reads and sequencing errors. To control for the first factor, only uniquely mapped reads were used for SNP calling. To estimate the effect of sequencing error, we calculated the expected proportion of reads supporting an erroneous allele assuming that the "allele" is introduced by sequencing errors at a rate of 1%, 10-folds higher than that reported for Illumina Hiseq2500 Genome Analyzer (0.1%). In our study, the expected proportion is obtained by

$$r = \frac{C_{30}^{n} \times m^{n} \times N}{N} = C_{30}^{n} \times m^{n},$$

where n is the number of individuals that have the erroneous allele; m the sequencing error rate; 30 is the sample size in the falcon population; N is the total number of reads that could be mapped to this allele. The expected proportion simulated with varying n is shown in supplementary figure S15, Supplementary Material online.

Next, for each heterozygous RNA variant at the α and β globin loci we calculated the actual proportion of reads supporting each new allele, which is different from that derived from the reference genome (Zhan *et al.* 2013). The proportion is estimated by n/N, where n is the number of reads supporting the new allele and N is the total number of reads that could be mapped to this locus in the falcon population. Results are shown in supplementary figure S15, Supplementary Material online.

Finally, we used a Poisson test to examine whether the observed proportion was significantly higher than expected for each heterozygous RNA variant.

Genomic Germline SNP Calling

A key feature of a genomic germline SNP is that the ratio of reads supporting its two alleles should not significantly deviate from 1:1. To identify germline SNPs, we mapped the NGS DNA sequencing reads from the small insert libraries (170–800 bp: accession SRP018394) onto the saker falcon reference genome (Zhan *et al.* 2013) using *SOAP* v2.21 and then called SNPs using *SOAPsnp* v1.04, in which a germline SNP is determined when the ratio of reads supporting the two alleles is 1:1.

PCR Validation of Heterozygous RNA Variants Identified in the Hemoglobin Genes

To examine whether the heterozygous RNA variants identified in α^D, α^A, β^H, and β^A genes originate from heterozygous genomic SNPs, we designed appropriate primers using *Primer* 3.0 and PCR was performed using the genomic DNA of three falcons (QH1, MN1, and SK1). The products were sequenced using an ABI 3130XL DNA Sequencer following the manufacturer's protocols. Genomic heterozygous SNPs were checked by eye and determined using *Chromas* v2.1.1. Since Sanger sequencing will rarely detect somatic mutations, heterozygous SNPs obtained here are expected to have originated in the germline.

Comparison of SNP Frequencies between Highly Expressed Genes and Intergenic Regions

Previously sequenced genomic reads (Zhan *et al.* 2013) were mapped to the falcon reference genome using *SOAP* v2.21, followed by the SNP calling using *SOAPsnp* v1.04 as described previously (Pan *et al.* 2017). SNP frequencies for intergenic regions (f) were obtained using n/N where n is the number of SNPs identified in an intergenic region and N is the total length of that region with gaps and repetitive sequences excluded. SNP frequencies for highly expressed genes ($RPKM \geq 20{,}000$) were estimated using the same equation. A Poisson test implemented in *R* was used to test for significance between the two estimates.

We conducted the same analysis with the mouse data ($N = 7$). However, as for chicken, both transcriptome and genome resequencing data are required for a comparison between highly expressed genes and intergenic regions. Therefore, we used a data set of 11 tissues that were different from that used for Splice Junction Prediction. The data were downloaded from GenBank, namely, ileum (ERR1298577), liver (ERR1298599), ovary (ERR1298617), proventriculus (ERR1298632), spinal neural tube (SRR5000696), kidney (ERR1298585), lung (ERR1298607), pancreas (ERR1298625), skin (ERR1298634), pituitary (SRR4478742), and liver under heat stress (SRR5273288). RNA variants, gene expression estimation, and mutation frequency calculation were performed using the same pipelines as for the falcons.

Strand Bias of TAM Occurrence in Hemoglobin Genes

Among the six mutation types: $C \rightarrow T$ ($G \rightarrow A$ on the opposite strand), $A \rightarrow G$ ($T \rightarrow C$), $G \rightarrow T$ ($C \rightarrow A$), $A \rightarrow T$ ($T \rightarrow A$), $G \rightarrow C$ ($C \rightarrow G$), $A \rightarrow C$ ($T \rightarrow G$), four types, $C \rightarrow T$, $A \rightarrow G$, $G \rightarrow T$, and $A \rightarrow T$ were the most frequent in TAM. To estimate whether TAM occurred preferentially on nontranscribed DNA strands in the four hemoglobin genes, we calculated the proportion of TAM events on both nontranscribed and transcribed DNA strands. A χ^2 test was used to evaluate the difference between the two estimates. A TAM induced mutation was considered to occur on nontranscribed strands if a mutation of $A \rightarrow T$, $C \rightarrow T$, $G \rightarrow T$, or $A \rightarrow G$ was observed. We also calculated the mutation proportions for intergenic regions on both DNA strands, followed by a χ^2 test for strand differences.

Somatic Mutation Identification

We carried out deep resequencing of the genomes of six falcon individuals ($\geq$30-fold coverage) using an Illumina Hiseq2500 Genome Analyzer and aligned the reads to the reference genome (Zhan *et al.* 2013). For each individual, we identified the heterozygous genomic SNPs with the *Phred* quality values >20 in each individual using an in-house *PERL* script. Identification of heterozygous genomic SNPs resulting from somatic mutation is described in supplementary note, Supplementary Material online. A heterozygous genomic SNP with the ratio of reads supporting its two alleles significantly deviating from 1:1 was considered a somatic mutation.

cDNA Cloning Experiments for TAM Validation

To verify the TAM mutations, we obtained cDNA fragments from RT-PCR RNA extracts of three falcon individuals (QH13-6, KZ4, MD2) using RT-PCR (Promega). The primers for β^H were designed from its cDNA sequence using *Primer* 3.0 (Forward: GCACTGGACAGCTGAAGAGA; Reverse: GTCC TTAGCAAAGTGGGCAG). PCR products were purified using the MiniBEST Agarose Gel DNA Extraction Kit v4 (TaKaRa) and cloned into pMD 19-T vector (TaKaRa) (Sambrook and Russell 2011). The products were sequenced using an ABI 3130XL DNA Sequencer following the manufacturer's protocols. Sequences were aligned using *ClustalW* implemented in *MEGA* v6.0.6 (Tamura *et al.* 2013) and allele variants were eye checked.

DNA Methylation Profiling

To investigate the influence of DNA methylation on TAMs, we profiled DNA methylation for five falcons (QH1, QH2, QH13-1, MN1, and MN2). Approximately 30-Gb clean data were generated for each individual (supplementary table S19, Supplementary Material online). Genomic DNA was fragmented into 200- to 300-bp fragments using Covaris S220, followed by end-repair by adding a single adenosine moiety. Adapters in which all of cytosines were methylated were ligated onto the DNA repaired end and the DNA product was treated using the EZ DNA Methylation Gold Kit (Zymo Research), during which, the unmethylated *C* was converted to *T* while methylated *C* remained unchanged. The treated product was amplified with specific methylation primers, followed by sequencing using an Illumina Hiseq2500 Analyzer.

The DNA methylation profiles of each sample were obtained using *Bismark* (Krueger and Andrews 2011), a flexible aligner, and methylation caller for Bisulfite-Seq. DNA methylation was classified into three types (CpG, CHH, and CHG), where H represents any deoxynucleotide (*A*, *T*, *C*, or *G*). When analyze methylation in the promoter region, we defined the promoter region as 1.1-kb upstream and 500-bp downstream of the transcription start site (Li *et al.* 2015). When analyze the methylation in hemoglobin genes, we calculated the methylation rate of each window (size = 20 bp) for each hemoglobin gene. It was estimated as (the number of methylated sites)/(window size).

Biological Significance of High Transcript Diversity Caused by TAM

To obtain the distribution pattern of reads supporting splice sites in hemoglobin identified by RNA-seq, we mapped the RNA-seq data to the saker genome using *Bowtie* 0.12.7, and counted the number of reads spanning each splice site individually.

To determine the number of reads supporting splice sites in the lowest abundance isoform that can be detected by Iso-Seq, we first BLASTed each isoform against the saker genome to identify the splice sites. We then extracted 200-bp flanking each splice site, and mapped RNA-seq reads to these fragments using *Bowtie* 0.12.7, and counted the number of reads supporting each splice site.

For RNA-seq data, the expression levels of hemoglobin genes were quantified using *RPKM* by mapping the RNA-seq reads of each sample to the target genes. For Iso-Seq data, the expression levels of each hemoglobin isoform were estimated by mapping the RNA-seq reads to the isoforms identified in each hemoglobin gene by the third-generation sequencing. We also performed the analysis of differential expression of Iso-Seq isoforms identified between QH and KZ–MN populations using *DESeq* (Anders and Huber 2010).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This study was supported by the Strategic Priority Program of the Chinese Academy of Sciences (XDB13000000), the National Key Programme of Research and Development, Ministry of Science and Technology (2016YFC0503200), the National Natural Science Foundation of China (No. 31471993, 31522052), the Recruitment Program of Global Youth Experts of China to X.Z., and by the Royal Society to X.Z. and M.W.B. Saker sample collection was supported by the Environment Agency Abu Dhabi and we thank the Wildlife Science and Conservation Centre of Mongolia, A. Levin, D. Ragyov, L. Deutschova, and T. Zhang for assistance. We thank X. Li for

help in modeling simulations. We acknowledge BGI-Shenzhen and Novogene for data generation.

Author Contributions

X.Z. and S.P. designed the study; S.P., X.Z., and Z.G. analyzed the data; A.D. and X.Z. sampled the falcons; X.D., X.H., and Z.L. sampled the chicken; X.Z. extracted the RNA of saker falcons; X.H. extracted the RNA of chicken; Y.W. extracted the DNA of chicken; Y.W. and Z.L. performed the experimental validation; S.P., X.Z., M.W.B., and J.A.M.G. wrote the paper. All authors read and provided input into the manuscript and approved the final version.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol*. 11(10):R106.

Bohr VA, Smith CA, Okumoto DS, Hanawalt PC. 1985. DNA repair in an active gene: removal of pyrimidine dimers from the *DHFR* gene of CHO cells is much more efficient than in the genome overall. *Cell* 40(2):359–369.

Brennicke A, Marchfelder A, Binder S. 1999. RNA editing. *FEMS Microbiol Rev*. 23(3):297–316.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the *Drosophila* transcriptome. *Nature* 512(7515):393–399.

Burset M, Seledtsov IA, Solovyev VV. 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res*. 28(21):4364–4375.

Chen L, Tovar-Corona JM, Urrutia AO. 2012. Alternative splicing: a potential source of functional innovation in the eukaryotic genome. *Int J Evol Biol*. 2012:596274.

DuBay SG, Witt CC. 2014. Differential high-altitude adaptation and restricted gene flow across a mid-elevation hybrid zone in Andean tit-tyrant flycatchers. *Mol Ecol*. 23(14):3551–3565.

Francino M, Ochman H. 1997. Strand asymmetries in DNA evolution. *Trends Genet*. 13(6):240–245.

Frésard L, Leroux S, Roux PF, Klopp C, Fabre S, Esquerré D, Dehais P, Djari A, Gourichon D, Lagarrigue S, et al. 2015. Genome-wide characterization of RNA editing in chicken embryos reveals common features among vertebrates. *PLoS One* 10(5):e0126776.

Gaillard H, Herrera-Moyano E, Aguilera A. 2013. Transcription-associated genome instability. *Chem Rev*. 113(11):8638–8661.

Glass J, Lavidor LM, Robinson SH. 1975. Studies of murine erythroid cell development. Synthesis of heme and hemoglobin. *J Cell Biol*. 65(2):298–308.

Gordon SP, Tseng E, Salamov A, Zhang J, Meng X, Zhao Z, Kang D, Underwood J, Grigoriev IV, Figueroa M, et al. 2015. Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. *PLoS One* 10(7):e0132628.

Gott JM, Emeson RB. 2000. Functions and mechanisms of RNA editing. *Annu Rev Genet*. 34:499–531.

Gout J, Kahn D, Duret L. Paramecium Post-Genomics Consortium. 2010. The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet*. 6(5):e1000944.

Graveley BR. 2005. Mutually exclusive splicing of the insect *Dscam* pre-mRNA directed by competing intronic RNA secondary structures. *Cell* 123(1):65–73.

Grishkevich V, Yanai I. 2014. Gene length and expression level shape genomic novelties. *Genome Res*. 24(9):1497–1503.

Hanawalt PC, Donahue BA, Sweder KS. 1994. Repair and transcription. Collision or collusion? *Curr Biol*. 4(6):518–521.

He T, Lei W, Ge C, Du P, Wang L, Li F. 2015. Large-scale detection and analysis of adenosine-to-inosine RNA editing during development in *Plutella xylostella*. *Mol Genet Genomics* 290(3):929–937.

Jackson IJ. 1991. A reappraisal of non-consensus mRNA splice sites. *Nucleic Acids Res*. 19(14):3795–3798.

Jin Y, Tian N, Cao J, Liang J, Yang Z, Lv J. 2007. RNA editing and alternative splicing of the insect *nAChR* subunit alpha6 transcript: evolutionary conservation, divergence and regulation. *BMC Evol Biol*. 7:98.

Keegan LP, Gallo A, O'Connell MA. 2001. The many roles of an RNA editor. *Nat Rev Genet*. 2(11):869–878.

Keren H, Lev-Maor G, Ast G. 2010. Alternative splicing and evolution: diversification, exon definition and function. *Nat Rev Genet*. 11(5):345–355.

Kim N, Robertson SJ. 2012. Transcription as a source of genome instability. *Nat Rev Genet*. 13(3):204–214.

Koressaar T, Remm M. 2007. Enhancements and modifications of primer design program *Primer*3. *Bioinformatics* 23(10):1289–1291.

Krueger F, Andrews SR. 2011. *Bismark*: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* 27(11):1571–1572.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 10(3):R25.

Lev Maor G, Yearim A, Ast G. 2015. The alternative role of DNA methylation in splicing regulation. *Trends Genet*. 31(5):274–280.

Li Q, Wang Y, Hu X, Zhao Y, Li N. 2015. Genome-wide mapping reveals conservation of promoter DNA methylation following chicken domestication. *Sci Rep*. 5:8748.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009. SNP detection for massively parallel whole-genome resequencing. *Genome Res*. 19(6):1124–1132.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. *SOAP2*: an improved ultrafast tool for short read alignment. *Bioinformatics* 25(15):1966–1967.

Li S, Li R, Li H, Lu J, Li Y, Bolund L, Schierup MH, Wang J. 2013. *SOAPindel*: efficient identification of indels from short paired reads. *Genome Res*. 23(1):195–200.

Li S, Tighe SW, Nicolet CM, Grove D, Levy S, Farmerie W, Viale A, Wright C, Schweitzer PA, Gao Y, et al. 2014. Multi-platform assessment of transcriptome profiling using RNA-seq in the ABRF next-generation sequencing study. *Nat Biotechnol*. 32(9):915–925.

Lin J, Cheng Z, Xu M, Huang Z, Yang Z, Huang X, Zheng J, Lin T. 2015. Genome re-sequencing and bioinformatics analysis of a nutraceutical rice. *Mol Genet Genomics* 290(3):955–967.

Mellon I, Hanawalt PC. 1989. Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. *Nature* 342(6245):95–98.

Mellon I, Spivak G, Hanawalt PC. 1987. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian *DHFR* gene. *Cell* 51(2):241–249.

Mugal CF, von Grünberg HH, Peifer M. 2009. Transcription-induced mutational strand bias and its effect on substitution rates in human genes. *Mol Biol Evol*. 26(1):131–142.

Nellore A, Jaffe AE, Fortin JP, Alquicira-Hernández J, Collado-Torres L, Wang S, Phillips RA III, Karbhari N, Hansen KD, Langmead B. 2016. Human splicing diversity and the extent of unannotated splice junctions across human RNA-seq samples on the Sequence Read Archive. *Genome Biol*. 17(1):266.

Ng B, Yang F, Huston DP, Yan Y, Yang Y, Xiong Z, Peterson LE, Wang H, Yang XF. 2004. Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes. *J Allergy Clin Immunol*. 114(6):1463–1470.

Ong CT, Corces VG. 2014. *CTCF*: an architectural protein bridging genome topology and function. *Nat Rev Genet*. 15(4):234–246.

Opazo JC, Hoffmann FG, Natarajan C, Witt CC, Berenbrink M, Storz JF. 2015. Gene turnover in the avian globin gene families and evolutionary changes in hemoglobin isoform expression. *Mol Biol Evol*. 32(4):871–887.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet*. 40(12):1413–1415.

Pan S, Zhang T, Rong Z, Hu L, Gu Z, Wu Q, Dong S, Liu Q, Lin Z, Deutschova L, et al. 2017. Population transcriptomes reveal

synergistic responses of DNA polymorphism and RNA expression to extreme environments on the Qinghai-Tibetan Plateau in a predatory bird. *Mol Ecol*. 26(11):2993–3010.
Park C, Qian W, Zhang J. 2012. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep*. 13(12):1123–1129.
Petschek JP, Scheckelhoff MR, Mermer MJ, Vaughn JC. 1997. RNA editing and alternative splicing generate mRNA transcript diversity from the *Drosophila 4f-rnp* locus. *Gene* 204(1–2):267–276.
Reuter JS, Mathews DH. 2010. *RNAstructure*: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11:129.
Rosenthal JJ. 2015. The emerging role of RNA editing in plasticity. *J Exp Biol*. 218(12):1812–1821.
Rueter SM, Dawson TR, Emeson RB. 1999. Regulation of alternative splicing by RNA editing. *Nature* 399(6731):75–80.
Sambrook J, Russell DW. 2001. Molecular cloning: a laboratory manual. New York: Cold Spring Harbor Laboratory Press.
Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S. 2011. *CTCF*-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* 479(7371):74–79.
Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H. 2005. Function of alternative splicing. *Gene* 344(3):1–20.
Storz JF, Moriyama H. 2008. Mechanisms of hemoglobin adaptation to high altitude hypoxia. *High Alt Med Biol*. 9(2):148–157.
Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 9(6):465–476.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585–595.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. *MEGA6*: molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol*. 30(12):2725–2729.
Taqi MM, Wärmländer SK, Yamskova O, Madani F, Bazov I, Luo J, Zubarev R, Verbeek D, Gräslund A, Bakalkin G. 2012. Conformation effects of CpG methylation on single-stranded DNA oligonucleotides: analysis of the opioid peptide dynorphin-coding sequences. *PLoS One* 7(6):e39605.
Tian N, Yang Y, Sachsenmaier N, Muggenhumer D, Bi J, Waldsich C, Jantsch MF, Jin Y. 2011. A structural determinant required for RNA editing. *Nucleic Acids Res*. 39(13):5669–5681.
Trapnell C, Pachter L, Salzberg SL. 2009. *TopHat*: discovering splice junctions with RNA-seq. *Bioinformatics* 25(9):1105–1111.
Tümer Z. 2013. An overview and update of *ATP7A* mutations leading to Menkes disease and occipital horn syndrome. *Hum Mutat*. 34(3):417–429.
Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. *Primer*3-new capabilities and interfaces. *Nucleic Acids Res*. 40(15):e115.
Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. 2016. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. *Nat Commun*. 7:11708.
Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.
Xu G, Zhang J. 2014. Human coding RNA editing is generally nonadaptive. *Proc Natl Acad Sci U S A*. 111(10):3769–3774.
Yearim A, Gelfman S, Shayevitch R, Melcer S, Glaich O, Mallm JP, Nissim-Rafinia M, Cohen AH, Rippe K, Meshorer E, et al. 2015. *HP1* is involved in regulating the global impact of DNA methylation on alternative splicing. *Cell Rep*. 10(7):1122–1134.
Zhan X, Pan S, Wang J, Dixon A, He J, Muller MG, Ni P, Hu L, Liu Y, Hou H, et al. 2013. Peregrine and saker falcon genome sequences provide insights into evolution of a predatory lifestyle. *Nat Genet*. 45(5):563–566.
Zhang J, He X. 2005. Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol*. 22(4):1147–1155.
Zhang Z, Qian W, Zhang J. 2009. Positive selection for elevated gene expression noise in yeast. *Mol Syst Biol*. 5(1):229.