Non-neutral evolution of H3.3-encoding genes occurs without alterations in protein sequence

Brejnev Muhire[1], Matthew Booker[1,2] and Michael Tolstorukov[1,2,*]

[1]Department of Molecular Biology, Massachusetts General Hospital and Harvard Medical School, Boston, MA 02114

[2]Dana-Farber Cancer Institute, Boston, MA 02215

*correspondence: tolsotrukov@molbio.mgh.harvard.edu

Abstract

Histone H3.3 is a developmentally essential variant encoded by two independent genes in human (*H3F3A* and *H3F3B*). While this two-gene arrangement is evolutionarily conserved, its origins and function remain unknown. Phylogenetics, synteny and gene structure analyses of the H3.3 genes from 32 metazoan genomes indicate independent evolutionary paths for *H3F3A* and *H3F3B*. While *H3F3B* bears similarities with H3.3 genes in distant organisms and with canonical H3 genes, *H3F3A* is sarcopterygian-specific and evolves under strong purifying selection. Additionally, *H3F3B* codon-usage preferences resemble those of broadly expressed genes and 'cell differentiation-induced' genes, while codon-usage of *H3F3A* resembles that of 'cell proliferation-induced' genes. We infer that *H3F3B* is more similar to the ancestral H3.3 and is likely evolutionarily adapted for broad expression pattern in diverse cellular programs, while H3F3A adapted for a subset of gene expression programs. Thus, the arrangement of two independent H3.3 genes facilitates fine-tuning of H3.3 expression across cellular programs.

Introduction

In eukaryotic cells genomic DNA is packaged into chromatin, which plays a dual role of genome compaction and regulation [1]. Basic repeating units of chromatin, called nucleosomes, comprise 147bp of DNA wrapped around a core that is formed by histone proteins of four types (H2A, H2B, H3, and H4), which are conserved from yeast to human [2,3]. The histones fall into two major types: replication-dependent (RD) canonical histones and replication-independent (RI) non-canonical variants. The RI histone variants have diverse biological roles and are part of epigenetic regulation of genome function [4–6]. Unlike the canonical histones that are encoded by co-regulated gene clusters (histone loci) [3], RI variants are encoded by individual genes that are regulated similarly to other protein coding genes.

One of the most studied histone variants is H3.3, which replaces canonical histone H3 and functionally can be associated with both gene activation [7,8] and silencing [9–11]. H3.3 variant is expressed and deposited throughout the cell-cycle independent of DNA replication [12–14]. In human genome H3.3 can be transcribed from either of two independent genes (*H3F3A* and *H3F3B*), which are located at different chromosomes, 1 and 17 respectively. These genes differ at the nucleotide level both within introns and exons, even though both of them encode exactly the same amino-acid sequence. Presence of multiple independent genes encoding H3.3 is also conserved in other organisms, including distant species such as fruit fly [15]. Moreover, despite absolute conservation at the protein level, the mutational profiles of *H3F3A* and *H3F3B* genes in human cancers differ substantially. For instance, mutation K27M was reported in only in *H3F3A* in brainstem gliomas [16], while mutation K36M is predominantly observed in *H3F3B* in bone cancers, such as chondroblastoma [17,18]. The regulatory genomic elements associated with these genes are also distinct, and the over-expression of *H3F3A* but not *H3F3B* is

implicated in lung cancer through aberrant H3.3 deposition [19]. Taken together, these observations indicate that while *H3F3A* and *H3F3B* encode the same protein product, they are under different regulatory mechanisms and play distinct roles.

Evolution of H3.3 encoding genes was analyzed in *Drosophila* species [20], however, on a larger scale, the biological function and evolutionary history of such two-gene organization remains unclear, despite its biomedical significance [21,22]. To approach these questions, we compared the sequences and genomic arrangements of the H3.3 genes from 32 metazoan genomes. Using phylogenetics, sequence identity, gene structure and synteny analyses we infer that *H3F3A* is sarcopterygian-specific (tetrapod and lobe-finned fish) gene, while *H3F3B* is of more ancient origin. Furthermore, analysis of codon-usage preferences in each of the H3.3 genes revealed that *H3F3B* is evolutionarily adapted for broad expression patterns across diverse cellular programs, including cell differentiation, while *H3F3A* is more fine-tuned for a specific transcriptional program associated with cell proliferation. This observation of coding sequence optimization for distinct transcriptional programs provides insight into why both *H3F3A* and *H3F3B* have been maintained in course of evolution, even though they produce identical proteins.

Results

Phylogenetic analyses of H3.3-encoding genes in metazoa

We identified the H3.3 coding sequences from the genomes of 32 metazoa organisms, primarily vertebrates, and used them in our analysis. We observed that two ‘independent’ genes (i.e. located in different genomic loci and controlled by distinct, non-overlapping promoters) encode histone H3.3 in all analyzed organisms except for actinopterygii (ray-finned fish lineage) and

73 coelacanth where H3.3 is encoded by three or five genes (Table S1). The high number of H3.3
74 genes in actinopterygians most likely resulted from whole genome duplication events [23–26]
75 and partial chromosome duplication events [27–29] that occurred in this lineage during
76 evolution. With this exception, the arrangement of two H3.3 genes is widespread among
77 vertebrates and is observed even in more distant metazoa such as flies, nematodes, and some
78 plants [30]. Remarkably, the encoded protein sequence is identical in all vertebrates and
79 *Drosophila melanogaster* (Fig S1). The existence of two independent genes that encode an
80 identical protein allows us to focus on analysis of the evolutionary pressure acting on these
81 genes at the nucleotide rather protein level.
82
83 Next, we analyzed phylogenetic relationship of the H3.3 genes in metazoa. The coding
84 sequences of these genes form several distinct groups in the phylogenetic tree, including two
85 major groups (clades 1 and 3), one minor group (clade 2) and outgroups of lamprey and fly
86 H3.3 genes (Fig. 1A). Clade 1 (shown in brown) consists exclusively of sarcopterygian *H3F3A*
87 genes (the lobe-finned fish lineage, including all tetrapods and coelacanth). Clade 3 comprises
88 all sarcopterygian *H3F3B* genes (blue) along with the majority of actinopterygian H3.3 genes
89 and the third coelacanth H3.3 gene (gray). We note that this clade also includes a 'hominid-
90 specific' gene H3F3C (green), which emerged as a recent retro-transposition of *H3F3B* [31].
91 *H3F3C* encodes another replacement histone from H3 family, H3.5, that differs from the histone
92 H3.3 by several amino-acids, and it was included in this analysis for further comparison. The
93 confident assignment of *H3F3C* to clade 3 that contains H3F3B genes (branch support=1),
94 highlights that the distinction between the coding sequences (CDS) of the genes forming clades
95 1 (*H3F3A*) and 3 (*H3F3B*) is substantial and evolutionary stable even though these genes
96 encode the same protein H3.3 (no amino-acid difference). Finally, clade 2 contains remaining

97 actinopterygian H3.3 genes that cluster neither with sarcopterygian *H3F3A* nor with
98 sarcopterygian *H3F3B*. This analysis gives the first evidence that, compared to sarcopterygian
99 *H3F3A*, sarcopterygian *H3F3B* is likely more evolutionarily related to actinopterygian H3.3
100 genes.
101
102 The observed relation between *H3F3B* and actinopterygian genes was confirmed by
103 comparison of the intron-exon structure of all H3.3-encoding genes throughout the species. In
104 sarcopterygian genomes *H3F3B* is generally shorter, spanning ~2-4kb with a total length of
105 introns ~0.16-1kb (Fig. 1B). *H3F3B* structure is similar to that of actinopterygian H3.3 (gene
106 length is approximately ~2-6kb and total intron length is ~0.16-4kb; Fig. 1C). The *H3F3A* gene
107 structure is noticeably different, with gene length spanning ~9-13kb and total intron length being
108 ~4.5-10kb (Fig. 1D). Thus, the intron-exon structure of sarcopterygian *H3F3B*, and not
109 sarcopterygian *H3F3A* is more similar to the actinopterygian H3.3 genes and H3.3 genes in
110 more distant vertebrates actinopterygians, lamprey, fly and worm, consistent with our previous
111 observations.
112
113 To further support these results, we carried out synteny analysis to determine whether genes
114 around *H3F3A* or *H3F3B* are evolutionary conserved in non-tetrapod organisms. We first used
115 Genomicus 80.01, a web-based synteny visualization tool that uses Ensembl database
116 comparative genomic data [32]. Comparison between human and actinopterygii shows no
117 syntenic genes conserved around human *H3F3A* and H3.3 genes in actinopterygian species
118 (Fig 2A), but at least six syntenic genes can be identified around human *H3F3B* and H3.3 genes
119 in four actinopterygian species (fugu, platyfish, spotted gar, and tetraodon) (marked with a blue
120 star, Fig 2B).

121

122 We extend this analysis to all tetrapods and distant metazoa (lamprey, fly and worm), by
123 implementing a flexible synteny detection method allowing the user to quantitatively measure
124 the degree of gene conservation around loci of interest in two genomes (see Methods).
125 Specifically, we compared 30 genes upstream and downstream of each of the H3.3 genes and
126 the degree of gene conservation was determined by sequence identities computed
127 independently for both coding sequences and translated amino-acid sequences. While we
128 found clear evidence of synteny conservation around both H3.3 genes in tetrapods, it was
129 consistently higher around *H3F3A* than *H3F3B*. For instance, the ratios of syntenic genes
130 around *H3F3A* to those around *H3F3B* were 25/17, 12/6, 12/6 for the human-mouse, human-
131 lizard and human-zebra finch comparisons respectively (Fig S2A). At the same time, we found
132 no synteny conservation around tetrapod *H3F3A* and actinopterygians H3.3. In contrast, for
133 *H3F3B* we found the same six genes conserved between tetrapods and one of the tetraodon
134 H3.3 genes, which were detected by Genomicus, and a weak conservation of these genes in
135 zebrafish and medaka (marked with a blue star Fig S2B and Fig 1A).

136

137 From these observations, we conclude that orthologs of mammalian *H3F3A* and *H3F3B* are
138 present in the coelacanth genome (i.e. throughout the sarcopterygian lineage). Sarcopterygian
139 *H3F3B* is evolutionarily related to many actinopterygian H3.3 genes while sarcopterygian
140 *H3F3A* seems to have no counterpart in actinopterygian lineage (Fig 1A). We infer that the
141 sarcopterygian-specific *H3F3A* clade with a long and well-supported branch (branch support=1,
142 Fig 1A) is consistent with one of the following scenarios: (i) the counterpart of *H3F3A* was lost
143 in actinopterygian lineage soon after actinopterygian-sarcopterygian split, or (ii) since the
144 actinopterygian/sarcopterygian split either an existing or a newly emerged H3.3 gene

145 underwent rapid evolution towards the current *H3F3A* form. We aimed to distinguish these
146 possibilities by the analysis described below.
147
148 **Comparison of H3.3 genes between sarcopterygians and distant metazoa**
149 One can expect that if *H3F3A* were lost in actinopterygians, both *H3F3A* and *H3F3B* would
150 exhibit roughly equal similarity to H3.3 genes in more distant metazoa. Thus, to resolve the
151 scenarios described above we directly compared the similarity of sarcopterygian *H3F3A* and
152 *H3F3B* to the H3.3 genes of actinopterygians and distant organisms (lamprey and fly) (Fig. 3).
153 We also included in this analysis genes encoding the RD canonical histones H3.1 and H3.2
154 because these genes emerged from ancient gene duplication event that resulted in a separation
155 of replication-dependent and replication-independent histones [33]. As sarcopterygian genes in
156 this analysis, we used coelacanth *H3F3A* and *H3F3B*. Coelacanth can be expected to show
157 more similarity with distant organisms than other sarcopterygians, in part because its protein-
158 coding genes evolved twice as slow as such genes in tetrapods [34], which makes it especially
159 suitable for this comparison.
160
161 This analysis revealed that most of the actinopterygian H3.3 genes and RD H3.1 and H3.2-
162 encoding gene of bony vertebrates (tetrapods and zebrafish) are more similar to sarcopterygian
163 *H3F3B* than to *H3F3A* (Fig. 3). This trend further extends to both lamprey H3.3 genes and one
164 fly H3.3 (chr2L) gene. In addition, *H3F3C* is also more similar to coelacanth *H3F3B* than *H3F3A*
165 as expected. Overall, only tetrapod *H3F3A* genes can be confidently 'assigned' to coelacanth
166 H3F3A. As a control, we have repeated this analysis using tetrapods (human, mouse and zebra
167 finch) *H3F3A* and *H3F3B* genes instead of coelacanth genes and observed similar trends (Fig.
168 S3). Overall, these results reveal that in comparison to *H3F3A*, sarcopterygian *H3F3B* is more

169 similar to the H3.3 genes in distant metazoa and to RD H3 genes, pointing to a possibility that
170 *H3F3B* is more similar to the ancestral form of the H3.3 gene.
171
172 Additional evidence supporting the hypothesis formulated above comes from the comparison
173 of the 3' untranslated regions (3'UTRs) of the H3.3 genes (Fig S4). UTRs are among the most
174 conserved non-coding sequences in eukaryotes [35,36], and the 3'UTRs of H3.3 genes are
175 similarly evolutionarily conserved (~60-80%) among tetrapods and actinopterygians. We
176 validated this approach by confirming that it produces results consistent with the phylogenetic
177 analysis of the coding H3.3 sequences when applied to genes from clades 1 and 3 (Fig. 1A),
178 containing sarcopterygian H3.3 genes. When we applied this approach to genes from other
179 clades, we observed that in every analyzed non-sarcopterygian organism (actinopterygian
180 species, lamprey, fly and worm), at least one H3.3 gene has higher similarity of its 3'UTRs to
181 that of tetrapod *H3F3B* (~75% identity) compared to tetrapod *H3F3A* (~60% identity) (Fig. S4A-
182 B). These organisms are marked with blue star in Fig 1A. There were no instances of a non-
183 tetrapod H3.3 3'UTR being more similar to the 3'UTR of tetrapod *H3F3A*.
184
185 Collectively, our results indicate that gene *H3F3A* is sarcopterygii-specific, while gene *H3F3B*
186 is evolutionary related to actinopterygian H3.3 genes as well as to the H3.3 genes in more
187 distant metazoans. Furthermore, our results suggest that *H3F3B* is more directly related to the
188 ancestral form of the H3.3 gene. We find that the possibility of a lineage-specific loss of *H3F3A*
189 in the actinopterygians is less plausible than the hypothesis of an existing or newly emerged
190 H3.3 gene copy that underwent rapid evolution to become *H3F3A* in sarcopterygian lineage.
191
192 **Distinct selection pressures within tetrapod *H3F3A* and *H3F3B* CDS**

193 The conservation of the arrangement of two distinct genes encoding the same protein suggests
194 functional significance. To investigate how potential functional differences between these two
195 genes may be reflected in their genomic sequences, we measured selection pressure operating
196 at the nucleotide level in *H3F3A* and *H3F3B*. Due to lack of variation among H3.3 protein
197 sequences in analyzed organisms, the methods based on non-synonymous and synonymous
198 substitution rates often used for detection of natural selection [37–39] are not suitable. Instead,
199 we investigate purifying selection operating on *H3F3A* and *H3F3B* genes based on the degree
200 of conservation of coding nucleotide-sequence in tetrapod organisms.
201
202 We started with calculating pairwise genetic distances between the tetrapod H3.3 genes,
203 defined here as the numbers of the observed nucleotide substitutions divided by the CDS length
204 (i.e. the "nucleotide substitution scores"). As a control, we also included in this analysis the
205 *H2AFZ* gene, which encodes the conserved replacement histone H2A.Z. Overall, we observed
206 that while *H3F3B* is not conserved significantly stronger than *H2AFZ* (P = 0.244, Mann-
207 Whitney's test), *H3F3A* is under a stronger selection pressure as compared to both *H3F3B* and
208 *H2AFZ* ($P = 2*10^{-7}$, $P = 3*10^{-6}$ respectively, Fig. 4A). Also, the distributions of the nucleotide
209 substitution scores are bimodal for all three genes, revealing that they are particularly
210 conserved within two distinct groups: (i) mammals and (ii) reptiles, birds and amphibians (Fig.
211 4A). This trend is especially pronounced for *H3F3A*, and we further confirmed a stronger
212 conservation of this gene within each individual group of organisms (Fig. S5A-B).
213
214 To rule out that the difference in sequence conservation of H3.3-encoding genes is determined
215 by the conservation of entire loci encompassing *H3F3A* or *H3F3B*, rather than these genes
216 themselves, we extended the analysis described above to six genes around each of the H3.3-

217 encoding genes. We found no significant difference between conservation of the genes around
218 *H3F3A* or *H3F3B* (Fig. 4B). At the same time, both *H3F3A* and *H3F3B* are significantly more
219 conserved than the neighboring genes ($P = 3*10^{-12}$ and $P = 10^{-6}$ respectively), with *H3F3A*
220 exhibiting highest level of conservation among the analyzed genes. This indicates that tetrapod
221 *H3F3A* evolves under stronger purifying selection at nucleotide level than *H3F3B*, *H2AFZ* or
222 neighboring genes.
223
224 Given that the H3.3 genes encode the same amino-acid sequence, not surprisingly most
225 substitutions were observed in the 3rd position of the codon. Interestingly, we found that
226 sarcopterygian *H3F3B* have generally higher GC-content at 3rd codon position (GC3) as
227 compared to sarcopterygian *H3F3A* (Fig. S6). The high GC3 in *H3F3B* genes mirrors
228 actinopterygian H3.3 and RI H3.1/H3.2-encoding genes while the *H2AFZ* genes have low GC3
229 that close to that of *H3F3A* (Fig S6). Thus, based on this metric *H3F3B* is more similar to
230 ancestral H3.3 and RI H3 histone genes, hence these results are in agreement with our
231 previous phylogenetic analyses.
232
233 To refine this analysis further, we compared the degree of nucleotide conservation at wobble
234 positions (i.e. 3rd codon positions where synonymous nucleotide substitutions are commonly
235 detected) between *H3F3A* and *H3F3B* gene alignments made of (i) all tetrapods, (ii) mammals,
236 and (iii) primates (Fig. 4C). We also separately considered a special case of wobble positions,
237 so-called 'fourfold degenerate' sites, i.e. 3rd codon positions at which all possible nucleotide
238 substitutions can occur without changing the encoded amino-acid; hence such fourfold
239 degenerate sites are under no selection pressure for amino-acid maintenance. A wobble

240 position was considered “absolutely conserved” if the nucleotide at that site is conserved in the
241 whole alignment (i.e. in all organisms).
242
243 In all groups, we consistently observed that there are more absolutely conserved 3rd codon
244 positions in *H3F3A* than *H3F3B* in all analyzed groups of species (Fig. 4C). This trend is most
245 pronounced for the fourfold degenerate sites (cf. horizontal bars in Fig. 4C). In addition, such
246 an over-representation is more pronounced for groups containing evolutionary distant
247 organisms e.g. FreqA/FreqB ratio for fourfold degenerate sites is 1.21, 2.10, 3.58 for primates,
248 mammals, and tetrapods respectively. This observation suggests that stronger selection on
249 synonymous sites in *H3F3A* than *H3F3B* is a stable phenomenon, deeply rooted in the tetrapod
250 lineage.
251
252 These findings revealed that there is a layer of selection pressure against nucleotide
253 substitutions operating on both *H3F3A* and *H3F3B* CDSs, driven not by the maintenance of
254 amino-acid sequence but maintenance of specific codons. Thus, our results suggest that codon
255 usage is under selection pressure among H3.3 genes. While this selection pressure is stronger
256 in *H3F3A* than in *H3F3B*, we infer that both genes have evolutionary adapted for distinct codon
257 usage preferences, and we investigate this phenomenon in more detail below.
258
259 **Differences in codon usage between H3.3 encoding genes**
260 The expression and abundance of transfer RNA (tRNA) vary substantially in human cell types
261 [40]. This variation correlates with codon usage preferences and plays a role in translational
262 control [41–43]. Furthermore, codon usage may differ between genes specialized in different
263 cellular processes such as cell proliferation and cell differentiation [41]. Thus, an analysis of the

264 codon usage in H3.3 genes can provide information on their functional specialization among
265 cellular gene expression programs.
266
267 To this end, we estimated the correlation between codon usage frequencies in each of the H3.3
268 genes and the genome-wide codon usage frequencies from each tetrapod genome. Similar to
269 a previously published study [41], we define these codon usage frequencies (hereby referred
270 to as "amino-acid specific codon frequencies") so that they represent the probability that a
271 codon is used when the amino-acid encoded by this codon appears in the protein product
272 sequence (see Methods). Since different genes are expressed in different cell types, we expect
273 that the codon usage frequencies computed for the entire genome ('genome-wide codon usage
274 frequencies') would correlate strongly with the codon usage frequencies of the genes showing
275 broad expression patterns. In line with this hypothesis, codon usage frequencies in a set of
276 human genes specifically selected for their ubiquitous expression in multiple cell types [44]
277 correlated with genome-wide frequencies with the Pearson's correlation coefficient equal about
278 0.70 (Fig. 5A). Application of this approach to the H3.3 genes revealed that the correlation
279 estimated for the human *H3F3B* gene (r=0.69) is close to benchmark value observed for the
280 ubiquitously expressed genes (UEG), while the correlation for the *H3F3A* gene is considerably
281 lower (r=0.54). Furthermore, all tetrapod *H3F3B* genes, actinopterygian H3.3 genes, and RC
282 H3.1/H3.2 genes (the latter are expressed in all dividing cells) show higher correlation with
283 genome-wide frequencies than either *H3F3A* or *H2AFZ* genes do (Fig 5A). We confirmed that
284 similar results are observed when codon usage is defined directly as the frequency of every
285 codon in a gene, without accounting for amino-acid abundance in the product ("codon
286 frequencies" in Fig. S7A). Based on these findings, we conclude that, as compared to *H3F3A*,
287 *H3F3B* is evolutionarily more optimized for a broad expression pattern.

288

289 To gain further insight on the evolutionary adaptation of the H3.3 genes, we compared their
290 codon usage frequencies to those estimated for the two groups of genes shown to be involved
291 in different transcriptional programs ('cell proliferation' and 'cell differentiation' genes; data from
292 [41]). Specifically, we computed pairwise correlation between the amino-acid specific codon
293 frequencies of the H3.3 genes and the individual genes associated with each of transcriptional
294 program (orange and green dots in Figs. 5B, S7B). This analysis showed that, by this metric,
295 *H3F3A* shares greater similarity with the 'proliferation' genes, while *H3F3B* is more similar to
296 the 'differentiation' genes ($P = 6.91*10^{-12}$ and $P = 8.3*10^{-12}$ respectively, Mann-Whitney's test;
297 Fig S7C-D). As previously, we confirmed these results in a similar analysis based on direct
298 codon frequencies which are not corrected for amino-acid abundance (Fig. S7E-F).

299

300 To benchmark the similarity between the codon usage of an individual gene and the codon
301 usage profiles associated with different transcriptional programs, we correlated codon usages
302 of individual proliferation- and differentiation-induced genes to both codon usage profiles (Fig.
303 5C). Comparison of the H3.3 genes with these benchmarks showed that *H3F3A* falls within 25th
304 percentile of the proliferation-associated genes when they are evaluated against codon usage
305 profile of their own group (r=0.58). The similarity of this gene to the differentiation group is low
306 and it is on par with the average similarity observed for the proliferation-induced genes when
307 they are compared to the codon usage profile of the differentiation group. In line with our
308 previous results, *H3F3B* exhibits an opposite trend: its codon usage correlates better with
309 differentiation gene profile (r=0.71 vs. r=0.35 for differentiation and proliferation profiles
310 respectively). We note however, that the *H3F3B* ranks relatively lowly among the differentiation-
311 induced in terms of their similarity to the group profile.

312

313 Based on these results, we conclude that *H3F3A* and *H3F3B* were evolutionary optimized for
314 distinct transcriptional programs. In this analysis we tested two programs that have been
315 described in literature [41]. While other programs may exist, our observations indicate better
316 fitness of *H3F3A* for the proliferation program and, arguably to a lesser extent, better fitness of
317 *H3F3B* for differentiation program. We also found that, similar to *H3F3B* (but not *H3F3A*), the
318 differentiation-induced genes correlate strongly with the genome-wide codon usage (r=0.88),
319 which suggests a broad expression profile. Thus, while *H3F3B* does not rank high within the
320 differentiation-induced genes, taken together our findings show that this gene is broadly
321 expressed in cell types, including differentiated cells. Overall, we report that despite encoding
322 identical protein sequence, *H3F3A* and *H3F3B* have distinct evolutionary histories and are
323 optimized for distinct transcriptional programs at the codon usage level, as illustrated in Figure
324 5D.

325

326 **Discussion**

327 The H3.3 histone is currently a subject of intense research due to its biological and biomedical
328 significance [21,22]; however, evolution of the genes encoding this protein is not fully
329 understood. In this study, we addressed this issue and studied the evolutionary history of the
330 H3.3-encoding genes from a diverse set of metazoan genomes. All analyzed genomes harbor
331 multiple genes (two in most cases, *H3F3A* and *H3F3B*) that encode an identical protein
332 sequence. We have shown that, despite being highly conserved at the protein level, H3.3-
333 encoding genes are subject to selection pressure at DNA sequence level, which is related to
334 their cellular function. Several lines of evidence stemming from phylogenetic analysis, as well
335 as analyses of the gene structure, synteny and codon usage (Figs 1, 2, 3 and 5) indicate that

336 *H3F3A* is specific for sarcopterygian (lobe-finned fish) lineage, whereas *H3F3B* exist in all
337 sarcopterygians and bears similarity to H3.3 genes in actinopterygians (ray-finned fish) and
338 jawless fish and with the vertebrate RD H3.1/H3.2 genes that diverged much earlier. These
339 results suggest that *H3F3B* is more similar to ancestral form of H3.3 gene than *H3F3A*, which
340 could be a product of a duplication event occurring after actinopterygian-sarcopterygian split.
341 However, we cannot completely exclude that *H3F3A* could have been lost in actinopterygians
342 and other lineages and additional studies are required to exactly trace the origin of each H3.3
343 gene.
344
345 Despite absolute conservation at the protein level in both genes, tetrapod *H3F3A* and *H3F3B*
346 are under varying degrees of purifying selection at the codon synonymous sites, resulting in
347 distinct codon usage profiles in these genes (Fig. 5). Our analysis revealed that codon usage
348 in *H3F3B* is similar that of the 'cell differentiation-induced' genes, in contrast to the codon usage
349 in *H3F3A*, which is similar to that of 'cell proliferation-induced genes' [41]. We note that while
350 proliferation-induced genes are active in a specific pathway, one can expect that the
351 'differentiation-induced' genes would show a broad expression profile as a group, because they
352 can be associated with various pathways in different cell types. This is also in line with our
353 observation that codon usage of *H3F3B*, but not of *H3F3A*, is similar to that of UEGs which are
354 active throughout cell types (Fig. 5B). Furthermore, similarly to the UEGs, *H3F3B* genes feature
355 a compact structure, with short introns (Fig. 1B) [45,46]. Given that we analyzed only two
356 transcriptional programs, it is possible that *H3F3A* and/or *H3F3B* would show similar or even
357 better fit for other programs. However, our results allow us to conclude that *H3F3A* and *H3F3B*
358 genes are evolutionary optimized for different transcriptional programs through codon usage
359 preferences and intron-exon organization.

360 In summary, the H3.3 genes provide a unique 'study case', in which protein sequence remain
361 constant in course of evolution for an extended time period, allowing analysis of the selection
362 operating at nucleotide level. Such analysis reveals an evolutionary mechanism of nucleotide
363 sequence optimization for fine-tuning of gene expression in specific cellular programs.
364

365 **Methods**

366 ***Phylogenetics analysis***

367 Sequences and annotations of the genes encoding histone variant H3.3 in different species, as
368 well as other genes used in this study were obtained from Ensembl and NCBI-RefSeq
369 databases. A phylogenetic tree was constructed using PHYML3.1 software [47], with
370 approximate likelihood ratio test (Chi2-based) for branch supports and GTR nucleotide
371 substitution model.
372

373 ***Synteny analysis***

374 Synteny around *H3F3A* and *H3F3B* genes in selected set of vertebrate genomes was detected
375 using a web application Genomicus version 80.01, that uses Ensembl comparative genomic
376 data (http://genomicus.biologie.ens.fr/genomicus) [32]. To supplement Genomicus-based
377 analysis and test for synteny between tetrapods and distant organisms such as fly and lamprey,
378 an additional method was used. This method measures the degree of conservation of the genes
379 neighboring H3.3-encoding genes based on comparison of their CDS and translated amino-
380 acid sequences. The annotated chromosome sequences were downloaded from Ensembl
381 (http://www.ensembl.org/info/data/ftp/index.html). Biopython (www.biopython.org) was used to
382 extract CDS of 30 genes upstream and downstream of all tetrapod *H3F3A* and *H3F3B*, and of
383 the H3.3 genes in distant organisms: actinopterygians (tetraodon, zebrafish and medaka),

384 lamprey and fly. Pairwise comparison of nucleotide and protein sequences was done by
385 aligning two sequences using MUSCLE [48] and computing sequence identity scores.
386
387 ***3'UTRs comparison***
388 3'UTR sequences of actinopterygian H3.3 genes were compared to those of tetrapod *H3F3A*
389 and *H3F3B* to find similarities. Comparison was performed through alignment of each pair of
390 3'UTR sequences and computing their sequences identity scores using MUSCLE. Since gaps
391 (indels) in alignments can substantially influence final identity scores [49], we excluded them
392 from calculations to insure that high UTR sequence variability due to insertions and deletions
393 does not deflate the scores and affect comparisons.
394
395 ***Codon usage analysis***
396 Two metrics of codon usage were used, the 'amino-acid specific codon frequencies' and 'codon
397 frequencies'. The amino-acid specific codon frequencies represent codon occurrences
398 normalized for amino-acids abundance [41], i.e. divided by the number of times the
399 corresponding amino-acid appears in the protein sequence. This metric corrects for potential
400 amino-acid usage biases and represents the probability that a codon will be used given that the
401 corresponding amino-acid is used. The second metric, 'codon frequencies', were computed by
402 dividing the codon occurrences by the total number of codons in the gene (i.e. normalized by
403 the length of the encoded amino-acid sequence). The codon usage profiles were computed for
404 different gene sets (proliferation-induced [41], differentiation-induced [41]). Genome-wide
405 codon counts were obtained from (http://www.kazusa.or.jp/codon).
406
407 **References**

408 1. Li B, Carey M, Workman JL. The Role of Chromatin during Transcription. Cell.
409 2007;128: 707–719. doi:10.1016/j.cell.2007.01.015
410 2. Hereford L, Fahrner K, Woolford J, Rosbash M, Kaback DB. Isolation of yeast histone
411 genes H2A and H2B. Cell. 1979;18: 1261–71. doi:10.1016/S0022-2836(83)80164-8
412 3. Marzluff WF, Gongidi P, Woods KR, Jin J, Maltais LJ. The human and mouse
413 replication-dependent histone genes. Genomics. 2002;80: 487–98. doi:10.1016/S0888-
414 7543(02)96850-3
415 4. Banaszynski LA, Allis CD, Lewis PW. Histone variants in metazoan development. Dev
416 Cell. 2010;19: 662–74. doi:10.1016/j.devcel.2010.10.014
417 5. Weber CM, Henikoff S. Histone variants: dynamic punctuation in transcription. Genes
418 Dev. 2014;28: 672–82. doi:10.1101/gad.238873.114
419 6. Wenderski W, Maze I. Histone turnover and chromatin accessibility: Critical mediators
420 of neurological development, plasticity, and disease. BioEssays. 2016;38: 410–419.
421 doi:10.1002/bies.201500171
422 7. Mito Y, Henikoff JG, Henikoff S. Genome-scale profiling of histone H3.3 replacement
423 patterns. Nat Genet. 2005;37: 1090–1097. doi:10.1038/ng1637
424 8. Jin C, Felsenfeld G. Nucleosome stability mediated by histone variants H3.3 and
425 H2A.Z. Genes Dev. 2007;21: 1519–1529. doi:10.1101/gad.1547707
426 9. Akiyama T, Suzuki O, Matsuda J, Aoki F. Dynamic replacement of histone H3 variants
427 reprograms epigenetic marks in early mouse embryos. PLoS Genet. 2011;7: e1002279.
428 doi:10.1371/journal.pgen.1002279
429 10. Santenard A, Ziegler-Birling C, Koch M, Tora L, Bannister AJ, Torres-Padilla M-E.
430 Heterochromatin formation in the mouse embryo requires critical residues of the histone
431 variant H3.3. Nat Cell Biol. 2010;12: 853–62. doi:10.1038/ncb2089

432 11. Voon HPJ, Wong LH. New players in heterochromatin silencing: histone variant H3.3
433 and the ATRX/DAXX chaperone. Nucleic Acids Res. 2016;44: 1496–1501.
434 doi:10.1093/nar/gkw012
435 12. Tagami H, Ray-Gallet D, Almouzni G, Nakatani Y. Histone H3.1 and H3.3 Complexes
436 Mediate Nucleosome Assembly Pathways Dependent or Independent of DNA
437 Synthesis. Cell. 2004;116: 51–61. doi:10.1016/S0092-8674(03)01064-X
438 13. Ahmad K, Henikoff S. The histone variant H3.3 marks active chromatin by replication-
439 independent nucleosome assembly. Mol Cell. 2002;9: 1191–200. doi:10.1016/s1097-
440 2765(02)00542-7
441 14. Ray-Gallet D, Quivy JP, Scamps C, Martini EMD, Lipinski M, Almouzni G. HIRA is
442 critical for a nucleosome assembly pathway independent of DNA synthesis. Mol Cell.
443 2002;9: 1091–1100. doi:10.1016/S1097-2765(02)00526-9
444 15. Akhmanova AS, Bindels PC, Xu J, Miedema K, Kremer H, Hennig W. Structure and
445 expression of histone H3.3 genes in Drosophila melanogaster and Drosophila hydei.
446 Genome. 1995;38: 586–600. Available: http://www.ncbi.nlm.nih.gov/pubmed/7557364
447 16. Sturm D, Witt H, Hovestadt V, Khuong-Quang D-A, Jones DTW, Konermann C, et al.
448 Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological
449 subgroups of glioblastoma. Cancer Cell. 2012;22: 425–37.
450 doi:10.1016/j.ccr.2012.08.024
451 17. Cleven AHG, Höcker S, Briaire-de Bruijn I, Szuhai K, Cleton-Jansen A-M, Bovée
452 JVMG. Mutation Analysis of H3F3A and H3F3B as a Diagnostic Tool for Giant Cell
453 Tumor of Bone and Chondroblastoma. Am J Surg Pathol. 2015;39: 1576–83.
454 doi:10.1097/PAS.0000000000000512
455 18. Behjati S, Tarpey PS, Presneau N, Scheipl S, Pillay N, Van Loo P, et al. Distinct H3F3A

456 and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. Nat
457 Genet. 2013;45: 1479–82. doi:10.1038/ng.2814
458 19. Park S-M, Choi E-Y, Bae M, Kim S, Park JB, Yoo H, et al. Histone variant H3F3A
459 promotes lung cancer cell migration through intronic regulation. Nat Commun. Nature
460 Publishing Group; 2016;7: 12914. doi:10.1038/ncomms12914
461 20. Matsuo Y, Kakubayashi N. Epigenetics Evolution and Replacement Histones:
462 Evolutionary Changes at Drosophila H3.3A and H3.3B. J Phylogenetics Evol Biol.
463 2016;04. doi:10.4172/2329-9002.1000174
464 21. Lan F, Shi Y. Histone H3.3 and cancer: A potential reader connection. Proc Natl Acad
465 Sci. 2015;112: 6814–6819. doi:10.1073/pnas.1418996111
466 22. Mohammad F, Helin K. Oncohistones: drivers of pediatric cancers. Genes Dev.
467 2017;31: 2313–2324. doi:10.1101/gad.309013.117
468 23. Glasauer SMK, Neuhauss SCF. Whole-genome duplication in teleost fishes and its
469 evolutionary consequences. Mol Genet Genomics. 2014;289: 1045–1060.
470 doi:10.1007/s00438-014-0889-2
471 24. Schartl M, Walter RB, Shen Y, Garcia T, Catchen J, Amores A, et al. The genome of
472 the platyfish, Xiphophorus maculatus, provides insights into evolutionary adaptation and
473 several complex traits. Nat Genet. Nature Publishing Group; 2013;45: 567–72.
474 doi:10.1038/ng.2604
475 25. Crow KD, Smith CD, Cheng JF, Wagner GP, Amemiya CT. An independent genome
476 duplication inferred from Hox paralogs in the American paddlefish-a representative
477 basal ray-finned fish and important comparative reference. Genome Biol Evol. 2012;4:
478 937–953. doi:10.1093/gbe/evs067
479 26. Alexandrou MA, Swartz BA, Matzke NJ, Oakley TH. Genome duplication and multiple

480 evolutionary origins of complex migratory behavior in Salmonidae. Mol Phylogenet Evol.
481 Elsevier Inc.; 2013;69: 514–523. doi:10.1016/j.ympev.2013.07.026
482 27. Volff J-N. Genome evolution and biodiversity in teleost fish. Heredity (Edinb). 2005;94:
483 280–94. doi:10.1038/sj.hdy.6800635
484 28. Volff JN, Körting C, Altschmied J, Duschl J, Sweeney K, Wichert K, et al. Jule from the
485 fish Xiphophorus is the first complete vertebrate Ty3/Gypsy retrotransposon from the
486 Mag family. Mol Biol Evol. 2001;18: 101–11. Available:
487 http://www.ncbi.nlm.nih.gov/pubmed/11158369
488 29. Postlethwait JH, Woods IG, Ngo-Hazelett P, Yan YL, Kelly PD, Chu F, et al. Zebrafish
489 comparative genomics and the origins of vertebrate chromosomes. Genome Res.
490 2000;10: 1890–902. Available: http://www.ncbi.nlm.nih.gov/pubmed/11116085
491 30. Cui J, Zhang Z, Shao Y, Zhang K, Leng P, Liang Z. Genome-wide identification,
492 evolutionary, and expression analyses of histone H3 variants in plants. Biomed Res Int.
493 2015;2015. doi:10.1155/2015/341598
494 31. Schenk R, Jenke A, Zilbauer M, Wirth S, Postberg J. H3.5 is a novel hominid-specific
495 histone H3 variant that is specifically expressed in the seminiferous tubules of human
496 testes. Chromosoma. 2011;120: 275–285. doi:10.1007/s00412-011-0310-4
497 32. Louis A, Thi N, Nguyen T, Muffato M, Crollius HR, Genomicus T. Genomicus update
498 2015 : KaryoView and MatrixView provide a genome-wide perspective to multispecies
499 comparative genomics. 2015;43: 682–689. doi:10.1093/nar/gku1112
500 33. Waterborg JH. Evolution of histone H3: emergence of variants and conservation of
501 post-translational modification sites. Biochem Cell Biol. 2012;90: 79–95.
502 doi:10.1139/o11-036
503 34. Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, et al. The African

504 coelacanth genome provides insights into tetrapod evolution. Nature. 2013;496: 311–6.
505 doi:10.1038/nature12027
506 35. Spieth J, Hillier LW, Wilson RK. Evolutionarily conserved elements in vertebrate , insect
507 , worm , and yeast genomes. 2005; 1034–1050. doi:10.1101/gr.3715005
508 36. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, et al. Systematic
509 discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of
510 several mammals. Nature. 2005;434: 338–345. doi:10.1038/nature03441
511 37. Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, et al.
512 FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Mol Biol
513 Evol. 2013;30: 1196–205. doi:10.1093/molbev/mst030
514 38. Delport W, Scheffler K, Botha G, Gravenor MB, Muse S V, Kosakovsky Pond SL.
515 CodonTest: modeling amino acid substitution preferences in coding sequences. PLoS
516 Comput Biol. 2010;6. doi:10.1371/journal.pcbi.1000885
517 39. Pond SLK, Frost SDW. Datamonkey: rapid detection of selective pressure on individual
518 sites of codon alignments. Bioinformatics. 2005;21: 2531–3.
519 doi:10.1093/bioinformatics/bti320
520 40. Dittmar KA, Goodenbour JM, Pan T. Tissue-specific differences in human transfer RNA
521 expression. PLoS Genet. 2006;2: 2107–2115. doi:10.1371/journal.pgen.0020221
522 41. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, et al. A
523 Dual Program for Translation Regulation in Cellular Proliferation and Differentiation.
524 Cell. Elsevier Inc.; 2014;158: 1281–1292. doi:10.1016/j.cell.2014.08.011
525 42. Plotkin JB, Robins H, Levine AJ. Tissue-specific codon usage and the expression of
526 human genes. Proc Natl Acad Sci U S A. 2004;101: 12588–91.
527 doi:10.1073/pnas.0404957101

528 43. Quax TEF, Claassens NJ, Söll D, van der Oost J. Codon Bias as a Means to Fine-Tune
529 Gene Expression. Mol Cell. 2015;59: 149–161. doi:10.1016/j.molcel.2015.05.035
530 44. Eisenberg E, Levanon EY. Human housekeeping genes, revisited. Trends Genet.
531 Elsevier Ltd; 2013;29: 569–574. doi:10.1016/j.tig.2013.05.010
532 45. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet.
533 2003;19: 362–365. doi:10.1016/S0168-9525(03)00140-9
534 46. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin E V, Kondrashov F a. Selection for
535 short introns in highly expressed genes. Nat Genet. 2002;31: 415–418.
536 doi:10.1038/ng940
537 47. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New
538 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the
539 performance of PhyML 3.0. Syst Biol. 2010;59: 307–21. doi:10.1093/sysbio/syq010
540 48. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
541 throughput. Nucleic Acids Res. 2004;32: 1792–7. doi:10.1093/nar/gkh340
542 49. Muhire BM, Varsani A, Martin DP. SDT: a virus classification tool based on pairwise
543 sequence alignment and identity calculation. PLoS One. 2014;9: e108277.
544 doi:10.1371/journal.pone.0108277
545

546 **Figure legends**

547

548 **Figure 1. Phylogenetic analyses H3.3-encoding genes**

549 **A.** Maximum likelihood tree illustrating the evolution of H3.3 genes in vertebrates. Three clades
550 were distinguished. Clade 1 comprises sarcopterygian *H3F3A* genes (brown); Clade 3
551 comprises sarcopterygii *H3F3B* (blue) which cluster together with actinopterygian H3.3 (gray).

552 Clade 2 consists of other actinopterygian H3.3 genes that cluster with neither clade 1 nor clade
553 3. Numbers along tree branches represent approximate log-likelihood ratio test values for
554 branch support. Blue stars mark non-tetrapod genes with syntenic relation to tetrapod *H3F3B*,
555 and blue asterisks mark non-tetrapod genes whose 3'UTRs are more similar to 3'UTRs of
556 tetrapod *H3F3B* than those of tetrapod *H3F3A*. **B, C, D.** Intron-exon structure of sarcopterygian
557 *H3F3B*, actinopterygian H3.3 genes and sarcopterygian *H3F3A*. All genes are drawn from 5' to
558 3' and are aligned at the start codon, position 0. The blue and red lines represent the 5' UTRs
559 and 3' UTRs respectively, and the squares in the middle represent the locations of protein-
560 coding exons.
561
562 **Figure 2: Synteny around H3F3A and H3F3B genes**
563 **A, B.** Synteny conservation analysis around human H3F3A (A) and H3F3B (B) genes
564 performed using selected actinopterygian genomes. Human *H3F3A* and *H3F3B* and
565 actinopterygian H3.3 are placed at the center of each plot (green block). A black outline
566 represents an ortholog of a gene in the same color, while a white outline represents a paralog
567 of gene in the same color. A blue star indicates actinopterygian organisms in which syntenic
568 genes around the H3.3 gene are also conserved around the human H3.3 gene.
569
570 **Figure 3. Comparison of coelacanth H3.3 genes to related genes in sarcopterygian and**
571 **non-sarcopterygian lineages.**
572 Sequence similarity was estimated for the CDS of coelacanth H3.3 genes (*H3F3A,* x-axis and
573 *H3F3B,* y-axis) and CDS of H3.3 genes from other sarcopterygian and more distant organisms
574 (actinopterygian, lamprey, fly). Additionally, CDS of tetrapod and zebrafish H3.1 and H3.2
575 genes were included in this analysis. Each point represents a gene and the organism name is

576 written in the matching color. The sequence similarity represents percentage of the identical
577 nucleotides in the sequence.
578
579 **Figure 4. Conservation of coding sequences in tetrapod histone variant genes**
580 **A.** Pairwise nucleotide substitution scores (genetic distances) computed for two H3.3 genes
581 (*H3F3A, brown* and *H3F3B,* blue), and *H2AFZ* gene (red) which was included in this analysis
582 for comparison. The analysis was performed for tetrapod genomes. Distribution shifting to the
583 left (smaller genetic distances) indicates higher conservation of a corresponding gene. (i) marks
584 the peak of the bimodal distribution corresponding to pairwise scores involving mammalian
585 organisms, while (ii) shows the distribution corresponding to pairwise scores involving
586 exclusively *H3F3A* genes in non-mammals. **B.** Pairwise nucleotide substitution scores for
587 *H3F3A* in tetrapod genomes (box plot filled in brown), *H3F3B* in tetrapod genome (filled in blue),
588 and their neighboring genes (brown and blue borders respectively). Both *H3F3A* and *H3F3B*
589 are significantly highly conserved relative to their surrounding genes (Wilcox sum rank test P =
590 2.9^{-12} and P = 1.01^{-6} respectively). No significant difference in conservation between gene
591 around *H3F3A* and those around *H3F3B* (P = 0.13). **C.** Absolute nucleotide conservation in
592 CDS of the H3.3 genes in tetrapod lineages. Top panel: all 3rd codon position; bottom panel:
593 the fourfold degenerate sites (i.e. sites where any possible nucleotide substitution is
594 synonymous). Columns show the number of absolutely conserved sites for a given group of
595 organisms, the total number of 3rd codon positions or fourfold degenerate sites, and the
596 corresponding frequencies of absolutely conserved sites. The horizontal bar represents the
597 H3F3A/H3F3B over-representation of absolutely conserved sites.
598 **Figure 5. Distinct codon usage preferences in the H3.3 genes (based on 'amino-acid**
599 **specific codon frequencies')**

600 **A.** Correlation between codon usage in the genes specified at x-axis and the genome-wide
601 codon usage. The box plots represent the lineage distributions of the correlation coefficients
602 calculated for the 'amino-acid specific codon frequencies' of a corresponding gene with those
603 estimated genome-wide (e.g. all tetrapod H3F3A genes vs. genome-wide frequencies). The
604 brown and blue diamonds provide reference for human *H3F3A* and *H3F3B* respectively. The
605 pink dashed line represents average correlation computed for human ubiquitously expressed
606 genes (UEG) [44]. **B.** Correlation of human H3F3A and H3F3B codon usage frequencies with
607 those computed for the genes associated with cell proliferation (orange) and cell differentiation
608 (green) [41]. Each dot represents individual gene from the corresponding group. The dotted
609 lines indicate the correlation coefficient medians for each group and the H3.3 gene. **C.**
610 Benchmarking of the codon usage frequencies in the H3.3 genes relative to the frequencies
611 estimated for the genes from the cell proliferation and differentiation groups. Boxplots represent
612 correlation values for the amino-acid specific codon frequencies of the individual cell
613 proliferation genes or cell differentiation genes with the overall profiles estimated for their own
614 groups. Dashed lines show the mean values of the correlations of individual genes from one
615 group with the opposite group profile (e.g. mean for the correlations of the codon usages of the
616 proliferation genes with the overall differentiation profile). The brown and blue diamonds
617 indicate correlation values for the human H3F3A and H3F3B genes. **D.** A model illustrating the
618 possible role of the evolutionary conserved arrangement of the two genes (H3F3A and H3F3B)
619 encoding the same protein (H3.3) in fine-tuning of this protein expression.

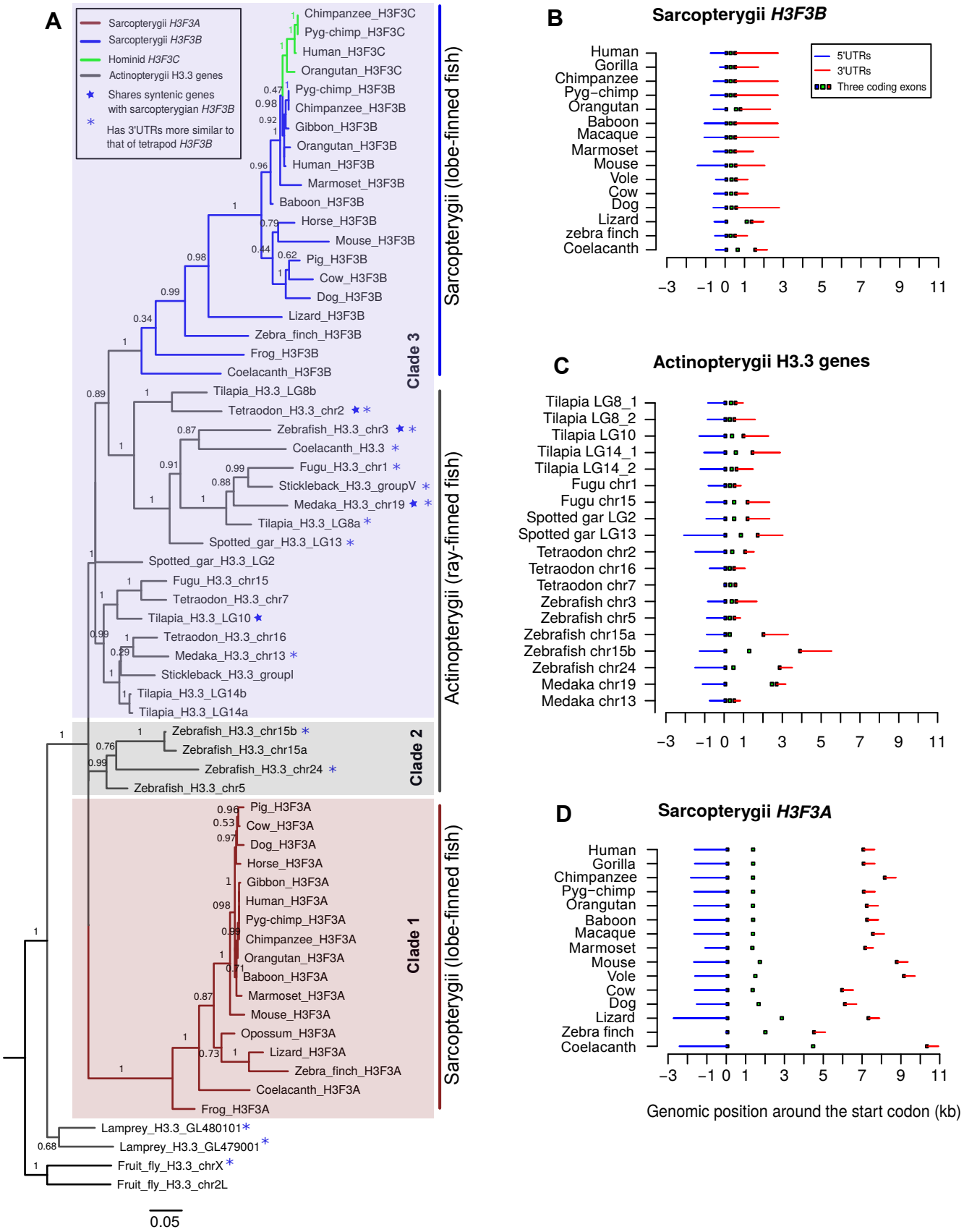

A
Sarcopterygii H3F3A
Sarcopterygii H3F3B
Hominid H3F3C
Actinopterygii H3.3 genes
Shares syntenic genes with sarcopterygian H3F3B
Has 3'UTRs more similar to that of tetrapod H3F3B
Chimpanzee_H3F3C
Pyg-chimp_H3F3C
Human_H3F3C
Orangutan_H3F3C
Pyg-chimp_H3F3B
Chimpanzee_H3F3B
Gibbon_H3F3B
Orangutan_H3F3B
Human_H3F3B
Marmoset_H3F3B
Baboon_H3F3B
Horse_H3F3B
Mouse_H3F3B
Pig_H3F3B
Cow_H3F3B
Dog_H3F3B
Lizard_H3F3B
Zebra_finch_H3F3B
Frog_H3F3B
Coelacanth_H3F3B
Clade 3
Sarcopterygii (lobe-finned fish)
Tilapia_H3.3_LG8b
Tetraodon_H3.3_chr2
Zebrafish_H3.3_chr3
Coelacanth_H3.3
Fugu_H3.3_chr1
Stickleback_H3.3_groupV
Medaka_H3.3_chr19
Tilapia_H3.3_LG8a
Spotted_gar_H3.3_LG13
Spotted_gar_H3.3_LG2
Fugu_H3.3_chr15
Tetraodon_H3.3_chr7
Tilapia_H3.3_LG10
Tetraodon_H3.3_chr16
Medaka_H3.3_chr13
Stickleback_H3.3_groupI
Tilapia_H3.3_LG14b
Tilapia_H3.3_LG14a
Actinopterygii (ray-finned fish)
Zebrafish_H3.3_chr15b
Zebrafish_H3.3_chr15a
Zebrafish_H3.3_chr24
Zebrafish_H3.3_chr5
Clade 2
Pig_H3F3A
Cow_H3F3A
Dog_H3F3A
Horse_H3F3A
Gibbon_H3F3A
Human_H3F3A
Pyg-chimp_H3F3A
Chimpanzee_H3F3A
Orangutan_H3F3A
Baboon_H3F3A
Marmoset_H3F3A
Mouse_H3F3A
Opossum_H3F3A
Lizard_H3F3A
Zebra_finch_H3F3A
Coelacanth_H3F3A
Frog_H3F3A
Clade 1
Sarcopterygii (lobe-finned fish)
Lamprey_H3.3_GL480101
Lamprey_H3.3_GL479001
Fruit_fly_H3.3_chrX
Fruit_fly_H3.3_chr2L
0.05
B
Sarcopterygii H3F3B
5'UTRs
3'UTRs
Three coding exons
Human
Gorilla
Chimpanzee
Pyg−chimp
Orangutan
Baboon
Macaque
Marmoset
Mouse
Vole
Cow
Dog
Lizard
zebra finch
Coelacanth
−3 −1 0 1 3 5 7 9 11
C
Actinopterygii H3.3 genes
Tilapia LG8_1
Tilapia LG8_2
Tilapia LG10
Tilapia LG14_1
Tilapia LG14_2
Fugu chr1
Fugu chr15
Spotted gar LG2
Spotted gar LG13
Tetraodon chr2
Tetraodon chr16
Tetraodon chr7
Zebrafish chr3
Zebrafish chr5
Zebrafish chr15a
Zebrafish chr15b
Zebrafish chr24
Medaka chr19
Medaka chr13
−3 −1 0 1 3 5 7 9 11
D
Sarcopterygii H3F3A
Human
Gorilla
Chimpanzee
Pyg−chimp
Orangutan
Baboon
Macaque
Marmoset
Mouse
Vole
Cow
Dog
Lizard
Zebra finch
Coelacanth
−3 −1 0 1 3 5 7 9 11
Genomic position around the start codon (kb)

Figure 1.

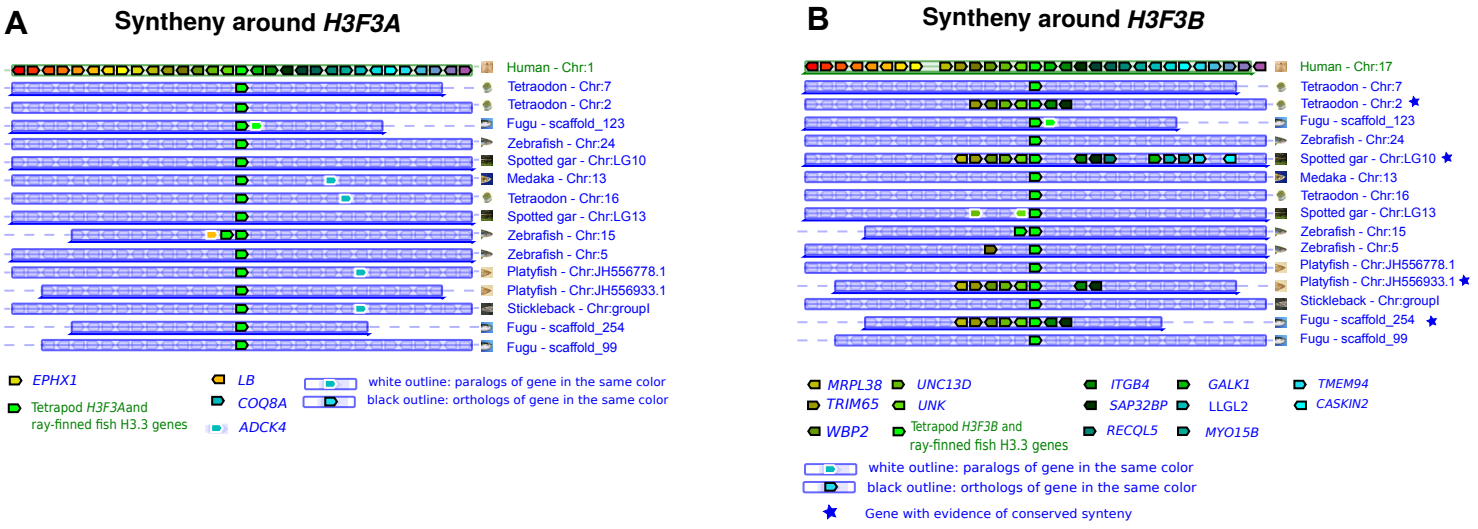

A
Syntheny around H3F3A
Human - Chr:1
Tetraodon - Chr:7
Tetraodon - Chr:2
Fugu - scaffold_123
Zebrafish - Chr:24
Spotted gar - Chr:LG10
Medaka - Chr:13
Tetraodon - Chr:16
Spotted gar - Chr:LG13
Zebrafish - Chr:15
Zebrafish - Chr:5
Platyfish - Chr:JH556778.1
Platyfish - Chr:JH556933.1
Stickleback - Chr:groupI
Fugu - scaffold_254
Fugu - scaffold_99
EPHX1
Tetrapod H3F3Aand ray-finned fish H3.3 genes
LB
COQ8A
ADCK4
white outline: paralogs of gene in the same color
black outline: orthologs of gene in the same color
B
Syntheny around H3F3B
Human - Chr:17
Tetraodon - Chr:7
Tetraodon - Chr:2
Fugu - scaffold_123
Zebrafish - Chr:24
Spotted gar - Chr:LG10
Medaka - Chr:13
Tetraodon - Chr:16
Spotted gar - Chr:LG13
Zebrafish - Chr:15
Zebrafish - Chr:5
Platyfish - Chr:JH556778.1
Platyfish - Chr:JH556933.1
Stickleback - Chr:groupI
Fugu - scaffold_254
Fugu - scaffold_99
MRPL38
UNC13D
ITGB4
GALK1
TMEM94
TRIM65
UNK
SAP32BP
LLGL2
CASKIN2
WBP2
Tetrapod H3F3B and ray-finned fish H3.3 genes
RECQL5
MYO15B
white outline: paralogs of gene in the same color
black outline: orthologs of gene in the same color
Gene with evidence of conserved synteny

Figure 2.

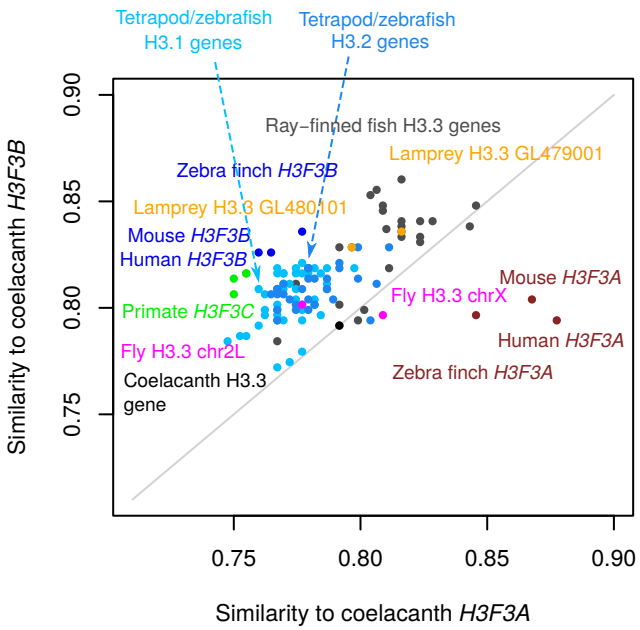
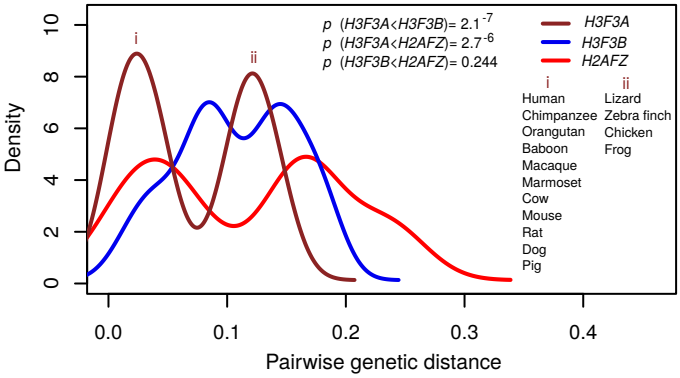

Tetrapod/zebrafish
H3.1 genes
Tetrapod/zebrafish
H3.2 genes
Ray−finned fish H3.3 genes
Lamprey H3.3 GL479001
Zebra finch H3F3B
Lamprey H3.3 GL480101
Mouse H3F3B
Human H3F3B
Primate H3F3C
Fly H3.3 chr2L
Coelacanth H3.3
gene
Mouse H3F3A
Fly H3.3 chrX
Human H3F3A
Zebra finch H3F3A
0.75
0.80
0.85
0.90
Similarity to coelacanth H3F3B
Similarity to coelacanth H3F3A

Figure 3.

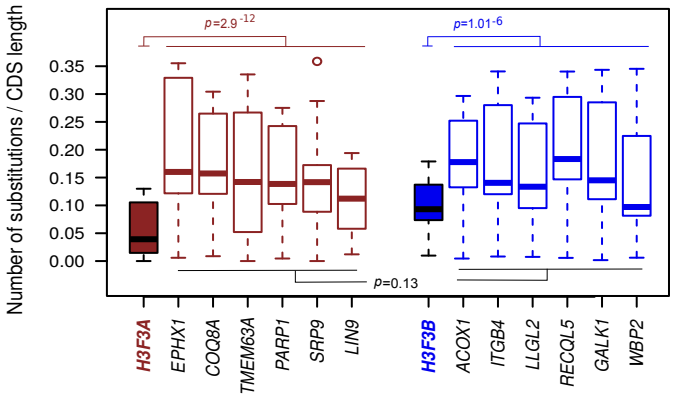
A

Conservation of H3.3 CDS in tetrapods



B

Purifying selection at nucleotide level for tetrapod *H3F3A, H3F3B* and neighboring genes



Six genes closest to *H3F3A* and *H3F3B* which are concerved in all tetrapods

C

Conservation of synonymous sites in H3.3 genes in tetrapods

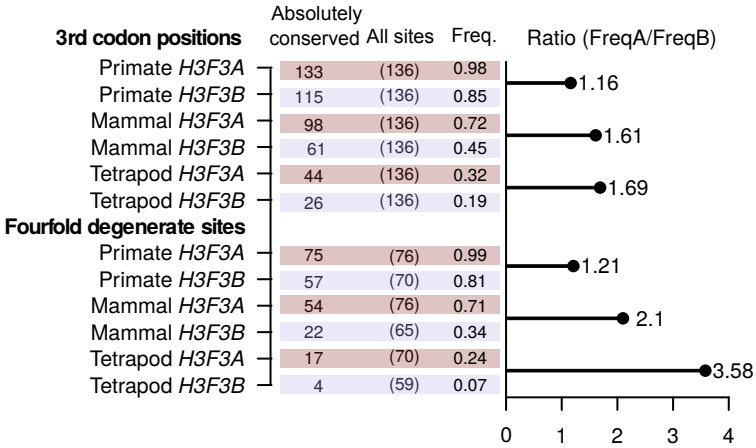


Figure 4.

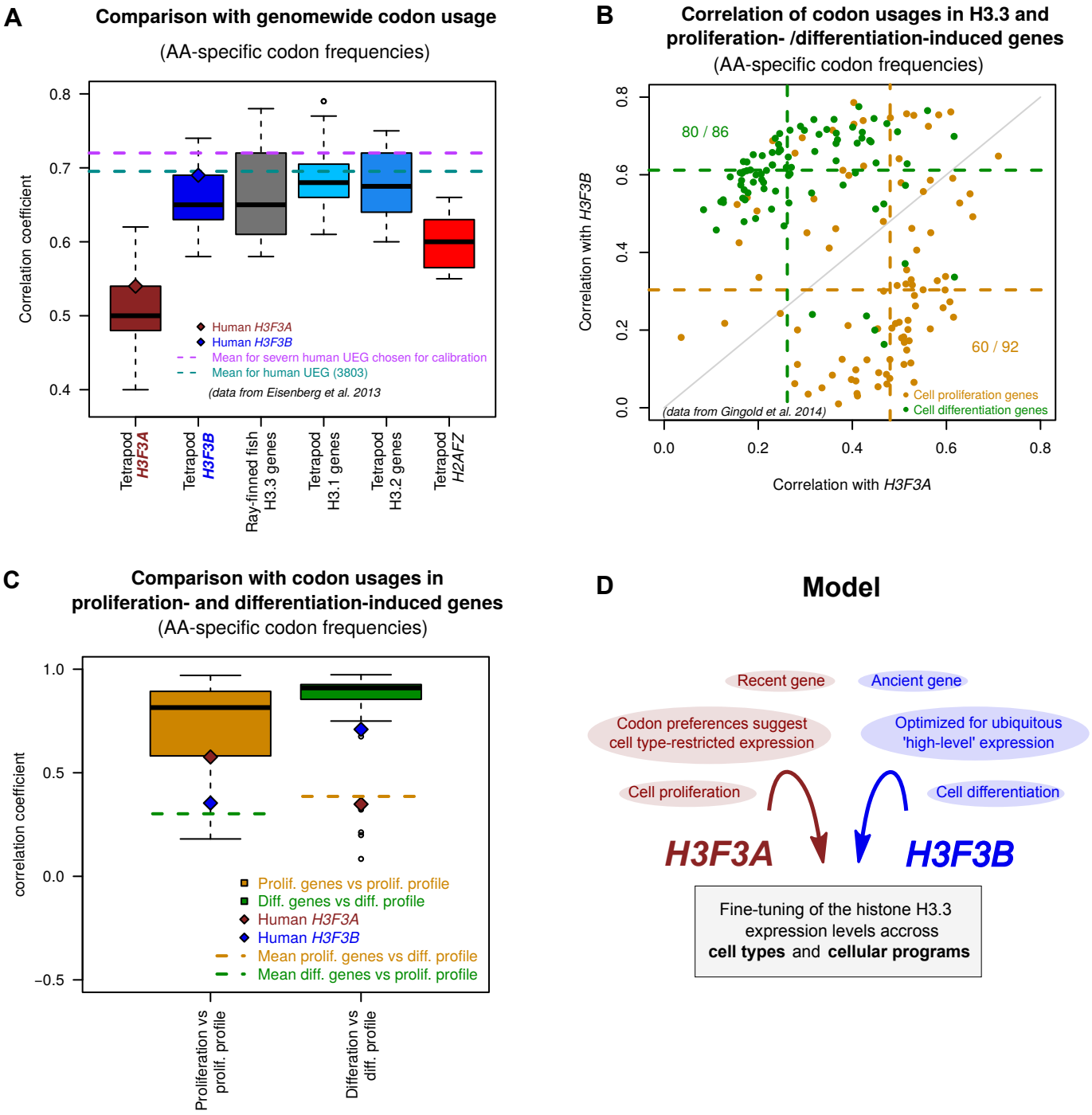

A
Comparison with genomewide codon usage
(AA-specific codon frequencies)
Correlation coefficient
0.8
0.7
0.6
0.5
0.4
Human H3F3A
Human H3F3B
Mean for severn human UEG chosen for calibration
Mean for human UEG (3803)
(data from Eisenberg et al. 2013
Tetrapod H3F3A
Tetrapod H3F3B
Ray-finned fish H3.3 genes
Tetrapod H3.1 genes
Tetrapod H3.2 genes
Tetrapod H2AFZ
B
Correlation of codon usages in H3.3 and proliferation- /differentiation-induced genes
(AA-specific codon frequencies)
Correlation with H3F3B
Correlation with H3F3A
0.0
0.2
0.4
0.6
0.8
80 / 86
60 / 92
(data from Gingold et al. 2014)
Cell proliferation genes
Cell differentiation genes
C
Comparison with codon usages in proliferation- and differentiation-induced genes
(AA-specific codon frequencies)
correlation coefficient
1.0
0.5
0.0
−0.5
Prolif. genes vs prolif. profile
Diff. genes vs diff. profile
Human H3F3A
Human H3F3B
Mean prolif. genes vs diff. profile
Mean diff. genes vs prolif. profile
Proliferation vs prolif. profile
Differation vs diff. profile
D
Model
Recent gene
Ancient gene
Codon preferences suggest cell type-restricted expression
Optimized for ubiquitous 'high-level' expression
Cell proliferation
Cell differentiation
H3F3A
H3F3B
Fine-tuning of the histone H3.3 expression levels accross cell types and cellular programs

Figure 5.