



**HAL**  
open science

## Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome

Sophie Fouchécourt, Valérie Fillon, Christelle Marraud, Caroline Callot, Sarah Ronsin, Floriane Picolo, Cécile Douet, Benoit Piegu, Philippe Monget

### ► To cite this version:

Sophie Fouchécourt, Valérie Fillon, Christelle Marraud, Caroline Callot, Sarah Ronsin, et al.. Expanding duplication of the testis PHD Finger Protein 7 (PHF7) gene in the chicken genome. *Genomics*, 2022, 114 (4), pp.1-9. 10.1016/j.ygeno.2022.110411 . hal-03746220

**HAL Id: hal-03746220**

**<https://hal.inrae.fr/hal-03746220v1>**

Submitted on 22 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License



29 **ABSTRACT**

30 Gene duplications increase genetic and phenotypic diversity and occur in complex genomic  
31 regions that are still difficult to sequence and assemble. PHD Finger Protein 7 (PHF7) acts  
32 during spermiogenesis for histone-to-histone protamine exchange and is a determinant of  
33 male fertility in *Drosophila* and the mouse. We aimed to explore and characterise in the  
34 chicken genome the expanding family of the numerous orthologues of the unique mouse *Phf7*  
35 gene (highly expressed in the testis), observing the fact that this information is unclear and/or  
36 variable according to the versions of databases. We validated nine primer pairs by *in silico*  
37 PCR for their use in screening the chicken bacterial artificial chromosome (BAC) library to  
38 produce BAC-derived probes to detect and localise *PHF7*-like loci by fluorescence *in situ*  
39 hybridisation (FISH). We selected nine BAC that highlighted nine chromosomal regions for a  
40 total of 10 distinct *PHF7*-like loci on five *Gallus gallus* chromosomes: Chr1 (three loci), Chr2  
41 (two loci), Chr12 (one locus), Chr19 (one locus) and ChrZ (three loci). We sequenced the  
42 corresponding BAC by using high-performance PacBio technology. After assembly, we  
43 performed annotation with the FGENESH program: there were a total of 116 peptides,  
44 including 39 PHF7-like proteins identified by BLASTP. These proteins share a common  
45 exon-intron core structure of 8–11 exons. Phylogeny revealed that the duplications occurred  
46 first between chromosomal regions and then inside each region. There are other duplicated  
47 genes in the identified BAC sequences, suggesting that these genomic regions exhibit a high  
48 rate of tandem duplication. We showed that the *PHF7* gene, which is highly expressed in the  
49 rooster testis, is a highly duplicated gene family in the chicken genome, and this phenomenon  
50 probably concerns other bird species.

51

52

53 **Key words:** chicken genome, gene duplication, evolution, testis

54

55

56

57

58

59

## 60 INTRODUCTION

61 Gene families correspond to genes clustered by sequence similarity. The members often  
62 exhibit similar functions and evolve in a dynamic context of genomic rearrangements,  
63 including duplications within a single genome [1]. Moreover, it is well known that copy  
64 number variation (CNV) in loci with numerous paralogous genes has the potential to increase  
65 phenotypic diversity [2]. Indeed, CNV has a central role in explaining innovations across  
66 phyla, including the emergence of novel functions. We have analysed several duplicated gene  
67 families involved in reproduction in the mouse such as *Oogenesins* [3, 4] and *Nlrp*  
68 (Nucleotide-binding oligomerization domain, Leucine rich Repeat and Pysin domain  
69 containing Proteins) [5, 6]. In the chicken genome, several genes are duplicated, most of  
70 which are simple tandem duplications, such as *DEFENSIN6/7* [7]. However, some genes  
71 present several paralogues – for example, 10 for the Toll-like receptor gene family [8], 16 for  
72 the *TCRb* (T cell receptor beta) locus [9] and 23 for Free Fatty Acid Receptor-2  
73 (*FFAR2/GPR43*) characterised previously in our laboratory [10].

74 Numerous genes are required for the formation of fertile spermatozoa (at least 2,000  
75 different genes in humans), in particular genes involved in the successive stages of  
76 spermatogenesis (germ cell differentiation) [11]. Despite a huge number of studies, the causes  
77 of fertility defects (genetic or environmental perturbations) are not well understood,  
78 suggesting that there are still regulatory pathways to decipher, especially in species like birds  
79 in which testicular data are scarce compared with mammals. At the cellular and physiological  
80 levels, spermatogenesis represents a relatively well-conserved process among  
81 phylogenetically distant animal species [12, 13].

82 In a previous study published in 2019 [14], we were interested in identifying testis  
83 genes conserved between invertebrates and vertebrates and exhibiting high relative mRNA  
84 expression in the testis of vertebrates, with a focus on chicken species. Indeed, this study  
85 highlighted a substantial list of uncharacterised genes for testis function in vertebrates, in  
86 particular in the chicken. Among the chicken testis-specific genes highlighted in this previous  
87 study, PHD Finger Protein 7 (*Phf7*) is indispensable for mouse and *Drosophila* male fertility  
88 [15, 16], encodes an actor in histone-to-protamine exchange during spermiogenesis and is  
89 highly expressed in the mammalian and chicken testis [14, 17, 18]. We have been intrigued by  
90 the numerous bird orthologues of *Phf7* gene [14], as shown in its phylogenetic tree of  
91 Ensembl database (63 paralogues in Ensembl release 88 [March 2017]), whereas the gene  
92 is present in a single copy in mammals. Wang et al. [18] described at least two chromosomal  
93 loci in the chicken genome (with no precision about the number of paralogues). To our

94 knowledge, such a high number of gene duplicates (63) in chicken has not been described in  
95 the literature. Moreover and of note, we observed that Ensembl chicken *PHF7* paralogues in  
96 releases following 88 have fluctuated (68, then 0, currently four paralogues), suggesting that  
97 these predictions are incomplete and/or unreliable. Indeed, while mouse and human genomes  
98 are high quality, this is not the case for more recently sequenced genomes, in particular the  
99 chicken genome, because it contains numerous GC-rich regions. It is known that gene  
100 duplications, especially when they are multiple and in tandem, correspond to complex  
101 genomic regions that are difficult to sequence, assemble and annotate. The scientific  
102 community has made an effort to fill these lacunae due to such technological and  
103 bioinformatic challenges. Thus, our objective was to explore and characterise the existence of  
104 this predicted expanding family of numerous *PHF7* orthologues in the chicken genome, using  
105 the high-performance PacBio technology [19].

106

107

108

## 109 **RESULTS**

### 110 **Statement of the existence of an expanding family of *PHF7* genes in the chicken genome**

111 In our 2019 study using data from the Ensembl 88 database released in March 2017, we  
112 noted that the mouse *Phf7* gene (present in a single copy in mammalian genomes) exhibited  
113 111 avian orthologs for only five bird species, as schematised in the trees in Supplemental Fig  
114 1(A/B) (see also Supplemental Table 1), with 63 homologs in the chicken, and between 2 and  
115 21 in the four other birds (flycatcher: 2; duck: 6; zebra finch: 19; turkey: 21). As seen in the  
116 tree in Supplemental Fig 1B, the 63 predicted chicken genes were phylogenetically regrouped  
117 in nine subtrees. We arbitrarily named these subtrees: groups (Gr) A to I, for better  
118 clarification and further investigation. According to Ensembl 88, these genes are located on  
119 nine distinct loci on five chromosomes: Chr19 (GrA), ChrZ (GrB and GrC), Chr1 (GrD, GrE  
120 and GrI), Chr2 (GrF and GrH) and Chr12 (GrG). Surprisingly, we have noticed that the  
121 number of chicken *PHF7* genes annotated in Ensembl has fluctuated (from 0 to 68 genes)  
122 according to database versions (Supplemental Table 2), with only four orthologues described  
123 in the current version (release 105, [December 2021]) (Supplemental Table 2, column K).  
124 Thus, we decided to evaluate precisely the number of *PHF7* copies in the chicken genome by  
125 *in silico* PCR and genomic sequencing of bacteria artificial chromosomes (BAC)  
126 corresponding to genomic regions identified by fluorescence *in situ* hybridisation (FISH).

### 127 **Characterisation of the chicken *PHF7* loci by *in silico* PCR**

128 Primers designed in our previous study [14] and corresponding to the nine GrA to GrI *PHF7*  
129 subtrees/groups described above are listed in Supplemental Table 3 (column B). We first  
130 verified that these primers amplified unique amplicons with the expected sizes (Supplemental  
131 Table 3, column C) when used in genomic PCR with chicken DNA (Supplemental Fig 2). We  
132 then used these primer pairs in *in silico* PCR (UCSC database). They highlighted nine distinct  
133 genomic loci described in Supplemental Table 3 (column D). These loci are on the same  
134 chromosome for each respective group of genes as those described in the Ensembl 88  
135 release: two loci on Chr19 corresponding to GrA; two loci on ChrZ corresponding to GrB and  
136 GrC; three distinct loci on Chr1 corresponding to GrD, GrE, and GrI; two loci on Chr2  
137 corresponding to GrF and GrH; and one locus on Chr12 corresponding to GrG. Thus, these  
138 results confirm the existence of at least nine loci on five chromosomes with *PHF7* genes in  
139 the chicken genome as suggested by Ensembl version 88, in contrast with the data available  
140 in the current Ensembl version 105 that describes four loci on two chromosomes (Chr12 and  
141 Chr1, see Supplemental Table 2, column L). This allowed us to validate the relevance of the  
142 primer pairs for their use in BAC screening further followed by FISH.

143 **Chromosomal localisation of chicken *PHF7* loci by FISH and comparison with *in silico***  
144 **PCR localisation**

145 For the nine groups, BAC clones were selected by PCR with primer pairs characterised above  
146 and as described in Materials and methods. We then used these specific BAC clones, listed in  
147 Table 1, in FISH to map the nine groups of genes (GrA to GrI), after we validated the  
148 nucleotide sequence of each probe by sequencing (sequences obtained are in Supplemental  
149 Table 3, column E) and submitted them to BLASTN analysis (Supplemental Table 3, column  
150 F). The GrA and GrB sequence probes were associated with an accession number with the  
151 name 'PHD finger protein 7\_like' (GrA: NC\_006106.5; GrB NC\_006127.5), this is why we  
152 use '*PHF7*-like' for further designation in the text. The coordinates in BLASTN results were  
153 coherent with the loci delivered by *in silico* PCR (Supplemental Table 3, column D). The  
154 FISH localisations are shown in Fig 1A (pictures) and drawn in Fig 1B (chromosomal  
155 schemes), and listed in Table 1 (with their measures) and in Supplemental Table 3. FISH  
156 localisations specified a unique chromosomal region with *PHF7*-like genes for each BAC,  
157 except GrG that brightened two loci, one on Chr12 and one on ChrZ (this latter was not  
158 predicted by *in silico* PCR). GrD and GrE brightened the same chromosomal region on Chr1  
159 (p26). There were a total of nine hybridisation signals dispatched on **five *G. gallus***  
160 **chromosomes: Chr1** (two signal), **ChrZ** (three signal), **Chr2** (two signal), **Chr12** (one  
161 signal) and **Chr19** (one signal).

162 **Sequencing, assembling and gene annotation of BAC clones**

163 To further characterise the loci identified on the chicken genome by FISH, we sequenced the  
164 nine BAC clones targeted by the *PHF7* screen (sequences are available in Genbank<sup>1</sup>). The  
165 sequences contained between 87,395 bases for the smallest BAC (E) to 177,594 for the largest  
166 (I) (Table 2). For each of the nine nucleotide sequences, we performed BLASTN by using the  
167 NCBI database (GCRg7w -white Leghorn race- version 106). The BLASTN match result was  
168 unique for each BAC, except for BAC G that first matched with Chr12 for almost its entire  
169 length and then with ChrZ for only 19% of its length (see Supplemental Table 3, columns H  
170 and I). As expected, the loci defined by these coordinates included the loci identified by *in*  
171 *silico* PCR for all BAC. We noted that BAC E and BAC D, which both hybridised on  
172 GGA1p26 in FISH as stated above, are 2.5 Mb apart and thus define two distinct *PHF7*-like  
173 loci on Chr1. Finally, these results confirm the chromosomal localisation and coordinates

---

<sup>1</sup> GenBank accession numbers: ON022098 (BAC A), ON022099 (BAC B), ON022100 (BAC C), ON022101 (BAC D), ON022102 (BAC E), ON022103 (BAC F), ON022104 (BAC G), ON022105 (BAC H), ON022106 (BAC I)

174 described above for the nine BAC as well as the existence of a tenth locus on ChrZ (GGAZ  
175 p21-22 corresponding to BAC G) that was not predicted by *in silico* PCR.

176 We performed gene annotation of the nine sequences by using the FGENESH program  
177 (in Softberry) as described in the Materials and methods. We identified 116 predicted genes  
178 dispatched on the nine BAC, from 4 for BAC F to 28 for BAC I (Table 2). Details of each  
179 BAC annotation FGENESH outputs are listed in Supplemental Table 4. We submitted the  
180 corresponding 116 peptides to BLASTP to determine functional homology. We retrieved a  
181 name and accession number for each of the 116 peptides (Supplemental Table 5). The  
182 annotation of each BAC is illustrated by gene maps in Fig 2 (BAC I) and Supplemental Fig 3  
183 (BAC A to H). There were a total of 39 PHF7-like proteins (in yellow) from 1 in BAC C,  
184 BAC F and BAC H to 10 in BAC I (see Supplemental Table 5 and Table 2). The BLASTP  
185 results for these 39 PHF7-like proteins (sequences in Supplemental Table 6) corresponded to  
186 an E-value equal or very close to zero (using ‘by default’ parameters) (Supplemental Table 5).  
187 Based on using ESPript for obtention of a consensus sequence, their sequence similarity was  
188 30%–55%, with a consensus sequence of 265 amino acids (Supplemental Data; see also the  
189 phylogenetic tree Fig 3).

## 190 **Genomic organisation and phylogenetic link of the PHF7-like protein family**

### 191 Exon-intron structure

192 We retrieved the exon sequence from each *PHF7*-like gene from FGENESH outputs and then  
193 used BLASTN to compare the sequences against one another. Comparative analysis of the  
194 exon/intron structure of the 39 *PHF7*-like genes revealed a frequent common ‘core-structure’  
195 containing from 8 (1 gene) to 11 exons (11 genes), with the most prevalent configuration (16  
196 genes) exhibiting 9 exons (Supplemental Table 7 and Supplemental Fig 4). This common  
197 ‘core-structure’ exhibits various changes: exon deletion (for example, in BAC B gene 20,  
198 exon 8 is deleted compared with other neighbour genes); reverse duplication on the other  
199 strand (for example, genes 3 and 4 in BAC G); split of an exon, that is, one exon gives two  
200 exons (for example, genes 3 and 6 in BAC E); and the presence of specific exons (genes 7 and  
201 8 in BAC E). Finally and intriguingly, four genes (genes 4 and 5 in BAC E; gene 1 in BAC F;  
202 gene 5 in BAC H) were predicted to be longer because of several duplications of the ‘core-  
203 structure’. Gene 5 in BAC H was particularly intriguing (70 exons: 7 repetitions of 10 exon  
204 structures) and was also predicted by the annotation obtained with two other classical  
205 programs, Augustus and GENESCAN (not shown). On the contrary, the three other genes  
206 (predicted with 30, 29, and 28 exons, respectively for genes 4 and 5 in BAC E, and gene 1 in



207 BAC F) were split in 'normal-sized' *PHF7*-like genes with these programs (see the  
208 Discussion for more details).

### 209 Phylogeny

210 We aligned the predicted protein sequences to construct a phylogenetic tree that exhibited  
211 reliable bootstrap values (Fig 3). The phylogenetic link designates the peptides in BAC A as  
212 the sequences closest to the common ancestor. The phylogeny showed that the duplications  
213 occurred first between chromosomal regions and then inside each region. For example, on  
214 ChrZ, B and C loci are around 4 Mb apart and the tree suggested that one B peptide probably  
215 duplicated in the unique C peptide before its own multiple tandem duplications. The same  
216 scenario is observed for D and E peptides, which are separated by around 2 Mb on Chr1.

### 217 Other duplicate gene families in the *PHF7*-like loci

218 In each BAC, we noticed several groups of genes for which BLASTP results gave the same  
219 name and/or Genbank ID (Supplemental Table 5, Fig 2 and Supplemental Fig 3), suggesting  
220 that they are phylogenetically related and are members of the same family. There are 1)  
221 families with two or three members: ras GTPase-activating protein 1 in BAC B, Cadherin-18  
222 in BAC F, scm-like with four MBT domains protein 2 in BAC D, testis-expressed protein  
223 264, metabotropic glutamate receptor 2 in BAC G and centrosomal protein of 126 kDa in  
224 BAC I; or 2) families with many members: chemokine in BAC A (6 genes), reverse  
225 transcriptase dispatched in BAC B, BAC C and BAC I (16 genes); and translation initiation  
226 factor IF-2-like (9 genes) in BAC I. Interestingly, the genomic organisation of this latter BAC  
227 suggests tandem duplication of 'PHF7-like/translation initiation factor' genes. Overall, this  
228 confirms that these chicken genomic regions are complex, with numerous tandem  
229 duplications.

230

231 **Table 1.** Results of the fluorescence *in situ* hybridisation (FISH) localisation for the nine gene  
 232 groups (GrA to GrI), identifying nine chromosomal regions with *PHF7*-like genes on *G.*  
 233 *gallus* chromosomes with the selected BAC.

Gene group	Selected BAC	FISH localisation	Measures
A	WAG-038I15	GGA19	With WAG-062P02
B	WAG-119F08	GGAZ q12-13	Flcen: 38.8 ± 5.3
C	WAG-041O23	GGAZ p12-21	Flcen: 50.9 ± 7.0
D	WAG-023G20	GGA1 p26	Flpter: 3 ± 1.8
E	WAG-038A02	GGA1 p26	No measure
F	WAG-037B06	GGA2 q11-21	Flpter: 51.4 ± 2.5
G	WAG-035C03	GGAZ p21-22 GGA12	Flcen GGA Z: 34.7 ± 5.9 With WAG-033L02
H	WAG-034E06	GGA2 p11-12	Flpter: 25.6 ± 1.7
I	WAG-119J04	GGA1 q35-41	Flpter: 90.5 ± 2.2

234 *FLpter* refers to the fractional length of the chromosome from the telomere of the p arm (%).

235 *FLcen* refers to the fractional length of the chromosome from the centromere (%).

236

237

238 **Table 2.** Details of the sequenced bacterial artificial chromosomes (BAC A to I).

BAC	Size (bp)	Total number of genes	Number of <i>PHF7</i> -like genes	Numbering of <i>PHF7</i> -like genes
A	103,706	19	3	13, 15, 17
B	122,743	23	9	5, 7, 8, 10, 12, 14, 16, 18, 20
C	147,352	9	1	6
D	147,939	9	2	1, 3
E	88,262	8	8	1, 2, 3, 4, 5, 6, 7, 8
F	96,796	4	1	1
G	86,864	10	4	1, 2, 3, 4
H	154,838	6	1	5
I	177,594	28	10	10, 12, 14, 16, 18, 20, 22, 24, 26, 28
<i>Total</i>	-	<i>116</i>	<i>39</i>	-

239 For each sequenced BAC A to I, the total gene number was identified by FGENESH in  
 240 Softberry. For the details of each BAC annotation (positions of predicted genes) see  
 241 Supplemental Table 4. The number and numbering\* of *PHF7*-like genes is based on BLASTP  
 242 results (as described in Supplemental Table 5).

243 bp= base pair

244 \*The numbering is as described on the BAC maps in Fig 2 (BAC I) and Supplemental Fig 3  
 245 (BAC A to H).

246

247 **DISCUSSION**

248 In the present study, using FISH we identified 10 loci with *PHF7*-like genes in *G. gallus*,  
249 dispatched on five chromosomes: **Chr1** with three distinct loci, **ChrZ** with three distinct loci,  
250 **Chr2** with two distinct loci and **Chr12** and **Chr19** each with a unique locus. We also  
251 observed nine of these loci by using *in silico* PCR, and some of them had also been referenced  
252 in former Ensembl version 88 (then ‘disappeared’ in the genome annotation, as shown in  
253 Supplemental Table 2). According to our phylogenetic tree, the first duplication concerned the locus  
254 on Chr 19 (BAC A) into 2 ancestors, one at the root of BAC F/G (Chr 2 and 12, respectively) and the  
255 other at the root of BAC H/D/E/C/B/I. In this latter group, BAC D and E (Chr1) have a common  
256 ancestor, and BAC B, C (both on Chr Z) and I (Chr 1) have a common ancestor, these two ancestors  
257 sharing a common older ancestor with BAC H (Chr 2).

258 Currently in Ensembl (release 105, [December 2021]), only four loci are described  
259 corresponding to four distinct genes (Supplemental Table 2, columns K and L), with one  
260 being on Chr12 with coordinates included in BAC G, and three on Chr1, with two genes  
261 included in BAC I and one (ENSG00000048616) on a different locus that we have not  
262 characterised. Thus, it is likely that we missed this locus and that the number of *PHF7*-like  
263 genes whose existence we demonstrated in the present study is underestimated. Another point  
264 of putative under-estimation is that we did not target any BAC corresponding to the second  
265 hybridisation of BAC G on ChrZ (with a different locus from BAC B and BAC C also on  
266 ChrZ), which may contain several other *PHF7*-like genes. Moreover, one can also imagine  
267 that there are *PHF7*-like genes present in genomic regions present in the upstream 5'  
268 extremity and the downstream 3' extremity of the sequenced BAC, especially for BAC that  
269 contain *PHF7*-like genes at their extremities, as is the case for BAC I and BAC E. A limit of  
270 our approach is that it does not allow sequencing BAC that were not detected by the screen  
271 with the initial primers (which are dependent on sequences available in the databases).  
272 Nevertheless, the chicken genome is relatively ‘young’ compared with the well-sequenced  
273 mouse genome (in which we have characterised massive duplications of reproductive genes  
274 [3-6, 10]) or the human genome. The quality of the chicken genome assembly may be  
275 optimised further in the future. Moreover, the chicken genome exhibits microchromosomes  
276 that are very difficult to sequence (GC-rich sequences). In the present study, we used the  
277 PacBio method to sequence the BAC [20]. The advantage of using PacBio’s Circular  
278 Consensus Sequence (CCS) method is that it provides very high-fidelity, quality reads and  
279 allows obtaining a unique contig for each BAC clone. The technique allows correcting the  
280 sequence reads – the greater the number of repeated passes, the higher the Phred quality value

281 (QV) – and it allows considerably reducing polymorphisms that could be due to the  
282 technology itself (polymerase bias). Coupled with the evolution of PacBio chemistry, which  
283 allows reads between 15 and 20 kb (compared with ~7–8 kb with the old chemistry), it  
284 presents the great advantage to go beyond the repeated zones (whose average size is 10 kb)  
285 and to obtain a good quality assembly [21-23].

286 The technical strength of PacBio sequencing lies in new tools allowing for greater  
287 sequencing depth, thus better alignment and high-quality assembly. Thus, complex genomic  
288 regions that are difficult to access with more classical sequencing/assembly methods are now  
289 easier to access. Coupled with FISH to target specific genomic regions of the chicken bank of  
290 BAC available in our lab, it is very efficient to characterise a massive duplication family in  
291 the chicken genome, as we have done for *PHF7* in the present study.

292 Another source of difficulty and thus variability in results is the annotation process.  
293 High duplication is relatively rare, and thus the more recent EnSEMBL versions have  
294 probably simplified annotation with automatic algorithmic processes that eliminate massive  
295 duplications in the sequence, aiming to avoid putative false-positive gene redundancies. We  
296 have already observed this phenomenon with the expanding *FFAR2* family: a version of  
297 EnSEMBL described a family of 23 paralogues, then in the following versions (including the  
298 current) only one *FFAR2* gene was present, whereas we experimentally found that the chicken  
299 genome contains 22 ( $\pm 2$ ) paralogues [10].

300 Concerning our ‘own’ *ab initio* annotation of BAC sequences, we had the choice  
301 between three classical annotation programs available: Augustus, GENESCAN and Softberry,  
302 the last one based on FGESH program that is the subject of a number of publications (as  
303 reviewed previously [24]). One inconvenience of GENESCAN is that no species can be  
304 targeted – only the vertebrate class. A convenient advantage of Softberry compared with the  
305 other two programs is the availability of a large amount of data – in particular, we retrieved  
306 each exon sequence so that we could align them. Annotation of predicted genes was almost  
307 identical for the three programs, with few differences. One concerns gene 1 in BAC F:  
308 Augustus and GENSCAN instead predicted three and two genes, respectively (the long gene 1  
309 is split into three and two smaller genes, that are also *PHF7*-like genes). Subtleties in the  
310 annotation process/program may explain such variations [25], which may be a further source  
311 of underestimation of *PHF7*-like genes in the present results. Of note, all three programs  
312 predicted the strange long gene 5 in BAC H exhibiting 70 exons, corresponding to a repetition  
313 of seven *PHF7* ‘core-structures’. We performed several trials by RT-PCR dedicated to long  
314 RNA, but we were unable to find a trace of this long mRNA (10 kb) that could correspond to

315 gene 5 in BAC H. Moreover, we have our own NGS data from chicken testis (data not  
316 published but deposited in GEO<sup>2</sup>) but could not find long reads matching BAC H. Thus, it is  
317 still unclear whether such large mRNA (and its corresponding protein) exists or is an artefact  
318 and a consequence of imperfect annotation with existing tools. According to their exon  
319 structures (see suppl Fig 4), these long genes may be split in smaller “conventional” *PHF7*-  
320 like genes with 8-10 exons, and would thus correspond to: 7 genes in BAC H (instead of one  
321 70-exons gene), 6 genes in BAC E (instead of two 30-exons genes) and 3 genes in BAC F  
322 (instead of one 28-exons gene). Thus, instead of 4 long genes, there may be 12 smaller *PHF7*-  
323 like genes (for a putative total of 52 *PHF7*-like genes instead of 39).

324 *PHF7* protein is expressed in male germ cells during spermiogenesis and involved in  
325 histone-to-protamine exchange. In *Drosophila melanogaster*, deletion mutants of *Phf7* have  
326 demonstrated the important role of this gene for male fertility [16, 26]. Male infertility in mice  
327 with *Phf7* deletion is due to aberrant histone retention and impaired protamine replacement in  
328 elongated spermatids [27]. In the chicken, as in the rat and human, *PHF7* mRNA expression  
329 is much higher, if not even exclusive, in the testis compared with other tissues [14, 18]. In this  
330 species, we observed an increase in the mRNA level with the animal’s age (data not shown),  
331 suggesting germ cell expression in the chicken as in other species. Additional studies would  
332 be needed to better characterise the protein expression and function in the chicken testis. At  
333 the evolutionary level, a previous study established that one copy of the gene is present in the  
334 *Drosophila* and mammalian genomes, whereas several copies (but not characterised/counted)  
335 are present on two loci in the chicken genome [18]. Moreover, these authors showed that *Phf7*  
336 has a common ancestor with *G2e3* (G2/M-phase specific E3 ubiquitin protein ligase). Both  
337 genes arose from a duplication before the divergence of vertebrates, and non-vertebrates have  
338 only one gene of this family. These two proteins possess three zinc fingers (PHD domains and  
339 RING fingers, respectively, for *Phf7* and *G2e3*) in their N-terminus. In their study, Wang et  
340 al. [18] showed that *G2e3* is present in all metazoan genomes whereas *Phf7* is absent in fish  
341 and reptile genomes. Currently, however, there are reptile orthologues of the mouse *Phf7* gene  
342 that can be found in Ensembl release 105 (a unique orthologue in each of the Goodes thorn  
343 scrub tortoise, the painted turtle, the Abingdon Island giant tortoise; two in the three-toed box  
344 turtle). Because there are numerous predicted duplications of *PHF7* in other birds (Ensembl  
345 release 105: turkey, 24; Japanese quail, 9; zebra finch, 2; collared flycatcher, 2; duck, 5), we

---

<sup>2</sup> <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE133401> (available in 2023).

346 hypothesise that the massive *PHF7* duplication is restricted to birds (at least until better  
347 annotation is available in reptiles), with no hypothesis of their biological sense (if any).

348 In the chicken genome, we previously characterised the *FFAR2* [10] gene massive  
349 duplication (22 paralogues, whereas a unique gene is present in the mouse and human) and  
350 exhibiting a high testicular level of expression. In our more recent study, we identified, in  
351 addition to *PHF7*, two other chicken testicular family genes with many paralogues (whereas a  
352 unique gene in the mouse): *SUN3* (11 paralogues) and *SPAG4/SUN5* (20 paralogues) [14]. As  
353 for *PHF7* and for reasons discussed above, the presence of these numerous paralogous genes  
354 fluctuates according to databases and their successive versions (for example, there is a unique  
355 *FFAR2* gene in the chicken genome according to current EnSEMBL version 105). Studies  
356 similar to the one we have conducted here for *PHF7* would be of interest to characterise such  
357 families; however, such work is time-consuming and expensive. Better annotation of the *G.*  
358 *gallus* genome (Galgal7) may improve identification of massively duplicated regions. Another  
359 interesting point would be to improve annotation in other bird species, as well as in other  
360 sauropsids, aiming to conduct evolutionary studies to date massive duplications.

361 From a functional point of view, the significance of the presence of several, and  
362 sometimes many, paralogues in animal genomes is still unclear. In the mouse, when  
363 paralogues encode proteins with similar sequences and tissue expression, as is this the case for  
364 *PHF7* and other genes mentioned above, they are quite often, but not always, able to  
365 compensate for the loss (in mutants) of their paralogues. Indeed, there are cases of  
366 dispensability of paralogous genes and there are also cases of non-redundancy of paralogues  
367 [28]. For example, for *Nlrp5* (or *Mater*) and its paralogue *Nlrp4e* characterised in the lab [3,  
368 6], individual invalidation of each of them leads in both cases to a drastic phenotype of  
369 sterility with early embryonic death [29, 30]. Guschanski et al. [31] showed that, during  
370 vertebrate evolution, the contribution of paralogues to specific organ functions differs  
371 according to the organ, with paralogues expressed in young testis putatively involved in  
372 lineage-specific biology consistently with their reproductive function.

373 Overall, we have characterised a new expanding germ cell-specific gene family in the  
374 chicken genome. The *PHF7*-like proteins and genes exhibit a strong level of similarity (as  
375 stated by rakes in their phylogenetic tree) that may be explain by gene conversion process; this  
376 hypothesis would need future specific studies. Also, its functional and evolutionary  
377 significance (as for other germinal specific gene families as *Oogenesisin* or *Nlrp5* in the mouse)  
378 remain to be investigated further, given the lack/absence of information in the genome  
379 database due to the technical problems mentioned above. We need to clarify and illuminate

380 'the dark side' of the chicken genome, such regions with recent duplications still being the  
381 blind spot of genomic sequencing programmes.

382

383 **LEGENDS**

384 **Figure 1.**

385 (A) Fluorescence *in situ* hybridisation (FISH) localisations (white arrows) of bacterial  
386 artificial chromosome (BAC) clones (red signals) screened for groups of genes (GrA to GrI)  
387 based on the international standard of the chicken karyotype [32]. Green signals are for  
388 microchromosome detection (GrA and GrG). GrA (A), WAG-038I15 (red) together with  
389 WAG-062P02 (green) on GGA19; GrB (B), WAG-119F08 on GGAZq12-13; GrC (C), WAG-  
390 041O23 on GGAZp12-21; GrD (D), WAG-023G20 on GGA1p26; GrE (E), WAG-038A02 on  
391 GGA1p26; GrF (F), WAG-037B06 on GGA2q11-21; GrG (G), WAG-035C03 on GGAZp21-  
392 22 and together with WAG-033L02 (green) on GGA12; GrH (H), WAG-034E06 on  
393 GGA2p11-12; GrI (I), WAG-119J04 on GGA1q35-41.

394 (B) Fluorescence *in situ* hybridisation (FISH) localisations of the gene groups (GrA to GrI)  
395 obtained in Fig 1A drawn on the chicken standard idiograms [32] from the measures  
396 summarised in Table 1.

397 **Figure 2.**

398 BAC I gene map obtained after annotation by FGENESH and BLASTN/BLASTP. The grey box  
399 represents the BAC I with its first nucleotide in position 1. The coloured boxes represent the 22 genes  
400 predicted by FGENESH (with their relative coordinates on the BAC); the colours of the text 'Gen#'  
401 refer to BLASTP results described in Supplemental Table 5, with families in colour (uncoloured genes  
402 are unique): Yellow = PHF7 family; red = translation initiation factor IF-2-like isoform X2 family;  
403 pink = reverse transcriptase family; green = centrosomal protein of 126 kDa family. Other BAC maps  
404 are in Supplemental Fig 3.

405 **Figure 3.**

406 Phylogenetic tree of the 39 PHF7-like proteins (sequences are in Supplemental Table 6). See  
407 the Materials and methods for details on its construction. The bootstrap values are in red.

408

409



## 410 MATERIALS AND METHODS

### 411 *In silico* and genomic PCR to characterize chicken *PHF7* genes

412 We used *PHF7* primers (nine pairs) listed in Supplemental Table 3 (column B) that we had  
413 designed in our previous work [14] (to study mRNA expression). These primers were  
414 designed using NCBI “primer-blast” tool. It was not possible to design a unique primer pair  
415 for the 63 EnSEMBL genes, but a pair was obtained for each of the 9 subtrees/groups (suppl  
416 Fig 1 and suppl Table 3), allowing to cover all sequences of chicken paralogs. In the present  
417 study, we used them to perform *in silico* genomic PCR with the tool available at  
418 <https://genome.ucsc.edu/cgi-bin/hgPcr>. Genomic PCR (30 cycles: 95°C for 10 s, 60°C for 10 s  
419 and 72°C for 30 s), with chicken DNA extracted from the blood of Leghorn chicken (pool of  
420 three animals; kindly provided by Amélie Juanchich, BOA INRAE F-37380 Nouzilly)  
421 allowed us to verify their specificity on agarose gel (single amplicons at their theoretical sizes),  
422 before being used for probe production as described below.

### 423 FISH

424 For each of the nine groups (GrA to GrI), BAC clones were selected from the Wageningen  
425 chicken (White Leghorn breed) library by two-dimensional PCR screening of super-pools and  
426 pools arranged in microplates as described by Crooijmans et al. [33]. PCR amplifications  
427 were carried out for each group by using the primers listed in Supplemental Table 3 (column  
428 B) as follows: 35 cycles with denaturation at 95°C for 30 s, specific annealing at 60°C for 30  
429 s and elongation at 72°C for 30 s. Each 20 µl reaction contained 2 mM MgCl<sub>2</sub>, 0.2 mM  
430 dNTPs, 0.5 µM primers and 0.625 units Taq polymerase (Go Tad Flexi DNA polymerase  
431 Promega™ M3005). The reactions were run on an Applied Biosystems™ 2720 Thermal  
432 Cyclor.

433 After isolating a single colony on a Petri dish to avoid any risk of contamination, BAC  
434 clones were grown in 25 ml of LB medium with 34 µg/ml chloramphenicol. The DNA was  
435 extracted based on alkaline lysis using the Qiagen Plasmid Midi Kit. The presence of each  
436 group of genes in the corresponding BAC clone was checked by PCR as described previously.  
437 PCR products were sequenced by using the Sanger technique on the Get-Plage Genotoul  
438 Platform (GeT-PlaGe INRAE Auzeville F-31326 Castanet-Tolosan Cedex France) to confirm  
439 the gene identities. The sequenced were visualised with Chromas software and aligned by  
440 Blast.

441 FISH was carried out on metaphase spreads obtained from fibroblast cultures of 7-day-  
442 old chicken and duck embryos, arrested with 0.05 µg/ml colcemid (Sigma). After a 10 min  
443 hypotonic treatment (1:5 foetal calf serum hypotonic solution mixed equal parts with 0.075 M

444 KCl), the cell suspension was fixed overnight in a 3:1 ratio of ethanol to acetic acid solution  
445 and stored at -20°C until spreading.

446 The single-colour FISH protocol is based on Yerle et al. [34]. Briefly, 150 ng of DNA  
447 of each BAC clone was biotin labelled (biotin 16-dUTP) by random priming using the  
448 Bioprime Kit (Invitrogen). The probes were purified using MicroSpin G-50 columns (GE  
449 Healthcare Life sciences) to remove the non-incorporated nucleotides. Probes were ethanol  
450 precipitated and resuspended in 50% formamide hybridisation buffer. After denaturation of  
451 probes (7 min at 100°C) and chromosomes (2 min at 72°C in 70% formamide), slides were  
452 hybridised *in situ* for 17 h at 37°C in the presence of 5 µg chicken cot1 competitor DNA on a  
453 humid plate (Dako Hybridizer). After hybridisation, slides were washed 2 × 30 min in 2X  
454 SSC then 4 min at 73°C in 0.4X SSC. The biotin was detected with Alexa568-Streptavidin  
455 (from Invitrogen).

456 For group A (WAG-038I15) and G (WAG-035C03), expected to be located on a  
457 microchromosome pair, the corresponding BAC were co-hybridised with specific FISH  
458 markers (WAG-062P02 for GGA19 and WAG-033L02 for GGA12) used as references to  
459 identify precisely the microchromosome pairs involved [35-37]. Two-colour FISH was  
460 performed according to Trask et al. (1991) [38]. One probe was labelled with digoxigenin  
461 (digoxigenin-11-dUTP, Roche) and the other with biotin (biotin 16-dUTP) using the  
462 BioPrime Kit (Invitrogen). The two labelled probes were ethanol precipitated together before  
463 hybridisation. The biotin-labelled probe was detected with Alexa568-streptavidin and the  
464 digoxigenin labelled probe was detected with Alexa488-anti-digoxigenin (from Invitrogen).

465 Chromosomes were counterstained with DAPI in antifade solution (Vectashield with  
466 DAPI, Vector Laboratories-H-1200). The hybridised metaphases were screened with a Zeiss  
467 fluorescence microscope; a minimum of 20 spreads were analysed for each experiment. Spot-  
468 bearing metaphases were captured and analysed with a cooled CCD camera using Cytovision  
469 software (Leica Biosystem).

470 We defined the precise localisations for macrochromosomes by the fractional length  
471 from the p arm telomere (Flpter) after measurement of 10 chromosomes (Cytovision  
472 software), except for ChrZ (GGAZ), for which we used the fractional length from the  
473 centromere (Flcen) because this chromosome is difficult to orientate. We used Flpter and  
474 Flcen to determine the FISH localisation on the G-banded chicken standard karyotype as  
475 shown on Fig 1B [32].

476 **BAC clone sequencing: PacBio library preparation, sequencing and data assembly**

477 Individual BAC clone DNA was extracted by using the Nucleobond Xtra Midi Kit  
478 (Macherey-Nagel). Two micrograms of each sample was used to construct a multiplexed  
479 SMRTbell® library by the INRAE-CNRGV. We followed the PacBio recommendations for  
480 Multiplexed Microbial Library preparation (PN 101-696-100) with some adjustments by  
481 using the SMRTbell Express Prep kit v2.0 (Pacific Biosciences, Menlo Park, CA, USA). The  
482 first enzymatic steps consist of removing single-stranded overhangs, repairing any DNA  
483 damage and polishing the ends of the double-stranded fragments and tailing with an A-  
484 overhang. Ligation with specific barcoded hairpin T-overhang adapters to both ends of the  
485 targeted double-stranded DNA (dsDNA) molecule creates a closed, single-stranded circular  
486 DNA. Each individual sample was treated with nuclease by using SMRTbell Enzyme Clean-  
487 up kit (Pacific Biosciences). The Blue-Pippin size-selection system (Sage Science, Beverly,  
488 MA, USA) was used to remove fragments < 15 kb from pooled sample previously purified  
489 with 0.45X AMPure PB beads (Pacific Biosciences). The size and concentration of the final  
490 library were assessed using the FemtoPulse system and the Qubit Fluorometer and Qubit  
491 dsDNA HS reagents Assay kit (Thermo Fisher Scientific, Waltham, MA, USA), respectively.

492 Sequencing primer v2 and Sequel DNA Polymerase 2.0 were annealed and bound,  
493 respectively, to the SMRTbell library. The library was loaded onto one SMRTcell at an on-  
494 plate concentration between 50 and 85 pM by using a diffusion loading. Sequencing was  
495 performed on the Sequel II system with a run movie time of 30 h with 120 min pre-extension  
496 step and Software v9.0 (PacBio) by Gentyane Genomic Platform (INRAE-Clermont-Ferrand,  
497 France).

498 We corrected the PacBio raw reads by using SMRTLink\_v9.0.0 with eight passes, then  
499 demultiplexed the data. We identified residual *Escherichia coli* reads by using BLAST+  
500 2.10.0 and removed them by using Seqfilter. We filtered the HiFi reads by identifying the  
501 vector sequences using cross\_match and removed them with custom Perl scripts. We filtered  
502 HiFi reads < 15 kb by using Seqfilter, and then subsampled with SeqKit to obtain an  
503 estimated average assembly depth of 50X. We assembled the reads with hifiasm-0.12. To  
504 validate the result, we checked the length of the obtained contig and mapped BAC end  
505 sequences with the extremities on the assembly using BLAST+ 2.10.0. We remapped HiFi  
506 reads to the assembly and obtained the depth with samtools-1.8.

507 ***In silico* analyses (except *in silico* PCR)**

508 [Orthology link](#)

509 We obtained *in silico* data concerning *PHF7* orthologues in birds (and other species) from  
510 Ensembl <https://www.ensembl.org/> (from version 88 [March 2017] to the current version  
511 105 [December 2021]).

#### 512 Gene annotation

513 For gene structure prediction of the BAC nucleotide sequences, we performed *ab initio*  
514 annotation by using the commonly used FGENESH program on the Softberry site [39], which  
515 relies on hidden Markov model (HMM) statistical models to identify promoters, coding or  
516 noncoding regions, and intron–exon junctions (available at <http://www.softberry.com/>) [24].  
517 When needed, we consulted two other classical programs: Augustus ([http://bioinf.uni-  
518 greifswald.de/augustus/submission.php](http://bioinf.uni-<br/>518 greifswald.de/augustus/submission.php)) and GENESCAN  
519 (<http://hollywood.mit.edu/GENSCAN.html>).

#### 520 Similarity with sequences in databases using NCBI BLAST

521 We searched for similarity and/or functional homology by using the BLAST tool of NCBI  
522 (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>). We performed BLASTN with the nucleotide  
523 sequence of each BAC on the galGal7 genome (GCRg7w, white Leghorn layer, NCBI 106).  
524 We performed BLASTP with the 116 peptides (sequences in Supplemental Table 6) obtained  
525 with FGENESH. The criteria of the BLASTP interrogation were ‘by default’ or, when no  
526 results were obtained, the threshold was upgraded (we arbitrarily chose 1000; in this case, E-  
527 value may be high, i.e. > 0).

#### 528 Phylogeny and sequence alignment and similarity

529 We constructed the phylogenetic tree from the PHF7-like protein sequences with  
530 <http://www.phylogeny.fr/alacarte.cgi> (MUSCLE for Multiple Alignment; Gblocks for  
531 Alignment curation; construction of phylogenetic tree with PhyML; visualisation of  
532 phylogenetic tree with TreeDyn). The bootstrap values were estimated with 1000 replications  
533 and the tree was rooted using midpoint rooting method. We created a representation of the  
534 PHF7 sequence alignment with the program ESPript (Easy Sequencing in PostScript,  
535 available at <https://esprict.ibcp.fr/ESPript/ESPript/>) [40], which displays sequence similarities  
536 from aligned sequences.

537 **Acknowledgments**

538 SF and PM are very grateful to Charlotte Lécureuil for her careful re-reading of the  
539 manuscript. We thank Amélie Juanchich (INRAE BOA Nouzilly) for providing chicken  
540 DNA. The manuscript was proofread and corrected by Proof-Reading-Service.com.

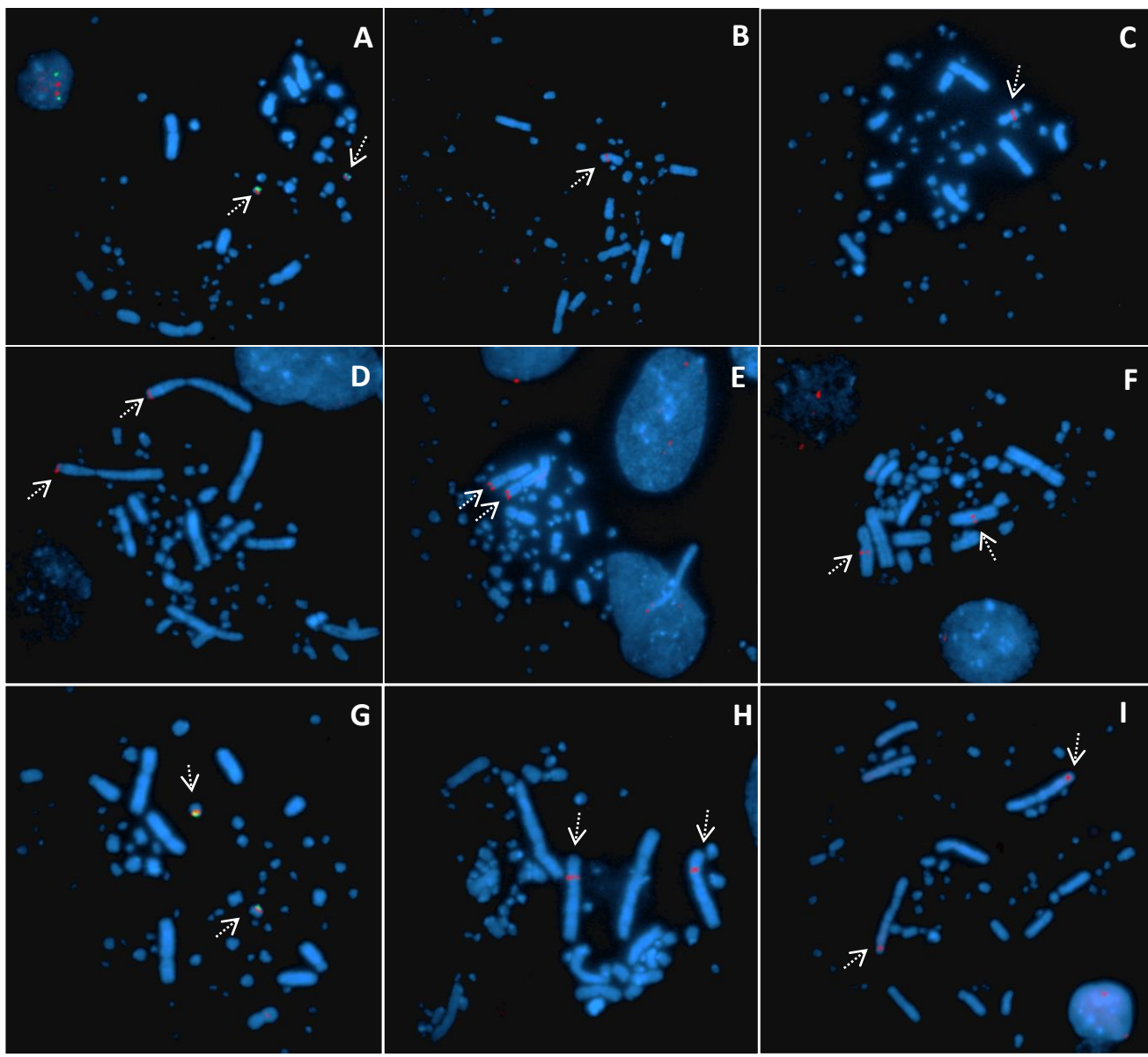
541

542 **References**

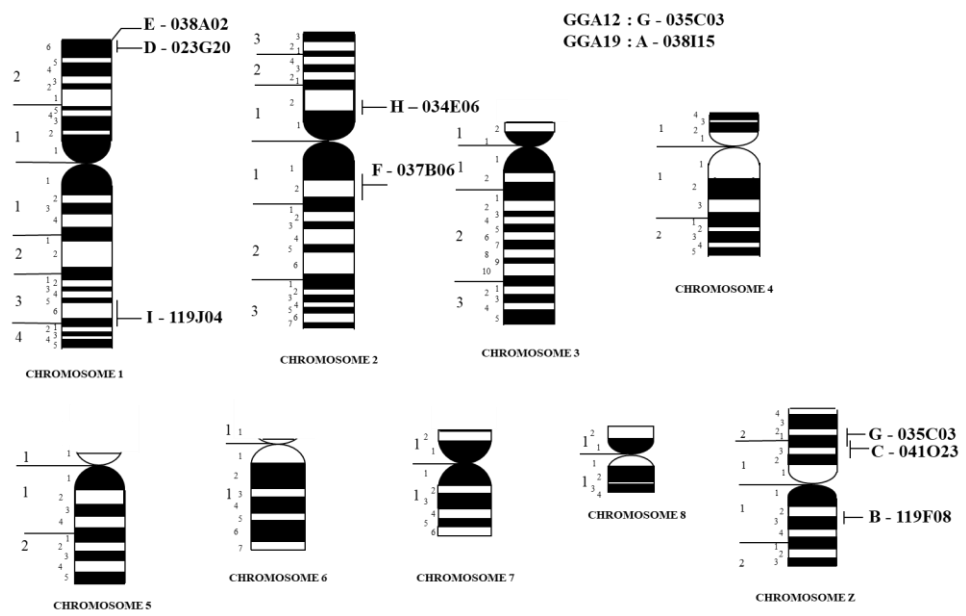
- 543 1. Demuth, J.P. and M.W. Hahn, *The life and death of gene families*. Bioessays, 2009. **31**(1): p.  
544 29-39.
- 545 2. Taylor, J.S. and J. Raes, *Duplication and divergence: the evolution of new genes and old ideas*.  
546 Annu Rev Genet, 2004. **38**: p. 615-43.
- 547 3. Dade, S., et al., *Identification of a new expanding family of genes characterized by atypical*  
548 *LRR domains. Localization of a cluster preferentially expressed in oocyte*. FEBS Lett, 2003.  
549 **555**(3): p. 533-8.
- 550 4. Dade, S., et al., *In silico identification and structural features of six new genes similar to*  
551 *MATER specifically expressed in the oocyte*. Biochem Biophys Res Commun, 2004. **324**(2): p.  
552 547-53.
- 553 5. Paillisson, A., et al., *Identification, characterization and metagenome analysis of oocyte-*  
554 *specific genes organized in clusters in the mouse genome*. BMC Genomics, 2005. **6**: p. 76.
- 555 6. Tian, X., G. Pascal, and P. Monget, *Evolution and functional divergence of NLRP genes in*  
556 *mammalian reproductive systems*. BMC Evol Biol, 2009. **9**: p. 202.
- 557 7. Lee, M.O., et al., *Duplication of chicken defensin7 gene generated by gene conversion and*  
558 *homologous recombination*. Proc Natl Acad Sci U S A, 2016. **113**(48): p. 13815-13820.
- 559 8. Temperley, N.D., et al., *Evolution of the chicken Toll-like receptor gene family: a story of gene*  
560 *gain and gene loss*. BMC Genomics, 2008. **9**: p. 62.
- 561 9. Zhang, T., et al., *Genomic organization of the chicken TCRB locus originated by duplication of*  
562 *a Vβ segment combined with a trypsinogen gene*. Vet Immunol Immunopathol, 2020. **219**: p.  
563 109974.
- 564 10. Meslin, C., et al., *Expanding Duplication of Free Fatty Acid Receptor-2 (GPR43) Genes in the*  
565 *Chicken Genome*. Genome Biol Evol, 2015. **7**(5): p. 1332-48.
- 566 11. Bhasin, S., C. Mallidis, and K. Ma, *The genetic basis of infertility in men*. Baillieres Best Pract  
567 Res Clin Endocrinol Metab, 2000. **14**(3): p. 363-88.
- 568 12. Bonilla, E. and E.Y. Xu, *Identification and characterization of novel mammalian*  
569 *spermatogenic genes conserved from fly to human*. Mol Hum Reprod, 2008. **14**(3): p. 137-42.
- 570 13. Rodgers-Melnick, E.B. and R.K. Naz, *Male-biased genes of Drosophila melanogaster that are*  
571 *conserved in mammalian testis*. Front Biosci (Elite Ed), 2010. **2**: p. 841-8.
- 572 14. Fouchecourt, S., et al., *An evolutionary approach to recover genes predominantly expressed*  
573 *in the testes of the zebrafish, chicken and mouse*. BMC Evol Biol, 2019. **19**(1): p. 137. doi:  
574 10.1186/s12862-019-1462-8.
- 575 15. Kim, C.R., et al., *PHF7 Modulates BRDT Stability and Histone-to-Protamine Exchange during*  
576 *Spermiogenesis*. Cell Rep, 2020. **32**(4): p. 107950.
- 577 16. Yang, S.Y., E.M. Baxter, and M. Van Doren, *Phf7 controls male sex determination in the*  
578 *Drosophila germline*. Dev Cell, 2012. **22**(5): p. 1041-51.
- 579 17. Merkin, J., et al., *Evolutionary dynamics of gene and isoform regulation in Mammalian*  
580 *tissues*. Science, 2012. **338**(6114): p. 1593-9.
- 581 18. Wang, X.R., et al., *Evidence for parallel evolution of a gene involved in the regulation of*  
582 *spermatogenesis*. Proc Biol Sci, 2017. **284**(1855). doi: 10.1098/rspb.2017.0324.
- 583 19. English, A.C., et al., *Mind the gap: upgrading genomes with Pacific Biosciences RS long-read*  
584 *sequencing technology*. PLoS One, 2012. **7**(11): p. e47768.
- 585 20. Athanasopoulou, K., et al., *Third-Generation Sequencing: The Spearhead towards the Radical*  
586 *Transformation of Modern Genomics*. Life (Basel), 2021. **12**(1). doi: 10.3390/life12010030.

- 587 21. Hon, T., et al., *Highly accurate long-read HiFi sequencing data for five complex genomes*. Sci  
588 Data, 2020. **7**(1): p. 399.
- 589 22. Murigneux, V., et al., *Comparison of long-read methods for sequencing and assembly of a*  
590 *plant genome*. Gigascience, 2020. **9**(12).
- 591 23. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant*  
592 *detection and assembly of a human genome*. Nat Biotechnol, 2019. **37**(10): p. 1155-1162.
- 593 24. Ejigu, G.F. and J. Jung, *Review on the Computational Genome Annotation of Sequences*  
594 *Obtained by Next-Generation Sequencing*. Biology (Basel), 2020. **9**(9). doi:  
595 10.3390/biology9090295.
- 596 25. Mudge, J.M. and J. Harrow, *The state of play in higher eukaryote gene annotation*. Nat Rev  
597 Genet, 2016. **17**(12): p. 758-772.
- 598 26. Yang, S.Y., et al., *Control of a Novel Spermatocyte-Promoting Factor by the Male Germline Sex*  
599 *Determination Factor PHF7 of Drosophila melanogaster*. Genetics, 2017. **206**(4): p. 1939-  
600 1949.
- 601 27. Wang, X., et al., *PHF7 is a novel histone H2A E3 ligase prior to histone-to-protamine exchange*  
602 *during spermiogenesis*. Development, 2019. **146**(13).
- 603 28. Roth, C., et al., *Evolution after gene duplication: models, mechanisms, sequences, systems,*  
604 *and organisms*. J Exp Zool B Mol Dev Evol, 2007. **308**(1): p. 58-73.
- 605 29. Chang, B.H., et al., *Developmental expression and possible functional roles of mouse Nlrp4e in*  
606 *preimplantation embryos*. In Vitro Cell Dev Biol Anim, 2013. **49**(7): p. 548-53.
- 607 30. Tong, Z.B., et al., *Mater, a maternal effect gene required for early embryonic development in*  
608 *mice*. Nat Genet, 2000. **26**(3): p. 267-8.
- 609 31. Guschanski, K., M. Warnefors, and H. Kaessmann, *The evolution of duplicate gene expression*  
610 *in mammalian organs*. Genome Res, 2017. **27**(9): p. 1461-1474.
- 611 32. Ladjali-Mohammed, K., et al., *International system for standardized avian karyotypes*  
612 *(ISSAK): standardized banded karyotypes of the domestic fowl (Gallus domesticus)*. Cytogenet  
613 Cell Genet, 1999. **86**(3-4): p. 271-6.
- 614 33. Crooijmans, R.P., et al., *Two-dimensional screening of the Wageningen chicken BAC library*.  
615 Mamm Genome, 2000. **11**(5): p. 360-3.
- 616 34. Yerle, M., et al., *Localization of the pig luteinizing hormone/choriogonadotropin receptor*  
617 *gene (LHCGR) by radioactive and nonradioactive in situ hybridization*. Cytogenet Cell Genet,  
618 1992. **59**(1): p. 48-51.
- 619 35. Fillon, V., et al., *Identification of 16 chicken microchromosomes by molecular markers using*  
620 *two-colour fluorescence in situ hybridization (FISH)*. Chromosome Res, 1998. **6**(4): p. 307-13.
- 621 36. Masabanda, J.S., et al., *Molecular cytogenetic definition of the chicken genome: the first*  
622 *complete avian karyotype*. Genetics, 2004. **166**(3): p. 1367-73.
- 623 37. Douaud, M., et al., *Addition of the microchromosome GGA25 to the chicken genome*  
624 *sequence assembly through radiation hybrid and genetic mapping*. BMC Genomics, 2008. **9**:  
625 p. 129.
- 626 38. Trask, B.J., et al., *Mapping of human chromosome Xq28 by two-color fluorescence in situ*  
627 *hybridization of DNA sequences to interphase cell nuclei*. Am J Hum Genet, 1991. **48**(1): p. 1-  
628 15.
- 629 39. Solovyev, V., et al., *Automatic annotation of eukaryotic genes, pseudogenes and promoters*.  
630 Genome Biol, 2006. **7** Suppl 1(Suppl 1): p. S10.1-12.
- 631 40. Gouet, P., X. Robert, and E. Courcelle, *ESPrpt/ENDscript: Extracting and rendering sequence*  
632 *and 3D information from atomic structures of proteins*. Nucleic Acids Res, 2003. **31**(13): p.  
633 3320-3.
- 634

**A**



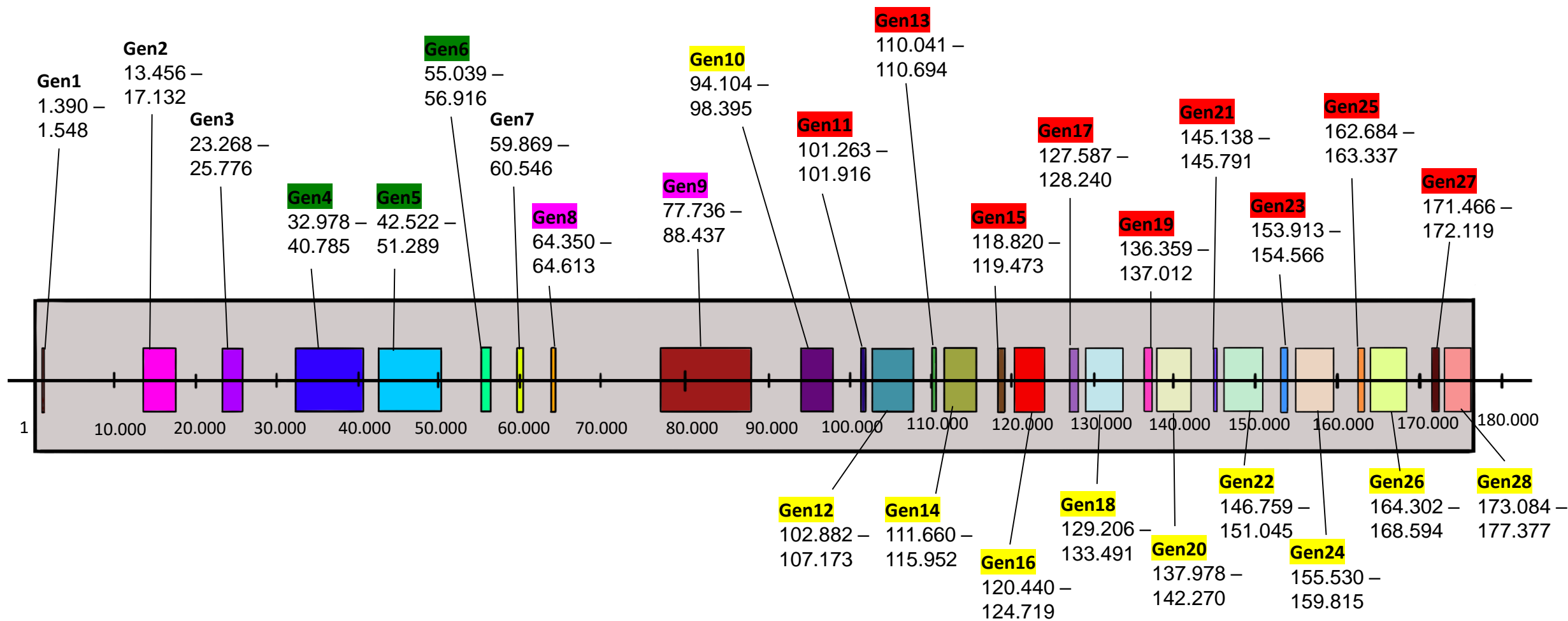
**B**



BAC I : 177.594 bp

Blastn result: Chromosome 1 with following coordinates

NCBI version 106 (GRCg7w): 182.685.165 - 182.769.800

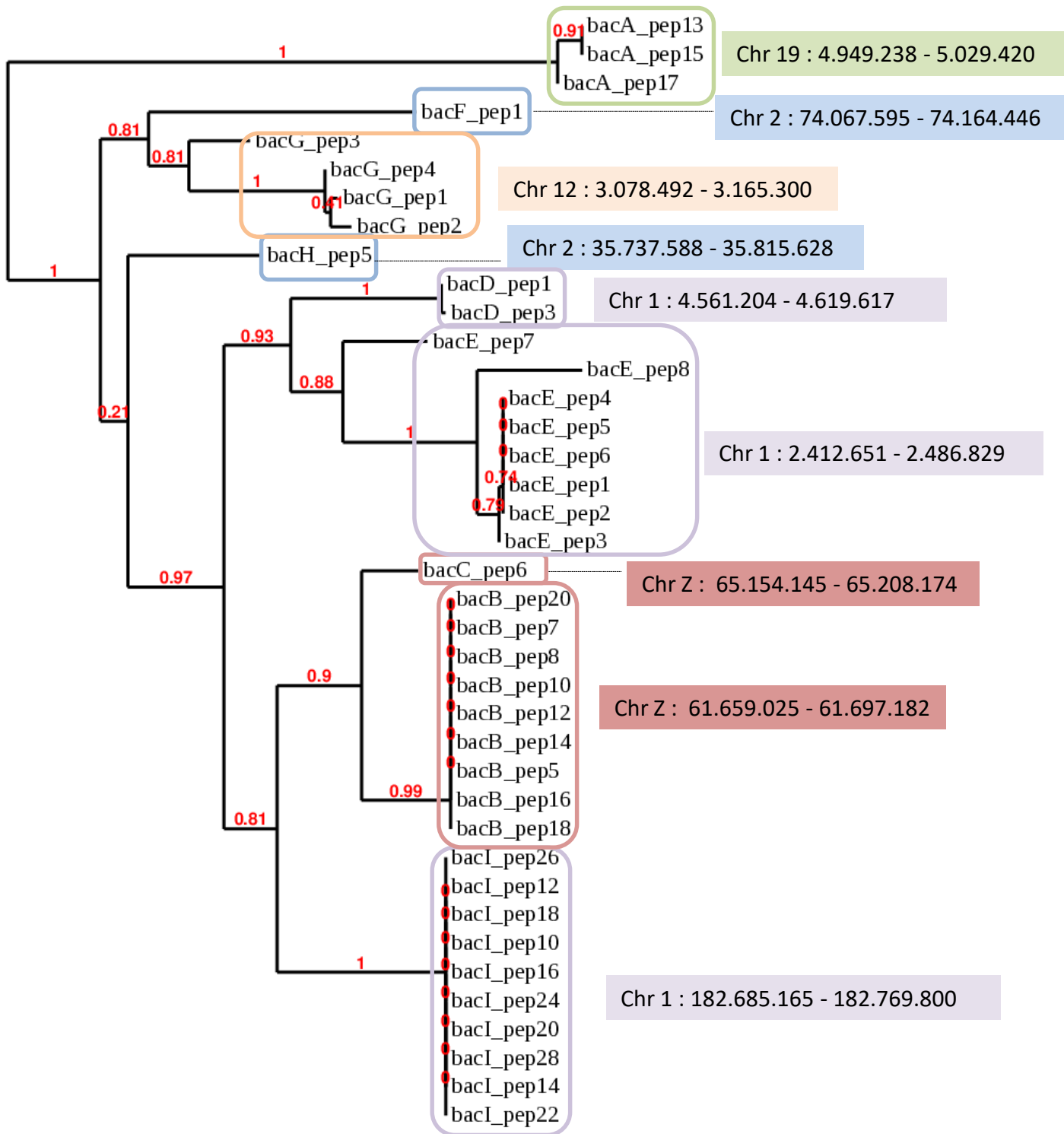


Gen4/5/6: centrosomal protein of 126 kDa

Gen10/12/14//16/18/20/22/24/26/28: PHF7-like

Gen11/13/15/17/19/21/23/25/27: translation initiation factor family





0.5

Chr 1 Chr 2 Chr 12 Chr 19 Chr Z