

RESEARCH ARTICLE

Open Access



# Expansion of the *HSFY* gene family in pig lineages

Benjamin M Skinner<sup>1</sup>, Kim Lachani<sup>1</sup>, Carole A Sargent<sup>1</sup>, Fengtang Yang<sup>2</sup>, Peter Ellis<sup>1,3</sup>, Toby Hunt<sup>2</sup>, Beiyuan Fu<sup>2</sup>, Sandra Louzada<sup>2</sup>, Carol Churcher<sup>2,4</sup>, Chris Tyler-Smith<sup>2</sup> and Nabeel A Affara<sup>1\*</sup>

## Abstract

**Background:** Amplified gene families on sex chromosomes can harbour genes with important biological functions, especially relating to fertility. The Y-linked heat shock transcription factor (*HSFY*) family has become amplified on the Y chromosome of the domestic pig (*Sus scrofa*), in an apparently independent event to an *HSFY* expansion on the Y chromosome of cattle (*Bos taurus*). Although the biological functions of *HSFY* genes are poorly understood, they appear to be involved in gametogenesis in a number of mammalian species, and, in cattle, *HSFY* gene copy number may correlate with levels of fertility.

**Results:** We have investigated the *HSFY* family in domestic pig, and other suid species including warthog, bushpig, babirusa and peccaries. The domestic pig contains at least two amplified variants of *HSFY*, distinguished predominantly by presence or absence of a SINE within the intron. Both these variants are expressed in testis, and both are present in approximately 50 copies each in a single cluster on the short arm of the Y. The longer form has multiple nonsense mutations rendering it likely non-functional, but many of the shorter forms still have coding potential. Other suid species also have these two variants of *HSFY*, and estimates of copy number suggest the *HSFY* family may have amplified independently twice during suid evolution.

**Conclusions:** The *HSFY* genes have become amplified in multiple species lineages independently. *HSFY* is predominantly expressed in testis in domestic pig, a pattern conserved with cattle, in which *HSFY* may play a role in fertility. Further investigation of the potential associations of *HSFY* with fertility and testis development may be of agricultural interest.

**Keywords:** Evolution, Sex chromosomes, Gene sequencing

## Background

Sex chromosomes, and Y chromosomes in particular, are sites of frequent evolutionary change, due in part to the smaller population size of these chromosomes compared to autosomes, the lack of recombination on the Y, frequent and dramatic remodelling of the Y chromosome, and the accumulation of ampliconic sequences. During our collaborative project sequencing the pig X and Y chromosomes (*in submission*; preprint in [1]), we became interested in a sequence that appeared repeatedly in the data being produced: the pig Y-linked heat shock transcription factor, *HSFY*. The structure and organisation of the pig X and Y chromosomes are described in previous

papers [1–3]. Briefly, the Y chromosome long arm is highly repetitive, with all known single copy genes on the short arm. A central band of repetitive material is also found on the short arm at cytogenetic band Yp1.2.

Little is known about the biological function of *HSFY*. In humans, two functional *HSFY* copies are found in the azoospermia factor b (AZFb) region [4, 5]. Deletions in the AZFb region of the human Y are usually linked to problems with fertility in patients; however, microdeletions in the AZFb region affecting only *HSFY* do not seem to impair fertility [6]. The gene encodes a heat shock transcription factor, but it appears not to function as such in humans; the DNA binding region does not bind DNA, and no promoters have been identified that *HSFY* targets specifically [5, 6]. Yet, earlier reports have suggested alterations to *HSFY* expression are associated with maturation arrest of spermatogenic cells [7].

\* Correspondence: na106@cam.ac.uk

<sup>1</sup>Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK  
Full list of author information is available at the end of the article

Consequently, the biological functions of *HSFY* remain something of a black box in humans, let alone other species.

Looking across mammalian genomes, *HSFY* orthologues can be found from marsupial mammals to eutherian mammals, and the gene appears to be identifiable even in birds, with the inference that it has important roles in at least some species [8]. Mammalian *HSFY* seems present in low but variable copy number across many species, with between two and eight copies in cats [9, 10], at least one retroposed active copy in mice [8] and the two active copies in humans [4] plus several pseudogene copies, one of which is found on chromosome 22. In cattle, the gene family has amplified to at least 70 copies [11, 12]. Recent work suggests that the amplification in cattle occurred after their divergence with sheep, and so is likely an independent amplification to that in pigs [13].

Gene amplifications on the sex chromosomes are particularly interesting; recent work in mice has linked the amplification of gene families on both the X and Y chromosomes to an ongoing genomic conflict affecting X gene expression and ultimately sex ratio skewing [14]. Given the evolutionary pressures on the sex chromosomes, and the homogametic chromosome in particular, genomic conflicts are likely to be widespread across species with sex chromosome systems. Deletions on the mouse Y chromosome are known to generate reduced fertility and sperm head shape abnormalities, linked to an ongoing genomic conflict between the sex chromosomes [15]. This phenotype could be recapitulated by targeted deletion of the mouse autosomal gene *Hsf2* [16], demonstrating that these classes of transcription factors can adopt key roles in chromatin organisation. The human autosomal homologue of *HSF2* is also associated with defects in spermatogenesis [17]. The classical description of the heat shock family of transcription factors is that they bind heat-shock response elements in gene promoters and activate transcription in conditions of heat or other stresses. It is clear though that the activity of heat-shock genes is not limited to stress responses, and they have important roles in development and gametogenesis [18].

The potential independent amplification of *HSFY* in both pigs and cattle is intriguing, as this suggests the gene is 'prone' to amplification on mammalian Y chromosomes, either driving or carried along with a genomic conflict. This, plus the association with fertility, makes the *HSFY* genes an important family to characterise. We show that in pigs, *HSFY* is amplified on the short arm of the chromosome, in two variant forms, with at least 100 copies combined. Both forms are expressed in testis, though only one is likely to produce a functional product. We also find both variants in other suid species, with differing copy numbers suggesting independent amplifications during suid diversification.

## Results

### Pigs carry at least two forms of HSFY

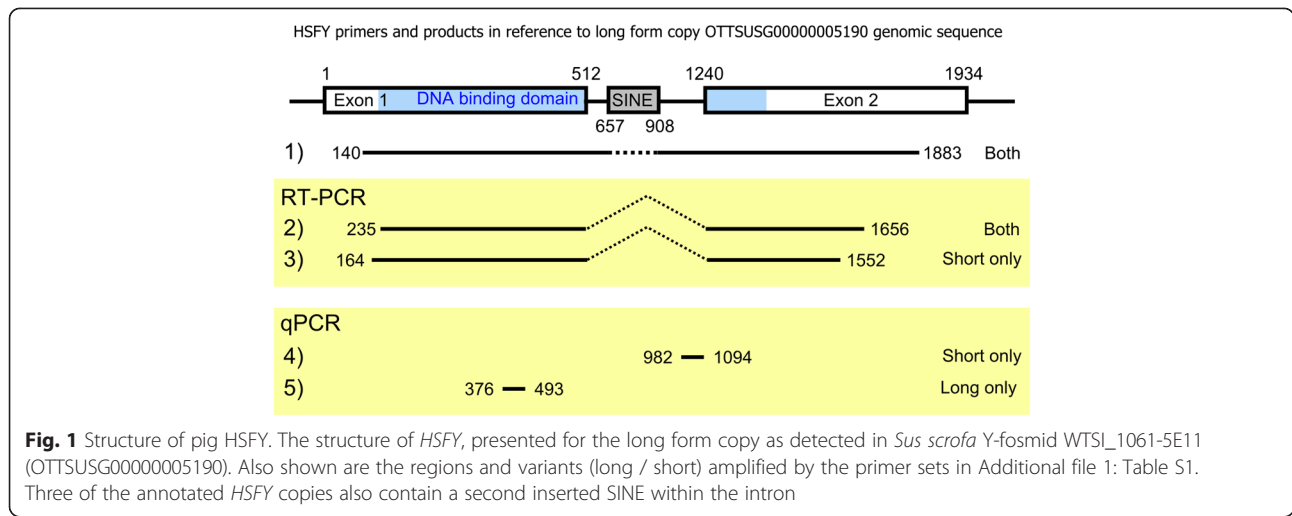
As the sequencing of pig Y chromosome fosmid clones progressed [1], it became clear that some sequences were present at a high copy number. BLAST searches of these sequences against the NCBI nucleotide database suggested that one class of repetitive sequence involved the pig homologues of the *HSFY* genes. Based on the alignment of fosmid clones identified with *HSFY* copies, we designed primers to amplify fragments across both exons and the intron over all these known copies. These are shown in Additional file 1: Table S1 as primer set 1; the schematic diagram of the *HSFY* gene and the regions amplified are also given in Fig. 1. Using these primers for PCR on genomic DNA from male and female domestic pigs, we obtained distinct sequences of two lengths in the male DNA, dubbed the 'long' and 'short' forms (see Fig. 2 genomic controls). The PCR products were subcloned and sequenced, and the products showed 99 % sequence identity to sequence annotated within the fosmid clones. Representative examples of short and long forms can be found in the Vega database with the accessions OTTSUSG00000005615 and OTTSUSG00000005190 respectively. We also found three instances of a third form of the sequence, longer still, supported by a faint band in the gels (not shown), exemplified by accession OTTSUSG00000002741.

All the genomic sequences from domestic pig were aligned. The alignment showed the sequences fell primarily into two distinct clusters (*OTTSUSG*- sequences in Fig. 3), with a key differentiator being an insertion of sequence within the intronic region of the longer form. Examination of the inserted sequence shows it to be a pig-lineage short interspersed nuclear element (SINE), Pre0\_SS [19]. The third longer form contained two SINES within the intron.

### Sus scrofa has more than 100 copies of HSFY

In order to estimate the number of copies of *HSFY* in the domestic pig genome, we designed quantitative PCR (qPCR) primers that would amplify specifically from the 'long' copies or from the 'short' copies (Primer set 2 and 3, Additional file 1: Table S2). The results were normalised against *SRY* copy number. This has long been assumed to be a single copy gene in the domestic pig, but we have recently found it to be present in two copies [1]. Based on an *SRY* copy number of two, and with strong caveats that this assumes ideal PCR efficiency and equal signal from all amplicons, we detected around 68 long form and 42 short form copies of *HSFY*.

Multi-copy genes are a common feature of Y chromosomes, either tandemly repeated or dispersed across the chromosome. In order to determine the physical organisation of *HSFY* on the pig Y, we performed fluorescence



*in-situ* hybridisation (FISH) using four fosmids known to contain HSFY copies. These fosmids co-localised on the short arm of the Y in metaphase chromosome spreads (Fig. 4a). FISH on extended DNA fibres showed that the fosmids (and the HSFY sequences they contain) are dispersed within this region amongst other, as-yet unidentified, sequences (Fig. 4b). The total size of the HSFY-containing block, estimated from cytogenetic measurements, is about 5 Mb. The annotated copies in Vega contain 37 short form and 27 long form copies. These were derived from 21 fosmid clones that end-to-end would span a little under one megabase. Comparing to our estimated copy numbers, there is scope for many other sequences within this region.

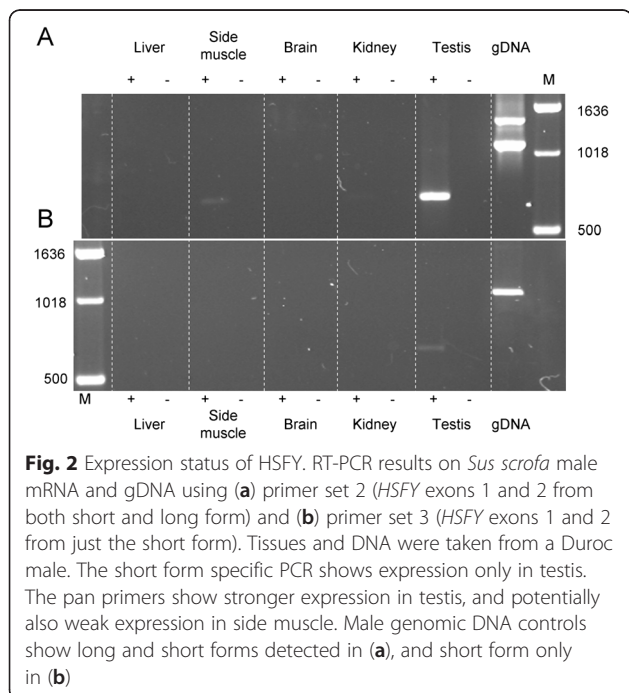
**Expression of HSFY**

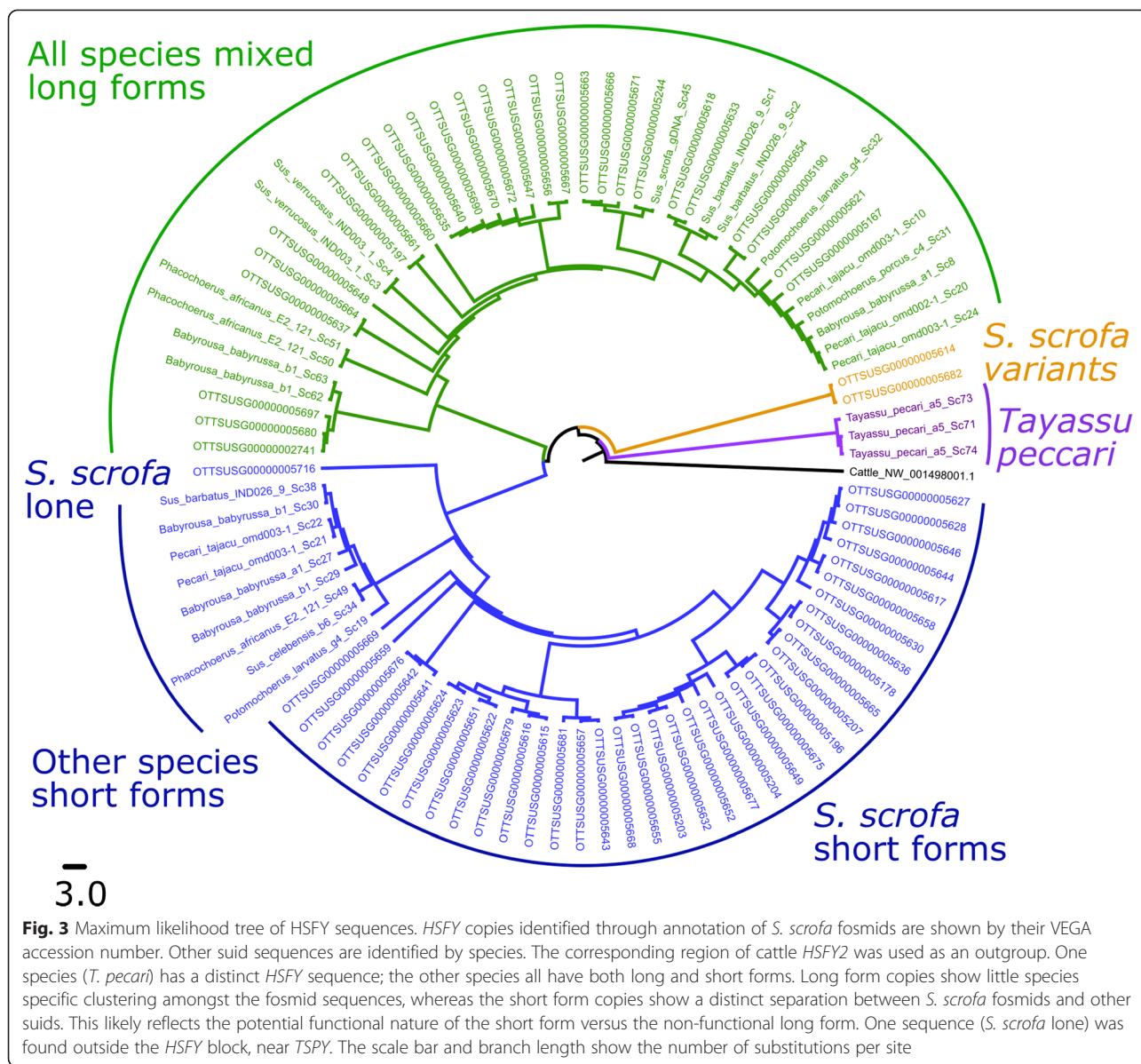
Multi-copy genes are of particular interest if they are expressed, especially if their expression is restricted to a certain tissue or cell type. We used reverse transcriptase PCR (RT-PCR) to characterise expression patterns of the long and short forms of HSFY in a range of tissues (liver, side muscle, brain, kidney and testis) from the same animal. The primers amplified the bulk of exons 1 and 2, and the two sets designed were able to amplify either the short variants specifically, or both long and short variants together (Primer sets 2 and 3, Additional file 1: Table S1, Fig. 1). The RT-PCRs showed some expression from the short form in testis. More expression in testis was seen with the ‘both form’ primers, suggesting higher expression from the long form. The ‘both form’ primers also suggest some low levels of expression in side (loin) muscle.

At the sequence level, the long form copies all have multiple sequence changes disrupting the open reading frame, making it unlikely that a functional protein product is produced. In contrast, most of the short form copies appear to retain an open reading frame, and all but one of the identified pig expressed sequence tags (ESTs) matching our HSFY sequences cluster with the short form copies (see Additional file 2: Figure S1).

**HSFY variants in other suiforms**

In order to provide information on the date of the amplification of HSFY, we investigated a range of related species (Table 1) with the primer sets we had available. Primer set 1 was able to amplify products from all the suids we tested, and sequencing of these products revealed both long and short forms were present in all individual animals we tested of each species, with the exception of *T. pecari*. However, current assessment of suid phylogeny [20] has the *T. pecari/P. tajacu* split occurring after the divergence from the *Sus* lineage. Consequently, it seems more likely that





these sequences represent a different evolutionary history in the *T. peccari* lineage, and not an ancestral state. This places the initial divergence between the long and short forms - i.e. the insertion of the SINE - before the diversification of modern suids (Fig. 3). The peccary sequences also cluster outside the two variant sequences in *Sus scrofa*.

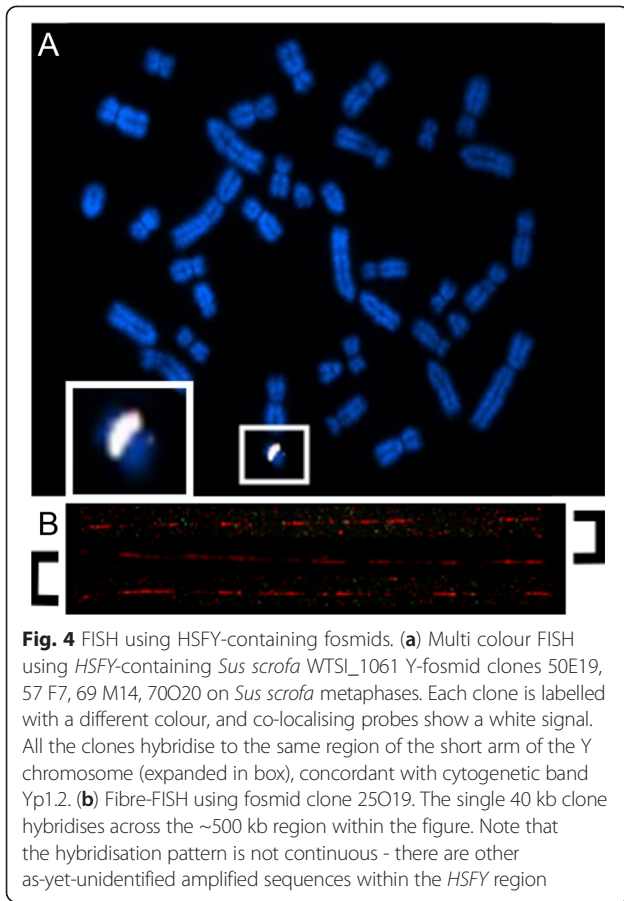
The qPCR primers were also used to estimate the copy number of the detectable forms of HSFY in each species. Again, normalisation was performed in reference to *SRY*. Although we know *Sus scrofa* has two copies of *SRY*, we do not know when this duplication occurred. Hence the error bars on the cross-species qPCR results include both possibilities (Fig. 5). Conservatively considering all possible combinations of single or dual-copy *SRY*, some patterns emerge:

- 1) The *Sus* genus has a consistent high copy number (~100 copies) of HSFY.
- 2) *B. babirusa* has a low copy number due to lack or amplification, or amplification of variants undetectable here, and may represent the ancestral state.
- 3) Warthog (*P. africanus*) and bushpig (*P. larvatus*) have considerable difference between them that may be attributable to different levels of ongoing expansion of HSFY in each lineage.

**Selection within potentially coding HSFY copies**

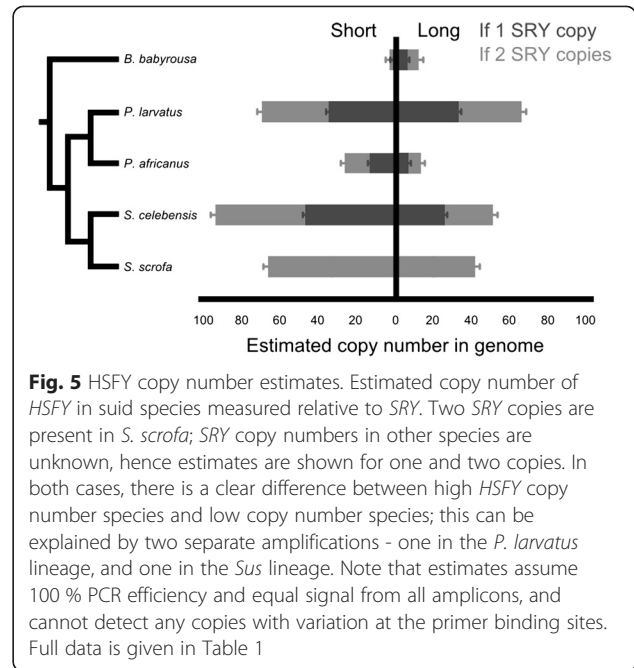
None of the long form copies of HSFY found in *Sus scrofa* or any of the other suids had coding potential. They all have frameshift and nonsense mutations





occurring within the DNA-binding domain of the product rendering them likely non-functional. Given that there is considerable gene expression from the long forms in domestic pig, it is possible that the transcripts are simply noise, or that the transcripts do not produce a protein product, but the RNAs have acquired a new function. Since there appears to be no sequence preservation in the different suid lineages (as evidenced by the ‘mixed’ clustering of the long form products), transcriptional noise seems the most likely explanation.

We therefore looked at the short form copies with a viable open reading frame, and tested for evidence of



positive or purifying selection. The results of these tests are given in Additional file 1: Tables S3 and S4, and revealed that there is little evidence for positive selection amongst *HSFY* copies. There was more significant evidence for purifying selection, both within *Sus scrofa*, and between species, as summarised in Table 2. Multi-gene families are often subject to gene conversion; we tested for evidence of regions of gene conversion with the coding *HSFY* copies using GENECONV, and found no significant regions.

### Discussion

#### HSFY genes are amplified on pig Yp

Amplified genes and gene families are a common feature of Y chromosomes in mammals - indeed, of sex chromosomes in general. The *HSFY* genes are an example of this in pigs. We have shown here that there are two forms of *HSFY*, long and short. Both forms are present at high copy number on the Y chromosome, almost entirely located within a single cytogenetic band on the

**Table 1** qPCR results

Species	Copy number relative to <i>SRY</i> (+/- SEM)		Absolute <i>HSFY</i> copy number estimate	
	Short	Long	Short	Long
<i>Sus scrofa</i>	34 (1.09)	21 (1.25)	68	42
<i>Sus celebensis</i>	47 (1.11)	25 (1.18)	47/94	25/50
<i>Potamochoerus larvatus</i>	35 (1.15)	33 (1.18)	35/70	33/66
<i>Phacochoerus africanus</i>	13 (1.09)	6 (1.05)	13/26	6/12
<i>Babyrousa babyrousa</i>	1 (1.09)	6 (1.09)	1/2	6/12

Results of qPCR on five species of long and short form copy number relative to *SRY* (primer sets 4–6 in Additional file 1: Table S1). Since *SRY* copy number is uncertain outside *S. scrofa*, absolute values for *HSFY* are given for one and two *SRY* copies. See also Figure 5

**Table 2** Purifying selection test

	<i>Sus scrofa</i>	<i>Sus barbatus</i>	<i>Sus celebensis</i>	<i>Phacochoerus africanus</i>	<i>Babyrousa babyrussa</i>	<i>Pecari tajacu</i>	<i>Tayassu pecari</i>
<i>Sus scrofa</i>	9/27	11/27	16/27	27/27	15/27	18/27	27/27
<i>Sus barbatus</i>			Sig	Sig	N/S	N/S	Sig
<i>Sus celebensis</i>				Sig	N/S	N/S	Sig
<i>Phacochoerus africanus</i>					Sig	Sig	Sig
<i>Babyrousa babyrussa</i>						N/S	2/2
<i>Pecari tajacu</i>							2/2
<i>Tayassu pecari</i>							

Test for purifying selection, summarising the pairwise comparisons detected as significant at the 0.05 level or not significant (N/S). Where multiple sequences were available for a species, the number of significant comparisons are given. Full data are given in Additional file 1: Table S3

short arm. Expression analysis reveals that both forms are expressed, though evidence from EST libraries and our sequencing suggests that only the short forms have coding potential. Nonetheless, pseudogenes can acquire biological functions, for example as regulatory long non-coding RNAs [21], and thus there remains a possibility for functionality to be identified in future.

The major structural difference between the long and short forms is the presence or absence of a SINE within the intron. This SINE - *PreO\_SS* - is annotated in Repbase as being a still active pig lineage specific tRNA SINE [19]. Given that we can find long and short forms in all the suiform species in this study, it is probable that the SINE originally inserted when there were a small number of *HSFY* copies in the ancestral genome, and subsequently both long and short copies underwent amplification. Given that we found two copies that do not cluster with long or short form, it is likely that there are other variants of *HSFY* not detected by our primer sets.

Estimates of the overall copy number of *HSFY* (Table 2, Fig. 5) suggest that there are about 100 copies in the domestic pig genome, split between long and short forms, with a bias toward the short form. This number is based on comparison to *SRY*, which until recently was believed to be a single copy gene in suids (see [1]). The other four species presented may also have two *SRY* copies, but the estimates we generate consider both possibilities. The other *Sus* member, *S. celebensis*, has 70 or 140 copies; again with a bias towards the short form. This suggests there has been only limited expansion of the *HSFY*s in either lineage from their common ancestor. The babirusa (*B. babirusa*) is an outgroup to the other species, and only a small number of copies were detected (1-2 short, 6-12 long); either *B. babirusa* has significant copy number loss or sequence divergence, or the *HSFY* amplifications predominantly occurred after the *B. babirusa* lineage diverged from other suids. The remaining two species tested provide tentative support to the latter scenario: the warthog *P. africanus* has a low number of copies (20–40), compared to the bushpig *P. larvatus* (70–140). Based on

the phylogeny of these species, the most plausible explanation for this pattern is an amplification within the *P. larvatus* lineage. Consequently, even with the caveats of *SRY* copy number and broadness of primer coverage, there is evidence supporting two independent bursts of *HSFY* amplification within the suids.

The study here has focussed on the *HSFY* genes. However, the FISH analysis has demonstrated that the ~5 Mb *HSFY* region of the Y chromosome is not solely composed of *HSFY* copies. The full extent of the other sequences within this region is not known, due to the difficulty of assembling highly repetitive sequences reliably. Still, there are two other identified genes close to *HSFY* copies that are also amplified, thought to be pseudogenes (*RPS2* and *XKR3-like*; see also [1] for the complete context of the pig Y).

Amplified genes on the sex chromosomes have been associated with genomic conflicts in mice (e.g. [14]). These genes generally act by favouring the transmission of the chromosome on which they reside, or by suppressing the transmission of their opposite gametologue [15]. The situation in pig is different to known genomic conflict models, however, in that there are no observed gene family expansions on the X chromosome that might be responding to the expansion on the Y (see [1]), and we therefore consider that a similar mechanism of genomic conflict is unlikely. The X-chromosome homologue of *HSFY*, *HSFX*, was previously predicted (Genbank: XM\_005654314.1). As with *HSFX/HSFY* comparisons in other species (e.g. [11]), there is little sequence identity between the X and Y copies. Indeed, the only alignable region is the DNA binding domain. It is clear that if there is any biological role for *HSFY*, it has been distinct from *HSFX* for the majority of mammalian Y chromosome evolution.

A further possibility is that the expansion is evolutionarily neutral - a concentration of repetitive material provided a substrate for process such as non-allelic sister chromatid exchange, causing sequence amplification, but without any selective pressure, or a biological function associated with the increase. This seems less likely; if

there were no functional role for the extra *HSFY* copies, we would expect to see an accumulation of mutations within both short and long forms, abolishing the open reading frame. However, the short form copies appear to be predominantly translatable. The status of the promoters is not clear, given we have sequences for only a subset of the total *HSFY* complement: weakened or disrupted promoter activity could 'normalise' expression to the level of a single gene copy (and this would be consistent with the apparently lower expression levels we found from the short form (Fig. 3). Further work is required to distinguish between these possibilities.

Our tests for evidence of selection for rapid amino acid change suggested no evidence for such positive selection. However, there was strong evidence for purifying selection amongst the coding copies between *Sus scrofa* and the other species, and between the *Sus scrofa* copies themselves. This again supports the idea that the copy number of these genes is functionally relevant, and that this function is maintained amongst the suid species studied here.

#### Further *HSFY* variants may be present

Two *S. scrofa* non-coding *HSFY* variants lack a SINE, but also do not cluster with the short form copies (OTT-SUSG00000005614 and 5682; orange in Fig. 3). These variants have nucleotide differences within the binding regions for the primer sets 1, and could not be detected in expression or copy number studies in *S. scrofa* or any other species. It is thus possible that these are two representatives of a further diversification of the *HSFY* family; the sampled fosmids cover only a small portion of the complete ~5 Mb *HSFY*-block.

One species showed a different organisation to the others: *Tayassu pecari*, the white-lipped peccary. Neither the consensus short form nor the long form was identified. Instead, three similar variant species-specific forms were seen (purple sequences in Fig. 4). None of these appear to have coding potential, nor is it known what the copy number of *HSFY* is in any of the peccary species. Both peccary species share a common ancestor after the divergence with the suids approximately 40 million years ago. Since *P. tajacu* has at least one each of long and short forms, it is most likely that there has been little amplification in the peccary lineage, and species-specific diversification of the *HSFY* copies in *T. pecari*. Previous comparative chromosome painting studies have suggested that the peccaries have higher rates of chromosomal rearrangement than suids [22]. Of the two peccary species in this study, the *T. pecari* karyotype appears the more derived [23, 24], and this may contribute to the differences seen in *T. pecari*.

#### A single *HSFY* pseudocopy lies outside the main block near *TSPY*

One *HSFY* copy (OTT-SUSG00000005716) in domestic pig lies outside the *HSFY*-block, close to *TSPY* [1]. It has a premature stop codon within the DNA binding domain of the first exon, and thus cannot form a valid *HSFY* product, nor do we have evidence it is expressed. The sequence is similar to the short form, but clusters distinctly outside the other short forms (Fig. 3; *S. scrofa* lone). Its presence could be attributable to (1) an ancestral *HSFY* copy (many other species have multiple *HSFY* copies, and perhaps one of these copies gave rise to the long and short forms while the other remained unamplified; or (2) this is derived from another short form copy that relocated from the *HSFY* block during the evolution of the pig Y chromosome. We reconstructed the series of rearrangements on the Y chromosome from the ancestral mammalian Y as described in our associated X and Y sequencing paper [1], but see no obvious opportunity for an *HSFY* copy to be relocated to the vicinity of *SRY* or *RBMY*. This does not preclude more complex undocumented rearrangements. Further cross-species cytogenetics will be able to investigate this possibility.

#### Comparison with cattle suggests independent amplifications

Cattle also have a documented expansion of *HSFY* [11, 12], also with no apparent corresponding *HSFX* expansion on the X chromosome. This opened the possibility that the amplification predated the bovine/suid divergence, and was then maintained in each lineage. Recent evidence from sheep has suggested that this is not the case, the cattle *HSFY* expansion occurring after sheep and cattle diverged about 22 million years ago [13], with variation in *HSFY* copy number between different cattle breeds [12]. Accordingly, our alignments of pig *HSFY* sequences to documented cattle *HSFY* sequences show no evidence for the intronic SINE that distinguishes the long and short forms, and which must predate the initial amplification of the copies in pig. As a result, there are multiple lines of evidence pointing to independent amplifications of *HSFY* in these two lineages.

Further to this, our qPCR data provide tentative support for at least two separate amplifications of *HSFY* within the suids: once within the *P. larvatus* lineage, and again in the *Sus* lineage. However, this is subject to uncertainties of *SRY* copy number in each species and variation in qPCR primer binding sites; full confirmation of the copy numbers will require a more detailed sequencing approach to detect all variants of *HSFY* in each species.

From an evolutionary perspective, recurrent amplifications are very interesting; we do not know if the *HSFY* expansion is neutral, driven by chance and the genomic

landscape within which they occur, or subject to selection for increased copy number, with an important biological role. In humans, the active *HSFY* genes are expressed in Sertoli cells and spermatogenic cells, potentially with a different role in each [5]; in cattle, *HSFY* is expressed in spermatogonial and spermatocyte cells [11]. Some evidence from cattle breeds has suggested an inverse correlation between *HSFY* copy number and testicular size, and a positive correlation with conception rate [12]. It is possible that similar phenotypes will be associated with the *HSFY* genes in pigs. Testicle size in pigs is correlated with the levels of the hormone androstenedione in body fats [25], which is predominantly genetically determined [26]. High levels of androstenedione contribute to an unpleasant odour in male carcasses called boar taint; currently male piglets are often castrated to reduce the risk of this taint developing. Consequently, understanding any associations of *HSFY* genes with fertility and testis development will be of particular interest to the animal breeding industries. However, it remains to be determined whether *HSFY* copy number is variable between individuals or breeds of domestic pigs.

## Conclusion

Y chromosomes are hotspots of evolutionary change and diversity, and it is becoming clear with increasing number of sequenced Y chromosomes that the evolutionary pressures on the sex chromosomes can drive the amplification of particular genes with dramatic functional consequences. It remains to be seen whether *HSFY* has a functional role driving its expansion, or if it has been carried as a by-product of some other process in pig and other species. Nonetheless, it appears that some genes are predisposed to amplification by their roles, locations or both.

## Methods

### Animal ethics statement

Duroc tissue samples were provided by Genus PLC from a boar culled in a routine process. DNA samples from other suid species were provided by Genus PLC. No ethical approval was required for this study.

### Amplification of *HSFY* from genomic DNA

Primers were designed against *HSFY* sequence from pig Y fosmids. The fosmid sequences were previously deposited in Genbank as part of the pig X and Y sequencing efforts described elsewhere [1]. Primer set 1 (Additional file 1: Table S1) was tested on male and female gDNA from *Sus scrofa*, and on male genomic DNA from the other species in Table 1; the sampled species have been previously described ([27], save *S. verrucosus* and *P. tajacu* DNA, which were provided by Lawrence Schook (University of Illinois). PCR followed standard protocols

using Taq polymerase and buffer kit (Roche). Cycling conditions were 95 °C for 3mins, 25 cycles of 95 °C/ 56 °C/ 72 °C 30s for 30s/30s/60s, followed by 72 °C for 10mins.

### Subcloning and sequencing

PCR products were ethanol precipitated and subcloned into pGEM<sup>®</sup>-T Easy Vector System I (Promega) following the manufacturer's instructions. Ligations were carried out at 16 °C overnight. Ligated products were transformed into XL1-Blue Competent cells (Agilent) following the manufacturer's protocols, plated on LB-ampicillin plates supplemented with x-gal and IPTG, and incubated overnight at 37 °C. White colonies were selected and used for PCR with SP6 and T7 primers to confirm product insertion. PCR products were purified following agarose gel electrophoresis and then sequenced at the sequencing facility in the Department of Genetics, University of Cambridge. Prior to the sequencing reactions, the PCR products were purified using ExoSAP-IT (USB Corporation, USA) following the manufacturer's recommendations (samples were held at 37 °C for 30 min, 80 °C for 15 min, and chilled at 4 °C until removed from the machine). Amplicons were sequenced using Big Dye version 3.1 (Applied Biosystems). The sequencing program consisted of 30 cycles of: 96 °C/55 °C/60 °C for 10s/5s/4 min. Products were run on an ABI 3100 capillary sequencer. Traces were edited using Chromas version 2.2 (Technelysium Pty Ltd). Sequences from different species were viewed using the MultAlin program (<http://multalin.toulouse.inra.fr/multalin/>), and within the ClustalW2 program (<http://www.ebi.ac.uk/Tools/clustalw2/>). Repetitive content within sequences was analysed using RepeatMasker [28]. Sequences have been uploaded to GenBank under accessions KP211992-KP212018. *HSFY* copies from fosmids in the domestic pig were annotated as part of the pig X and Y chromosome sequencing project [1].

### Alignments and evolutionary analysis

Analysis of *HSFY* sequences was performed in MEGA6 [29]. The maximum likelihood nucleotide substitution model was identified as the Tamura 3-parameter model [30], and the evolutionary history was inferred using the Maximum Likelihood method based on this model. The tree with the highest log likelihood (-13948.7067) is shown in Fig. 3. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach, and then selecting the topology with superior log likelihood value. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.9667)). Branch lengths are measured in the number of substitutions per site. The analysis involved 94 nucleotide sequences. All



positions with less than 95 % site coverage were eliminated. That is, fewer than 5 % alignment gaps, missing data, and ambiguous bases were allowed at any position. There were a total of 1433 positions in the final dataset.

Selection tests were conducted using the Nei-Gojobori method [31]. The analysis involved 35 nucleotide sequences. All ambiguous positions were removed for each sequence pair, for a total of 339 positions in the final dataset.

ESTs corresponding to *HSFY* were identified in the NCBI EST database using BLAST (CV866737, CV873904, CX058656, CX063068, EW632312, EW633910, EW636148). Cattle *HSFY* sequence copies were identified from BLAST searches of cattle genome sequences, GenBank: NC\_016145.1 [32], using the *HSFY* sequence determined by [11], and trimmed to match the regions of *HSFY* covered in our analysis. All sequences were aligned as above, and a phylogenetic tree was constructed using the same parameters. Newick format files for both trees are provided in Additional files 3 and 4.

The 35 *HSFY* nucleotide sequences with a potentially valid open reading frame were examined separately for evidence of selection. The intronic sequences were removed, and all sequences were trimmed to the region amplified by primer set 1. The test statistics ( $d_S - d_N$ ) were calculated for each pairwise sequence comparison, and are given in Additional file 1: Tables S3 and S4. Analyses were conducted using the Nei-Gojobori method [30].

Tests for gene conversion amongst *S. scrofa* coding *HSFY* copies were performed using GENECONV (<http://www.math.wustl.edu/~sawyer>) to identify silent polymorphic sites. (full parameters/w123 /lp /sp /r).

### Expression analysis of *HSFY*

*HSFY* expression was tested only in *Sus scrofa*. Five tissues (liver, side muscle, brain, kidney and testis) were acquired from the same boar used in the pig Y chromosome sequencing project [1] and stored at -80°C in RNA-Later (Qiagen). Tissues were homogenised in Trizol, nucleic acids were extracted with phenol-chloroform and DNase treated. RNA was precipitated with isopropanol and stored at 1 µg/µl in ddH<sub>2</sub>O at -80 °C. RT-PCR was carried out using a OneStep RT-PCR kit (Qiagen) on 25 ng of total RNA. PCRs used primer sets 2 and 3, given in Table 3.

### qPCR

Copy number estimates were generated by quantitative PCR (qPCR). Primers (sets 4 and 5, Table 3) were designed to amplify ~100 bp fragments of the long form or the short form as wide a range of species as possible based on our sequenced products. Five species were covered: *S. scrofa*, *S. celebensis*, *P. larvatus*, *P. africanus* and

**Table 3** Suiform species in this study

Binomial name	Common name
<i>Sus scrofa</i>	Domestic pig (Duroc)
<i>Sus celebensis</i>	Sulawesi pig
<i>Sus verrucosus</i>	Java warty pig
<i>Sus barbatus</i>	Bornean bearded pig
<i>Potamochoerus larvatus</i>	Bushpig
<i>Potamochoerus porcus</i>	Red river hog
<i>Phacochoerus africanus</i>	Warthog
<i>Babyrousa babyrussa</i>	Buru babirusa
<i>Tayassu pecari</i>	White-lipped peccary
<i>Pecari tajacu</i>	Collared peccary

Peccaries are members of the family Tayassuidae; all other species are of the family Suidae. For a recent phylogeny of the suids, see Gongora et al [20]. A single animal from each species was studied

*B. babyrussa*. Control primers (set 6, Table 3) were designed against *SRY*, presumed to be single copy, targeted to amplify from all these five species. qPCR was performed using an iCycler (BioRad) and SYBR-FAST qPCR kit (Kapa Biosystems) on male gDNA. Cycling conditions were 95 °C for 3mins, followed by 40 cycles of 95 °C/ 57 °C/72 °C 30s for 10s/20s/30s. To enable use of consistent reference genes, an annealing temperature of 57 °C was used for all qPCR reactions. The fluorescent signal threshold crossing point (Ct) was normalized to the (presumed single-copy) *SRY* signal to produce a normalised  $\Delta C_t$  and an estimate of the absolute *HSFY* copy number as  $2^{\Delta C_t}$ . During the course of analysis it was determined that *SRY* is dual copy in *S. scrofa*; its status in other species remains undetermined. Results are presented for both possibilities in light of this.

### Preparation of single DNA-molecule fibres by molecular combing and fibre-FISH

Single-molecule DNA fibres were prepared by molecular combing [33] according the manufacturer's instructions (Genomic Vision) using fibroblast cells of a Duroc boar. Briefly, the cells were embedded in a low-melt-point agarose plug (1 million cells per 90 µl plug), followed by proteinase K digestion, washing in 1 × TE (10 mM Tris, 1 mM EDTA, pH8.0) & beta-agarase digestion steps. The DNA fibres were mechanically stretched onto saline-coated coverslips using a Molecular Combing System (Genomic Vision). To make FISH probes, purified fosmid DNAs were first amplified using a GenomePlex® Whole genome Amplification (WGA) kit (Sigma-Aldrich) following the manufacturer's protocols, then labelled using a WGA reamplification kit (Sigma-Aldrich) using a custom-made dNTP mix as described before [34]. For the fibre-FISH approximately 500 ng of labelled DNA from each probe and 4 µg of porcine Hybloc DNA (Applied Genetics Laboratories) were precipitated

using ethanol, then resuspended in a mix (1:1) of hybridisation buffer [containing  $2 \times$  SSC, 10 % sarkosyl, 2 M NaCl, 10 % SDS and blocking aid (Invitrogen)] and deionised formamide (final concentration 50 %). Coverslips coated with combed DNA fibres were dehydrated through an 70 %, 90 % and 100 % ethanol series and aged in 100 % ethanol at 65 °C for 30 s, followed by denaturation in an alkaline denature solution (0.5 M NaOH, 1.5 M NaCl) for 1-3 min, three washes with  $1 \times$  PBS (Invitrogen) and dehydration through an 70 %, 90 % and 100 % ethanol series. The probe mix was denatured at 65 °C for 10 min before being applied onto the coverslips and the hybridisation was carried out in a 37 °C incubator overnight. The post-hybridisation washes consisted of two rounds of washes in 50 % formamide/ $2 \times$  SSC (v/v), followed by two additional washes in  $2 \times$  SSC. All post-hybridisation washes were done at 25 °C, for 5 min each. Digoxigenin-11-dUTP (Roche) labelled probes were detected using a 1:100 dilution of monoclonal mouse anti-dig antibody (Sigma-Aldrich) and a 1:100 dilution of Texas Red-X-conjugated goat anti-mouse IgG (Molecular Probes/Invitrogen); DNP-11-dUTP (PerkinElmer) labelled probes were detected using with a 1:100 dilution of Alexa 488-conjugated rabbit anti-DNP IgG and 1:100 dilution of Alexa 488-conjugated donkey anti-rabbit IgG (Molecular Probes/Invitrogen); biotin-16-dUTP (Roche) labelled probes were detected with one layer of 1:100 dilutions of Cy3-avidin (Sigma-Aldrich). After detection, slides were mounted with SlowFade Gold® mounting solution containing 4',6-diamidino-2-phenylindole (Molecular Probes/Invitrogen). Images were visualised on a Zeiss AxioImager D1 microscope. Digital image capture and processing were carried out using the SmartCapture software (Digital Scientific UK).

### Availability of supporting data

Domestic pig Y chromosome sequences are available from Vega ([http://vega.sanger.ac.uk/Sus\\_scrofa/Info/Index](http://vega.sanger.ac.uk/Sus_scrofa/Info/Index)) with the accessions for domestic pig *HSFY* loci provided in Additional file 1: Table S2. *HSFY* sequences generated here from other species are deposited in Genbank under accessions KP211992-KP212018.

### Additional files

**Additional file 1: Supplementary Tables.** The file contains the following tables: Table S1 - Primers. Table S2 - Annotated *HSFY* loci on the domestic pig Y. Table S3 - Test for purifying selection on coding *HSFY* loci. Table S4 - Test for positive selection on coding *HSFY* loci.

**Additional file 2: Figure S1.** Tree including cattle *HSFY* and pig EST sequences. *HSFY* sequences aligned as described in Fig. 3, with the inclusion of cattle *HSFY* sequences and pig *HSFY* ESTs. The clustering shows that the ESTs are almost all associated with the short form of *HSFY*, and that the cattle sequences are distinct from all suid copies, reflecting their independent amplification.

**Additional file 3: *HSFY* tree in Newick format.** The Newick format tree in Fig. 3, created using FigTree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

**Additional file 4: Newick file of complete *HSFY* tree.** The Newick format tree in Additional file 2: Figure S1, created using FigTree v1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

### Abbreviations

AZFB: Azoospermia Factor B; BLAST: Basic local alignment search tool; EST: Expressed sequence tag; FISH: Fluorescence *in-situ* hybridisation; PCR: Polymerase chain reaction; qPCR: Quantitative PCR; RT-PCR: Reverse transcriptase PCR; SINE: Short interspersed nuclear element.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

BS and CAS designed primers, conducted the evolutionary analyses and wrote the manuscript. KL carried out the PCR reactions and cloning. PJJE performed qPCRs. FY, BF and SL performed the FISH analysis. TH performed the annotation of *HSFY* loci. NA, CTS and CC conceived and designed the study and wrote the manuscript. All authors read and approved the final manuscript.

### Acknowledgements

We gratefully acknowledge the Wellcome Trust Sanger Institute core teams for fingerprinting, mapping, archiving, library construction, sequence improvement and sequencing and Genus for providing the Duroc boar samples. This work was funded by BBSRC grant BB/F021372/1. The Flow Cytometry and Cytogenetics Core Facilities at the Wellcome Trust Sanger Institute and Sanger investigators are funded by the Wellcome Trust (grant number WT098051). The authors report no conflict of interests.

### Author details

<sup>1</sup>Department of Pathology, University of Cambridge, Cambridge CB2 1QP, UK. <sup>2</sup>Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. <sup>3</sup>Present addresses: School of Biosciences, University of Kent, Canterbury CT2 7NJ, UK. <sup>4</sup>Present addresses: School of Clinical Medicine, University of Cambridge, Cambridge CB2 0SP, UK.

Received: 11 January 2015 Accepted: 20 May 2015

Published online: 09 June 2015

### References

- Skinner BM, Sargent CA, Churcher C, Hunt T, Herrero J, Loveland J, et al. The pig X and Y chromosomes: structure, sequence and evolution. *bioRxiv*. 2014;012914. <http://dx.doi.org/10.1101/012914>.
- Skinner BM, Lachani K, Sargent CA, Affara NA. Regions of XY homology in the pig X chromosome and the boundary of the pseudoautosomal region. *BMC Genet*. 2013;14:3.
- Groenen MAM, Archibald AL, Uenishi H, Tuggle CK, Takeuchi Y, Rothschild MF, et al. Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*. 2012;491:393–8.
- Tessari A, Salata E, Ferlin A, Bartoloni L, Slongo ML, Foresta C. Characterization of *HSFY*, a novel AZFB gene on the Y chromosome with a possible role in human spermatogenesis. *Mol Hum Reprod*. 2004;10:253–8.
- Shinka T, Sato Y, Chen G, Naroda T, Kinoshita K, Unemi Y, et al. Molecular characterization of heat shock-like factor encoded on the human Y chromosome, and implications for male infertility. *Biol Reprod*. 2004;71:297–306.
- Kichine E, Rozé V, Cristofaro JD, Taulier D, Navarro A, Streichenberger E, et al. *HSFY* genes and the P4 palindrome in the AZFB interval of the human Y chromosome are not required for spermatocyte maturation. *Hum Reprod*. 2012;27:615–24.
- Sato Y, Yoshida K, Shinka T, Nozawa S, Nakahori Y, Iwamoto T. Altered expression pattern of heat shock transcription factor, Y chromosome (*HSFY*) may be related to altered differentiation of spermatogenic cells in testes with deteriorated spermatogenesis. *Fertil Steril*. 2006;86:612–8.

8. Kinoshita K, Shinka T, Sato Y, Kurahashi H, Kowa H, Chen G, et al. Expression analysis of a mouse orthologue of HSFY, a candidate for the azoospermic factor on the human Y chromosome. *J Med Investig JMI*. 2006;53:117–22.
9. Pearks Wilkerson AJ, Raudsepp T, Graves T, Albracht D, Warren W, Chowdhary BP, et al. Gene discovery and comparative analysis of X-degenerate genes from the domestic cat Y chromosome. *Genomics*. 2008;92:329–38.
10. Murphy WJ, Pearks Wilkerson AJ, Raudsepp T, Agarwala R, Schaffer AA, Stanyon R, et al. Novel gene acquisition on carnivore Y chromosomes. *PLoS Genet*. 2006;2.
11. Hamilton CK, Revay T, Domander R, Favetta LA, King WA. A large expansion of the HSFY gene family in cattle shows dispersion across Yq and testis-specific expression. *PLoS ONE*. 2011;6, e17790.
12. Yue X-P, Dechow C, Chang T-C, DeJarnette JM, Marshall CE, Lei C-Z, et al. Copy number variations of the extensively amplified Y-linked genes, HSFY and ZNF280BY, in cattle and their association with male reproductive traits in Holstein bulls. *BMC Genomics*. 2014;15:1–12.
13. Chang T-C, Yang Y, Retzel EF, Liu W-S. Male-specific region of the bovine Y chromosome is gene rich with a high transcriptomic activity in testis development. *Proc Natl Acad Sci*. 2013;110:12373–8.
14. Cocquet J, Ellis PJI, Mahadevaiah SK, Affara NA, Vairman D, Burgoyne PS. A genetic basis for a postmeiotic X versus Y chromosome intragenomic conflict in the mouse. *PLoS Genet*. 2012;8, e1002900.
15. Ellis PJI, Bacon J, Affara NA. Association of Sly with sex-linked gene amplification during mouse evolution: a side effect of genomic conflict in spermatids? *Hum Mol Genet*. 2011;20:3010–21.
16. Åkerfelt M, Henriksson E, Laiho A, Vihervaara A, Rautoma K, Kotaja N, et al. Promoter ChIP-chip analysis in mouse testis reveals Y chromosome occupancy by HSF2. *Proc Natl Acad Sci*. 2008;105:11224–9.
17. Mou L, Wang Y, Li H, Huang Y, Jiang T, Huang W, et al. A dominant-negative mutation of HSF2 associated with idiopathic azoospermia. *Hum Genet*. 2013;132:159–65.
18. Abane R, Mezger V. Roles of heat shock factors in gametogenesis and development. *FEBS J*. 2010;277:4150–72.
19. De Sapio F, Jurka J, Schook LB, Archibald AL, Faulkner GJ. Non-LTR retrotransposons from pig. *Repbases Rep*. 2010;10:1800.
20. Gongora J, Cuddahee RE, do Nascimento FF, Palgrave CJ, Lowden S, Ho SYW, et al. Rethinking the evolution of extant sub-Saharan African suids (Suidae, Artiodactyla). *Zool Scr*. 2011;40:327–35.
21. Milligan MJ, Lipovich L. Pseudogene-derived lncRNAs: emerging regulators of gene expression. *RNA*. 2015;5:476.
22. Bosma AA, de Haan NA, Arkesteijn GJA, Yang F, Yerle M, Zijlstra C. Comparative chromosome painting between the domestic pig (*Sus scrofa*) and two species of peccary: the collared peccary (*Tayassu tajacu*) and the white-lipped peccary (*T. pecari*): a phylogenetic perspective. *Cytogenet Genome Res*. 2004;105:115–21.
23. Adegas F, Chaves R, Guedes-Pinto H. Chromosomal evolution and phylogenetic analyses in *Tayassu pecari* and *Pecari tajacu* (Tayassuidae): tales from constitutive heterochromatin. *J Genet*. 2007;86:19–26.
24. Adegas F, Chaves R, Guedes-Pinto H. Suiformes orthologous satellite DNAs as a hallmark of *Pecari tajacu* and *Tayassu pecari* (Tayassuidae) evolutionary rearrangements. *Micron*. 2008;39:1281–7.
25. Aldal I, Andresen Ø, Egeli AK, Haugen J-E, Grørdum A, Fjetland O, et al. Levels of androstenone and skatole and the occurrence of boar taint in fat from young boars. *Livest Prod Sci*. 2005;95:121–9.
26. Bonneau M. Use of entire males for pig meat in the European Union. *Meat Sci*. 1998;49(Supplement 1):S257–72 [Meat Consumption and Culture 44th International Congress of Meat Science and Technology].
27. Cliffe KM, Day AE, Bagga M, Siggins K, Quilter CR, Lowden S, et al. Analysis of the non-recombining Y chromosome defines polymorphisms in domestic pig breeds: ancestral bases identified by comparative sequencing. *Anim Genet*. 2010;41:619–29.
28. Smit A, Hubley R, Green P. RepeatMasker Open-3.0. 1996. Source: <http://www.repeatmasker.org/faq.html#faq3>.
29. Tamura K, Stecher G, Peterson D, Filipiński A, Kumar S. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 2013;30:2725–9.
30. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*. 1992;9:678–87.
31. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3:418–26.
32. Elsik CG, Tellam RL, Worley KC. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*. 2009;324:522–8.
33. Michalet X, Ekong R, Fougereousse F, Rousseaux S, Schurra C, Hornigold N, et al. Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science*. 1997;277:1518–23.
34. Gribble SM, Wiseman FK, Clayton S, Prigmore E, Langley E, Yang F, et al. Massively Parallel Sequencing Reveals the Complex Structure of an Irradiated Human Chromosome on a Mouse Background in the Tc1 Model of Down Syndrome. *PLoS ONE*. 2013;8:e60482.

**Submit your next manuscript to BioMed Central and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

