Transcript Isoform Diversity of Y Chromosome Ampliconic Genes of Great Apes Uncovered Using Long Reads and Telomere-to-Telomere Reference Genome Assemblies

Aleksandra Greshnova[1,2], Karol Pál[1], Juan Francisco Iturralde Martinez[4, 5], Stefan Canzar[3,4]*, Kateryna D. Makova[1]*

*Co-corresponding authors

1 Department of Biology, Penn State University, University Park, PA, USA
2 Current address: Max Planck Institute for Evolutionary Biology, Plön, Germany
3 Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
4 Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, United States
5 Huck Institutes of the Life Sciences. Pennsylvania State University, University Park, PA 16802, USA

Abstract

Y chromosomes of great apes harbor **A**mpliconic **G**enes (YAGs)—multi-copy gene families (*BPY2*, *CDY*, *DAZ*, *HSFY*, *PRY*, *RBMY*, *TSPY*, *VCY*, and *XKRY*) that encode proteins important for spermatogenesis. Previous work assembled YAG transcripts based on their targeted sequencing but not using reference genome assemblies, potentially resulting in an incomplete transcript repertoire. Here we used the recently produced gapless telomere-to-telomere (T2T) Y chromosome assemblies of great ape species (bonobo, chimpanzee, human, gorilla, Bornean orangutan, and Sumatran orangutan) and analyzed RNA data from whole-testis samples for the same species. We generated hybrid transcriptome assemblies by combining targeted long reads (Pacific Biosciences), untargeted long reads (Pacific Biosciences) and untargeted short reads (Illumina)and mapping them to the T2T reference genomes. Compared to the results from the reference-free approach, average transcript length was more than two times higher, and the total number of transcripts decreased three times, improving the quality of the assembled transcriptome. The reference-based transcriptome assemblies allowed us to differentiate transcripts originating from different Y chromosome gene copies and from their non-Y chromosome homologs. We identified two sources of transcriptome diversity—alternative splicing and gene duplication with subsequent diversification of gene copies. For each gene family, we detected transcribed pseudogenes along with protein-coding gene copies. We revealed previously unannotated gene copies of YAGs as compared to currently available NCBI annotations, as well as novel isoforms for annotated gene copies. This analysis paves the way for better understanding Y chromosome gene functions, which is important given their role in spermatogenesis.

Introduction

Sex chromosomes originated from autosomes independently in different taxa during evolution (Charlesworth 1996). The mammalian proto-Y chromosome gained a male-determining gene, *SRY*. This event resulted in the suppression of recombination with the proto-X chromosome followed by the accumulation of deleterious mutations and degradation on the Y chromosome. As a result, the Y lost most of its genes (Skaletsky et al. 2003; Waters et al. 2007) whereas the X chromosome largely preserved its original gene content (Mueller et al. 2013; Miga et al. 2020).

The human Y chromosome consists of two pseudoautosomal regions (PARs), which still recombine with the PAR regions of the X chromosome, and a large male-specific region (the MSY) (Bhowmick et al. 2007). The MSY contains three categories of regions: X-transposed (acquired through a recent duplication from the X chromosome), X-degenerate (originating from proto-sex chromosomes), and ampliconic (Skaletsky et al. 2003). The Y chromosome ampliconic genes (YAGs), which are located in the ampliconic region, are expressed in the testis, encode proteins functioning in spermatogenesis, and affect male fertility (Skaletsky et al. 2003; Ross et al. 2005). YAGs form families with high sequence identity among gene copies (Bhowmick et al. 2007). Although ampliconic genes were originally characterized as genes located exclusively on the Y chromosome, later work revealed the presence of their homologs on the X chromosome and autosomes (Vallender and Lahn 2004; Bhowmick et al. 2007). Indeed, some YAGs were shown to originate as a result of duplication of autosomal copies (*CDY*, *DAZ*) or of proto-X/Y gene pairs (*HSFY*, *VCY*, *XKRY, RBMY,* and *TSPY*). *BPY2* (also known as VCY2) originated as a result of the transposition of a non-functional homolog *BEYLA* on chromosome 8 to the ampliconic region of the Y chromosome (Cao et al. 2015). The origin of *PRY* remains unclear (Bhowmick et al. 2007).

In human, YAGs include nine gene families (Table 1): *BPY2*, *CDY*, *DAZ*, *HSFY*, *PRY*, *RBMY*, *TSPY*, *VCY*, and *XKRY* (Skaletsky et al. 2003). Except for *TSPY*, whose copies are organized in tandem repeats, human YAGs are located in palindromes, facilitating gene conversion (Skaletsky et al. 2003). YAG copies positioned at symmetrical arms of the same palindrome are nearly identical (Hallast et al. 2013). However, copies located at nonsymmetrical locations within one palindrome, as well as copies located at different palindromes, vary in sequence identity (Bhowmick et al. 2007). Copy numbers of YAGs vary substantially among species and even among individuals of the same species (Vegesna et al. 2019, 2020). Some YAG families, for example, *PRY*, underwent pseudogenization (Tomaszkiewicz et al. 2023).

The telomere-to-telomere (T2T) assembly of the human Y chromosome, generated with the use of long-read sequencing technologies (Pacific Biosciences, PacBio, and Oxford Nanopore Technologies), has allowed to resolve its complete repetitive structure, including YAG family copies (Rhie et al. 2023). The T2T Y and X chromosome assemblies of five non-human great ape species (bonobo, chimpanzee, gorilla, Sumatran orangutan, and Bornean orangutan) have just been generated (Makova et al. 2024) with the same approaches and their YAG families have been completely deciphered. There is currently an urgent need to characterize the function of genes on the Y chromosome of great apes, as many of them are important for reproduction. Studying the transcriptome of YAGs is an important step in this direction.

Technological advances in sequencing technologies, the invention of long-read sequencing in particular, revealed the tremendous diversity of the transcriptome (Ren et al. 2023), which translates into proteome diversity (Park et al. 2018; Soto et al. 2019). Transcriptome diversity

originates from alternative splicing and gene duplication with subsequent accumulation of mutations in duplicate copies (Goldtzvik et al. 2023). In the human genome, >90% of genes undergo alternative splicing (Wang et al. 2008). The contribution of gene duplication to isoform diversification is particularly important to decipher for YAGs due to their high copy number. Long sequencing reads, frequently spanning full-length transcripts, are essential to capture the differences between transcripts originating from highly similar copies within each YAG family.

Below we utilized the recently generated T2T assemblies of the Y chromosomes of great ape species (Rhie et al. 2023; Makova et al. 2024) to study transcriptome diversity of their YAGs. We used testis samples of six great ape species (human, bonobo, chimpanzee, gorilla, Bornean orangutan, and Sumatran orangutan) and assembled their transcriptomes using data generated with three different approaches: (1) targeted (YAG-specific) PacBio IsoSeq, in which individual YAG families were captured with sequence-specific probes, (2) untargeted (i.e. genome-wide) PacBio IsoSeq, and (3) untargeted Illumina short-read sequencing. The combination of long reads with high-quality short reads and T2T genome assemblies has allowed us to characterize transcript YAG isoform diversity in great apes in detail. In particular, we determined whether the observed diversity of individual transcripts is explained by alternative splicing or by differences between gene copies. The results of this study provide important information about the evolution of MSY of great apes, all of which are endangered species (except for human). They also represent a useful resource for functional studies aiming to characterize cellular functions of YAGs at the isoform level, which remain unknown to date.

Results

Combination of sequencing approaches yields a comprehensive list of YAG transcripts

To characterize transcript isoform diversity of YAGs in great apes, we combined targeted and untargeted long-read PacBio IsoSeq sequencing approaches with the untargeted short-read Illumina sequencing of RNA isolated from testis samples of six great ape species—bonobo, chimpanzee, gorilla, Bornean orangutan, Sumatran orangutan, and human (for human only targeted IsoSeq data were analyzed). Targeted PacBio IsoSeq data were available for two technical replicates (Tomaszkiewicz et al. 2023). For our analysis, we took advantage of the T2T assemblies available for great ape Y chromosomes (Rhie et al. 2023; Makova et al. 2024). Briefly, we performed reference-based assemblies of transcripts using StringTie (Shumate et al. 2022) and used short-read sequencing data to perform hybrid transcriptome assemblies (StringTie mix mode) to improve the quality of assembled transcripts (Shumate et al. 2022). For each non-human great ape species, we obtained three sets of YAG transcripts: those resulting from technical replicate 1 for targeted data, from technical replicate 2 for targeted data, and from untargeted data. For human, we obtained two sets of YAG transcripts—using technical replicates 1 and 2 of targeted data.

The transcripts resulting from two technical replicates of targeted data were in good agreement with each other (Fig. S1). The number of identical transcripts between replicates (across all gene families) varied from 59 for gorilla to 104 for bonobo. Compared to the targeted approach, the untargeted approach captured fewer YAG transcripts, likely due to the low level of expression of YAGs (Vegesna et al. 2019). For example, for bonobo, we assembled 140 and 138 YAG transcripts using the targeted approach (for technical replicate 1 and technical replicate 2, respectively) and only 31 YAG transcripts using the untargeted approach (Fig. S1). Only a small fraction of transcripts assembled using the untargeted approach were absent from

those assembled using the untargeted approach. For example, in bonobo, among 31 transcripts assembled from untargeted IsoSeq reads, 30 were also found in transcripts assembled from targeted IsoSeq reads. However, for other species, untargeted IsoSeq data still captured some YAG transcripts that were not present in the targeted IsoSeq YAG transcript set. For downstream analysis, we merged the three transcript sets (targeted technical replicate 1, targeted technical replicate 2, and untargeted) to create a non-redundant comprehensive set of YAG transcripts for each species. YAGs transcripts assemblies of great ape species are presented as additional files 1—6. The transcript was called to be replicate-supported if it was present in at least two of the three transcript sets.

Identification of novel YAG copies and isoforms

We performed YAG transcript structure classification against the available NCBI reference gene annotations with SQANTI3 (Pardo-Palacios et al. 2023). Among the assembled YAG transcripts, the majority originated from annotated gene copies represented by “full splice match” (FSM; similar to a transcript in the reference annotation) and “incomplete splice match” (ISM; similar to the reference, but containing fewer 5’ exons) categories (Fig. S2). For all species, the largest proportion of transcripts was classified as FSM, suggesting a good agreement of assembled transcripts with reference annotations. However, we identified novel isoforms of annotated genes for all species. New isoforms were classified into two categories: “novel in catalog” (NIC; transcripts with splice junctions between annotated donor/acceptor sites) and “novel not in catalog” (NNIC; transcripts with at least one novel donor or acceptor site). For all species, there were more NNIC than NIC transcripts.

Besides new transcripts, SQANTI3 identified potentially new gene copies for all species (Fig. S2). These included both coding and non-coding copies (transcribed pseudogenes) and suggested under-annotation of the reference assemblies. We refined the discovery of novel genes by SQUANTI3 by strictly excluding any overlap with annotated genes (see “Annotation of transcripts” in the Methods) and confirmed the presence of novel gene copies in the reference assemblies. Our estimates of novel gene copy numbers (Fig. 1A) were lower than those identified by SQANTI3 (category “Intergenic” in Fig. S2), suggesting that SQANTI3 overestimates the number of new copies. We present numbers for both replicate-supported and not replicate-supported datasets (Fig. 1A). Although replicate-supported transcript assemblies serve as more reliable evidence of novel gene copies, we found several cases when transcripts supported by a single technical replicate indicated the presence of additional gene copies (Fig. 1B). Thus, data not supported by replicates could also be useful to improve existing annotations and identify novel gene copies.

We next examined whether distinct isoforms originate from different gene copies or are produced as a result of alternative splicing from one gene copy (Fig. S3). We limited our analysis to (potential) protein-coding copies. Our results suggest that the majority of such copies encode multiple protein isoforms (237 and 143 gene copies encoding multiple and single isoforms across all species, respectively; Fig. S3). Many gene copies produced only two or three isoforms. However, for some gene families, e.g., *DAZ* and *BPY*2, we identified gene copies producing more than ten isoforms (Fig. S3). For example, a single *DAZ* gene copy in bonobo produced 11 different isoforms (Fig. S4). Gene families with a low abundance of protein-coding transcripts (*VCY* and *XKRY*) had gene copies each producing only one isoform.

X chromosome and autosomal homologs of YAGs in great apes

We identified some transcripts originating from homologs of YAGs on the X chromosome and/or on the autosomes (Fig. 2B). *VCY* had transcripts originating from X chromosomal homologs in all species and additionally from autosomal homologs in chimpanzee and gorilla. *HSFY* had transcripts originating from autosomal homologs in all species and X chromosomal homologs in gorilla and bonobo. We identified transcripts originating from autosomal homologs of *BPY2* in bonobo and orangutans; *XKRY* from bonobo and chimpanzee; *PRY* from orangutans. We detected *RBMY* transcripts originating from autosomal copies in gorilla and chimpanzee and X chromosomal homologs in bonobo. *PRY* had transcripts originating from an autosomal homolog in both species of orangutan and in gorilla. We did not detect any *CDY*, *DAZ*, and *TSPY* transcripts that originated from non-Y chromosome copies.

Transcripts vary in numbers and length across YAG families and species

In total, we identified 1,266 unique transcripts mapping to the Y chromosome across all species and YAG families (Table S2), among which 445 were replicate-supported (Table S3). On average, 211 (74 replicate-supported) transcripts were assembled per species with a minimum of 132 for chimpanzee (61 replicate-supported) and a maximum of 307 for human (93 replicate-supported). The numbers and the lengths of transcripts differed substantially among species and YAG families (Table S2 and Fig. S5). Across the data set, we found an abundance of *RBMY* (374 transcripts in total, 159 replicate-supported) and *TSPY* (230 transcripts in total, 77 replicate-supported) transcripts. In contrast, *VCY* had only 9 transcripts (all replicate-supported) across all species. The mean length of the assembled transcripts across all gene families and species was 1,937 bp (standard error, SE = 41). The longest assembled transcript was 9,838 bp and belonged to the *HSFY* gene family from Bornean orangutan. The shortest transcript was 202 bp and originated from a copy of the *XKRY* gene from Bornean orangutan. When we averaged transcript lengths per family (and considered all species together), *PRY* had the longest transcripts (mean length = 3,416 bp, SE = 268, range = 633–8,894) and *XKRY* the shortest (mean length = 512 bp, SE = 34, range = 202–896).

Coding potential of assembled YAG transcripts

To characterize the coding potential of assembled transcripts, we first used the threshold of a minimum of 50 amino acids (aa) to consider an ORF a complete ORF (cORF). Following Tomaszkiewicz et al. (2023), we next identified high-confidence cORFs as any cORFs that had significant homology to the known sequences of great ape YAG proteins (see “Identification of high-confidence complete open reading frames” in Methods). We used BLASTP to align all cORFs found in a transcript to the known protein sequences of great ape YAG proteins from the NCBI protein database (Fig. S6). For the majority of alignments, coverage was above 80% and sequence identity was above 70%. The alignments of *DAZ*, *RBMY*, *TSPY*, and *CDY* transcripts demonstrated a large variation in sequence identity (60-100%; Fig. S6A). Across all species and gene families, we identified 1063 transcripts with high-confidence cORFs (mean length = 267 aa, SE = 6, range = 50–951; Table 2, Fig. 3A) among which 445 (42%) were replicate-supported (mean length = 315 aa, SE = 9, range = 50–833; Table S4). On average We identified an average of 177 transcripts (74 replicate-supported) with high-confidence cORFs per species. The majority of high-confidence cORFs originated from *RBMY*, *DAZ*, *TSPY*, and *CDY* gene families. *VCY* and *XKRY* had the smallest number of protein-coding transcripts.

We compared the lengths of high-confidence cORFs with the lengths of previously published protein sequences when such sequences were available in the NCBI protein database (Fig. 3B

and Table 1). Overall, we found the lengths of the predicted and previously reported proteins to be in good agreement. For example, a significant proportion of the identified *BPY2* protein-coding transcripts encode 106 aa long proteins matching the length of the previously reported *BPY2* protein isoform. In agreement with previously published data (Table 1), in many cases we predicted protein isoforms of different lengths for the same gene family (Fig. 3B). For example, the lengths of *CDY* proteins have two distinct groups roughly matching the lengths of previously reported protein isoforms (470 and 540 aa). Similarly, the *HSFY* gene family with lengths of predicted isoforms group around previously reported lengths of *HSFY* proteins (323 and 401 aa). The remaining predicted *HSFY* proteins had lengths between 50 and ~250 aa. Interestingly, the two longer (323 and 401 aa) *HSFY* isoforms were identified in human and orangutans, whereas the shorter (50-250 aa) isoforms were found in gorilla, bonobo, chimpanzee, and human, suggesting the presence of species-specific isoforms. The *RBMY*, *CDY*, *TSPY*, and *DAZ* gene families, known to have numerous protein-coding copies across great ape species, demonstrated a large variation in the length of predicted proteins.

The characterization of the coding potential of YAG transcripts allowed us to differentiate between genes and transcribed pseudogenes. We identified that *CDY*, *DAZ*, *RBMY*, *TSPY*, and *HSFY* had multiple gene copies in all species. Except for *CDY*, we also identified numerous transcribed pseudogenes in these gene families (Fig. 4). *BPY2*, *XKRY*, *VCY*, and *PRY* were represented by a smaller number of gene copies, with protein-coding copies missing in some species (Fig. 4).

Diversity of structural and sequence isoforms of YAGs across species

To characterize the isoform diversity of YAG transcripts across great ape species, for each gene family, we mapped transcripts from all species to a single human protein-coding gene copy (Tomaszkiewicz et al., 2023). *Structural isoforms* were identified by comparing the exon-intron structure of transcripts within and across species. We limited our analysis to transcripts with high-confidence cORFs. *Sequence isoforms* were identified by comparing transcript protein sequences within and across species. This part of the analysis was limited to replicate-supported transcripts with high-confidence cORFs.

For every gene family, we observed a high diversity of structural and sequence isoforms (Fig. 5; the exon-intron structure of each transcript is presented in Figs. S12–S20). *RBMY* had the highest number of unique sequence isoforms (n = 87) and, together with *DAZ*, had the highest number of structural isoforms (n = 40; Fig. 5A). The lowest number of sequence isoforms was identified for *BPY2* (n = 4) and the lowest number of structural isoforms was identified for *VCY* (n = 2). For seven out of nine gene families (*CDY*, *HSFY*, *PRY*, *RBMY*, *TSPY*, *VCY, DAZ*), the diversity of sequence isoforms was higher than the diversity of structural isoforms (Fig. 5B). For *XKRY*, the number of unique structural and sequence isoforms was equal. For *BPY2*, the diversity of structural isoforms was higher than the diversity of sequence isoforms.

We identified isoforms shared between species and isoforms unique to each species (Figs. 20–22). *HSFY* and *TSPY* had structural isoforms shared across all species that have protein-coding copies of these genes, and *CDY* had a structural isoform that was shared by five out of six species (Fig. 6). Bornean and Sumatran orangutans had shared isoforms at both structural and sequence levels for each gene family (except for *DAZ* that was not observed in Sumatran orangutan; Fig. 6). Chimpanzee and bonobo shared common structural, but not sequence, isoforms for *CDY*, *PRY*, *VCY*, and *RBMY,* and common sequence and structural isoforms for *TSPY* and *DAZ* (Fig. 6). Interestingly, *TSPY* demonstrated remarkable conservation

at the level of transcript structure with the majority of transcripts presented by a single isoform shared across all species (Fig. 6).

Finally, we investigated whether transcript isoform diversity, as measured with the Shannon diversity index, is correlated with gene expression levels or gene copy number. We limited our analysis to *CDY*, *DAZ*, *HSFY*, *TSPY*, and *RBMY*—gene families that have a large number of protein-coding transcripts and are present in more than three species. Neither structural nor sequence diversity demonstrated a relationship with gene expression that was consistent across gene families (Fig. S24); correlation coefficients were not statistically significant. Gene copy number was mostly positively correlated with sequence isoform diversity in most cases, but only the correlation coefficient for *CDY* was statistically significant ($p = 0.018$; Fig. 7). There was no consistent relationship between structural isoform diversity and gene copy number (Fig. S25).

Discussion

To produce a comprehensive catalog of YAG transcripts in great apes, we used a combination of long- and short-read sequencing and took advantage of the recently released T2T assemblies of their sex chromosomes. Long-read sequencing was crucial to differentiate between transcripts originating from highly similar copies present within each YAG family. Additionally, we were able to identify both new gene copies and new isoforms of annotated copies of YAGs. The number of transcripts assembled, as well as their coding potential, varied substantially between species and gene families. Some gene families underwent pseudogenization; in each gene family, we identified at least one transcribed pseudogene. The observed transcript isoform diversity originated from alternative splicing, leading to structural diversity, and gene duplication and divergence, leading to sequence diversity. In general, sequence isoform diversity was higher than structural isoform diversity.

A refined catalog of YAGs in great apes

We present our YAGs transcript assemblies in comparison with previously reported findings. Consistent with Tomaszkiewicz et al. (2023), we report the presence of transcripts from all nine YAG families in all great ape species, except *BPY2* in bonobo (Fig. 2A). Vegesna et al. (2020) reported that five gene families (*BPY2*, *CDY*, *DAZ*, *RBMY*, *TSPY*) are shared across all species, three (*HSFY*, *PRY*, and *XKRY*) are shared between gorilla, both species of orangutans and human, and *VCY* is specific to chimpanzee and human. In contrast to Vegensa et al. (2020), we did not detect *BPY2* transcripts originating from the Y chromosome of bonobo. However, we identified transcripts from the autosomal homolog(s) of *BPY2* (Fig. 2B), which might have been misinterpreted as YAG transcripts by Vegesna et al. (2020) due to the lack of a high-quality reference genome at the time.

Consistent with other studies (Feng et al. 2019; Zhang et al. 2019; Leung et al. 2021), our results suggest that long-read sequencing and reference-based assembly using the T2T assemblies can improve assemblies of transcriptomes. The previous work that characterized transcript isoform diversity of YAGs in great apes used the *de novo* assembly approach as high-quality references were not available then (Tomaszkiewicz et al., 2023). In contrast, we utilized the recently produced T2T assemblies of Y chromosomes of great apes (Makova et al. 2024). Our results are largely consistent with those of Tomaszkiewicz et al. (2023), however, there are several differences. Most importantly, compared with Tomaszkiewicz et al. (2023), we identified fewer transcripts, and on average they were longer. Indeed, we identified only 445

replicate-supported transcripts, whereas Tomaszkiewicz et al. (2023) identified 1,510. On average, transcripts identified by the *de novo* assembly approach were more than two times shorter (mean length of 820 bp, Tomaszkiewicz et al., 2023), as compared to those in our reference-based approach (mean length = 1,893,). One possible explanation for this difference is that the *de novo* approach failed to assemble full transcripts due to the lack of the reference and thus some partial transcripts were considered as different (shorter) transcripts. In the case of *PRY*, for example, Tomaszkiewicz et al. (2023) assembled 444 replicate-supported transcripts across all species with an average length of <500 bp (Tomaszkiewicz et al., 2023). In contrast, we assembled only 12 replicate-supported transcripts with an average transcript length of 3,865 bp (Table S3).

Another major difference between the *de novo* and reference-based transcriptome assembly methods is that the *de novo* method does not differentiate YAG transcripts and transcripts originating from non-Y chromosome homologs of YAGs. This could have also contributed to the overestimation of YAG transcripts. Many YAGs have homologs on the X chromosome or autosomes (Vallender and Lahn 2004; Bhowmick et al. 2007) informing us about their origin (movement from autosomes or originating from the proto-X/Y pair). Bhowmick et al. (2007) reported homologs of *HSFY*, *CDY*, *XKRY*, *RBMY*, and *TSPY* on both X chromosome and autosomes, *VCY* homologs on the X, and *DAZ* homologs on an autosome. Our results in general are in agreement with these observations, although we could not detect any *DAZ, TSPY,* and *CDY* transcripts originating from non-Y chromosomes. This might be explained by the fact that we only used targeted data for this part of the analysis, therefore our hybridization probe did not capture transcripts originating from non-Y chromosome homologs (for example, *DAZ* homolog on chromosome 3; Saxena et al. 1996).

Isoform diversification of YAGs is explained by both gene duplication and alternative splicing

Our results suggest that both gene duplication and alternative splicing contributed to the observed diversification of isoforms. The potential to produce alternative isoforms from one gene copy, i.e. structural isoforms, varies substantially among YAG families but does not vary substantially among species. In particular, we found that many gene copies of *RBMY*, *DAZ*, and *HSFY* produce several isoforms. The functional effects of the identified protein isoforms of YAGs in great apes remain to be explored, especially in cases when they are species- or genus-specific. In general, alternative splicing has a functional effect on protein isoforms and could affect the localization of proteins, the binding affinity of transcriptional regulators, enzymatic properties, and protein-protein interactions (Kelemen et al. 2013).

Gene duplication and subsequent accumulation of mutations among duplicate copies, together with sequence diversification due to speciation, represent additional mechanisms contributing to generating protein isoforms (Goldtzvik et al. 2023). Duplicate copies mask the potential negative effect of new mutations and thus facilitate the accumulation of genetic diversity playing an important role in evolution. The potential suppression of phenotypic expression of deleterious effects of mutations is particularly important in YAGs, because YAGs encode proteins that have critically important functions in reproduction (Table 1). YAGs undergo species-specific turnover and their gene copy numbers differ substantially between species and even between individuals (Ye et al. 2018; Vegesna et al. 2020). Previous work shed light on the biological meaning of copy number variation of YAGs. For instance, a positive correlation between expression level and gene copy number suggests that copy number variation is among the mechanisms of regulation of expression level on the level of gene family (Vegesna et al. 2020). Our findings

complement the previously reported importance of copy number variation of YAG copies and suggest that the expansion of gene families also plays a role in the diversification of protein isoforms.

PRY undergoes pseudogenization

Consistent with previous studies (Tomaszkiewicz et al. 2023), our results suggest the pseudogenization of *PRY*. For *PRY*, the majority of the identified gene copies did not encode proteins and were marked as transcribed pseudogenes. Although the average length of transcripts was comparable with the length of transcripts of gene families with protein-coding sequences, we identified only a few high-confidence cORFs for *PRY*; in orangutans, we did not find any protein-coding copies. We identified several protein-coding *PRY* copies in human, bonobo, gorilla, and chimpanzee. However, from a comparison of the length of predicted proteins, we found that only a subset of predicted isoforms from human match the length of the previously reported *PRY* protein (147 aa) whereas the rest of the predicted proteins are more than two times shorter (60-70 aa). These short protein isoforms might represent non-functional truncated proteins and thus protein-coding potential of *PRY* in non-human apes remains unresolved. Additionally, *PRY* has the lowest expression across all YAGs in all great ape species (Vegesna et al. 2020), also suggesting that this gene family is on its way to becoming a pseudogene.

Interestingly, in orangutans we did not detect any protein-coding *PRY* transcripts, however, we found *PRY* transcripts originating from autosomal *PRY* homologs in these species. Protein-coding potential of the identified homologs needs further investigation, however, the presence of non-Y homologs might suggest the non-essentiality of Y chromosome *PRY* copies. Finally, the identified homologs might indicate the origin of PRY since little is currently known about it (Bhowmick et al. 2007).

BPY2 in bonobo and orangutans

Consistent with Tomaszkiewicz et al. (2023), we identified protein-coding *BPY2* copies in human, gorilla, and chimpanzee. In human and gorilla, the identified copies encoded multiple structural protein-coding isoforms In chimpanzee we identified one copy producing one isoform and one copy producing two isoforms. Additionally to protein-coding copies, we identified numerous non-coding *BPY2* copies in all species, except bonobo. We could not identify high-confidence cORFs in bonobo and the two orangutan species. Possible explanations include (1) pseudogenization of *BPY2* on the Y chromosome in these species, (2) low sequencing depth that prevented us from capturing the full repertoire of protein-coding transcripts, (3) independent evolutionary origin of *BPY2* in these species (hypothesized by Tomaszkiewicz et al., 2023), which would make the detection of statistically significant homologs unlikely due to high sequence divergence. Low expression levels of *BPY2* and the presence of transcribed *BPY2* pseudogenes on the Y chromosome in all species favor the hypothesis of *BPY2* pseudogenization.

In bonobo and orangutans we identified transcripts originating from chromosome 8 with high sequence similarity to *BPY2* transcripts originating from Y chromosome. Our results are in agreement with previous findings (Stuppia et al. 2005; Cao et al. 2015) reporting the presence of a non-functional autosomal BPY2 homolog *BEYLA*. *BPY2* originated as a result of the transposition of a large fragment of *BEYLA* pseudogene to the amplicon of the Y chromosome. The transposition was followed by an acquisition of protein-coding ability. The absence of

functional *BPY2* copies on the Y chromosome in bonobo and orangutans could suggest the non-essentiality of *BPY2* in these species.

Limitations of the present study and future directions

Our results should be interpreted considering several limitations. The isoforms identified here have reasonable support with some of them found by three independent experiments. However, the fact that in certain species we could not confirm the presence of some gene families or identify protein-coding genes does not necessarily mean that these genes/transcripts are absent. First, ampliconic genes are expressed at low levels (Fagerberg et al. 2014; Vegesna et al. 2019, 2020) and the sequencing depth might have been insufficient to capture the full diversity of YAG transcripts. To overcome this obstacle, targeted IsoSeq data was generated by Tomaszkiewicz et al. (2023) and used in our study. However, the hybridization probes for the targeted approach IsoSeq data were generated prior to the availability of high-quality references. Thus, some transcripts that did not anneal to the designed probes might have not been captured by the targeted IsoSeq approach. We attempted to overcome this potential limitation by complementing the transcript set obtained by targeted IsoSeq with untargeted IsoSeq. Finally, YAGs are expressed in testis at different stages of spermatogenesis and it is possible that in some samples we did not capture the correct stage to obtain a full repertoire of transcripts. The sequencing data used here originate from whole-testis samples without differentiation of different cell types. Thus, we could not investigate the relationship between isoforms and cell types present in the testis. Several studies (Hermann et al. 2018, Grive et al. 2019, Guo et al. 2021, Nie et al. 2022, Zhou et al. 2023) characterized the variability in expression levels at single-cell level in human testis. However, none of them targeted YAGs specifically. Future single-cell studies will provide a more detailed characterization of differential isoform usage across cell types in the testis in human and non-human great apes.

Gene copies within each YAG family are highly similar. We observed that a small proportion of reads mapped non-uniquely, i.e. to several copies of one YAG family. In such cases, it was not possible to unambiguously assign reads to one gene copy. One possible reason for this is that these reads covered transcripts only partially and thus could not capture the small difference between gene copies. This limitation could be potentially overcome with higher sequencing depth. Another potential pitfall lies in the fact that the *DAZ* gene includes almost identical repeats that could cause alignment issues, leading to artificial transcripts.

We predicted the protein-coding potential of the assembled YAG transcripts based on their homology to known proteins. However, this approach is biased toward protein sequences present in the databases; not all proteins were available for all species studied (Fig. S26). Thus, we could have missed some of the protein-coding transcripts. Alternative, alignment-free coding-potential prediction methods) could verify and potentially expand our findings. Another limitation associated with the NCBI protein database relates to the presence of entries of predicted proteins derived from transcriptomic analyses, alongside authentic protein entries originating from proteomics analysis. The results from transcriptomic studies are less reliable, however, due to the limited availability of the data on YAGs proteins of great apes we still used such entries. We predicted the presence of multiple protein-coding transcript isoforms for YAG families in testis. In general, transcriptomic studies report a high diversity of isoforms in contrast to proteomics studies, which typically report a much lower diversity of isoforms (Goldtzvik et al. 2023). Experimental validation of the predicted protein isoforms, for example using mass spectrometry proteomics (Ferrández-Peral et al. 2022), could verify our results.

Functions of proteins encoded by YAGs remain poorly characterized. Several studies analyzed the effect of deletions of YAG copies on spermatogenesis (Saut et al. 2000; Navarro-Costa et al. 2007) and could only make conclusions about the essentiality of these genes for male fertility. Our dataset with the characterization of isoform diversity provides a valuable resource for subsequent functional characterization of proteins encoded by YAGs. Rastegar et al. (2015) characterized differences in gene expression levels between testis biopsies from men with functional vs. impaired spermatogenesis and identified isoform-level signatures that could serve as markers for diagnosis of azoospermia. Only *XKRY*, *HSFY*, *PRY*, *BPY2*, and *DAZ* families in humans were analyzed though. Similar studies with additional YAG families and including non-human great apes could deepen our understanding of their role in ape spermatogenesis. Finally, little is known about several novel YAG families that were recently discovered by Zhou and colleagues (Zhou et al. 2023) and by Makova and colleagues (Makova et al. 2024). Targeted sequencing is required to characterize these gene families in future studies.

Methods

Sequencing data

We analyzed sequencing data previously generated in our laboratory as well as publicly available RNA-seq datasets (summarized in Table S1 and Fig. S7). Sequencing data was generated from testis samples of six great apes species, human, gorilla, chimpanzee, bonobo, Bornean orangutan, and Sumatran orangutan, using three different approaches: untargeted IsoSeq, the standard PacBio protocol that enables to capture full-length RNA isoforms; targeted IsoSeq, a modified IsoSeq protocol that was designed to enrich for transcripts of interest (YAGs, in our case), and short-read Illumina RNA sequencing data. For human, only targeted IsoSeq data was analyzed. For each species only one individual (biological replicate) was analyzed.

For targeted IsoSeq data, the full description of samples and experimental procedures is provided in Tomaszkiewicz et al. (2023). All six testis samples were available in two technical replicates. Hybridization to biotinylated oligo probes, designed specifically to capture YAGs cDNA, was performed to enrich YAG cDNA. Raw PacBio reads from both targeted and untargeted samples were preprocessed using the Iso-Seq3 pipeline (https://github.com/PacificBiosciences/IsoSeq3) to produce full-length non-chimeric (FLNC) reads.

Short-read RNA-seq datasets were obtained from paired-end stranded libraries (except for bonobo with single-end, unstrained library) and sequenced with Illumina MiSeq or HiSeq 2000 (Table S1).

Alignments

For IsoSeq data, FLNC reads were first converted from bam to fasta format using samtools (Danecek et al. 2021). Reads were then aligned to the corresponding full genome (accession numbers are provided in Table S1) using minimap2 (Li 2018) in splice-aware mode (`minimap2 -ax splice -uf -C5 genomic.fna input.bam > alignment.sam`). The alignments were converted from sam to bam format, sorted, and indexed using samtools. For Illumina data, adapters were removed using Trim Galore (https://github.com/FelixKrueger/TrimGalore). Reads were aligned using HISAT2 (Kim et al. 2015; scripts are available at https://github.com/agreshno/YAGs_great_apes/blob/main/short_read_alignment/trim_galore_his

at2.txt). We then filtered alignments to retain only reads aligned to the Y chromosome (our primary focus).

Transcriptome assembly

We used Illumina data to aid the transcriptome assembly from long PacBio reads. We separately assembled three transcriptomes: untargeted IsoSeq with Illumina short reads and each of two replicates of targeted IsoSeq with Illumina short reads. Reference-based transcriptome assembly was performed using StringTie (Shumate et al. 2022) in mix mode (`stringtie -G genomic.gff --mix short_reads.aligned.bam long_reads.aligned.bam -o output.gff`).

Comparisons of assemblies

We compared three assembled sets of transcripts using gffcompare (Pertea and Pertea 2020; `gffcompare -r reference.gff query.gff`). Gffcompare classifies transcripts present in a query gff file into several categories: (1) exact match, (2) unknown (absent from the reference), (3) transcript is not completely matched to the transcript in the reference (further split into 15 different categories that clarify how two transcripts are different, for example retained intron etc.). From these comparisons, we called transcripts *replicate-supported* if they were present in at least two sets of transcripts (for example, in both technical replicates of targeted or in one technical replicate of targeted IsoSeq and in untargeted IsoSeq) and non replicate-supported if they were present only in one set of transcripts.

For downstream analysis we merged three sets (`gffcompare untargeted.gff targeted_1.gff targeted_2.gff -o combined`) to compile a non-redundant set of transcripts for each species. As low expression of YAGs (Vegesna et al. 2020) complicates the comprehensive capture of isoform diversity, we included both replicate-supported transcripts and non-replicate-supported transcripts into our analysis, unless specified otherwise. Replicate-supported transcripts are expected to mitigate sequencing errors and potential artifacts introduced during computational inference of transcripts by StringTie. However, this approach inevitably reduces the diversity of identified isoforms.

Quality control with SQANTI3

We used SQANTI3 (Pardo-Palacios et al. 2023) to perform quality control of the assembled set of transcripts and filter out low-quality transcripts (`python sqanti3_qc.py combined.gtf reference.gtf reference.fna`).

Annotation of transcripts

By default, StringTie annotates transcripts by finding the overlapping genes in the reference annotation for each assembled transcript and then assigning this gene name to the transcript. However, we found that YAGs in the NCBI annotations are under-annotated and the default StringTie method did not work well in our case. Thus, to annotate the transcripts assembled by StringTie we used custom scripts. First, we searched for overlaps between ampliconic genes from the NCBI annotations and the transcripts assembled by StringTie (script annotation_manual_1.R). Second, we used an annotation-free approach and assigned gene families to assemble transcripts based on sequence identity to known transcripts. For this, we used lastz (Harris, 2007) to search for similarities between sequences of assembled transcripts

(query sequences) and all known sequences of YAGs transcripts of great apes (reference sequences). To create query sequences, we extracted transcript sequences assembled with StringTie using gffread (Pertea and Pertea 2020). To create reference sequences, we extracted the sequences of individual YAG transcripts based on NCBI annotations for great ape T2T assemblies and united these sequences into a single file (https://github.com/agreshno/YAGs_great_apes/blob/main/Snakemake-Targeted-IsoSeq/all_species_YAGs.fasta). Using lastz (Harris, 2007), we aligned the query sequences against the reference sequences (`lastz query.fasta[multiple] reference.fasta --format=general:name1,zstart1,end1,name2,strand2,zstart2+,end2+,id%,blastid% > output.lastz`). The gene family of the best matching query sequence from the lastz output was assigned as a reference gene family for transcripts produced with StringTie (script annotation_manual_2.R). To select the best matching sequence from the reference sequences for each query sequence from the lastz output, we filtered the output of the previous step to retain sequences with minimum percent identity of 80% and an alignment length of at least 100 bp and selected the reference sequence that has the highest percent identity and the longest alignment.

Identification of potential complete open reading frames

Following Tomaszkiewicz et al. (2023) to identify potential open reading frames (ORF), we used the getorf package from EMBOS suite (Rice et al. 2000). We used a minimum threshold of 50 amino acids to consider an ORF complete(cORF) (`getorf -sequence input.fasta -minsize 150 -find 1 -outseq output.orf`). We added this information to the gff files using the script annotation_final.R.

Identification of high-confidence complete open reading frames

Following Tomaszkiewicz et al. (2023) to identify high-confidence cORFs, we searched for homology between products of predicted open reading frames and coding sequences of YAGs of great apes. We created a database of primate YAGs by searching for protein sequences in the NCBI protein database of the gene families of YAGs and by specifying the species of interest (create_protein_db.txt). We used BLASTP (Camacho et al. 2009) to search for homology between predicted complete open reading frames and the database of great ape YAG proteins. We used thresholds for E-values of $<10^{-6}$ and for bit scores >50 (`blastp -query input.fasta -db protein_db/human_apes_YAGs_full -out homologs.txt -evalue 1e-30 -outfmt "6 qseqid sseqid qcovs length pident evalue bitscore mismatch gaps qstart qend sstart send qseq sseq qlen"`). The retrieved homologs are statistically significant according to Pearson (2013).

Comparison with reference annotation and isoform classification

We performed isoform classification using SQUANTI3 (Pardo-Palacios et al. 2023) against reference NCBI annotations (Table S1; `python sqanti3_qc.py combined.gtf reference.gtf reference.fna`).

Alignment to one human YAG copy

In order to compare alternative splicing patterns within one gene family across different species, following Tomaszkiewicz et al. (2023), we aligned all replicate-supported high-confidence complete open reading frames to one of the copies of a given gene family in the human genome

using uLTRA (version v0.0.4) (Sahlin and Mäkinen 2021). For this, we (1) filtered the gff file to include only replicate-supported transcripts with high-confidence cORF with the script subset_gtf_hc_cORF.R, (2) extracted fasta sequences of the transcripts with gffread, (3) concatenated fasta files for each gene family across all species, (4) used uLTRA to map transcripts to one copy in the human genome (`uLTRA pipeline human_YAGs.fa Human.vcf gene_family.fasta gene_family_ultra/ --t 40`), (5) converted sam files to bam format, sorted and indexed them using samtools, (6) transformed sorted bam files to bed files using bedtools, and finally (7) used the script plots_transcripts_all.R to visualize transcripts across different species for each gene family.

Identification of unique structural and sequence isoforms

Structural isoforms. First, we transformed a bed alignment from uLTRA to gff. From the obtained gff file we extracted information about each transcript into a separate gff file (both of these steps were done with script subset_gff_transcript.R) and compared against the full gff file using gffcompare (`gffcompare transcript_1.gff -r gene_family_transcripts.gff -o isoform_1`).

Sequence isoforms. Using script extract_orf_sequences.R we extracted protein sequences encoded by high-confidence cORFs into a separate file for each gene family. For each gene family we built a multiple sequence alignment of protein sequences using MEGA (version 11.0.13; Tamura et al. 2021).

Identification of unique isoforms. For each gene family we counted the number of unique structural isoforms and unique sequence isoforms and assigned the corresponding isoform id to each transcript (separate scripts for each gene family, for example BPY2_isoform_count.R). To identify unique structural isoforms we used tmap output files produced for each transcript on the step “Structural isoforms”; we assigned the same isoform id to all transcripts that were marked as reference match by gffcompare. To identify unique sequence isoforms we performed all possible pairwise comparisons of the sequences in the multiple sequence alignment file. For each pair of sequences we compared each position in the alignment file (including positions marked as gaps in sequences) and quantified the proportion of identical positions. We assigned sequences that were identical at every position to the same sequence isoform.

Quantification of diversity of isoforms. To quantify diversity of transcripts produced (script summary.R) we used two measures: (1) isoform count, (2) Sympson inverse diversity index

($1/D = \sum_i p_i^2$, where p_i is the proportional abundance of isoform i). The diversity index takes into account the number of identified isoforms and their relative abundances; the inverse diversity index is less biased by the sample size (https://github.com/vegandevs/vegan). All scripts are available in the GitHub folder “isoform_identification”.

Correlation of YAG isoform diversity with expression levels and gene copy number

We investigated the relationship between the measure of diversity of structural and sequence isoforms and gene expression levels (Vegesna et al. 2020) and gene copy number (only protein coding copies were counted) by calculating the pairwise Spearman correlation coefficient for each YAG family (script summary.R).

Full-genome analysis of targeted IsoSeq

We performed additional analysis of targeted IsoSeq data (for all species) by modifying the pipeline in, Fig. S28 (available on GitHub). In this analysis, we assembled transcripts from full-genome alignments (by omitting the filtering for Y chromosome specific alignments). This step was required to detect non-Y chromosome homologs of YAGs and their locations (X chromosome/autosomes). We restricted our analysis to targeted IsoSeq because we expected that the enrichment step will capture not only transcripts that originate from YAGs, but also transcripts from their non-Y chromosome homologs, due to high sequence identity between them (Ahmadi Rastegar et al. 2015).

Analysis of human sequencing data

For human, we analyzed only two technical replicates of targeted IsoSeq. The pipeline was similar to the non-human pipeline, with the exception of the StringTie assembly step (`stringtie -G genomic.gff -L long_reads.aligned.bam -o output.gff`).

Data Availability

All datasets used in this study, including untargeted IsoSeq, targeted IsoSeq, and Illumina sequencing data are publicly available and listed in Table S1 along with accession numbers.

Code Availability

For non-human great apes, we integrated data from three different sequencing approaches into one pipeline to assemble YAGs transcripts (Fig. 14). The pipeline is available at GitHub as a Snakefile (https://github.com/agreshno/YAGs_great_apes/blob/main/Snakemake-Targeted-IsoSeq/Snakefile).

ACKNOWLEDGMENTS

We are grateful to Katy Munson and Evan Eichler for generating the untargeted IsoSeq data from the great ape testis samples collected in the Makova lab. This study was supported by grants R01GM130691 and R35GM151945 to KDM.

References

1. Ahmadi Rastegar, D., M. Sharifi Tabar, M. Alikhani, P. Parsamatin, F. Sahraneshin Samani, M. Sabbaghian, M. A. Sadighi Gilani, A. Mohammad Ahadi, A. Mohseni Meybodi, A. Piryaei, N. Ansari-Pour, H. Gourabi, H. Baharvand, and G. H. Salekdeh. 2015. Isoform-Level Gene Expression Profiles of Human Y Chromosome Azoospermia Factor Genes and Their X Chromosome Paralogs in the Testicular Tissue of Non-Obstructive Azoospermia Patients. J. Proteome Res. 14:3595–3605. American Chemical Society.Bachtrog, D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nat. Rev. Genet. 14:113–124. Nature Publishing Group.
2. Bhowmick, B. K., Y. Satta, and N. Takahata. 2007. The origin and evolution of human ampliconic gene families and ampliconic structure. Genome Res. 17:441–450.
3. Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.
4. Charlesworth, B. 1996. The evolution of chromosomal sex determination and dosage compensation. Curr. Biol. 6:149–162. Elsevier.
5. Danecek, P., J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, and H. Li. 2021. Twelve years of SAMtools and BCFtools. GigaScience 10:giab008.
6. Fagerberg, L., B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C. A.-K. Szigyarto, M. Skogs, J. O. Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J. M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, Å. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Ponten, and M. Uhlén. 2014. Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics *. Mol. Cell. Proteomics 13:397–406. Elsevier.
7. Feng, S., M. Xu, F. Liu, C. Cui, and B. Zhou. 2019. Reconstruction of the full-length transcriptome atlas using PacBio Iso-Seq provides insight into the alternative splicing in Gossypium australe. BMC Plant Biol. 19:365.
8. Ferrández-Peral, L., X. Zhan, M. Alvarez-Estape, C. Chiva, P. Esteller-Cucala, R. García-Pérez, E. Julià, E. Lizano, Ò. Fornas, E. Sabidó, Q. Li, T. Marquès-Bonet, D. Juan, and G. Zhang. 2022. Transcriptome innovations in primates revealed by single-molecule long-read sequencing. Genome Res. 32:1448–1462.
9. Goldtzvik, Y., N. Sen, S. D. Lam, and C. Orengo. 2023. Protein diversification through post-translational modifications, alternative splicing, and gene duplication. Curr. Opin. Struct. Biol. 81:102640.
10. Grive, K. J., Y. Hu, E. Shu, A. Grimson, O. Elemento, J. K. Grenier, and P. E. Cohen. 2019. Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. PLOS Genet. 15:e1007810. Public Library of Science.
11. Guo, J., E. Sosa, T. Chitiashvili, X. Nie, E. J. Rojas, E. Oliver, K. Plath, J. M. Hotaling, J.-B. Stukenborg, A. T. Clark, and B. R. Cairns. 2021. Single-cell analysis of the developing human testis reveals somatic niche cell specification and fetal germline stem cell establishment. Cell Stem Cell 28:764-778.e4.
12. Hallast, P., P. Balaresque, G. R. Bowden, S. Ballereau, and M. A. Jobling. 2013. Recombination Dynamics of a Human Y-Chromosomal Palindrome: Rapid GC-Biased Gene Conversion, Multi-kilobase Conversion Tracts, and Rare Inversions. PLOS Genet. 9:e1003666. Public Library of Science.

13. Hermann, B. P., K. Cheng, A. Singh, L. Roa-De La Cruz, K. N. Mutoji, I.-C. Chen, H. Gildersleeve, J. D. Lehle, M. Mayo, B. Westernströer, N. C. Law, M. J. Oatley, E. K. Velte, B. A. Niedenberger, D. Fritze, S. Silber, C. B. Geyer, J. M. Oatley, and J. R. McCarrey. 2018. The Mammalian Spermatogenesis Single-Cell Transcriptome, from Spermatogonial Stem Cells to Spermatids. Cell Rep. 25:1650-1667.e8.
14. Kelemen, O., P. Convertini, Z. Zhang, Y. Wen, M. Shen, M. Falaleeva, and S. Stamm. 2013. Function of alternative splicing. Gene 514:1–30.
15. Kim, D., B. Langmead, and S. L. Salzberg. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat. Methods 12:357–360. Nature Publishing Group.
16. Leung, S. K., A. R. Jeffries, I. Castanho, B. T. Jordan, K. Moore, J. P. Davies, E. L. Dempster, N. J. Bray, P. O'Neill, E. Tseng, Z. Ahmed, D. A. Collier, E. D. Jeffery, S. Prabhakar, L. Schalkwyk, C. Jops, M. J. Gandal, G. M. Sheynkman, E. Hannon, and J. Mill. 2021. Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Rep. 37:110022.
17. Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.
18. Makova, K. D., B. D. Pickett, R. S. Harris, G. A. Hartley, M. Cechova, K. Pal, S. Nurk, D. Yoo, Q. Li, P. Hebbar, B. C. McGrath, F. Antonacci, M. Aubel, A. Biddanda, M. Borchers, E. Bornberg-Bauer, G. G. Bouffard, S. Y. Brooks, L. Carbone, L. Carrel, A. Carroll, P.-C. Chang, C.-S. Chin, D. E. Cook, S. J. C. Craig, L. de Gennaro, M. Diekhans, A. Dutra, G. H. Garcia, P. G. S. Grady, R. E. Green, D. Haddad, P. Hallast, W. T. Harvey, G. Hickey, D. A. Hillis, S. J. Hoyt, H. Jeong, K. Kamali, S. L. K. Pond, T. M. LaPolice, C. Lee, A. P. Lewis, Y.-H. E. Loh, P. Masterson, K. M. McGarvey, R. C. McCoy, P. Medvedev, K. H. Miga, K. M. Munson, E. Pak, B. Paten, B. J. Pinto, T. Potapova, A. Rhie, J. L. Rocha, F. Ryabov, O. A. Ryder, S. Sacco, K. Shafin, V. A. Shepelev, V. Slon, S. J. Solar, J. M. Storer, P. H. Sudmant, Sweetalana, A. Sweeten, M. G. Tassia, F. Thibaud-Nissen, M. Ventura, M. A. Wilson, A. C. Young, H. Zeng, X. Zhang, Z. A. Szpiech, C. D. Huber, J. L. Gerton, S. V. Yi, M. C. Schatz, I. A. Alexandrov, S. Koren, R. J. O'Neill, E. E. Eichler, and A. M. Phillippy. 2024. The Complete Sequence and Comparative Analysis of Ape Sex Chromosomes. bioRxiv.
19. Miga, K. H., S. Koren, A. Rhie, M. R. Vollger, A. Gershman, A. Bzikadze, S. Brooks, E. Howe, D. Porubsky, G. A. Logsdon, V. A. Schneider, T. Potapova, J. Wood, W. Chow, J. Armstrong, J. Fredrickson, E. Pak, K. Tigyi, M. Kremitzki, C. Markovic, V. Maduro, A. Dutra, G. G. Bouffard, A. M. Chang, N. F. Hansen, A. B. Wilfert, F. Thibaud-Nissen, A. D. Schmitt, J.-M. Belton, S. Selvaraj, M. Y. Dennis, D. C. Soto, R. Sahasrabudhe, G. Kaya, J. Quick, N. J. Loman, N. Holmes, M. Loose, U. Surti, R. ana Risques, T. A. Graves Lindsay, R. Fulton, I. Hall, B. Paten, K. Howe, W. Timp, A. Young, J. C. Mullikin, P. A. Pevzner, J. L. Gerton, B. A. Sullivan, E. E. Eichler, and A. M. Phillippy. 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585:79–84. Nature Publishing Group.
20. Mueller, J. L., H. Skaletsky, L. G. Brown, S. Zaghlul, S. Rock, T. Graves, K. Auger, W. C. Warren, R. K. Wilson, and D. C. Page. 2013. Independent specialization of the human and mouse X chromosomes for the male germ line. Nat. Genet. 45:1083–1087. Nature Publishing Group.
21. Navarro-Costa, P., L. Pereira, C. Alves, L. Gusmão, C. Proença, P. Marques-Vidal, T. Rocha, S. C. Correia, S. Jorge, A. Neves, A. P. Soares, J. Nunes, C. Calhaz-Jorge, A. Amorim, C. E. Plancha, and J. Gonçalves. 2007. Characterizing partial AZFc deletions of the Y chromosome with amplicon-specific sequence markers. BMC Genomics 8:342.
22. Nie, X., S. K. Munyoki, M. Sukhwani, N. Schmid, A. Missel, B. R. Emery, DonorConnect, J.-B. Stukenborg, A. Mayerhofer, K. E. Orwig, K. I. Aston, J. M. Hotaling, B. R. Cairns, and J. Guo. 2022. Single-cell analysis of human testis aging and correlation with elevated body mass index. Dev. Cell 57:1160-1176.e5. Elsevier.

23. Pardo-Palacios, F. J., A. Arzalluz-Luque, L. Kondratova, P. Salguero, J. Mestre-Tomás, R. Amorín, E. Estevan-Morió, T. Liu, A. Nanni, L. McIntyre, E. Tseng, and A. Conesa. 2023. SQANTI3: curation of long-read transcriptomes for accurate identification of known and novel isoforms. bioRxiv.
24. Park, E., Z. Pan, Z. Zhang, L. Lin, and Y. Xing. 2018. The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am. J. Hum. Genet. 102:11–26.
25. Pearson, W. R. 2013. An Introduction to Sequence Similarity (“Homology”) Searching. Curr. Protoc. Bioinforma. 42:3.1.1-3.1.8.
26. Pertea, G., and M. Pertea. 2020. GFF Utilities: GffRead and GffCompare. F1000Research.
27. Ren, Y., E. Tseng, T. P. L. Smith, S. Hiendleder, J. L. Williams, and W. Y. Low. 2023. Long read isoform sequencing reveals hidden transcriptional complexity between cattle subspecies. BMC Genomics 24:108.
28. Rhie, A., S. Nurk, M. Cechova, S. J. Hoyt, D. J. Taylor, N. Altemose, P. W. Hook, S. Koren, M. Rautiainen, I. A. Alexandrov, J. Allen, M. Asri, A. V. Bzikadze, N.-C. Chen, C.-S. Chin, M. Diekhans, P. Flicek, G. Formenti, A. Fungtammasan, C. Garcia Giron, E. Garrison, A. Gershman, J. L. Gerton, P. G. S. Grady, A. Guarracino, L. Haggerty, R. Halabian, N. F. Hansen, R. Harris, G. A. Hartley, W. T. Harvey, M. Haukness, J. Heinz, T. Hourlier, R. M. Hubley, S. E. Hunt, S. Hwang, M. Jain, R. K. Kesharwani, A. P. Lewis, H. Li, G. A. Logsdon, J. K. Lucas, W. Makalowski, C. Markovic, F. J. Martin, A. M. Mc Cartney, R. C. McCoy, J. McDaniel, B. M. McNulty, P. Medvedev, A. Mikheenko, K. M. Munson, T. D. Murphy, H. E. Olsen, N. D. Olson, L. F. Paulin, D. Porubsky, T. Potapova, F. Ryabov, S. L. Salzberg, M. E. G. Sauria, F. J. Sedlazeck, K. Shafin, V. A. Shepelev, A. Shumate, J. M. Storer, L. Surapaneni, A. M. Taravella Oill, F. Thibaud-Nissen, W. Timp, M. Tomaszkiewicz, M. R. Vollger, B. P. Walenz, A. C. Watwood, M. H. Weissensteiner, A. M. Wenger, M. A. Wilson, S. Zarate, Y. Zhu, J. M. Zook, E. E. Eichler, R. J. O’Neill, M. C. Schatz, K. H. Miga, K. D. Makova, and A. M. Phillippy. 2023. The complete sequence of a human Y chromosome. Nature 621:344–354. Nature Publishing Group.
29. Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16:276–277. Elsevier.
30. Rice, W. R. 1987. The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes. Evolution 41:911–914. [Society for the Study of Evolution, Wiley].
31. Ross, M. T., D. V. Grafham, A. J. Coffey, S. Scherer, K. McLay, D. Muzny, M. Platzer, G. R. Howell, C. Burrows, C. P. Bird, A. Frankish, F. L. Lovell, K. L. Howe, J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, M. E. Hurles, T. D. Andrews, C. E. Scott, S. Searle, J. Ramser, A. Whittaker, R. Deadman, N. P. Carter, S. E. Hunt, R. Chen, A. Cree, P. Gunaratne, P. Havlak, A. Hodgson, M. L. Metzker, S. Richards, G. Scott, D. Steffen, E. Sodergren, D. A. Wheeler, K. C. Worley, R. Ainscough, K. D. Ambrose, M. A. Ansari-Lari, S. Aradhya, R. I. S. Ashwell, A. K. Babbage, C. L. Bagguley, A. Ballabio, R. Banerjee, G. E. Barker, K. F. Barlow, I. P. Barrett, K. N. Bates, D. M. Beare, H. Beasley, O. Beasley, A. Beck, G. Bethel, K. Blechschmidt, N. Brady, S. Bray-Allen, A. M. Bridgeman, A. J. Brown, M. J. Brown, D. Bonnin, E. A. Bruford, C. Buhay, P. Burch, D. Burford, J. Burgess, W. Burrill, J. Burton, J. M. Bye, C. Carder, L. Carrel, J. Chako, J. C. Chapman, D. Chavez, E. Chen, G. Chen, Y. Chen, Z. Chen, C. Chinault, A. Ciccodicola, S. Y. Clark, G. Clarke, C. M. Clee, S. Clegg, K. Clerc-Blankenburg, K. Clifford, V. Cobley, C. G. Cole, J. S. Conquer, N. Corby, R. E. Connor, R. David, J. Davies, C. Davis, J. Davis, O. Delgado, D. DeShazo, P. Dhami, Y. Ding, H. Dinh, S. Dodsworth, H. Draper, S. Dugan-Rocha, A. Dunham, M. Dunn, K. J. Durbin, I. Dutta, T. Eades, M. Ellwood, A. Emery-Cohen, H. Errington, K. L. Evans, L. Faulkner, F. Francis, J. Frankland, A. E. Fraser, P. Galgoczy, J. Gilbert, R. Gill, G. Glöckner, S. G. Gregory, S. Gribble, C. Griffiths, R. Grocock, Y. Gu, R. Gwilliam, C. Hamilton, E. A. Hart, A. Hawes, P. D. Heath, K. Heitmann, S. Hennig, J. Hernandez, B. Hinzmann, S. Ho, M. Hoffs, P. J.

Howden, E. J. Huckle, J. Hume, P. J. Hunt, A. R. Hunt, J. Isherwood, L. Jacob, D. Johnson, S. Jones, P. J. de Jong, S. S. Joseph, S. Keenan, S. Kelly, J. K. Kershaw, Z. Khan, P. Kioschis, S. Klages, A. J. Knights, A. Kosiura, C. Kovar-Smith, G. K. Laird, C. Langford, S. Lawlor, M. Leversha, L. Lewis, W. Liu, C. Lloyd, D. M. Lloyd, H. Loulseged, J. E. Loveland, J. D. Lovell, R. Lozado, J. Lu, R. Lyne, J. Ma, M. Maheshwari, L. H. Matthews, J. McDowall, S. McLaren, A. McMurray, P. Meidl, T. Meitinger, S. Milne, G. Miner, S. L. Mistry, M. Morgan, S. Morris, I. Müller, J. C. Mullikin, N. Nguyen, G. Nordsiek, G. Nyakatura, C. N. O'Dell, G. Okwuonu, S. Palmer, R. Pandian, D. Parker, J. Parrish, S. Pasternak, D. Patel, A. V. Pearce, D. M. Pearson, S. E. Pelan, L. Perez, K. M. Porter, Y. Ramsey, K. Reichwald, S. Rhodes, K. A. Ridler, D. Schlessinger, M. G. Schueler, H. K. Sehra, C. Shaw-Smith, H. Shen, E. M. Sheridan, R. Shownkeen, C. D. Skuce, M. L. Smith, E. C. Sotheran, H. E. Steingruber, C. A. Steward, R. Storey, R. M. Swann, D. Swarbreck, P. E. Tabor, S. Taudien, T. Taylor, B. Teague, K. Thomas, A. Thorpe, K. Timms, A. Tracey, S. Trevanion, A. C. Tromans, M. d'Urso, D. Verduzco, D. Villasana, L. Waldron, M. Wall, Q. Wang, J. Warren, G. L. Warry, X. Wei, A. West, S. L. Whitehead, M. N. Whiteley, J. E. Wilkinson, D. L. Willey, G. Williams, L. Williams, A. Williamson, H. Williamson, L. Wilming, R. L. Woodmansey, P. W. Wray, J. Yen, J. Zhang, J. Zhou, H. Zoghbi, S. Zorilla, D. Buck, R. Reinhardt, A. Poustka, A. Rosenthal, H. Lehrach, A. Meindl, P. J. Minx, L. W. Hillier, H. F. Willard, R. K. Wilson, R. H. Waterston, C. M. Rice, M. Vaudin, A. Coulson, D. L. Nelson, G. Weinstock, J. E. Sulston, R. Durbin, T. Hubbard, R. A. Gibbs, S. Beck, J. Rogers, and D. R. Bentley. 2005. The DNA sequence of the human X chromosome. Nature 434:325–337. Nature Publishing Group.

32. Sahlin, K., and V. Mäkinen. 2021. Accurate spliced alignment of long RNA sequencing reads. Bioinformatics 37:4643–4651.
33. Saut, N., P. Terriou, A. Navarro, N. Lévy, and M. J. Mitchell. 2000. The human Y chromosome genes BPY2, CDY1 and DAZ are not essential for sustained fertility. Mol. Hum. Reprod. 6:789–793.
34. Saxena, R., L. G. Brown, T. Hawkins, R. K. Alagappan, H. Skaletsky, M. P. Reeve, R. Reijo, S. Rozen, M. B. Dinulos, C. M. Disteche, and D. C. Page. 1996. The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. Nat. Genet. 14:292–299. Nature Publishing Group.
35. Shumate, A., B. Wong, G. Pertea, and M. Pertea. 2022. Improved transcriptome assembly using a hybrid of long and short reads with StringTie. PLOS Comput. Biol. 18:e1009730. Public Library of Science.
36. Simpson, E. H. 1949. Measurement of Diversity. Nature 163:688–688. Nature Publishing Group.
37. Skaletsky, H., T. Kuroda-Kawaguchi, P. J. Minx, H. S. Cordum, L. Hillier, L. G. Brown, S. Repping, T. Pyntikova, J. Ali, T. Bieri, A. Chinwalla, A. Delehaunty, K. Delehaunty, H. Du, G. Fewell, L. Fulton, R. Fulton, T. Graves, S.-F. Hou, P. Latrielle, S. Leonard, E. Mardis, R. Maupin, J. McPherson, T. Miner, W. Nash, C. Nguyen, P. Ozersky, K. Pepin, S. Rock, T. Rohlfing, K. Scott, B. Schultz, C. Strong, A. Tin-Wollam, S.-P. Yang, R. H. Waterston, R. K. Wilson, S. Rozen, and D. C. Page. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423:825–837. Nature Publishing Group.
38. Soto, E. J. L., M. J. Gandal, T. Gonatopoulos-Pournatzis, E. A. Heller, D. Luo, and S. Zheng. 2019. Mechanisms of Neuronal Alternative Splicing and Strategies for Therapeutic Interventions. J. Neurosci. 39:8193–8199. Society for Neuroscience.
39. Tamura, K., G. Stecher, and S. Kumar. 2021. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol. Biol. Evol. 38:3022–3027.
40. Tomaszkiewicz, M., K. Sahlin, P. Medvedev, and K. D. Makova. 2023. Transcript Isoform Diversity of Ampliconic Genes on the Y Chromosome of Great Apes. bioRxiv.
41. Vallender, E. J., and B. T. Lahn. 2004. How mammalian sex chromosomes acquired their peculiar gene content. BioEssays 26:159–169.

42. Vegesna, R., M. Tomaszkiewicz, P. Medvedev, and K. D. Makova. 2019. Dosage regulation, and variation in gene expression and copy number of human Y chromosome ampliconic genes. PLoS Genet. 15:e1008369. Public Library of Science San Francisco, CA USA.
43. Vegesna, R., M. Tomaszkiewicz, O. A. Ryder, R. Campos-Sánchez, P. Medvedev, M. DeGiorgio, and K. D. Makova. 2020. Ampliconic Genes on the Great Ape Y Chromosomes: Rapid Evolution of Copy Number but Conservation of Expression Levels. Genome Biol. Evol. 12:842–859.
44. Veyrunes, F., P. D. Waters, P. Miethke, W. Rens, D. McMillan, A. E. Alsop, F. Grützner, J. E. Deakin, C. M. Whittington, K. Schatzkamer, C. L. Kremitzki, T. Graves, M. A. Ferguson-Smith, W. Warren, and J. A. M. Graves. 2008. Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res. 18:965–973.
45. Wang, E. T., R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. 2008. Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. Nature Publishing Group.
46. Waters, P. D., M. C. Wallis, and J. A. M. Graves. 2007. Mammalian sex—Origin and evolution of the Y chromosome and SRY. Semin. Cell Dev. Biol. 18:389–400.
47. Zhang, G., M. Sun, J. Wang, M. Lei, C. Li, D. Zhao, J. Huang, W. Li, S. Li, J. Li, J. Yang, Y. Luo, S. Hu, and B. Zhang. 2019. PacBio full-length cDNA sequencing integrated with RNA-seq reads drastically improves the discovery of splicing transcripts in rice. Plant J. 97:296–305.
48. Zhou, Y., X. Zhan, J. Jin, L. Zhou, J. Bergman, X. Li, M. M. C. Rousselle, M. R. Belles, L. Zhao, M. Fang, J. Chen, Q. Fang, L. Kuderna, T. Marques-Bonet, H. Kitayama, T. Hayakawa, Y.-G. Yao, H. Yang, D. N. Cooper, X. Qi, D.-D. Wu, M. H. Schierup, and G. Zhang. 2023. Eighty million years of rapid evolution of the primate Y chromosome. Nat. Ecol. Evol. 7:1114–1130. Nature Publishing Group.

Figure 1. Previously unannotated YAG copies identified from mapping the assembled YAG transcripts to the reference T2T genomes of great apes. **A**. Potential new gene copies with at least one replicate-supported transcript mapped (transparency differs between replicate-supported and not replicate-supported new gene copies).. **B**. An example of a potential new *BPY2* copy with seven transcripts (all not replicate-supported) mapping to it.

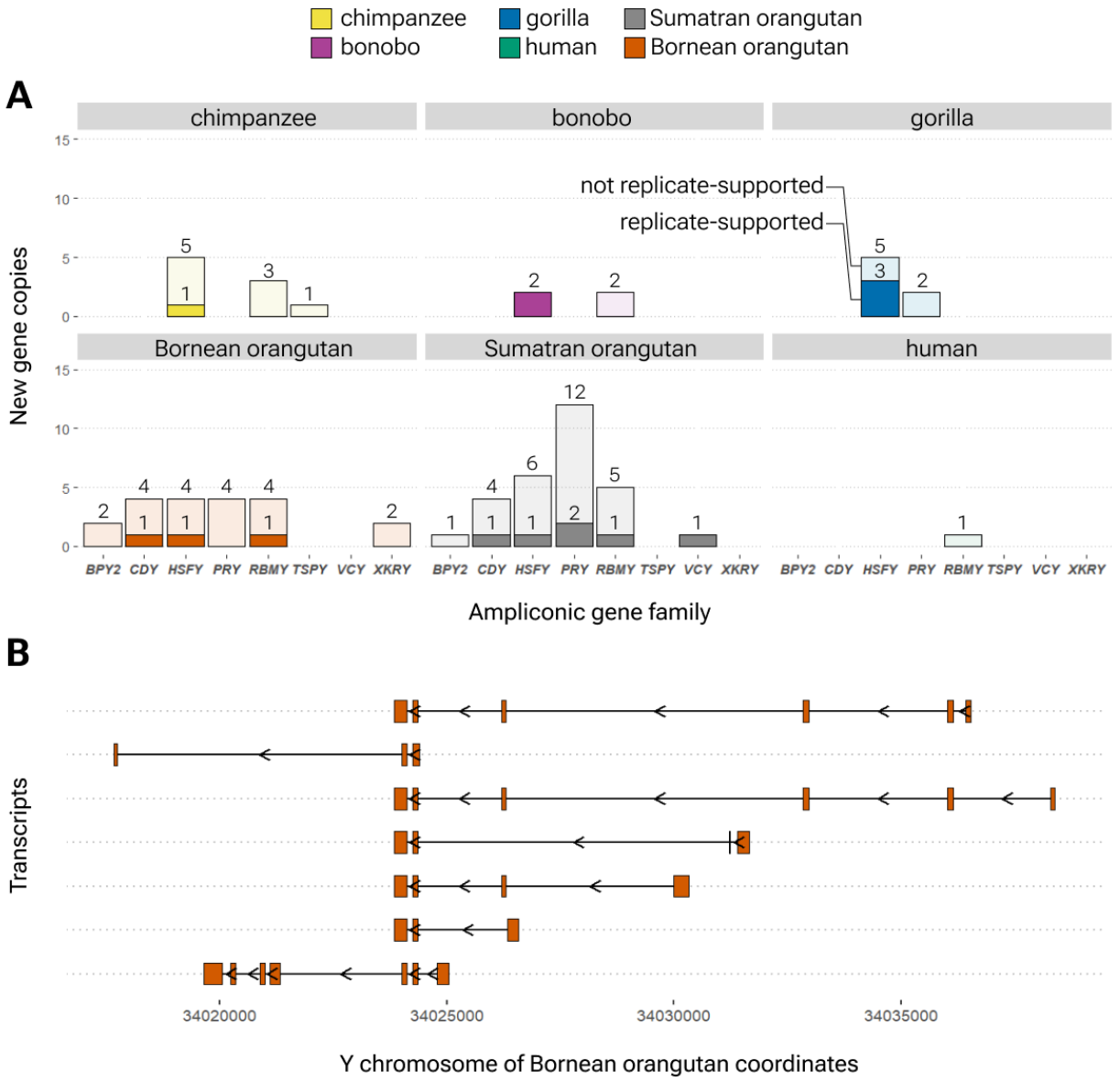

Figure 2. **A.** Overview of the presence of genes and transcribed pseudogenes for nine Y chromosome ampliconic gene families in great ape species based on the identified replicate-supported transcripts. A gene family in a species was marked as "protein-coding" if at least one replicate-supported transcript was identified and as "transcribed pseudogene" if no replicate-supported transcripts with high-confidence cORFS were found (but transcripts without open reading frames were identified). **B.** Overview of the presence of transcripts originating from X-chromosome and autosome homologs for nine ampliconic gene families in great ape species. These data were derived from the transcriptome assembly of targeted IsoSeq data only.

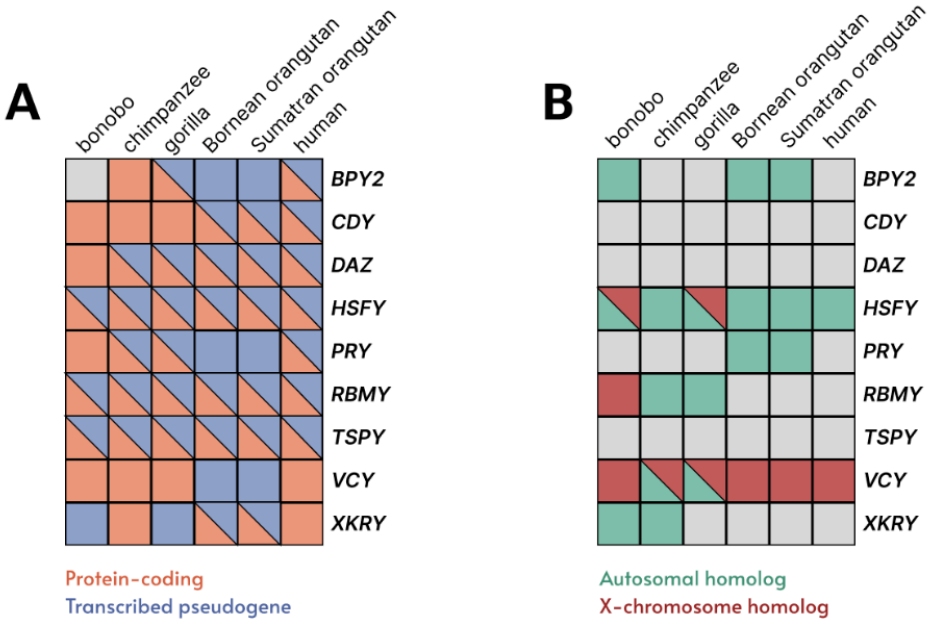

Figure 3. **A.** Total number of high-confidence cORFs per YAG family per species. **B.** Lengths of replicate-supported high-confidence cORFs and their comparisons to the lengths of known proteins (present in the NCBI protein database).

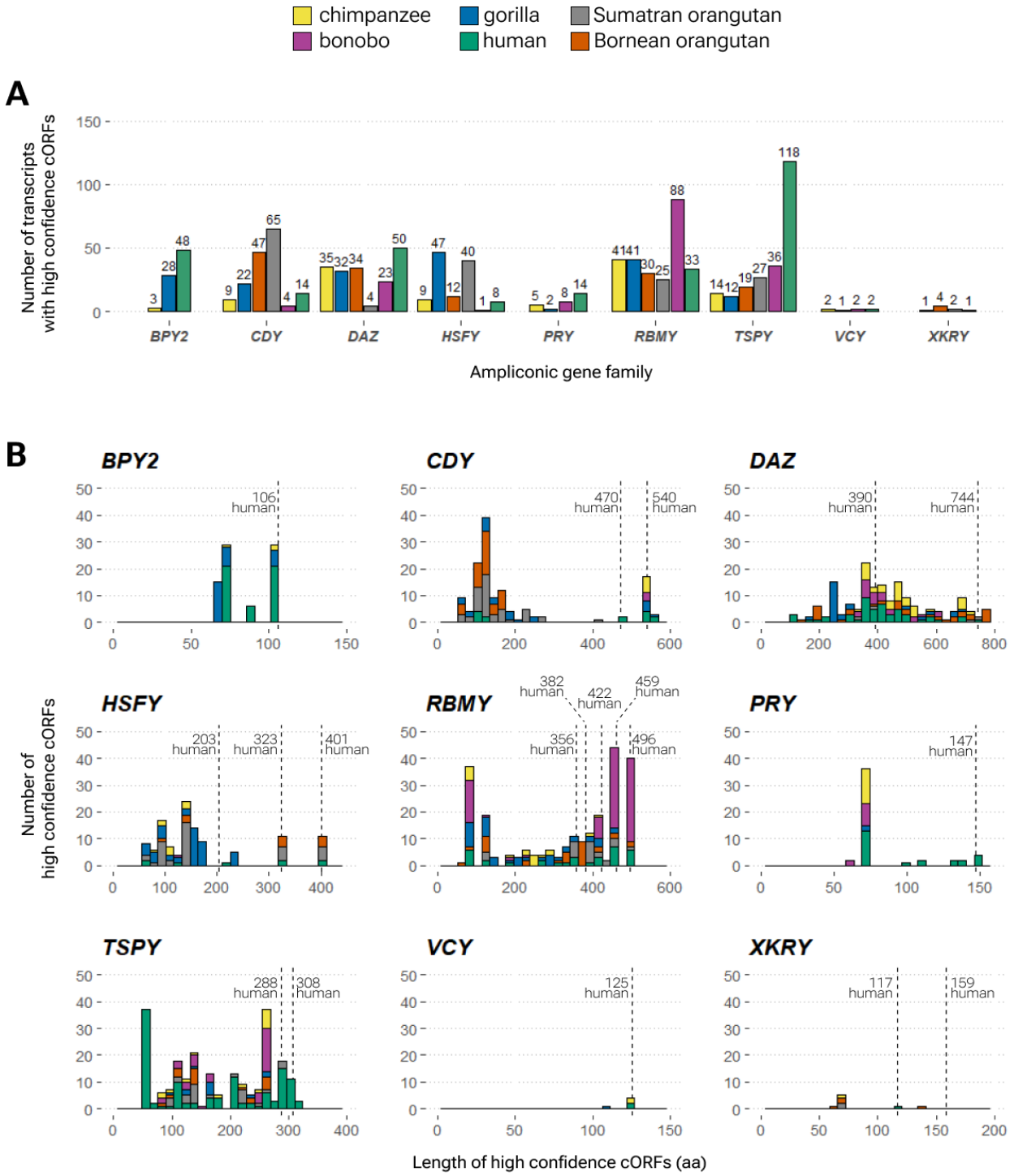

Figure 4. Summary of the number of identified gene copies and transcribed pseudogene copies of YAGs in great ape species. The transcribed pseudogenes are shown at the lower part of each graph.

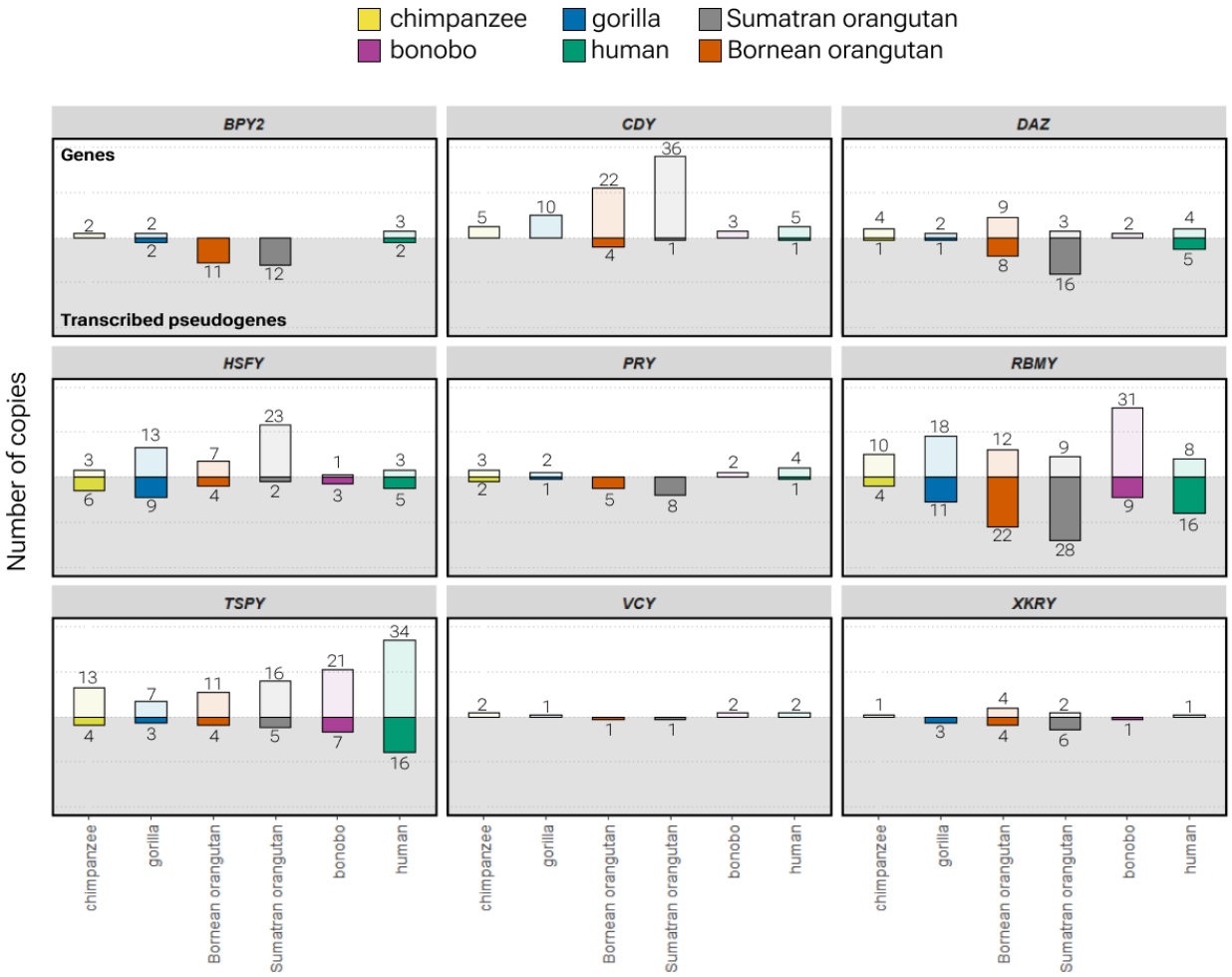

Figure 5. Characterization of YAG transcript isoforms diversity of great apes. Structural isoforms (differ in exon-intron structure, in orange) and sequence isoforms (differ in sequence, in blue) are shown. **A**. Total number of unique structural and sequence isoforms per gene family. **B**. Diversity measure (inverse Simpson; Simpson 1949) of transcript isoform diversity per gene family for each species.

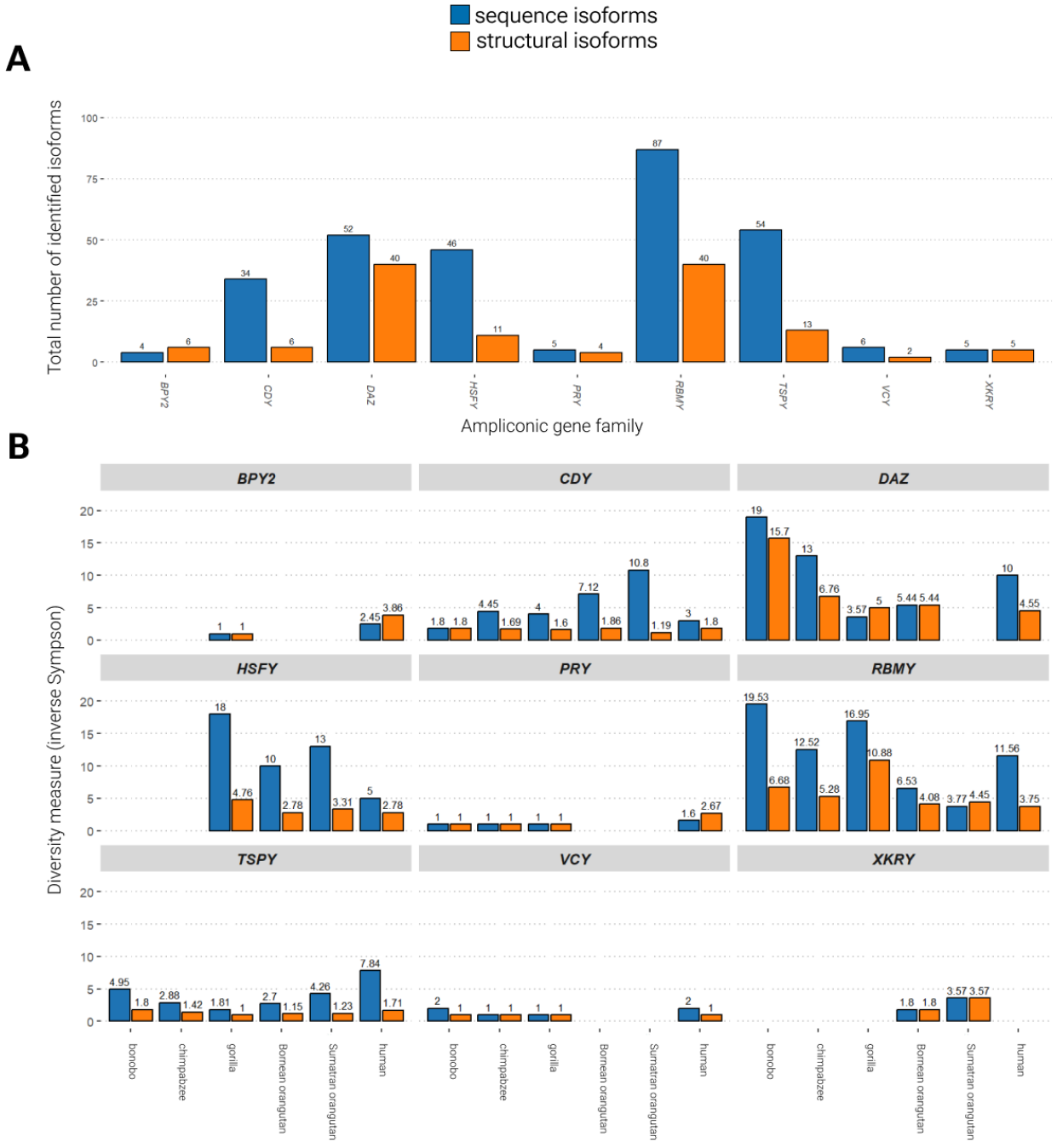

Figure 6. Sequence and structural transcript diversity for *CDY*, *TSPY,* and *DAZ* gene families across great ape species. Species are shown on the y-axis, isoform (structural in red and sequence in blue) ids on the x-axis; the number on the intersection corresponds to the number of transcripts representing a given isoform; data are presented for replicate-supported isoforms only. Figs. S21–S23 have similar plots for *BPY2*, *HSFY*, *PRY*, *RBMY*, *VCY*, and *XKRY* gene families.

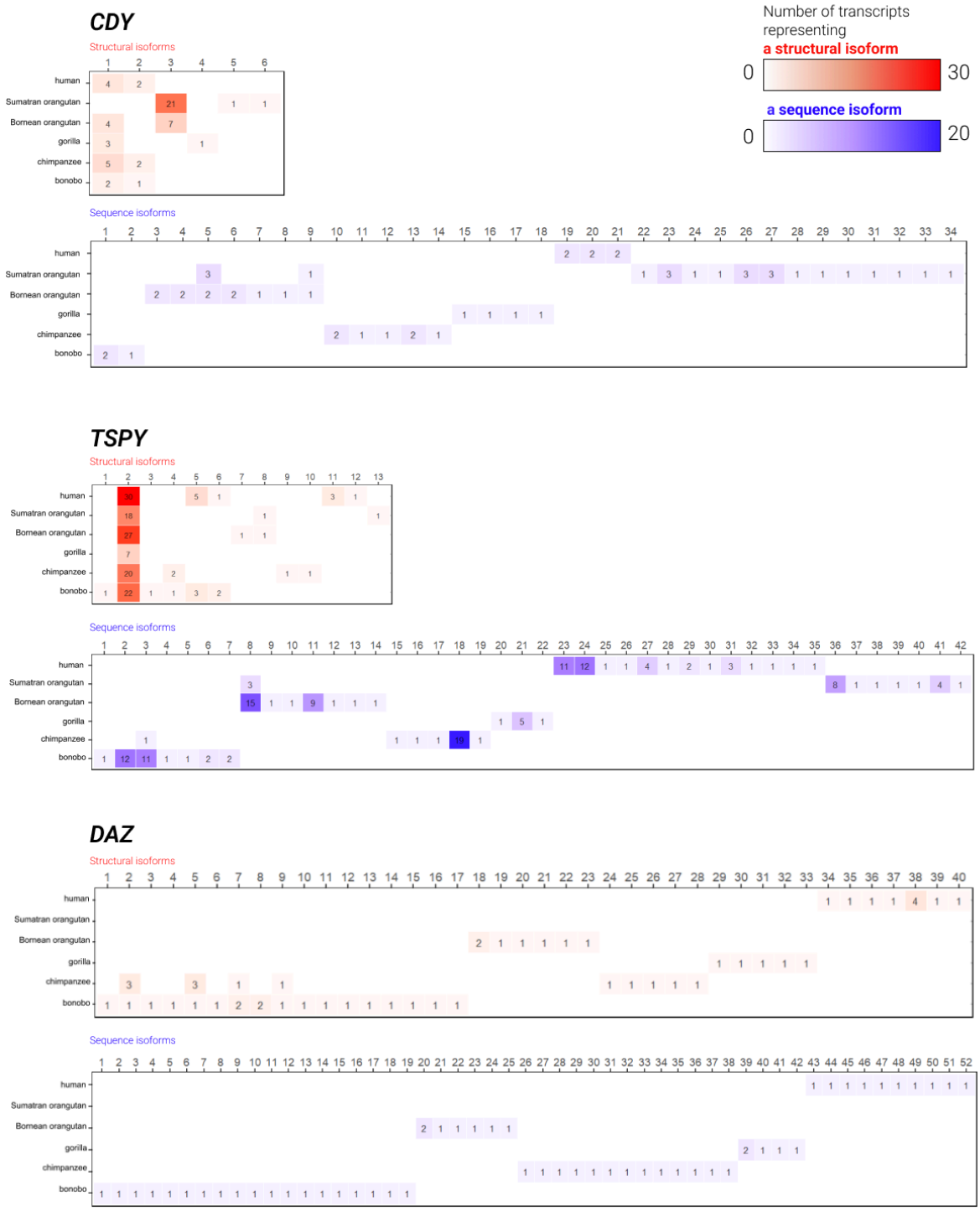
CDY
Structural isoforms
Sequence isoforms
TSPY
Structural isoforms
Sequence isoforms
DAZ
Structural isoforms
Sequence isoforms
human
Sumatran orangutan
Bornean orangutan
gorilla
chimpanzee
bonobo
Number of transcripts representing
a structural isoform
0
30
a sequence isoform
0
20

Figure 7. Relationship between sequence isoform diversity (inverse Simpson) and gene copy number (natural logarithm) across all great ape species per gene family. Black line represents a regression line fitted. Pearson correlation coefficient and *p*-value are shown at the upper part of each plot, sample size (i.e. the number of species that have a given gene family) is shown at the bottom .

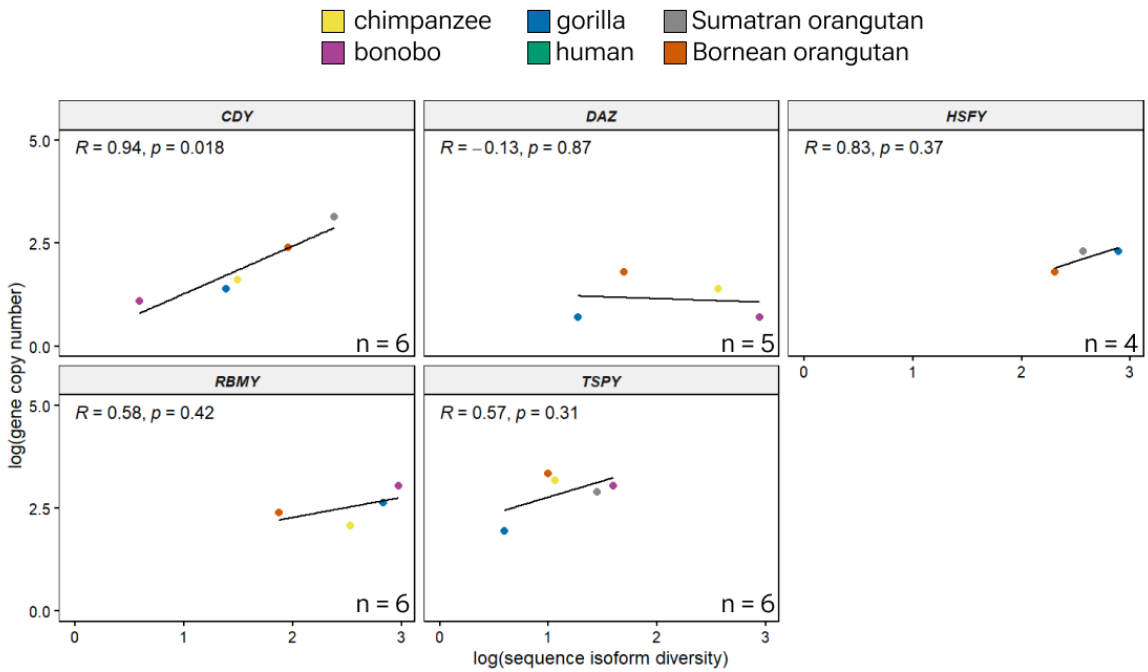

Table 1. The Y chromosome ampliconic gene (YAG) families and functions for nine known YAG families

Ampliconic gene family	Full name	Function	Protein length in amino acids (species; accession number)
BPY2 **(*VCY2*)**	Testis-Specific Basic Protein, Y-linked	Unknown; potentially involved in the cytoskeletal network (Wong et al. 2004); potentially involved in male germ-cells development and male infertility (Lahn and Page, 1997)	106 (human; NP_004669.2)
CDY	Testis-Specific Chromodomain Protein, Y-linked	Histone acetyltransferase (Lahn et al. 2002), regulates gene expression and post meiotic nuclear remodeling (Caron et al. 2003; Navarro-Costa et al. 2007)	541 (Gorilla; JAC06559), 540 (Human; NP_001003894), 470 (Human; XP_054184366), 541 (Chimpanzee; NP_001233424)
DAZ	Deleted in Azoospermia	Translational activator involved in spermatogonia differentiation (Schrans-Stassen et al. 2001), onset and progression of meiosis (Lin et al. 2008) through mRNA transportation and enhancement of translation (Collier et al. 2005); potentially regulates mRNA stability (Takeda et al. 2009). More potential functions are summarized in (VanGompel and Xu 2011)	390 (Human; NP_001375425.1), 744 (Human; NP_004072.3), 366 (Human; AAB02393.1)
HSFY	Heat Shock Transcription Factor, Y-Linked	Unknown, review in (Kovács et al. 2022); demonstrated transcription factor activity (Gaudet et al. 2011) and possibly regulate gametogenesis (Tessari et al. 2004; Widlak and Vydra 2017)	401, 203, and 213 (human; Tessari et al., 2004)
PRY	Testis-Specific PTP-BL-Related Y; PTPN13 Like, Y-Linked	Possible role in germ cell apoptosis (Stouffs et al. 2004)	147 (Sumatran orangutan; AJA38019.1), 143 (Gorilla; AJA38011.1), 147 (Human; NP_004667.2), 68 (Chimpanzee; PNI88546.1)
RBMY	RNA Binding Motif Protein, Y-Linked	Regulates nuclear splicing in the germline by regulation of expression of splicing factors (Dreumont et al. 2010)	356 (Human; NP_001307874.1), 459 (Human; NP_001307873.1), 496 (Human; NP_005049.1)
TSPY	Testis Specific Protein, Y-Linked	Regulation of spermatogonial cell replication and renewal; possible role in spermatogenesis as a catalyst/meiotic factor (Lau et al. 2011)	262 (Gorilla; JAC06569.1), 308 (Human; NP_003299.2), 294 (Human; NP_001184171.1), 221 (Human; NP_001307893.1)
VCY	Variable Charge, Y-Linked	Regulation of ribosome assembly in the male germline through modulation of expression of a regulator of ribosomal assembly (Zou et al. 2003)	125 (human; Lahn & Page, 2000)

XKRY	XK Related, Y-Linked	Multipass membrane transport protein involved in fertilization process (Lahn and Page 1997; Mahanta et al. 2012)	159 (Human; AAC51844.1)	

Table 2. The descriptive statistics for predicted YAG proteins across great ape species.

Species	YAG family	N		Length, bp	S.E. of length	Min length, bp	Max length, bp
Bonobo	***CDY***	4	162	554.50	13.50	595	541
	DAZ	23		420.00	17.10	611	318
	HSFY	1		130.00	NA	130	130
	PRY	8		71.00	0.76	73	69
	RBMY	88		392.18	16.61	497	87
	TSPY	36		203.06	11.38	264	76
	VCY	2		212.00	16.00	228	196
Bornean orangutan	***CDY***	47	146	193.47	25.41	601	56
	DAZ	34		542.71	40.92	951	128
	HSFY	12		284.42	35.85	404	88
	RBMY	30		304.07	24.44	497	66
	TSPY	19		175.63	15.58	264	79
	XKRY	4		86.50	17.91	140	64
Chimpanzee	***BPY2***	3	119	95.00	11.00	106	73
	CDY	9		559.00	9.00	595	541
	DAZ	35		477.57	18.07	720	342
	HSFY	9		108.67	7.21	134	78
	PRY	5		71.40	0.98	73	69
	RBMY	41		372.22	22.60	497	87
	TSPY	14		200.57	20.72	264	84
	VCY	2		124.00	0.00	124	124
	XKRY	1		70.00	NA	70	70
Gorilla	***BPY2***	28	185	76.46	2.78	104	68
	CDY	22		237.36	37.79	557	65
	DAZ	32		422.97	42.27	850	173
	HSFY	47		140.77	7.15	228	57
	PRY	2		73.00	0.00	73	73
	RBMY	41		221.61	19.72	460	81
	TSPY	12		174.58	15.53	264	96
	VCY	1		107.00	NA	107	107
Human	***BPY2***	48	288	89.44	2.25	106	73
	CDY	14		349.29	57.14	554	102
	DAZ	50		402.74	19.27	744	121
	HSFY	8		238.25	50.95	401	57
	PRY	14		73.57	1.88	98	71

	RBMY	33		333.88	27.43	496	81
	TSPY	118		168.71	9.41	314	50
	VCY	2		125.00	0.00	125	125
	XKRY	1		117.00	NA	117	117
Sumatran orangutan	***CDY***	65	163	271.22	26.34	601	63
	DAZ	4		549.25	108.65	741	330
	HSFY	40		175.40	18.16	404	63
	RBMY	25		362.92	20.39	497	114
	TSPY	27		173.04	12.41	290	90
	XKRY	2		71.00	0.00	71	71
All species	***BPY2***	79		85.05	1.86	106	68
	CDY	161		273.81	16.49	601	56
	DAZ	178		453.35	13.51	951	121
	HSFY	117		171.44	9.46	404	57
	PRY	29		72.45	0.95	98	69
	RBMY	258		341.36	9.56	497	66
	TSPY	226		177.57	5.83	314	50
	VCY	7		147.00	17.31	228	107
	XKRY	8		84.38	9.91	140	64
All species	All YAGs	1063		266.76	5.68	951	50