Copy Number Variation of Ampliconic Regions on Hominid X and Y chromosomes

Bachelor internship at BiRC and Bioscience, AU | 10 ECTS, Spring semester 2018

Written by Carl Mathias Kobel (201404379), kobel@pm.me

Supervised by Elise Lucotte and Mikkel Heide Schierup

Abstract

In order to describe the copy number of ampliconic genes on the hominid X chromosome, we assembled species specific artificial chromosomes (ACs) containing orthologs of genes which are known to be ampliconic in human. By mapping reads from several chimpanzee (Pan troglodytes) and gorilla (Gorilla gorilla) individuals onto these species specific ACs, we measured the copy number variation and argue that most of the genes have less copies in chimpanzee and gorilla than in humans, which indicates that they were amplified in the human lineage. We find that GAGE4 completely lacks ampliconic behavior in chimpanzee and gorilla.

Introduction

Sex determining systems

The modern XY sex determining system in mammals most probably emerged from a former environmental sex determining system. In an ancestor with the environmental sex determining system, a variant occurred on one of the homologous chromosomes. This variant disrupted the environmental factor such that all offspring with this variant would become male, and offspring without; female. As recombination between the homologous chromosomes stopped, sex specific genes accumulated, the chromosomes diverged to become what we know as modern sex chromosomes.

This emergence model was initially developed on the ZW sex determining system (S. Ohno 1967 [1]). Nonetheless, comparative mapping shows that it can be applied to the XY sex determining system as well.

In mammals, the SRY gene is defined as the Testis Determining Factor. It may be the variant that initially started the divergence between the homologous chromosomes.

Although the sex chromosomes have diverged to become very different, they still have pseudo-autosomal regions in the ends – PAR1 and PAR2 – The recombination activity in these areas is needed for successful cell division and thus are conserved from before the divergence.

Evolution is expected to be faster on the sex chromosomes because in a population, compared to autosomes, there are 3/4 X chromosomes and 1/4 Y chromosomes - a balanced sex ratio assumed. Another consequence is that the sex chromosomes are subjected to higher drift,

especially for the Y chromosome [2]. Additionally, the Y chromosome is subjected to a higher mutation rate because the male germline surpasses many more cell divisions per generation than does the female.

Ampliconic genes

Ampliconic genes are present only on sex chromosomes. They consist of very similar adjacent duplications with variable copy numbers. The mechanism by which they duplicate is not known. Most of the ampliconic genes are testis expressed and hypothesized to be involved in meiotic drive processes [2], [3]. Meiotic drive favors the segregation of specific genes, thus disturbing the mendelian segregation ratios. In mice there is evidence for an arms race between a pair of homologous ampliconic genes residing on each of the sex chromosomes, *Sly* and *Slx.* This pair of genes compete to be transmitted to the next generation, due to an intragenomic conflict. A deficiency of *Slx* distorts the sex ratio to have higher frequency of males, and *vice-versa* for Sly.

During meiosis, sex chromosome inactivation is crucial in order to avoid mechanisms that disturb the segregation of sex-chromosomes. This inactivation is often disrupted in hybrids, at least during formation of round spermatids, and evidence suggests that it is caused by an unbalanced copy number of *Sly* and *Slx*. [4]

Intragenomic conflicts lead to speciation because the arms-race processes triggers a rapid differentiation of the sex-chromosomes, which may become sufficiently divergent between sub-populations, inducing hybrid incompatibilities.

The genetic homology between human and mouse is much higher for single copy genes, than for ampliconic genes, it is therefore suggested that the evolutionary turnover is much faster for ampliconic genes.

Sperm competition in Hominids

There are different levels of sperm competition among the hominids. Indeed, Gorillas have a mating system where a dominant male monopolizes copulations with females and chimpanzees have a multi-male multi-female mating system, where many males copulate with each female. This differential behavior leads to higher selection on increased testes size in chimpanzee compared to Gorilla [5]. In humans most of the ampliconic genes are expressed in testis, therefore, sperm competition may have influenced the evolution of these genes differently across gorillas, chimpanzees and humans.

The goal of this study is therefore to compare the copy number of the ampliconic genes between gorillas, chimpanzees and humans, to learn more about the history of amplification of these genes.

Method

The pipeline (Figure 1) used in this study consists of assembling an artificial chromosome (AC) consisting of one copy of each ampliconic candidate gene. Then mapping reads from an individual onto this AC and assessing the copy number by relating the coverage of each gene to a control gene which is non-ampliconic in humans.

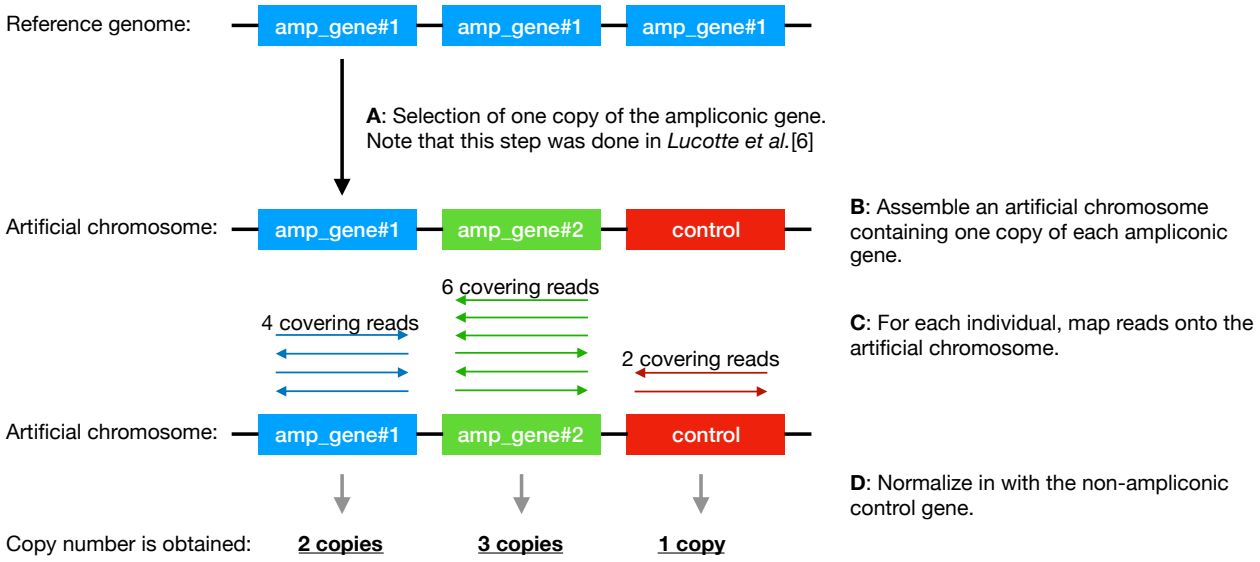


Figure 1: Overview of the pipeline.

Assembly of artificial chromosomes

In order to assemble the ACs, we adapted the method from Lucotte et al. 2018 [6] and selected the genes that exhibited copy number variation in human populations (X chromosome: GAGE4, CT47A4, CT45A5, SPANXB1 and OPN1LW; Y chromosome: BPY2, CDY, DAZ, HSFY, PRY, RBMY1A1, TSPY and XKRY). For each of the human ampliconic genes, we searched for orthologs in the chimpanzee and gorilla genomes. We used GRCh38.p12, Pan_tro_3.0 and gorGor4 as reference genomes for human, chimpanzee and gorilla, respectively.

To do so, we first used the Ensembl genome browser. Only orthologs with a subject/query identical factor of more than 0.5 and a length of at least half of the original, were included.

If orthologs were not found in the Ensembl genome browser, we used BLAST v2.7.1 to align the human genes against the chimpanzee and gorilla reference genomes. We selected regions with a subject/query identical factor of more than 0.5 and ortholog length of at least half of the original human gene.

See

Table 1 for the results of the ortholog search.

The orthologs found on Ensembl were downloaded directly as fasta files. The orthologs found with BLAST were extracted out of the reference. The isolated gene sequences were then merged into a complete AC. Because no orthologs with satisfying statistics were found for the gorilla Y chromosome except one (XKRY) we decided to omit that chromosome completely. Thus, we constructed ACs for chimpanzee X, chimpanzee Y and gorilla X. These were then used in the read-mapping method detailed in the next section.

Table 1: Table of the human ampliconic gene orthologs in chimpanzee and gorilla. The genes were assembled into ACs for each species. Query id% is the percentage of the human sequence matching the sequence of the ortholog. Length fraction is the query sequence length divided by the subject sequence length.

Chromosome	Gene	Ensembl Query id.	Ensembl length fraction	BLAST Identities	BLAST length fraction
chimpanzee X	CT45A5	0.96	1.81	-	-
chimpanzee X	CT47A4	0.92	1.15	-	-
chimpanzee X	GAGE4	0.82	2.27	-	-
chimpanzee X	OPN1LW	-	-	0.98	0.95
chimpanzee X	SPANXB1	0.68	0.85	-	-
chimpanzee X	DMD (control)	0.99	1.00	-	-
gorilla X	CT45A5	0.78	0.76	-	-
gorilla X	CT47A4	0.80	0.31	-	-
gorilla X	GAGE4	0.89	2.09	0.90	0.51
gorilla X	SPANXB1	0.67	1.75	-	-
gorilla X	DMD (control)	0.99	1.00	-	-
chimpanzee Y	BPY2	0.98	0.19	-	-
chimpanzee Y	CDY	0.97	0.77	-	-
chimpanzee Y	PRY	0.97	0.52	-	-
chimpanzee Y	RBMY1A1	0.91	0.32	-	-
chimpanzee Y	TSPY	0.90	1.43	-	-
chimpanzee Y	XKRY	-	-	0.98	1.00
chimpanzee Y	AMELY (control)	0.98	1.00	-	-

Mapping reads onto artificial chromosomes

For each individual:

- The reads from the fasta-files were mapped against the ACs using BWA [7] v0.7.5a
- The alignment was filtered using sambamba [8] v0.5.1 for a mapping quality >= 50 and cigar = 100M and NM < 3 These parameters were selected in order to be make the results comparable to Lucotte et al. 2018 [6]
- The read depth for each position of the AC was calculated using SAMtools [9] v1.3

We calculated the median read depth across all positions in the gene. We used the median instead of the mean because it is less sensitive to extreme outlier values. This median is what we subsequently define as the un-normalized copy number. In order to estimate the copy number of each ampliconic gene, we divided the median coverage of each gene by the median coverage of the control gene which is assumed to be non-ampliconic.

We selected DMD as the control gene for the X chromosome and AMELY for the Y chromosome, as they are both single copy in humans. In order to see how the species-specific ACs performed we also executed the pipeline with the human ACs built in Lucotte et al. 2018 [6]. Copy number estimations should be more accurate when using the species specific (chimpanzee and gorilla) ACs. Indeed, if the orthologs are very divergent between humans, chimpanzees and gorillas, using the human AC for our experiment would lead to a lower number of reads mapped to the chromosome, because of the filtering, and therefore to an underestimation of copy number. However, a comparison of both estimations is interesting because it can expose how different the

ampliconic gene-sequences are, between species. Also, because it is interesting to see if the human artificial sex-chromosomes are good enough for application on closely related species.

Results

Comparison: human vs. species-specific artificial chromosomes

We first validated the ACs assembled for chimpanzee and gorilla. We expect to see that the species-specific artificial chromosomes have a higher coverage compared to that of the human AC. For all species and sex, the mean difference of coverage between the species-specific AC and the human AC (Table 2) is higher when using the species-specific, over the human artificial chromosome. Because of the varying number of sex chromosomes present in different sexes, we expect to see double the amount of copies for each gene. Dosage compensation might counteract doubling though. Generally, females show a higher coverage (Figure 2 and Figure 3). As females don't have a Y chromosome, it is interesting that the coverage of AMELY and PRY is not exactly zero. This is likely because homologous genes from other chromosomes might have been mapped erroneously to the artificial chromosomes even though filtering has been applied.

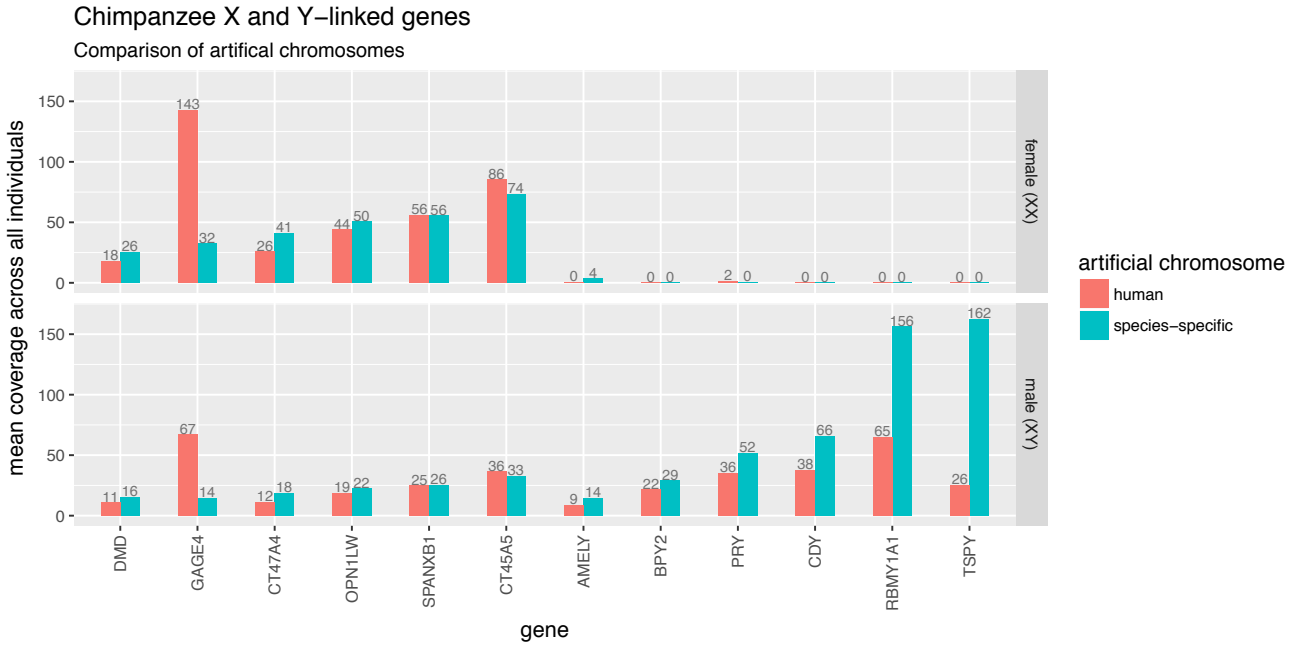


Figure 2: Comparison of the coverage for each gene, for the human AC or the species-specific ACs for chimpanzees. The mean is calculated by averaging the coverage over all the individuals.

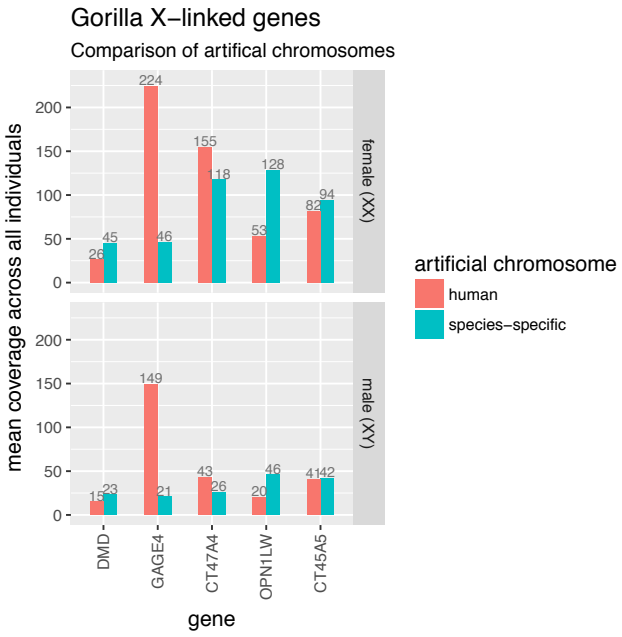
In chimpanzee, for most of the X-linked genes and for both males and females, the mean difference between the species-specific and the human AC coverage is small, except for GAGE4 where the human AC has a coverage that is approx. 4 fold higher than that of the species-specific AC (Figure 2).

The Y-linked genes show a higher coverage on the species-specific AC. This is consistent with the mean difference of coverage between the species-specific AC and the human AC, which is higher for the Y-linked genes (Table 2). The fact that the sensitivity of the artificial chromosome is bigger for the Y chromosome suggests that the Y chromosomes are more divergent between species than the X chromosomes.

Table 2: The mean individual-pair relative difference is calculated by taking one individual at a time, taking the relative difference of the species-specific and human AC coverage. Then taking the mean of a group (across each combination of species, sex and chromosome). This statistic gives an impression on how much the species-specific artificial chromosome performs better than the human AC, pairing the ACs one individual and gene at a time. A positive value means that the species-specific AC has higher coverage than the human AC.

Species	Chromosome	Sex	Mean individual-pair relative difference
chimpanzee	X	F	1.042
chimpanzee	X	M	1.060
chimpanzee	Y	F	-
chimpanzee	Y	M	2.504
gorilla	X	F	1.246
gorilla	X	M	1.121
		grand mean = 1.395	

Figure 3: Comparison of the coverage for each gene, for the human AC or the species-specific ACs for gorillas. The mean is calculated by averaging the coverage over all the individuals.



For gorilla the mean difference between the species-specific and human AC coverage is positive, indicating that the species-specific AC has higher sensitivity than the human AC (Table 2). Although, comparing the coverage across all individuals, the species-specific AC doesn't look generally more sensitive than the human (Figure 3). GAGE4 shows the highest relative difference in coverage between ACs here, as well as it does for chimpanzee.

Because the species-specific AC generally has a higher coverage, we decided to execute the rest of the pipeline with that.

X chromosome

For the X chromosome (Figure 4), most genes have shown a lower copy number in chimpanzee and gorilla, than in human. In both species, GAGE4 seems to be non-ampliconic, except in one individual. CT47A4, in both species, and OPN1LW, in chimpanzee, might also be single copy. The other genes seem to have more than one copy, however none of them show as high copy numbers in chimpanzee and gorilla as in human.

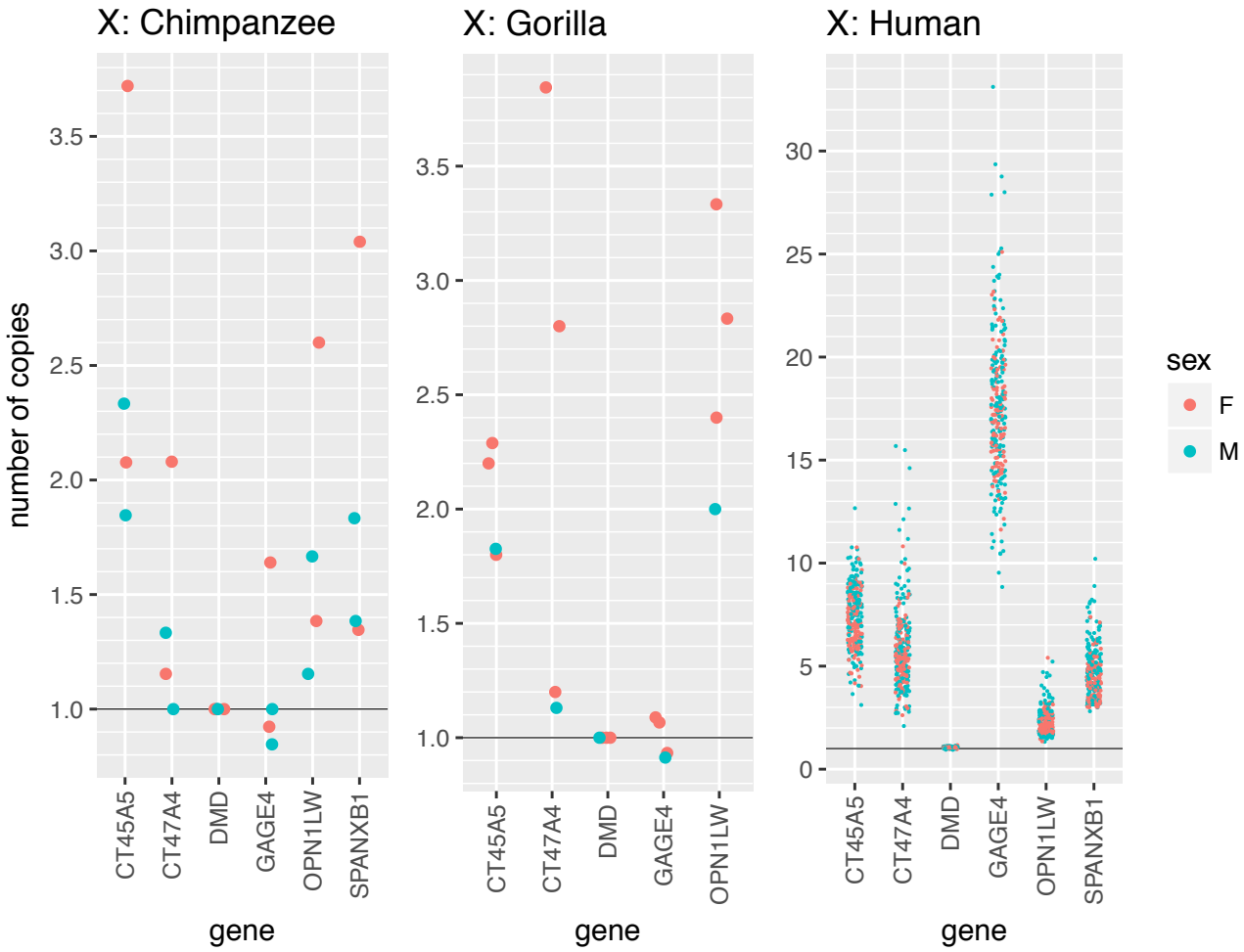


Figure 4: Copy number (normalized coverage) of X-linked genes. All individuals for each species. Note the differently scaled y-axis across species. Horizontal jitter applied. Horizontal line marked at copy number = 1.

Y chromosome

In human, the copy numbers of the Y-linked genes were normalized using a different control region: the X-degenerate region on the Y chromosome. Whereas, in chimpanzee, the copy numbers were normalized with AMELY. This may have an influence in the comparison of the results.

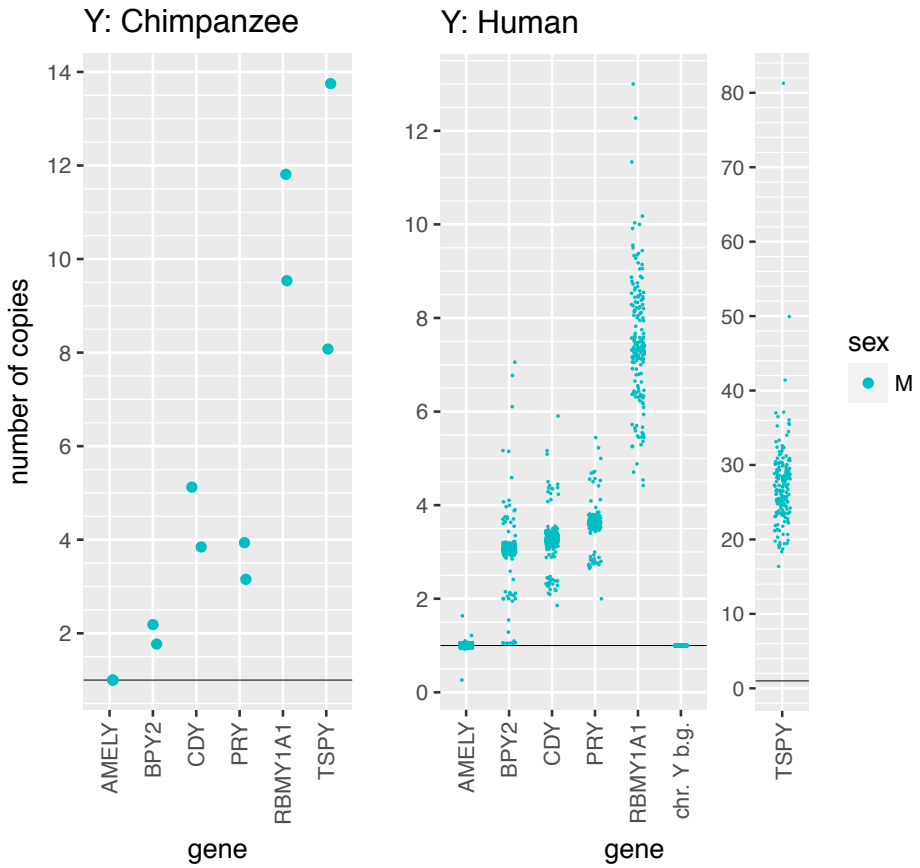


Figure 5: Copy number (normalized coverage) of Y-linked genes. All individuals for each species. Note the differently scaled y-axes across species. In human, TSPY has its own y-axis to allow for its extreme copy numbers. Horizontal jitter applied. Horizontal line marked at copy number = 1.

In chimpanzee BPY2 has a copy number close to 2, falling within the variation in human where the copy number for the gene is < 1, for 9/174 males.

TSPY seems to be ampliconic in chimpanzee as well, with 8 and 14 copies for each male, but in humans the main distribution is between 20 and 35 copies. The rest of the genes show signs of being ampliconic in chimpanzee: CDY and PRY are close to human copy numbers. RBMY1A1 in chimpanzee is above the median in human.

Table 3: Complete results from executing the pipeline with the species-specific AC. The human median copy number from Lucotte et al. 2018 [6] is included for comparison.

Species	Chrom	Sex	Gene	Min	Median	Max	SD	$n_{ind.s}$	Human median[6]
chimpanzee	X	F	CT45A5	2.1	**2.9**	3.7	1.2	2	6.9
chimpanzee	X	F	CT47A4	1.2	**1.6**	2.1	0.7	2	5.4
chimpanzee	X	F	DMD	1.0	**1.0**	1.0	0.0	2	1.1
chimpanzee	X	F	GAGE4	0.9	**1.3**	1.6	0.5	2	17.1
chimpanzee	X	F	OPN1LW	1.4	**2.0**	2.6	0.9	2	2.2
chimpanzee	X	F	SPANXB1	1.4	**2.2**	3.0	1.2	2	4.2
chimpanzee	X	M	CT45A5	1.9	**2.1**	2.3	0.3	2	7.9
chimpanzee	X	M	CT47A4	1.0	**1.2**	1.3	0.2	2	5.6
chimpanzee	X	M	DMD	1.0	**1.0**	1.0	0.0	2	1.1
chimpanzee	X	M	GAGE4	0.9	**0.9**	1.0	0.1	2	17.5
chimpanzee	X	M	OPN1LW	1.2	**1.4**	1.7	0.4	2	2.4
chimpanzee	X	M	SPANXB1	1.4	**1.6**	1.8	0.3	2	4.6
chimpanzee	Y	M	AMELY	1.0	**1.0**	1.0	0.0	2	1.0
chimpanzee	Y	M	BPY2	1.8	**2.0**	2.2	0.3	2	3.1
chimpanzee	Y	M	CDY	3.9	**4.5**	5.1	0.9	2	3.3
chimpanzee	Y	M	PRY	3.2	**3.6**	3.9	0.6	2	3.6
chimpanzee	Y	M	RBMY1A1	9.5	**10.7**	11.8	1.6	2	7.4
chimpanzee	Y	M	TSPY	8.1	**10.9**	13.8	4.0	2	26.6
gorilla	X	F	CT45A5	1.8	**2.2**	2.3	0.3	3	6.9
gorilla	X	F	CT47A4	1.2	**2.8**	3.8	1.3	3	5.4
gorilla	X	F	DMD	1.0	**1.0**	1.0	0.0	3	1.1
gorilla	X	F	GAGE4	0.9	**1.1**	1.1	0.1	3	17.1
gorilla	X	F	OPN1LW	2.4	**2.8**	3.3	0.5	3	2.2
gorilla	X	M	CT45A5	1.8	**1.8**	1.8	NA	1	7.9
gorilla	X	M	CT47A4	1.1	**1.1**	1.1	NA	1	5.6
gorilla	X	M	DMD	1.0	**1.0**	1.0	NA	1	1.1
gorilla	X	M	GAGE4	0.9	**0.9**	0.9	NA	1	17.5
gorilla	X	M	OPN1LW	2.0	**2.0**	2.0	NA	1	2.4

Discussion

Conclusion

In this study, we measured the copy number of human ampliconic genes in gorillas and chimpanzees.

For the X chromosome and for both species, genes CT45A5, CT47A4 and GAGE4 have a lower copy number than humans while gene OPN1LW falls within the range of humans and is lower in chimpanzee than in gorilla. If the variation of OPN1LW is lower in chimpanzee than in human, it could be non-ampliconic in chimpanzee. Ampliconic behavior in OPN1LW is interesting because

the gene is highly expressed in the testicles [10] though it codes for long-wavelength sensitive opsin which is important for vision in humans.

In chimpanzee SPANXB1 is lower than in humans.

For the Y chromosome, PRY, CDY and RBMY1A1 have copy numbers close to the human median. BPY falls within the human variation but is below the median. TSPY has a lower copy number than that of humans.

GAGE4 seems to be non ampliconic in chimpanzee and gorilla. This suggests that the ampliconic behavior of GAGE4 in human developed in the human lineage after the split of the human-chimpanzee last common ancestor, less than around 6.7 Mya [11].

However, the coverage of GAGE4 using the human AC was much higher than when using the species-specific AC, which suggests that maybe the ortholog chosen and included in the species-specific AC is not optimal.

Most genes show a lower copy number than that of humans in both chimpanzee and gorilla. This suggest that those genes were amplified recently in the human lineage, after the split with the chimpanzee lineage.

The difference in ampliconic behavior across genes among hominids might be influenced by sperm competition.

Criticism of the particular execution

This study represents a first attempt at estimating copy number and copy number variations, in chimpanzees and gorillas, of genes known to be ampliconic in humans. However, it is possible that some genes are ampliconic in chimpanzee and gorilla but not ampliconic in humans. In order to make the most comprehensible overview possible of the ampliconic genes in chimpanzee and gorilla, it is necessary to compile the list of candidate ampliconic genes with a method similar to what is used in Lucotte et al. 2018 [6]. After this list is created, it might be interesting to see if any of the genes are homologs between the sister species.

In the results it is mentioned that the coverage for the Y chromosome genes could be above zero for females, while they do not carry a Y chromosome. This is probably due to homolog genes present elsewhere in the genome (either on the X or on autosomes). In the human data from Lucotte et al. 2018 [6] the X chromosome was included as a decoy on the artificial Y chromosome, so that the reads containing X-linked genes would be aligned here, instead of being aligned to homologous genes on the artificial Y chromosome. Additionally, reads mapping on autosomes were removed. A perspective would therefore be to include the respective species X chromosomes on the ACs assembled in this experiment. However, the coverage on the Y chromosome genes for females is very low (2 and 4 reads), so the result should not be strongly affected.

AMELY might be a bad choice of control gene. As it has a homolog, AMELX, with high similarity. Because a decoy of the X chromosome is not included on the artificial Y chromosome, reads containing AMELX reads might have been mapped onto the AMELY gene on the AC.

A problem occurred in the assembly of the ACs: We checked coherence between Ensembl and BLAST results with GAGE4 in gorilla by comparing the results. The Query %id and BLAST Identities are on par (1 percent-point difference,

Table 1), but the length fraction was off, as Ensembl has an ortholog of 2 times (15'285bp) the size of the original human gene (-7'320bp), and the blast result from this experiment has yielded an ortholog size half of that of the original human gene's (3'726bp). This is likely due to the specifics of ortholog curation on Ensembl, and a sign that our blast ortholog search needs adjustment of its specificity in order to recognize longer orthologs as on Ensembl.

Proposals for continued studies

Because the sample size was small – 4 chimpanzees and 4 gorillas; the mean copy numbers obtained for each gene might not be representative for the species as a whole. Future studies should include more individual genomes to have a better overview of the copy number variations. Also, this would allow us to perform t-tests: to measure accurately to what extent genes have different copy number distributions between species.

The Y chromosome in placental mammals has palindromic repeats where non-allelic homologous recombination occurs [12]. Some of the genes that have been screened for ampliconic behavior in this experiment are residing in these palindromes. The limit for non-ampliconic behavior on the Y chromosome might be set at a higher threshold; i.e. two times that of the X-linked genes. My argument being, that copies in these palindromes do not indicate ampliconic behavior but are simply being kept similar because of non-allelic homologous recombination activity.

It would have been fitting to use more than a single control gene for each artificial chromosome. The controls, DMD and AMELY for the X and Y chromosomes, respectively, were chosen as controls because they are known to be non-ampliconic in human. This, together with the fact that they turn out to have a low coverage in the results/application, doesn't necessarily validate that they are non-ampliconic in chimpanzee and gorilla. By adding more control genes in future studies, it can be validated with a larger margin, that they indeed are non-ampliconic.

Genome reads from paralogs on other chromosomes might have been erroneously mapped to the artificial chromosomes. A way to combat this problem is to include the rest of the complete genome on the artificial chromosomes, such that the reads would map here instead of on the genes in the artificial chromosomes. A different more computationally friendly method would be to simply try and map the reads with lower mapping quality to the rest of the genome, and see what proportion of possible false positives would map elsewhere. As the genomic variation is differing between the hominid species, filtering parameters should be adjusted individually for each species.

References

[1] S. Ohno, *Sex Chromosomes and Sex-linked Genes*. Springer-Verlag, 1967.

[2] J. Y. Dutheil, K. Munch, K. Nam, T. Mailund, and M. H. Schierup, "Strong Selective Sweeps on the X Chromosome in the Human-Chimpanzee Ancestor Explain Its Low Divergence," *PLOS Genet.*, vol. 11, no. 8, pp. 1–18, 2015.

[3] K. Nam *et al.*, "Extreme selective sweeps independently targeted the X chromosomes of the great apes," *Proc. Natl. Acad. Sci.*, vol. 112, no. 20, pp. 6413–6418, 2015.

[4] J. Cocquet, P. J. I. Ellis, S. K. Mahadevaiah, N. A. Affara, D. Vaiman, and P. S. Burgoyne,

"A Genetic Basis for a Postmeiotic X Versus Y Chromosome Intragenomic Conflict in the Mouse," *PLOS Genet.*, vol. 8, no. 9, pp. 1–15, 2012.

[5] A. P. Møller, "Ejaculate quality, testes size and sperm competition in primates," *J. Hum. Evol.*, vol. 17, no. 5, pp. 479–488, 1988.

[6] E. A. Lucotte, L. Skov, J. M. Jensen, M. Coll Macià, K. Munch, and M. H. Schierup, "Dynamic Copy Number Evolution of X- and Y-Linked Ampliconic Genes in Human Populations," *Genetics*, 2018.

[7] H. Li and R. Durbin, "Fast and accurate short read alignment with Burrows–Wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–1760, 2009.

[8] A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins, "Sambamba: fast processing of NGS alignment formats," *Bioinformatics*, vol. 31, no. 12, pp. 2032–2034, 2015.

[9] H. Li *et al.*, "The Sequence Alignment/Map Format and SAMtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[10] L. Fagerberg *et al.*, "Analysis of the Human Tissue-specific Expression by Genome-wide Integration of Transcriptomics and Antibody-based Proteomics," *Mol. Cell. Proteomics*, vol. 13, no. 2, pp. 397–406, 2014.

[11] S. B. Hedges, J. Marin, M. Suleski, M. Paymer, and S. Kumar, "Tree of life reveals clock-like speciation and diversification," *Mol. Biol. Evol.*, vol. 32, no. 4, pp. 835–845, 2015.

[12] L. Skov, T. D. P. G. Consortium, and M. H. Schierup, "Analysis of 62 hybrid assembled human Y chromosomes exposes rapid structural changes and high rates of gene conversion," *PLOS Genet.*, vol. 13, no. 8, pp. 1–20, 2017.

Material and data

All code and resources except genome data are available at kortlink.dk/kobel-gitlab-X/u7pc