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

Supervisors: 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 indicate that they were amplified in the human lineage. We find that GAGE4 completely lacks ampliconic behavior in chimpanzee and gorilla. ??Either because there is a homolog or because it isn't ampliconic?

Introduction

??More technical, more metrics, examples/refs

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 which is defined as the Testis Determining Factor 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 are 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, assuming a balanced sex ratio. Another consequence is that the sex chromosomes are subjected to higher drift, especially for the Y chromosome [2].

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 in meiotic drive processes [2], [3]. Meiotic drive favors the segregation of specific genes, thus disturbing the mendelian segregation ratios. (reference??). 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 to avoid mechanisms that disturbs the segregation of sex-chromosomes. This inactivation is often disrupted in hybrids, at least in round spermatids, and evidence suggests that it is caused by unbalanced copy number of *Sly* and *Slx*.

Intragenomic conflicts lead to speciation because the arms-race process triggers a rapid differentiation of the sex-chromosomes, which may become sufficiently divergent between sub-populations to induce 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 yields higher selection on increasing testes size in chimpanzee compared to Gorilla [4]. In humans (…)

{Most of the ampliconic genes are expressed in testis, therefore, sperm competition may have influenced the evolution of these genes in different way in 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.}elise??

Method

The method used in this study consists of assembling an artificial chromosome (AC) consisting of one copy of each ampliconic candidate genes, 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.

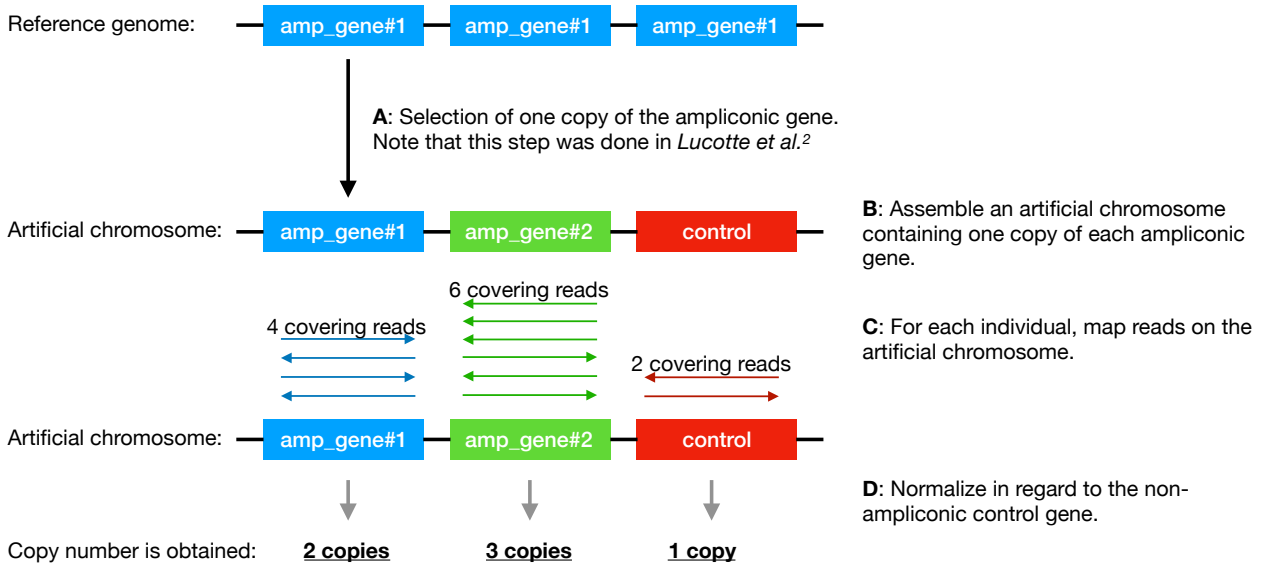


Figure 1: Overview of the method.

Assembly of artificial chromosomes

In order to assemble the ACs, we adapted the method from Lucotte et al. 2018 [5] 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 genes, we searched for orthologs in the chimpanzee and gorillas genome (Table 1). 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 version ?? to align the human genes against the chimpanzee and gorilla references genome. 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.

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 it completely. Thus, we constructed ACs for chimpanzee X, chimpanzee Y and gorilla X. These were then used in the read-mapping method presented in the next part. Several problems occurred in the assembly of the ACs. We checked coherence between Ensembl and BLAST results with GAGE4 in gorilla. The Query %id and BLAST Identities are on par (Table 1), but the length fraction was off, as Ensembl says that the ortholog is 2 times the size of the original human gene, and my blast result says that the ortholog is half the size of the original human gene.

Mapping reads onto artificial chromosomes

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	-	-

For each individual:

- The reads from the fastq-files were mapped against the ACs using BWA [6] v0.7.5a
- The alignment was filtered using sambamba [7] 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 [5]
- The read depth for each position of the AC was calculated using SAMtools [8] 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 normalize the coverage of each gene, we divided it by the coverage of the controls. 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, known to be single copy in humans.

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. We also performed the method with the human ACs built in Lucotte et al. 2018 [5]. 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 method would lead to a lower number of reads mapping to the chromosome, because of our filtering, and therefore to an under-estimation 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 applying this method in 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, though dosage compensation might counteract doubling. Generally, females show a higher coverage (Figure 2 and Figure 3)

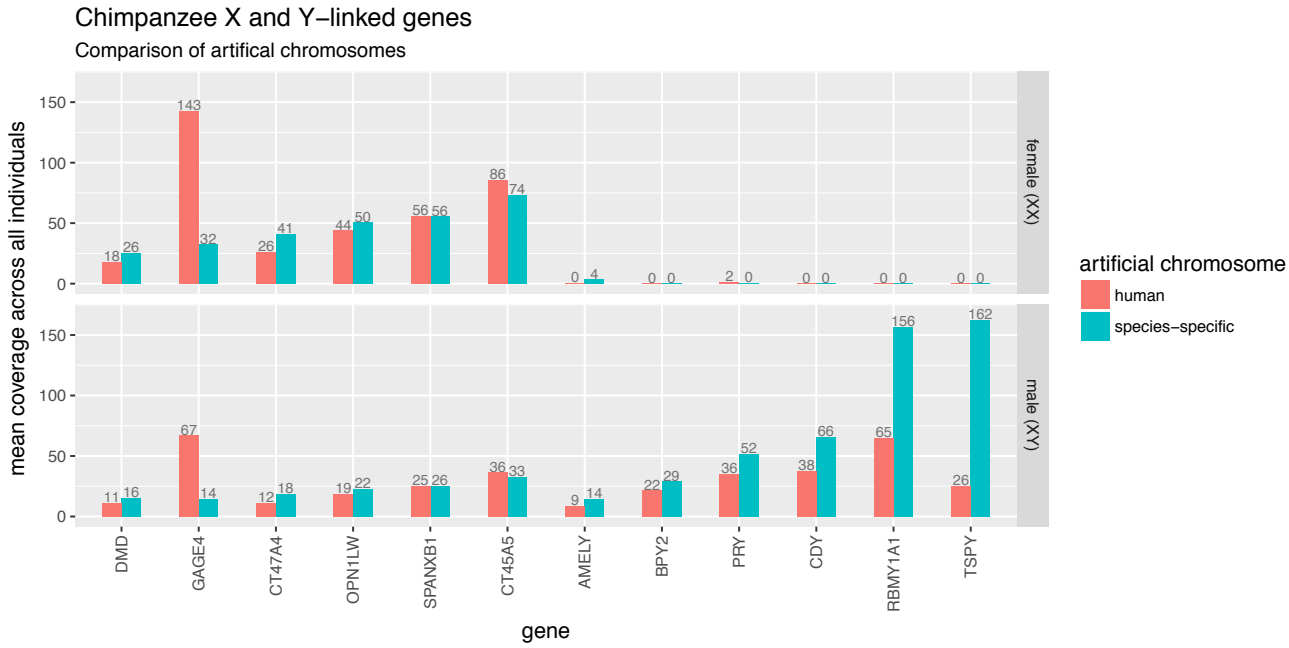


Figure 2: Comparison of the coverage for each gene, for the human AC or the species-specific ACs for chimpanzees. The mean is calculated by averaging the coverage over all the individuals.

For chimpanzee, there is a varying difference in the sensitivity of the ACs between the genes. For most X-linked genes, the difference is limited except for GAGE4 where the human AC has many fold higher coverage than the species specific. There is a negligible difference between female and male. The Y-linked genes show a higher coverage on the species-specific AC in general. The Y-linked genes show the highest individual-pair relative difference as well (Table 2). The fact that the sensitivity of the artificial chromosome is bigger for the Y chromosome, suggests that the Y chromosomes might be more divergent between species than the X is. Very limited copy numbers on the Y chromosome show that the method is not too sensitive. The small copy numbers for females in species-specific:AMELY and human AC:PRY might be because these two genes have

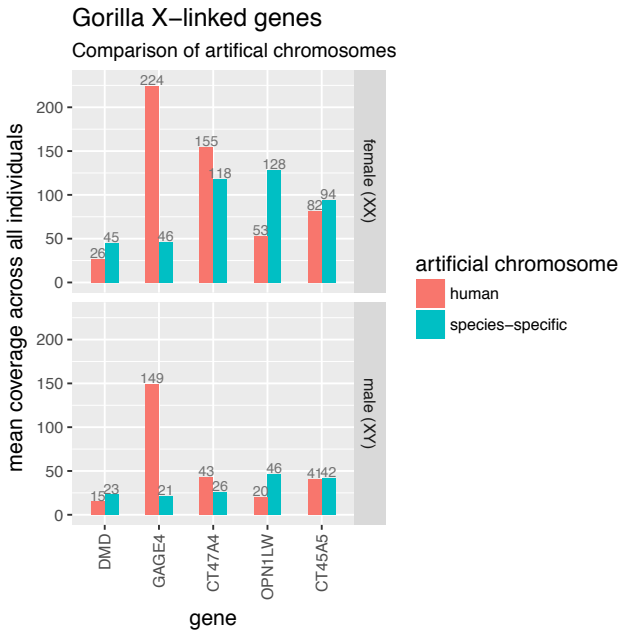
homologs on the X chromosome; AMELX AND PRX. ??It should be noted, that in human, many of these genes show presence in females as well.??

Table 2: The mean individual-pair relative difference is calculated by taking one individual at a time, taking the difference of species-specific and human AC coverage. Then taking the mean of a group (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 one individuals gene at a time.

For gorilla the mean difference between the species-specific and the 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 for chimpanzee.

Because the mean difference between the species-specific AC and the human AC coverage is positive and because species-specific ACs are less subjected to evolutionary turnover across genes – the following results are based on the method using the species-specific ACs. ?? hvad mener du med at være concise?

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.



X chromosome

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

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

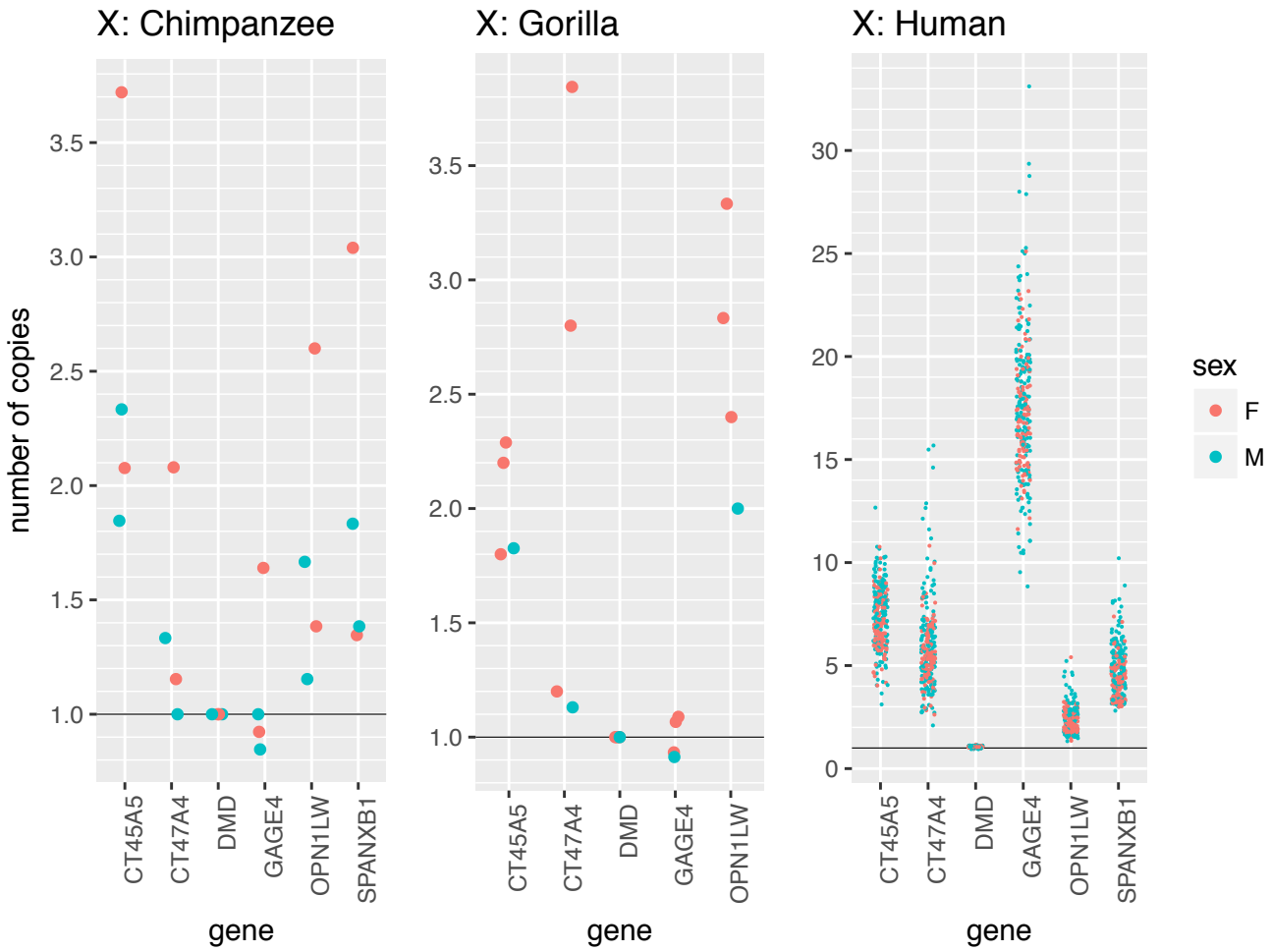


Figure 4: Copy number (normalized coverage) of X-linked genes. All individuals for each species. Note the differently scaled y-axes across species. Horizontal jitter applied.

Y chromosome

As females don't have Y chromosomes, there is theoretically no need to survey the copy numbers of theirs. However, the copy numbers are included in order to validate the sensitivity of the method. For unclear reasons, all female genes show copy numbers above 1. (?? Hi Elise, did you comment on this in your manuscript?)

Note that 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 using the gene AMELY. This may have an influence in the comparison of the results.

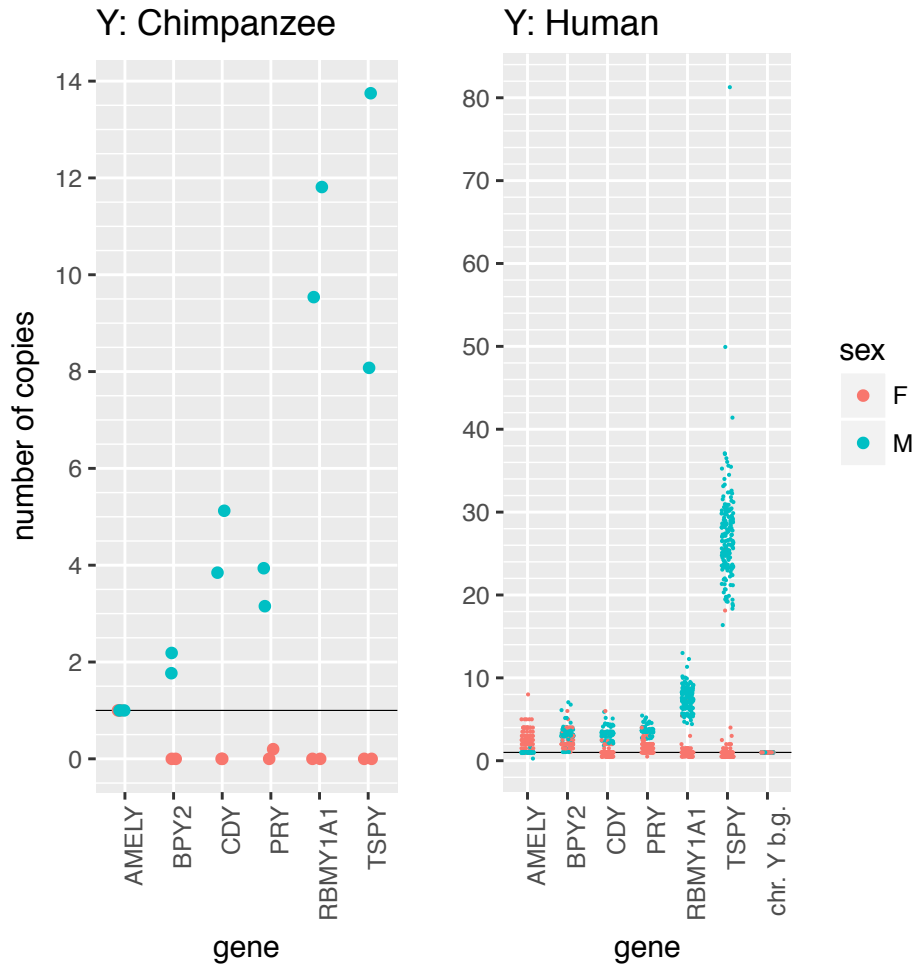


Figure 5: Copy number (normalized coverage) of Y-linked genes. All individuals for each species. Note the differently scaled y-axes across species. Horizontal jitter applied. ???fjern Female, og plot TSPY seperat

BPY2 has a copy number close to 2 and in humans, BPY2 resides in two palindromes, and the reference genome has 3 copies. In human, the copy number for BPY2 is below 1 for 9 out of 174 males.

??Hi Elise, do you think I should put a table with these results as well? I feel like it may be too much, especially since I can't do any meaningful significance tests?

TSPY seems to be ampliconic in chimpanzee as well, with 8 and 14 copies for each males, but in humans the main distribution is between 20 and 35 copies. The other genes have a slightly lower copy number than in human. ??(I would detail that a bit more, there are not that many genes left CDY, PRY and RBMY1A1)

Species	Chrom.	Sex	Gene	Min	Median	Max	SD
chimpanzee	X	F	DMD	1.00	1.00	1.00	0.00
chimpanzee	X	F	GAGE4	0.92	1.28	1.64	0.51
chimpanzee	X	F	CT47A4	1.15	1.62	2.08	0.65
chimpanzee	X	F	OPN1LW	1.38	1.99	2.60	0.86
chimpanzee	X	F	SPANXB1	1.35	2.19	3.04	1.20
chimpanzee	X	F	CT45A5	2.08	2.90	3.72	1.16
chimpanzee	X	M	GAGE4	0.85	0.92	1.00	0.11
chimpanzee	X	M	DMD	1.00	1.00	1.00	0.00
chimpanzee	X	M	CT47A4	1.00	1.17	1.33	0.24

chimpanzee	X	M	OPN1LW	1.15	1.41	1.67	0.36
chimpanzee	X	M	SPANXB1	1.38	1.61	1.83	0.32
chimpanzee	X	M	CT45A5	1.85	2.09	2.33	0.34
chimpanzee	Y	F	BPY2	0.00	0.00	0.00	0.00
chimpanzee	Y	F	CDY	0.00	0.00	0.00	0.00
chimpanzee	Y	F	RBMY1A1	0.00	0.00	0.00	0.00
chimpanzee	Y	F	TSPY	0.00	0.00	0.00	0.00
chimpanzee	Y	F	PRY	0.00	0.10	0.20	0.14
chimpanzee	Y	F	AMELY	1.00	1.00	1.00	0.00
chimpanzee	Y	M	AMELY	1.00	1.00	1.00	0.00
chimpanzee	Y	M	BPY2	1.77	1.98	2.19	0.30
chimpanzee	Y	M	PRY	3.15	3.55	3.94	0.55
chimpanzee	Y	M	CDY	3.85	4.49	5.12	0.90
chimpanzee	Y	M	RBMY1A1	9.54	10.68	11.81	1.61
chimpanzee	Y	M	TSPY	8.08	10.91	13.75	4.01
gorilla	X	F	DMD	1.00	1.00	1.00	0.00
gorilla	X	F	GAGE4	0.93	1.07	1.09	0.08
gorilla	X	F	CT45A5	1.80	2.20	2.29	0.26
gorilla	X	F	CT47A4	1.20	2.80	3.84	1.33
gorilla	X	F	OPN1LW	2.40	2.83	3.33	0.47
gorilla	X	M	GAGE4	0.91	0.91	0.91	-
gorilla	X	M	DMD	1.00	1.00	1.00	-
gorilla	X	M	CT47A4	1.13	1.13	1.13	-
gorilla	X	M	CT45A5	1.83	1.83	1.83	-
gorilla	X	M	OPN1LW	2.00	2.00	2.00	-

Discussion

Conclusion

??Here, I suggest this kind of structure??

In this study, we compared the copy number of human ampliconic genes with gorillas and chimpanzees. (??take from the introduction)

??For the X chromosome and for both species, gene X, X, X have a lower copy number than humans while gene X,X,X have a similar copy number compared to humans. In chimpanzee … (exception) and in gorilla… exception.

For the Y chromosome, PRY, CDY and RBMY1A1 have copy numbers close to the human median. TSPY, ??…,…,… have a lower copy number than that of humans

GAGE4 seems to be absent in chimpanzee and gorilla. This suggests that the ampliconic behavior of GAGE4 in human emerged after the split of the human-chimpanzee ancestor.

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 the best.

An interesting case in OPN1LW, which is the gene coding for optin ??lalalala.

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. Correlation with sperm competition (??to relate to your introduction).

Criticism of the particular execution

??delThe methodology of this experiment was the wrong way around.

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 [5]. After this list is created, it might be interesting to see if any of the genes are orthologs (??homologs) between the sister species.

??delWhat role does X-inactivation play?

In the results it was 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 [5] 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 be therefore to include the respective species X chromosomes on the ACs assembled in this experiment. However, the coverage on 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 (??Write something about this in the intro?) of the X chromosome is not included on the artificial Y chromosome here, reads containing AMELX sequences might have ended on the AMELY gene on the AC.

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 to perform t-tests between species to measure accurately of some genes have different copy number distributions.

The Y chromosome in placental mammals has palindromic repeats where non-allelic homologous recombination occurs [9]. 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 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. In future studies it might be interesting to look more into this matter.

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 validate that they are necessarily non-ampliconic in chimpanzee and gorilla. By adding more control genes in future studies, it can be validated with a larger margin, that they are indeed; non-ampliconic.

??Describe having a more unified statistic to compare orthologs from Ensembl and local Blast-alignments.

Reference

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

[5] 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.

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

[7] 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.

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

[9] 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.

Supplemental material and data

All code and resources except genome data is available at: http://kortlink.dk/github-kobel-X/u6y9