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

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

Supervisors: Elise Lucotte and Mikkel Heide Schierup

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

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 and gorilla individuals onto these species specific ACs, we measured the copy number variation and argue that GAGE4 – contrary to being ampliconic in human; might be non-ampliconic in chimpanzee and gorilla.

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 then 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 then 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 in general, because presence is lower; 3/4 X chromosomes and 1/4 Y chromosomes assuming equal sex ratio. The upshot is that there will be higher drift, especially on 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. Meiotic drive favors the segregation of specific genes, thus disturbing the mendelian segregation ratios. (reference??). In mice, there is an arms race between the pair of homologous ampliconic genes, SLY and SLX, residing on each of the sex chromosomes. Evidence suggests that these genes are involved in an intragenomic conflict where they compete to become transmitted to the next generation.

This mechanism leads to deleterious X-XY dosage disruption in hybrids. A deficiency of Slx distorts the sex ratio to have higher frequency of males. In the long run, this will be corrected with Sly deficiency.

Sex chromosome inactivation is crucial to avoid mechanisms that disturbs the segregation of sex-chromosomes during meiosis. This inactivation is often mis regulated, at least in round spermatids, depending on the copy number

This conflict, and intragenomic conflicts as whole leads to speciation, because the sex-chromosomes may become different between sub-populations.

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

Sperm quality is different among the hominids. There are differing levels of sperm competition among the hominids. Gorilla and orangutan males monopolizes copulation with females where chimpanzees instead have a multi-male breding system, where many males copulate with each female. This differential behavior yields higher selection on increasing testes size in chimpanzee compared to the sexual-partner-monopolizing species [3].

Method

The method used in this experiment consists of assembling an AC consisting of ampliconic candidate genes. Subsequently mapping reads from an individual onto this AC. Thereby measuring the copy number by relating the filtered coverage of maps to a control which is assumed to non-ampliconic.

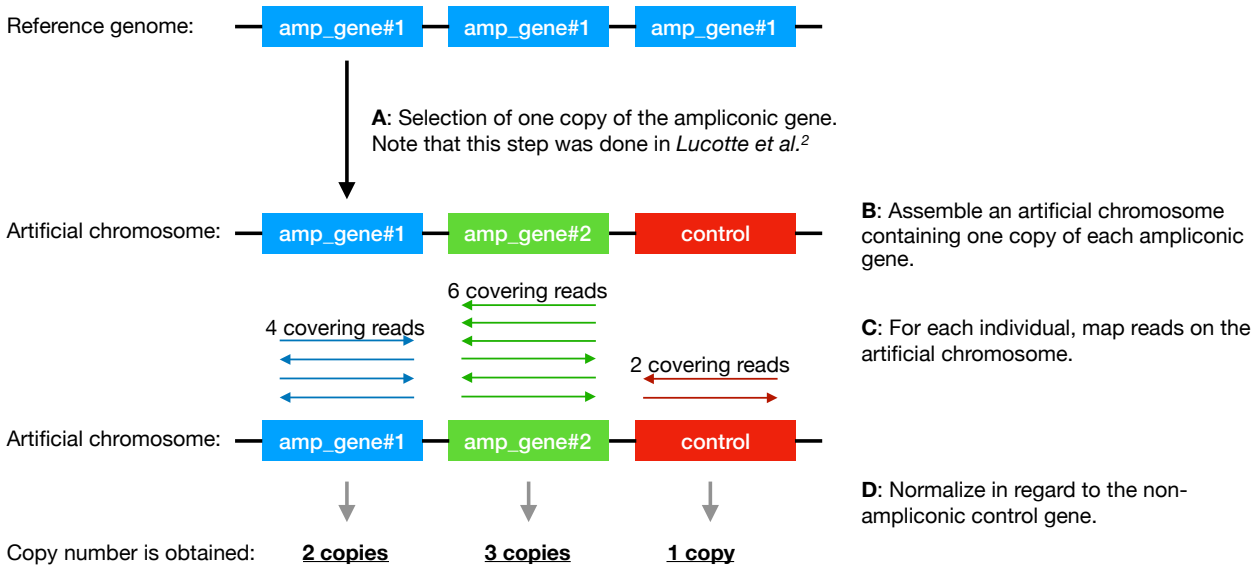


Figure 1: A complete overview of the method.

Assembly of artificial chromosomes

In order to assemble the ACs, we adapted the method from Lucotte et al. 2018 [4] 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 in chimpanzee and gorilla (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. 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 except one (XKRY) we decided to omit it completely. Thus, we ended up with the 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

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 following was done:

- The reads from the fastq-files were mapped against the ACs using BWA [5] v0.7.5a
- The alignment was filtered using sambamba [6] 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 [4]
- The read depth for each position of the AC was calculated using SAMtools [7] 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 number is what we subsequently regard as the un-normalized copy number. In order normalize the coverage of each gene, we divide it by the coverage of the controls. In order to estimate the copy number of each ampliconic gene, we divide the median coverage of each gene by the median coverage of the control gene, known to be single copy.

We selected DMD as the control gene for the X chromosome and AMELY for the Y chromosome, as they are both single copy. We also executed the method with the human

ACs built in Lucotte et al. 2018 [4]. 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.

Results

Comparison: human vs. species-specific artificial chromosomes

In order to validate the ACs assembled for chimpanzee and gorilla, respectively, we would expect to see that the species-specific artificial chromosomes have yielded a higher coverage in general, compared to that of the human AC. For all species and sex, the mean individual-pair relative difference in coverage (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 much higher copy number (Figure 2 and Figure 3)

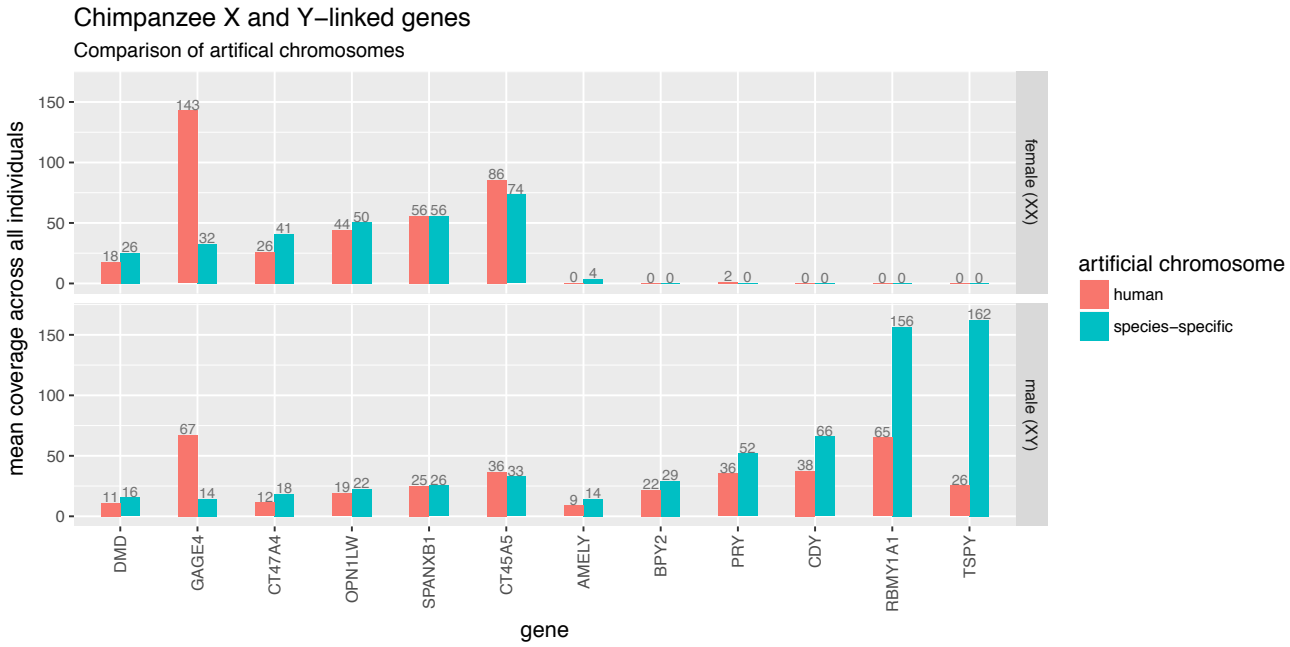


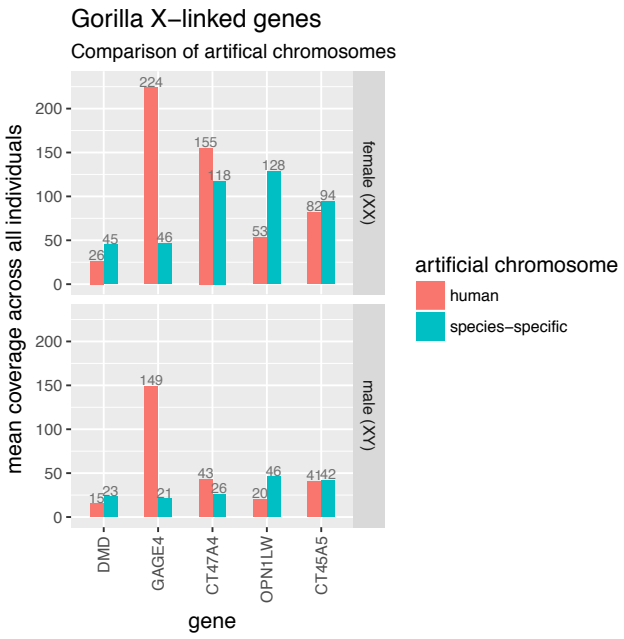
Figure 2: Comparison of using either the human AC (Lucotte et al. 2018) or the ACs assembled in this internship, to map the chimpanzee individuals onto. The mean is calculated by averaging the coverage of all the individuals

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 relative 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.

species	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 using either the human AC (Lucotte et al. 2018) or the ACs assembled in this internship, to map the chimpanzee individuals onto. The mean is calculated by averaging the coverage of all the individuals



For gorilla the mean individual-pair relative difference is positive, indicating that the species-specific AC has higher sensitivity than the human AC (Table 2). Though, 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 does for chimpanzee.

Because the individual-pair relative difference (species-specific over human) is positive and because species specific ACs are less sensitive to different evolutionary turnover across genes – the forthcoming results analysis is based on the data stemming from executing the method with the species-specific ACs.

X chromosome

For the X chromosome (Figure 4), most genes showed a lower copy number in chimpanzee and gorilla, than in human. GAGE4 is the only gene that looks to be non-ampliconic, this in both species. The next candidates to be non-ampliconic is CT47A4 (both species) and OPN1LW in chimpanzee. None of the other genes in chimpanzee and gorilla show as high copy numbers as in human.

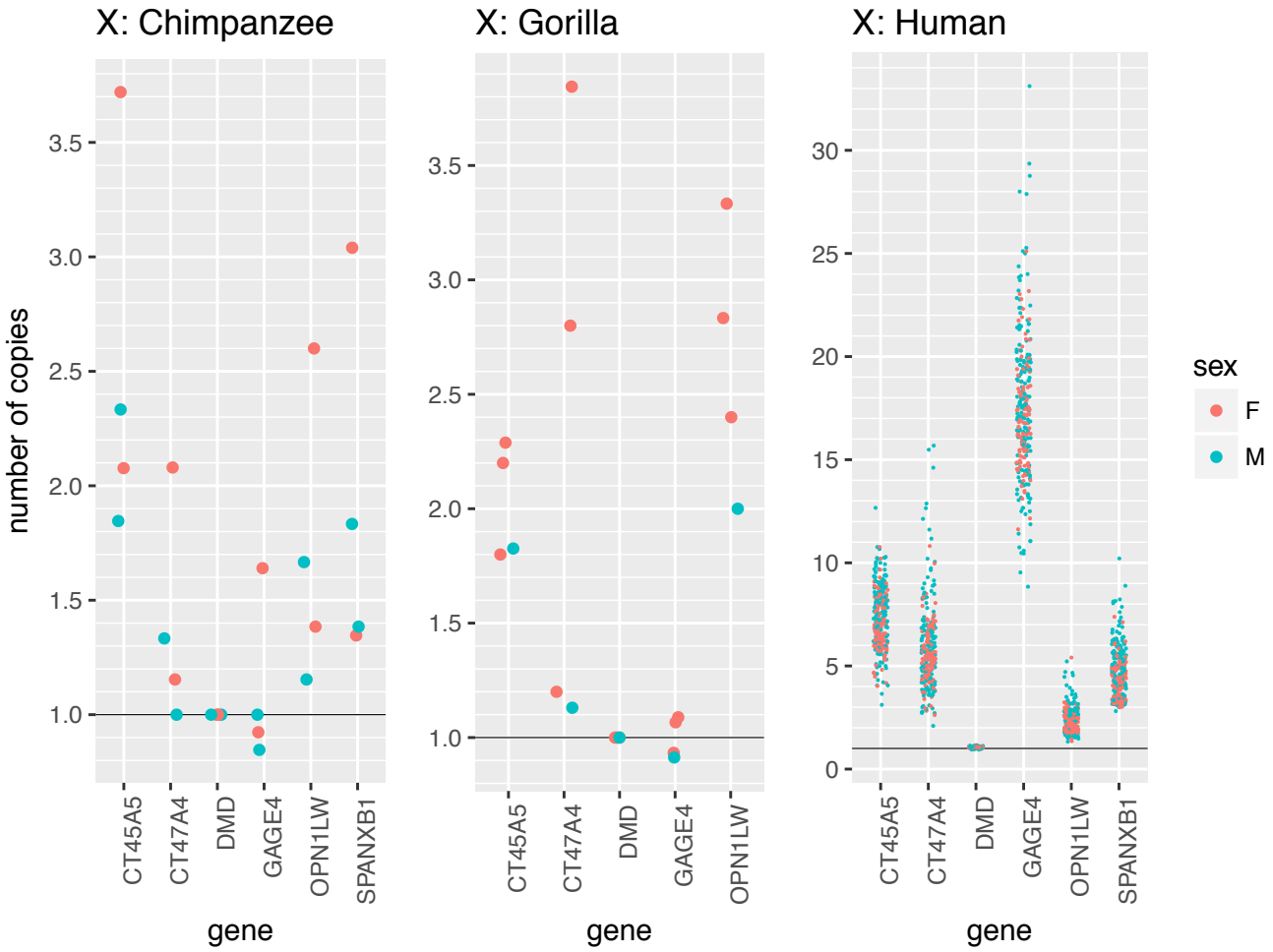


Figure 4: Copy number (normalized coverage) of X-linked genes. All individuals for each species. Note the differently scaled y-axes across species. Horizontal jitter applied.

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 in regard to chr. Y. b.g. which is a part of the Y chromosome without coding genes. (??I have a strong feeling that I'm just making something up here). Whereas, in chimpanzee, the copy numbers were normalized in regard to AMELY. This may complicate the comparison of the results.

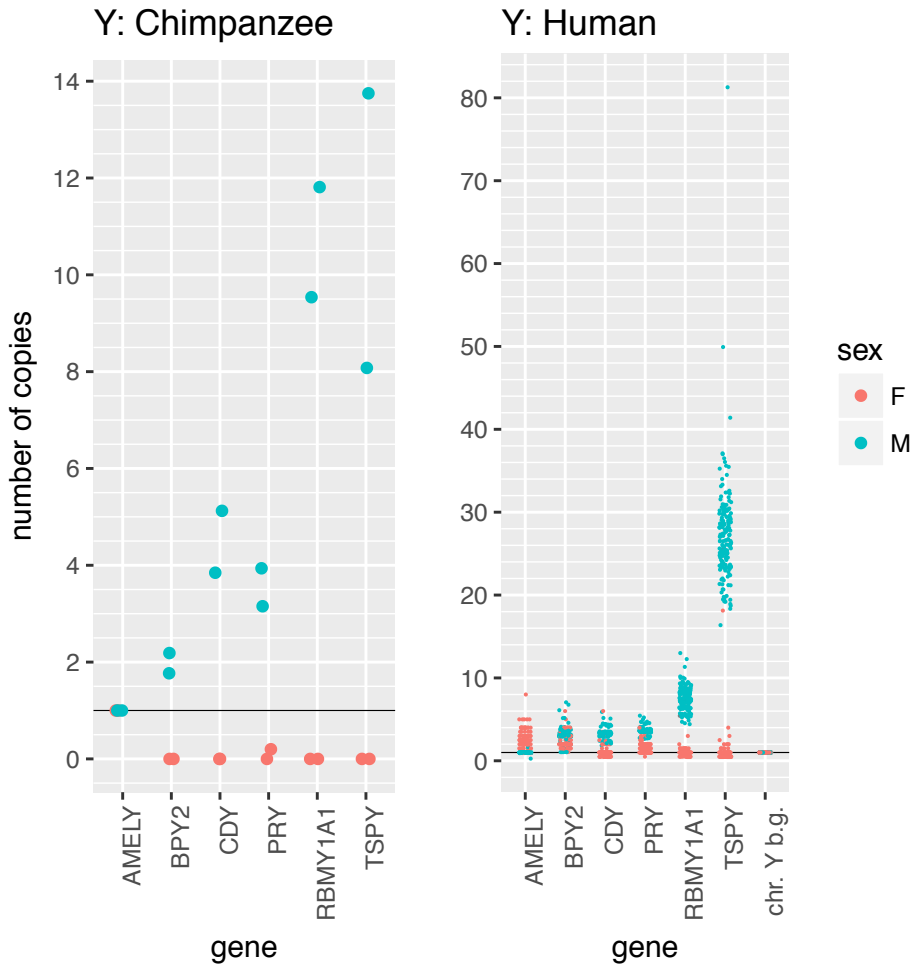


Figure 5: Copy number (normalized coverage) of Y-linked genes. All individuals for each species. Note the differently scaled y-axes across species. Horizontal jitter applied.

BPY2 has a copy number close to 2. If we assume that BPY2 resides in a palindrome, we can argue that it might be non-ampliconic (??). In human, the copy number for BPY2 is for 9/174 male individuals below 1. This makes the information about low copy numbers for chimpanzee less significant because looks like the method is picking up noise (??).

??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 looks to be in the same range, and the rest of the genes look like they have a lower copy number than in human, but not anything significant.

Discussion

Conclusion

If we assume that the method is valid, we can make the conclusion that GAGE4 is non-ampliconic in chimpanzee and gorilla. This suggests that the ampliconic behavior of GAGE4 in human emerged after the split of the human-chimpanzee ancestor. ??insert divergence times and ref. Gorilla:OPN1LW and Chimpanzee:TSPY, PRY, CDY have copy numbers close to the human median. Most other genes show a copy number smaller than that of human.

Criticism of the particular execution

The methodology of this experiment was the wrong way around. We looked at genes that were known to be ampliconic in human. There might be many genes that are ampliconic in chimpanzee and gorilla, which are not ampliconic in human. In order to make a comprehensible overview 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 [4]. After this list is created, it might be interesting to see if any of the genes are orthologs between the sister species.

??What role does X-inactivation play?

In the results it was mentioned that the copy numbers for the Y chromosome genes had copy numbers well above zero, which shouldn't be possible: because females don't have Y chromosomes. In the human data from Lucotte et al. 2018 [4] 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. If the respective species X chromosomes had been included as decoys on the ACs assembled in this experiment, the results might have been closer to reality.

AMELY might be a bad choice of non-ampliconic gene to normalize in regard to. As it has a homolog, AMELX, with high similarity. When a decoy of the X chromosome is not included on the artificial Y chromosome, 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 whole.

Future studies should include many more individual genomes, then t-tests between species could be used to measure accurately of some genes are different or not.

The Y chromosome in placental mammals has palindromic repeats where non-allelic homologous recombination occurs [8]. Some of the genes that have been screened for ampliconic behavior in this experiment might be residing in these palindromes. If so, their copy number should maybe have been interpreted differently. 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 fact, and the fact that they turn out to have a low coverage in the method, 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] A. P. Møller, "Ejaculate quality, testes size and sperm competition in primates," *J. Hum. Evol.*, vol. 17, no. 5, pp. 479–488, 1988.

[4] 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.

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

[6] 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.

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

[8] 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.