# Bachelor Project: Copy Number Variation of Ampliconic Regions on Hominid X and Y chromosomes

Carl Mathias Kobel*

*BiRC, AU*

E-mail: kobel@pm.me

**Abstract**

By mapping the genomes of a few Chimpanzee and Gorilla individuals, we measured the copy number variation of a few genes that were assumed to be ampliconic based on human trials.

## Introduction

### Sex determining systems

It is believed that the sex determining system from which the modern heterogametic male/homogametic female (XX/XY) arose, was the environmental sex determining system. The modern chromosome based sex determining system is believed to have arisen form the environmental sex determining system still seen in crocodiles and turtles. According to Ohno (??ref), a disruptive variant may have evolved on one of the chromosomes in a species with environmental sex determinaiton. This variant may have been the start of the divergence between

the sex chromosomes. This model is assumed to have been the background of the parallel evolution of both the ZW and the XY sex determining systems. In mammals (XY sex determining system) the male defining gene (testis determining factor) is the SRY gene. Since it resides on the Y chromosome, the Y chromosome is indeed the sex determining chromosome.

This system was disrupted by the evolution of the SRY gene (but then how did the Y chromosome arise??). Today the XX/XY system is present in mammals, and the SRY gene (present on the Y chromosome) is considered as the testis determining factor. No idea how ampliconic genes evolved. How did they diverge between human, chimp and gorilla? Did the ampliconic genes become amplified in the human lineage or in the great apes? What drives ampliconic gene evolution?

...We want to count copy number of these ampliconic genes.

# Method

In order to The pipeline (Figure 1) of this project consists of several parts, all originating from *Lucotte et al.*[1]. Firstly, an artificial chromosome with ampliconic candidates was assembled for each species, for each of their sex chromosomes. Later, reads of the genomes were mapped upon the artificial chromosome, and the copy number was estimated with a non-ampliconic control.

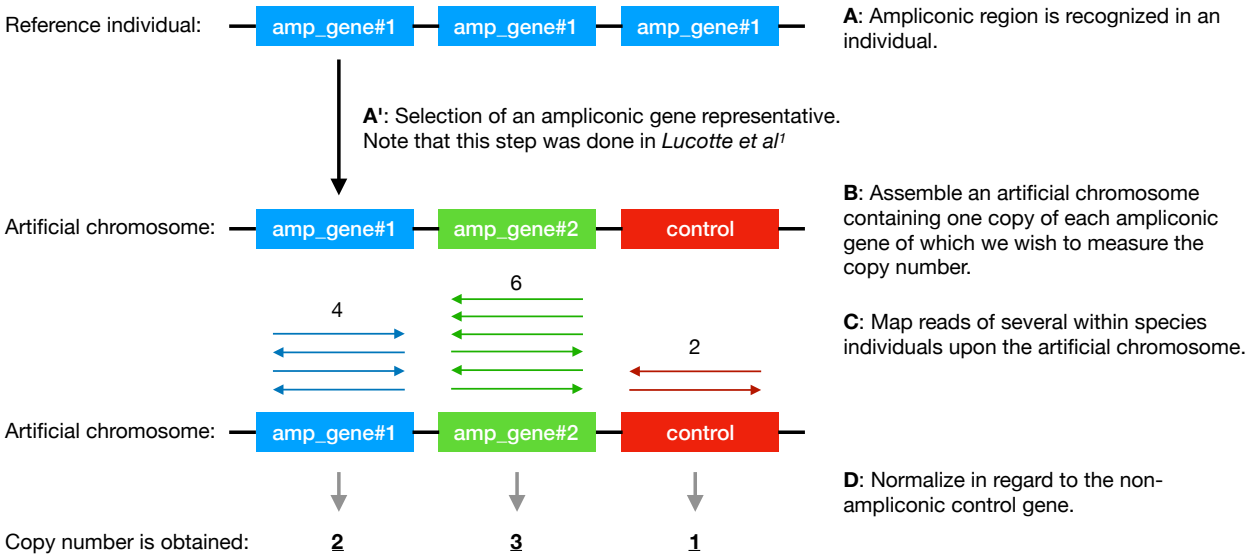



Figure 1: A complete overview of the pipeline.

## Assembly of artificial chromosome

For assembly of the artificial chromosomes (artificial chromosomes) we selected the human genes showing ampliconic behaviour from *Lucotte et al.*[1]. These genes are selected visually from dotplots. We searched for orthologs in chimpanzee (reference: Pan_tro_3.0) and gorilla (reference: gorGor4) for the associated human genes in the Ensembl genome browser. Only orthologs with a subject/query identical factor of more than 0.5 were included (see Table 1). When orthologs were not found in Ensembl, we blasted the human genes against the chimpanzee and gorilla references `blastn -subject <reference>.fa`

`-query <human_gene>.fa -out output.txt` and again only selected regions with a subject/query identical factor of more than 0.5. The orthologs found on ensembl were downloaded directly as fasta files. The orthologs found with blast were cut out of the reference with `samtools faidx <reference>.fa <chromosome>:<start>-<end>`. The files were later merged into complete artificial chromosomes with `cat *.fa > <artificial_chromosome>.fa`. These artificial chromosomes (chimpanzee X, chimpanzee Y and gorilla X) were then put into the read-mapping pipeline presented in the next part. Several problems occured in the assembly of the artificial chromosomes. We checked coherence between Ensembl and BLAST results with GAGE4 in gorilla. The Query %id and BLAST Identities are on par, but the length fraction is far off. This may be due to the way the human ampliconic regions were chosen. The human ampliconic regions from *Lucotte et al.*[1] has a length of 7330bp whereas the corresponding gene on Ensembl has a length of 15285bp. The remedy could be to blast the human genes from Ensembl against the chimpanzee and gorilla references.

## Mapping of artificial chromosomes upon genomes

The following part of the pipeline was written into a workflow with gwf-org[3] in order to support parametric inputs, easing reproducibility. The original code from *Lucotte et al.*[1] was rewritten from gwf v0.7 to v1.2.1 and python v2.7 to 3.6 in order to use newer functionality in the packages. In this process, much of the code was also reorganized in such a way that all input parameters were concentrated in one spot, such that the pipeline can easily be used for different batches of individuals, which may be needed in order to produce results with a higher statistical power. The takes in a number of individuals (grouped in species) and their genomes, and an artificial chromosome as reference. This is done by passing a dictionary with various elements to the workflow. When all the input parameters are parsed from various text files (??explain), it indexes the genomes and starts mapping each phase of the genome upon the artificial chromosome. Since the individuals might have been

Table 1: Table of orthologs chosen for assembly of the artificial chromosomes for the respective species. Query id% is the percentage of the human sequence matching the sequence of the orthologue. Length fraction denotes the length of the orthologue relative to the original (human) sequence. Identities denotes the fraction of similar nucleotides in the pairwise alignment. Orthologue candidate genes with a Query id% < 0.5 or length fraction < 0.5 were discarded. See the full table including discarded samples at google drive[2]

| **Species** | **Chr.** | **Gene** | **Ensemble 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 | - | - |

sequenced several times independently (??maybe from various tissues - who knows??), and these sequences are all included, we merge the sequence alignments (BAM-files) into one per individual. For quality control, we filter with Sambamba:[4] `(mapping_quality >= 50) and (cigar = /100M/) and ([NM] < 3)`. We then calculate the read depth with SAMtools.[5] Because the distribution is not necessarily symmetric, we take the median of the read depth across all positions in the gene. This median number is what we subsequently regard as the copy number. In order to normalize the copy number of each gene, we divide by the copy number of the controls. In case of the X chromosome, we selected DMD as a control gene, since it is known to be non-ampliconic (in humans). For the Y chromosome AMELY is selected for the same reason. Surely, it would have made a lot of sense to include many more control genes, as there is no reason to believe that DMD and AMELY might not be non-ampliconic, other than the fact that their distributions look very limited (See the discussion section??).

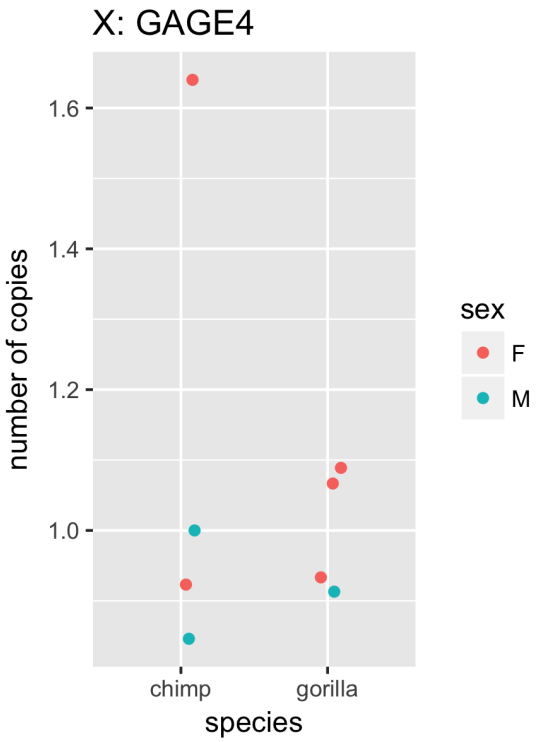
## Visualization

# Results

## Visual overview

Show many plots.

New float types are automatically set up by the class file. The means graphics are included as follows (scheme 1). As illustrated, the float is “here” if possible.




Scheme 1: An example graphics

# Discussion

What can we really conclude? Take basis in examples from the results. Try and make it sound like there is

# References


(1) Lucotte, E. A.; Skov, L.; Coll Macia, M.; Munch, K.; Schierup, M. H. *bioRxiv* **2017**,

(2) full compiled list of orthologs for assembly of AC kortlink.dk/tuyp, *Google Drive*

(3) grid workflow github.com/gwforg/gwf, *GitHub*

(4) Tarasov, A.; Vilella, A. J.; Cuppen, E.; Nijman, I. J.; Prins, P. *Bioinformatics* **2015**, *31*, 2032–2034.

(5) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. *Bioinformatics* **2009**, *25*, 2078–2079.