Copy Number Variation in The Chimpanzee X Chromosome

Project in Bioinformatics, 5 ECTS, Fall 2019 - Carl Mathias Kobel

ABSTRACT

Counting the copy number of genes in an individual can be done by read mapping to a reference, and measuring the coverage of each gene. Normalizing this measure, such that it can be compared across individuals can be done either by scaling with the coverage of a long single-copy gene (i.e. DMD) or by assuming that the copy number distribution for all linked genes is equal to one. In this study, both methods are discussed, and the latter method is used for a search of genes, on the Chimpanzee X chromosome, with high copy number variation. It is observed that a number of previously undescribed genes with high copy number and which show signs of testis-expression, might be part of a meiotic drive process. Anyhow, because expression data on Chimpanzee genes is in many occasions not readily available, it is not possible to immediately conclude that any of these genes are indeed related to meiotic drive.

Introduction

The primate X chromosome has undergone extensive adaptation to become part of the sexual differentiation mechanism. Its presence, and thereby possible displacement of the testis determining factor (SRY) defines the sex of the primate individual. This means that the genes on the X chromosome might interfere with the genes on the Y chromosome in a phenomenon termed meiotic drive. ??ref suggests that there is a link between copy number variation and meiotic drive. It is hypothesized that this tug of war, between the two sex-chromosomes, is played out by incrementing in the copy number of drive-related genes on each chromosome. This means, that if we can assess the copy number (CN) variation of the genes, we can hypothesize that the genes with a highly varying copy number, to be part of a meiotic drive process. Meiotic drive is studied in its relation to hybrid incompatibility ??ref.

Methods and Project Process

In order to measure the CNV of the genes on the Chimpanzee X chromosome, we aligned the reference chromosome X (Pan_Tro_3) to itself and created dotplots in partially overlapping windows of 500 Kbp. These dotplots depict internal duplicates inside and immediately surrounding the sequence in these windows. By manually browsing the catalog of 500 overlapping windows, we decided manually for each of the $\sim$1000 genes in the chromosome if they showed enough internal duplicates to be included in the downstream analysis. By concatenating these selected genes into an artificial chromosome (AC), and then mapping the reads from each individual to this AC, we were able to measure the relative CNV of the genes. The absolute coverage here was normalized using DMD, which is a long single-copy gene. Because the dotplot method takes only adjacent duplicates into account (limited window size), this method makes it possible to identify only the duplicated regions that reside inside and near by genes in the window. As a solution to this problem, we decided instead to compute the CNV of all annotated genes in the chromosome. We did this by mapping the reads from each individual to a well annotated reference chromosome (Pan_tro_3). To get a relative measure of CNV for each gene we again normalized using the DMD gene. Mapping to all X-linked genes might be more computationally intensive than mapping to only a few, but it has the advantage of eliminating selection bias. Another advantage is, that reads which map better in paralogous pseudogenes, outside the selected genes, will not affect the coverage of the annotated genes. We filtered to retain only the reads that satisfy the following conditians: only primary mapped reads, a mapping quality of at least 50, a consecutive mapping of at least 100 bp and an overall maximum nucleotide mismatch of at most 2. Because this project uses publicly available data from different projects, the quality and overall coverage might differ (see table 1). Because some individuals, namely Simliki and Julie-A959, showed large deletions in the DMD gene, we decided to assume that the median of the copy numbers of all genes in the chromosome should equal 1. Based on this assumption, it was now possible to keep Simliki and Julie-A959 in the analysis.

The method is inspired from Lucotte et al. 2019[1], where a similar study performed.

This study uses publicly available Chimpanzee genomes. See table 1 All data was sourced from the EMBL-EBI European Nucleotide Archive.

Results and discussion

Overall distribution

Normalizing the coverage of each gene with the coverage of the DMD gene, renders the overall distribution of copy numbers, across all genes and individuals, as shown in figure 1A.

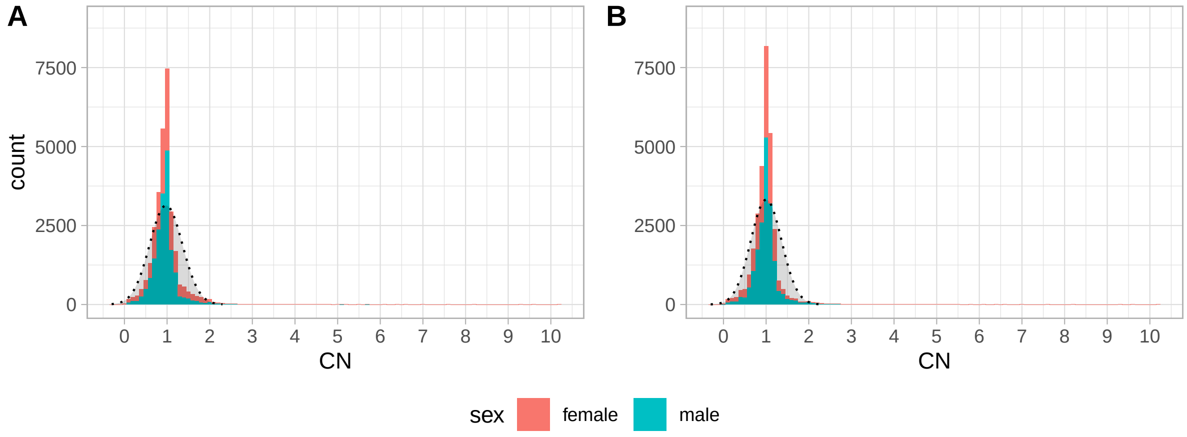

Figure 1. Distribution of median CN across all genes for all subjects. A normal distribution with an equal median, SD and area (grey) is overlapped as a visual aid. Colored by sexes which are stacked. **A**: Normalization is achieved through the coverage of the DMD gene. **B**: Normalization is achieved through shifting the distribution of each individual.

The distribution is near symmetrical, centered around a mean of 0.987. This observation lead us to conclude that the normalization of coverage might be achievable using the chromosome wide distribution CN instead of the DMD coverage. There is no known prior on the chromosome wide CN distribution. But a mean close to one makes good sense, as it suggests that linked paralogs have been described separately in the annotation. It also suggests, that the genes in the annotation have a length which is representative of the genomes used in this study.

We checked that all individuals obey this distribution (figure 2), and observed that most individuals do. But, Simliki shows a chromosome wide CN median of 1.64, which suggests that Simliki's DMD gene has a size of $\frac{100\%}{1.64} = 61\%$ compared to the reference annotation. If using DMD for normalization – failure to exclude Simliki from subsequent analysis might inflate the CN measurement.

Thus, we decided to carry out the downstream analysis, normalizing the coverage with the assumption that the distribution of all linked genes should have a mean equal to 1. The overall distribution is then as shown in figure 1B. This assumption carried forward effectively means that the vertical axis in 2 is negatively proportional with the length of the DMD gene in each individual.

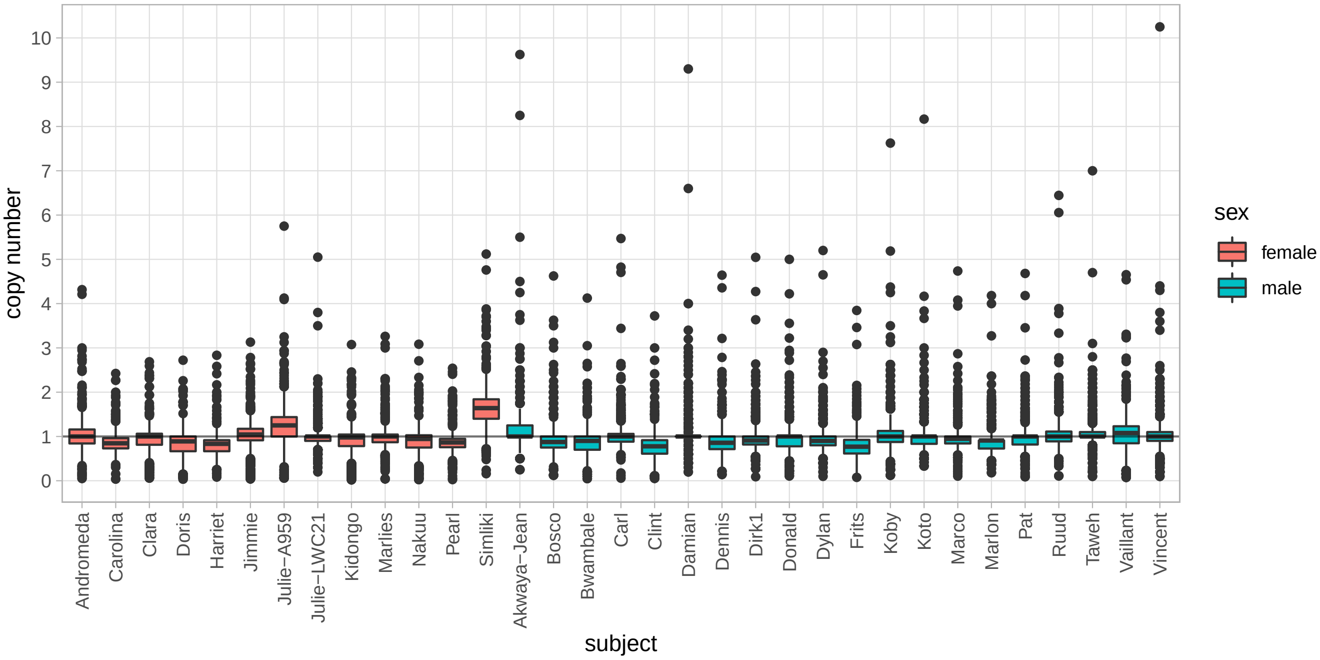

Figure 2. Distribution of median CN across all genes for each subject. Coverage normalization is achieved with the coverage of the DMD gene in each individual. Box edges denote quartiles. Points denote values beyond 1.5*IQR, where IQR is the interquartile range.

Of the 919 genes on the X chromosome, 64 ($\sim$7%) have an overlap with another gene. Performing a non-parametric (permutation) test on the difference (in means) of the distributions of medians for overlapping and non-overlapping genes, shows that the difference is not significantly different (p-value = 0.1927). This means that the copy number variation in the overlap of two genes does not behave differently than a piece of DNA without overlap.

Most copy number-variant genes

Many of the genes with high median copy number show a strong sex-grouping. This might be because the reads from the X degenerate region on the Y chromosome in these males have mapped to its counterpart on the X chromosome. Because the proportion of sexes is not equal amongst subspecies (table 1), we have chosen to not to delve further into this clustering, other than noting that something curious is going on.

We looked up the 13 genes (figure 3) with the highest median copy numbers. Unfortunately, many of the genes in the Pan_tro_3 reference genome are not named as exhaustively as in the human counterparts which means that many of the genes have serial numbers (ensembl gene ids) instead. These will be abbreviated as ENSPTRG00000049971 to E..49971.

The gene with the highest median E..49971 is required for the assembly of the 40S ribosomal subunit[2]. Unfortunately, no expression data is available, neither for orthologs.

E..42923 is a 40S ribosomal protein S2 pseudogene and E..46688 both are very sparsely described[2] and no expression data was possible to find.

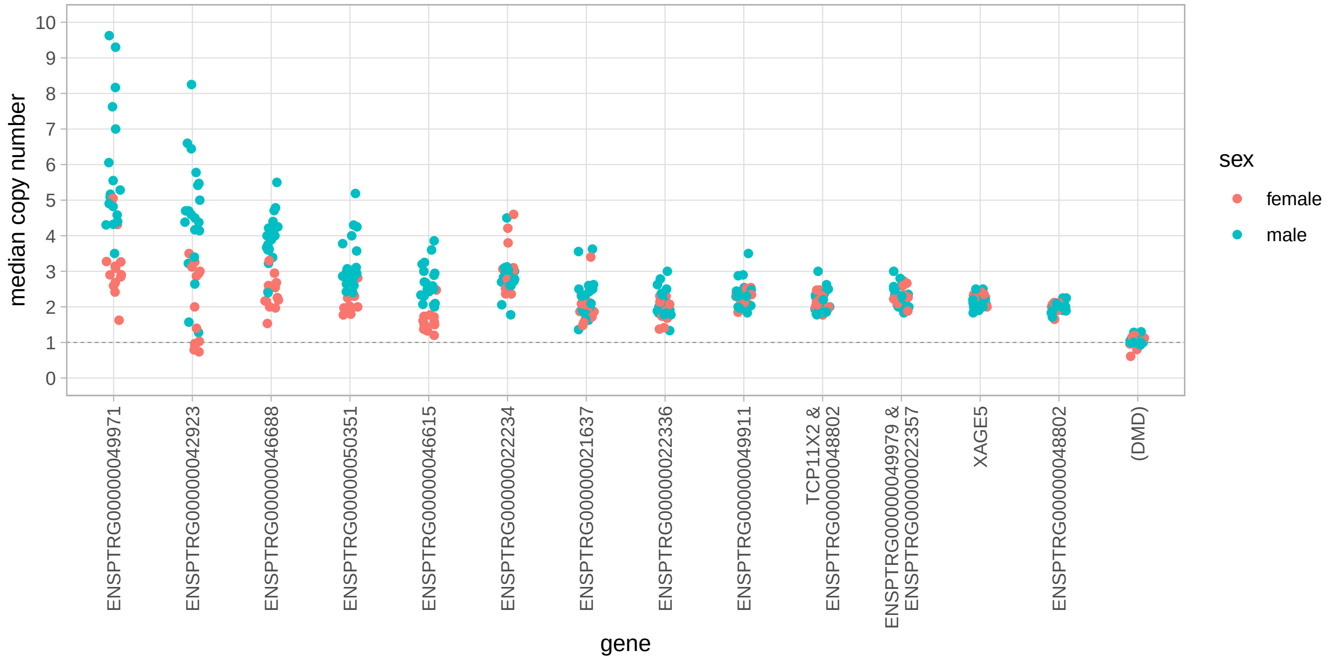

Figure 3. Copy number from all subjects, grouped by genes. These 13 genes or overlaps of genes have a median copy number ≥ 2. Sorted by descending standard deviation. '&' signifies the overlap of two genes. Horizontal jitter is applied. DMD is showed for comparison.

E..50351 has very sparse information and no ortholog exists in human. A blast search against the human genome identifies a 99% identical sequence of length $\sim$ 1900 bp which has expression in many tissues but highest in ovary and testis[2].

E..46615 and E..49911 have no description or information about orthologs.

E..22234 "rhox homeobox family member 2" has an ortholog in human: RHOXF2. In human, this gene is expressed in many tissues with the highest levels in testis. It is described as being expressed in early stage germ cells, type-B spermatogonia and early spermatocytes. It has variants that are linked to infertility[3].

E..21637 is an otherwise uncharacterized gene increasingly expressed in the following organs in order: cerebellum, kidney, prefrontal cortex and testis. In the adult Chimpanzee, this gene is 3 times more expressed in the testis than in any other organ[4].

E..22336 has no description, but is expressed in the heart and mostly in the testis[4].

TCP11X2 "t-complex" E..48802 is an overlap. TCP11X2 is present in Pan_tro_3 in two copies (orthologs), which might inflate its copy number. Interestingly the overlap with the gene E..48802 has a higher copy number than the TCP11X2 gene itself. TCP11X2 is not testis expressed[2], E..48802 is testis expressed.

The next region which is an overlap of E..49979 E..22357 might just be highly copied because it is fundamentally linked to E..49971 which is the gene with the highest copy numbers. Anyway, a look up informs that it is a kidney expressed gene which is otherwise undescribed.

XAGE5 "X antigen fam. 5" is expressed in cerebellum and mostly testis[4].

It seems like there is a relation between the high median CN and testis presence. As we don't know what proportion of the X-linked genes are expressed in the testis, it is hard to say if the genes with the highest variation are significantly more present in the testis. Nevertheless, for the genes where expression data

was available, all genes but ?? showed signs of presence in the testis.
??flyt højere op: Assuming that the genome wide CN median should be 1, suggests that it should be possible to convert the relative copy number measures to absolute by normalizing the subject-wise median to 1. Nonetheless, the DMD gene is a promising normalization candidate, as most subject get a median close to 1.

Copy number variation among relatives

The copy number of a gene in a female individual (XX) is the average of the copy number of the two X chromosomes in its genome. In male individuals (XY), the copy number is the number on the present X chromosome. This means that be investigating the copy number amongst relatives, we can infer the transmission of X chromosomes. Especially in a dad-daughter relation, we know what the copy number of the X chromosome passed to the daughter.

In order to investigate the copy number turnover between parents-offspring, we plotted the copy number over time measured in n'th degree relatives. For 7 of the 33 Chimpanzees included in this study, we know the parent relation, and for 3 we know the grandparent relation. This means that we can compare the progression of copy numbers, down through the pedigree. In figure 4, the copy numbers for the 5 most highly copied genes as well as OPN1LW is visualized. Dad-son relations are removed because no X chromosome is passed in this relation.

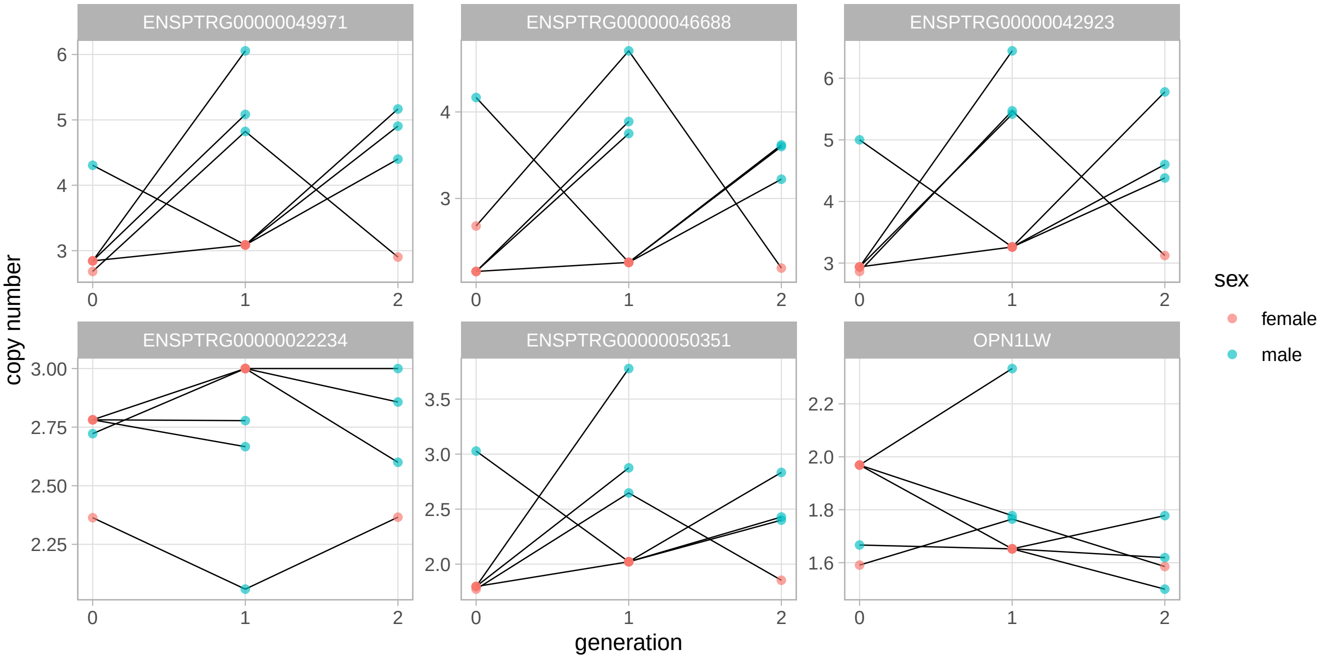

Figure 4. Copy number as a function of pedigree. Each straight line denotes the change in copy number for an X-linked gene, when the X chromosome is passed from parent to offspring. Dad-son relations are not included.

Conclusion

This study has searched for genes related to meiotic drive and found handful of good candidates. For the genes where expression data is present, it seems like there is a relation between high median copy number and testis presence. This might support the hypothesis that there is a relation between CNV and meiotic drive.

There is a clustering of median between sexes in some of the most ampliconic genes. The tendency is that males have a higher copy number than females. Though this might be a signal, nothing has been concluded or hypothesized from it. Whether it stems from reads from the X degenerate region on the Y chromosome can be investigated be concatenating the X and Y chromosomes together and checking whether the clustering vanishes. A different and much more interesting act, would be to propose that there is a link between high copy numbers of these genes, and the segregation of a Y chromosome to the sperm cell, leading to male individuals with high copy numbers. It might also be due to random fluctuations.

E..22234, E..21637 and ?? might be good candidates of genes involved in meiotic drive processes. Expression data is missing on the three most wildly copied genes (E..49971, E.46688, E..42923), which might also have a testis-relation.

For many of the genes where identical sequnces are found in other species (mostly human) where testis expression is present, it is not yet investigated whether the ampliconic region is identical to the ortholog or not.

This study has found a handful of genes with high copy number variation which are expressed in the testis. Though some of the genes had no description or expression data available, it seems like most of the most copy variant genes are testis related.

?? Relate to meiotic drive. Give ideas to further analysis.

?? Læs bachproj igen og få ideer.

?? fejlkilder

??Using decoy genome to get rid of paralogs.

References

1. Lucotte, E. A. *et al.* Dynamic copy number evolution of x- and y-linked ampliconic genes in human populations. *Genetics* **209**, 907–920, DOI: 10.1534/genetics.118.300826 (2018). https://www.genetics.org/content/209/3/907.full.pdf.

2. Hunt, S. E. *et al.* Ensembl variation resources. *Database* **2018**, DOI: 10.1093/database/bay119 (2018). Bay119, https://academic.oup.com/database/article-pdf/doi/10.1093/database/bay119/27329174/bay119.pdf.

3. Borgmann, J. *et al.* The human RHOX gene cluster: target genes and functional analysis of gene variants in infertile men. *Hum. Mol. Genet.* **25**, 4898–4910, DOI: 10.1093/hmg/ddw313 (2016). https://academic.oup.com/hmg/article-pdf/25/22/4898/10245950/ddw313.pdf.

4. Bastian, F. *et al.* Bgee: Integrating and comparing heterogeneous transcriptome data among species. In Bairoch, A., Cohen-Boulakia, S. & Froidevaux, C. (eds.) *Data Integration in the Life Sciences*, 124–131 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2008).

Appendix

Subject	Sex	Species	Source
Julie-LWC21	female	Pan troglodytes ellioti	Prado-Martinez et al. 2014
Akwaya-Jean	male	Pan troglodytes ellioti	Prado-Martinez et al. 2014
Damian	male	Pan troglodytes ellioti	Prado-Martinez et al. 2014
Koto	male	Pan troglodytes ellioti	Prado-Martinez et al. 2014
Taweh	male	Pan troglodytes ellioti	Prado-Martinez et al. 2014
Andromeda	female	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Harriet	female	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Kidongo	female	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Nakuu	female	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Bwambale	male	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Vincent	male	Pan troglodytes schweinfurthii	Prado-Martinez et al. 2014
Clara	female	Pan troglodytes troglodytes	Prado-Martinez et al. 2014
Doris	female	Pan troglodytes troglodytes	Prado-Martinez et al. 2014
Julie-A959	female	Pan troglodytes troglodytes	Prado-Martinez et al. 2014
Vaillant	male	Pan troglodytes troglodytes	Prado-Martinez et al. 2014
Jimmie	female	Pan troglodytes verus	Prado-Martinez et al. 2014
Bosco	male	Pan troglodytes verus	Prado-Martinez et al. 2014
Clint	male	Pan troglodytes verus	Prado-Martinez et al. 2014
Koby	male	Pan troglodytes verus	Prado-Martinez et al. 2014
Donald	male	Pan troglodytes verus x troglodytes	Prado-Martinez et al. 2014
Carolina	female	Pan troglodytes verus	Besenbacher et al. 2019
Simliki	female	Pan troglodytes verus	Besenbacher et al. 2019
Carl	male	Pan troglodytes verus	Besenbacher et al. 2019
Frits	male	Pan troglodytes verus	Besenbacher et al. 2019
Pearl	female	Pan troglodytes verus	Venn et al. 2014
Marlies	female	Pan troglodytes verus	Venn et al. 2014
Marco	male	Pan troglodytes verus	Venn et al. 2014
Dirk1	male	Pan troglodytes verus	Venn et al. 2014
Dennis	male	Pan troglodytes verus	Venn et al. 2014
Ruud	male	Pan troglodytes verus	Venn et al. 2014
Dylan	male	Pan troglodytes verus	Venn et al. 2014
Marlon	male	Pan troglodytes verus	Venn et al. 2014
Pat	male	Pan troglodytes verus	Venn et al. 2014

Table 1. Metadata for the 33 subjects used in this study.