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 mean of 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 show signs of testis-expression. These genes might be part of a meiotic drive process. Some of these genes also show a strong sex-grouping which might be due to the failure to include the X-degenerate region of the Y chromosome in the mapping-reference. 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 now present in primates[1]. It is proposed that the X and Y chromosomes are divergent copies of a once present autosome. The Y chromosome has lost most of its genes and now most importantly hosts the testis determining factor (SRY), which defines the sex of the primate individual. The other genes still present on the Y chromosome have mostly become inactivated and are now referred to as the X-degenerate region[1].

It is suggested that the X and Y chromosomes might interfere with one another during male meiosis where either the X or Y chromosome is passed down to the sperm cell, such that the corresponding sex chromosome has a higher probability of being transmitted to the offspring. This process termed meiotic drive, generally refers to an unequal segregation of sex chromosomes from the heterogametic sex. This results in biased sex-ratios in the population, and also implies that the fitness of these chromosomes is not optimal (??)[2]. 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 X chromosomes containing drive coding genes might have developed corresponding drive-suppressing genes on the Y chromosome.

The search for genes, or underlying regions, which are part of this meiotic drive process, can be obtained by looking for genes with a high copy number variation between individuals. If we can assess the copy number variation of the genes on the X chromosome, we can investigate these genes expression in testis, and possible further hypothesize that these genes are part of a meiotic drive process.

??Meiotic drive processes are studied in relation to hybrid incompatibility??ref

In a more or less recent study[3], ampliconic regions in human have been investigated in their relation to

testis expression, and here we present a smaller scaled but otherwise similar study, where we look for ampliconic regions on the Chimpanzee X chromosome.

?? Se hvad der linker til lucotte.

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 conditions: 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[3], 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.

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 of copy numbers instead of the DMD coverage. There is no known prior on the chromosome wide copy number distribution. But a mean close to one makes good sense, as it suggests that linked paralogs have been described separately in the annotation.

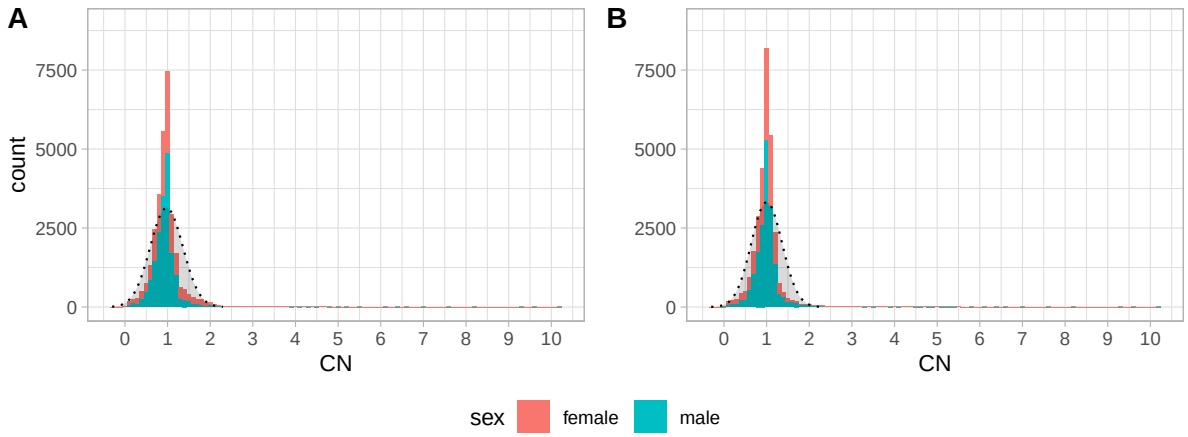

Figure 1. Distribution of median CN across all genes for all subjects. A normal distribution with an equal median, SD and area (grey) is overlapped as a visual aid. Colored by sexes which are stacked. **A**: Normalization is achieved through the coverage of the DMD gene. **B**: Normalization is achieved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

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.

Of the 919 genes on the X chromosome, 64 (~7%) have an overlap with another gene. One might suspect that regions where two genes overlap might have a slower turnover than a region with only one gene, because a mutation in one gene will most likely negatively affect the other gene. Nonetheless, 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 significant (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

By measuring the copy number of all genes in all individuals and focusing on the genes with a median above 2, we came up with the list of genes enumerated in figure 3.

Many of the genes with high median copy number show a strong grouping between sexes. This might be because the reads from the X degenerate region on the Y chromosome[1] in these males have erroneously mapped to its counterpart on the reference X chromosome. As a test, one of the most sexually grouped genes (E..46688) was blast-aligned to the Y chromosome with no matches, which suggests that the sex-grouping is not due to methodological errors. Because the proportion of sexes in each subspecies is varying (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 (listed in figure 3) with the highest SD median copy numbers, where the

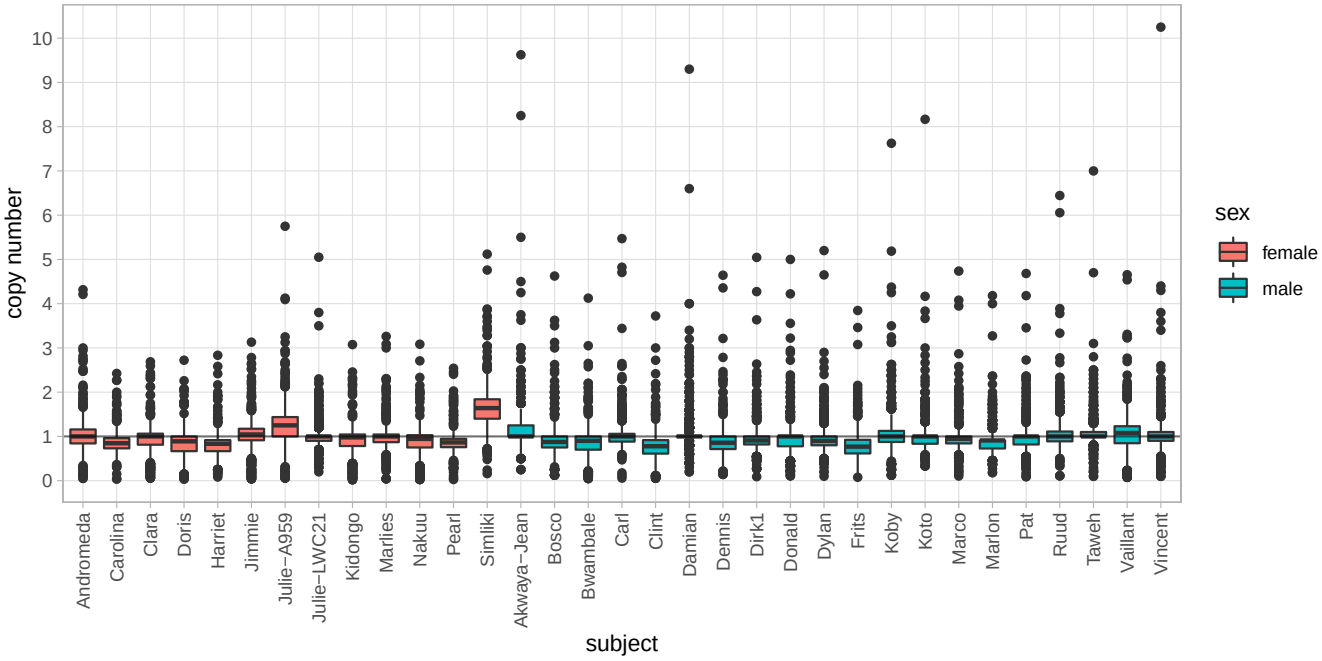

Figure 2. Distribution of median CN across all genes for each subject. Coverage normalization is achieved with the coverage of the DMD gene in each individual. Box edges denote quartiles. Points denote values beyond 1.5*IQR, where IQR is the interquartile range.

median was above 2. 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 genes with the highest median E..49971 and E..42923 are both related to the 40S ribosomal subunit[4]. Unfortunately, no expression data is available, neither for orthologs.

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[4].

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[5].

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[6].

E..22336 has no description, but is expressed in the heart and mostly in the testis[6].

TCP11X2 "t-complex" E..48802 is an overlap. TCP11X2 is present in the Pan_tro_3 reference genome 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[4] but E..48802 is.

The next region which is an overlap of E..49979 E..22357. A look up informs that it is an otherwise

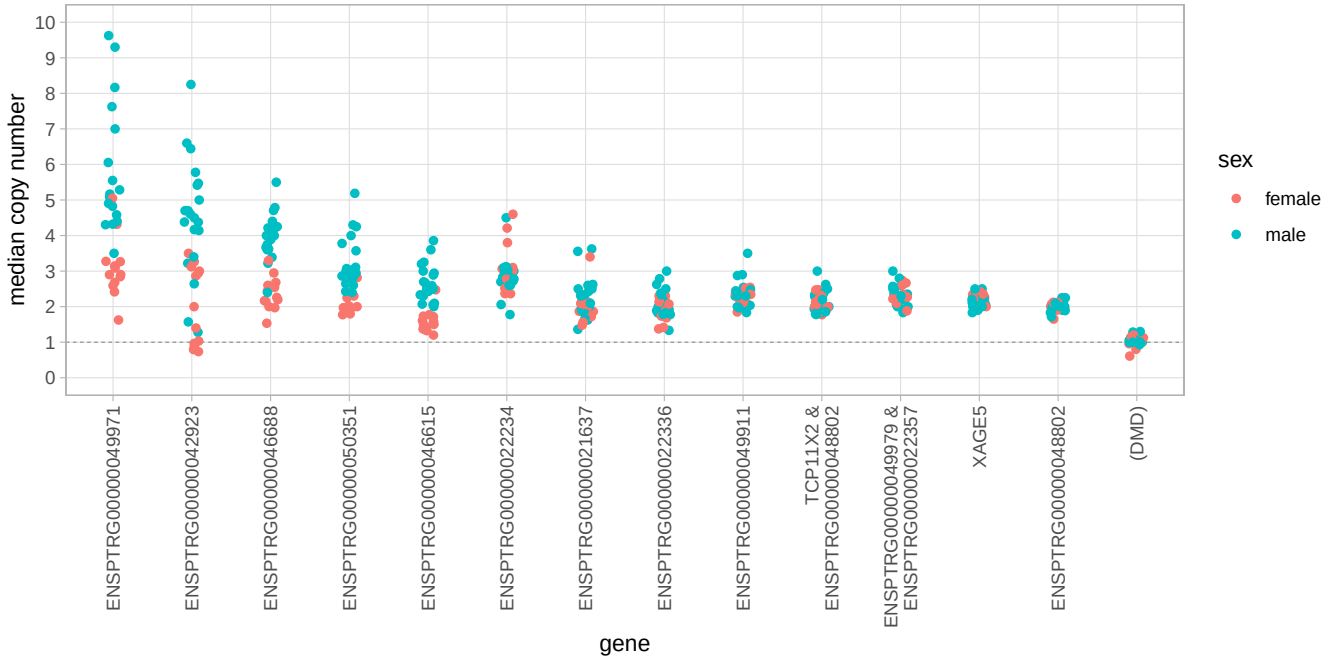

Figure 3. Copy number from all subjects, grouped by genes. These 13 genes or overlaps of genes have a median copy number ≥ 2. Sorted by descending standard deviation. '&' signifies the overlap of two genes. Horizontal jitter is applied. DMD is showed for comparison.

undescribed kidney expressed gene. E..49979 is a "heat shock transcription factor" and E..22357 is an otherwise undescribed gene with expression in kidneys.

XAGE5 "X antigen fam. 5" is expressed in cerebellum and mostly testis[6].

Overlapping the genes from this study with the ones in Lucotte *et al.* 2018[3] only two genes are present in the top 30 most copy number-variant genes. First one is OPN1LW which codes for long-wavelength opsin in the retina. Nonetheless, it has its highest expression (in Chimpanzee) in the testis[6]. The other gene, CT47A, presumably orthologous to ENSPTRG00000046894 "cancer/testis antigen 47A" has a max CN of 1.8 and a median CN of 0.9, and is thus not as variant in chimpanzee as in human (max CN: 15.07, median CN: 5.01). The other genes deemed copy number-variant in Lucotte et al. 2019 do not have identically named genes in the Chimpanzee annotation.

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 E..22357 showed signs of presence in the testis.

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 present 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 probable 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 and offspring, we plotted the copy number over time measured in generations. 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 2 generations. In figure 4, the copy numbers for the 6 most highly copied genes is visualized. Dad-son relations are removed because no X chromosome is passed in this relation.

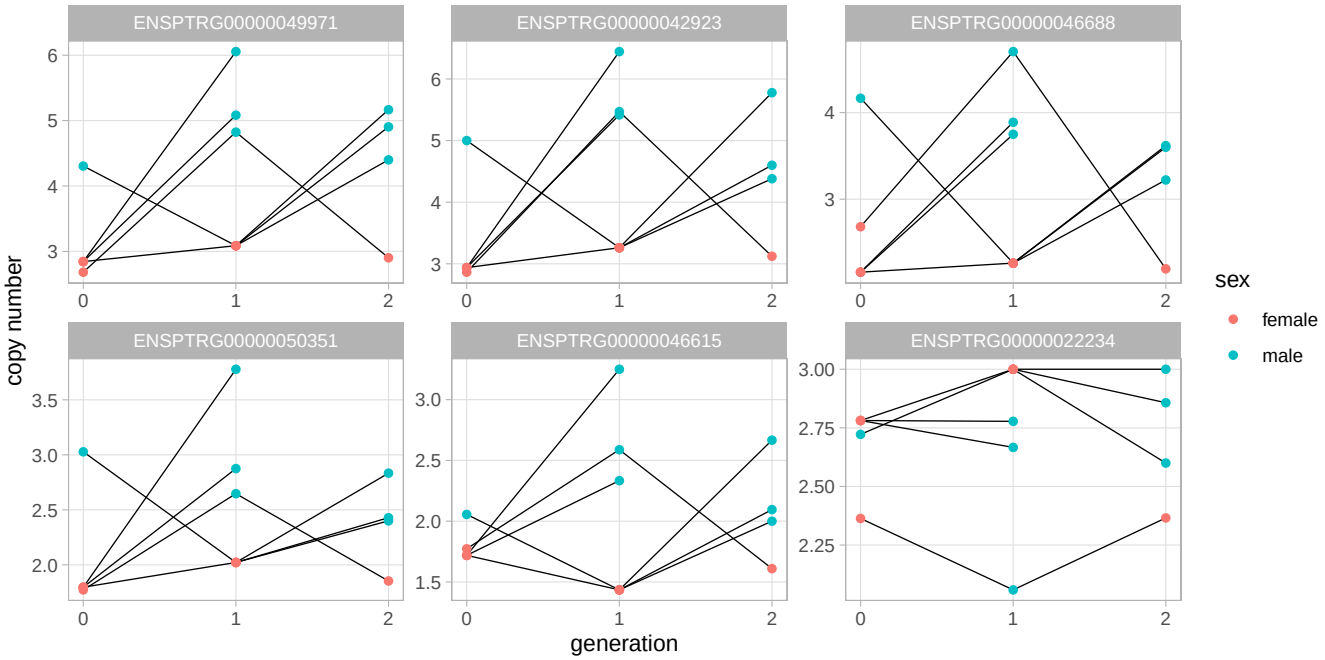

Figure 4. Copy number as a function of pedigree. Each straight line denotes the change in copy number for an X-linked gene, when the X chromosome is passed from parent to offspring. Dad-son relations are not included.

Because the sample size of each parent-offspring event is limited, it is impossible to conclude anything on the copy number variation between generations. For E..49971 there is a parents-daughter relation from generation 0 to 1. Because the daughter has one X from her dad (copy number = 4.3) and one copy from her mom (either copy number = 6 or copy number = 1, as evident from her other sons where the dad is not known and not relevant). Nonetheless, this daughter has three sons with copy numbers = 4.3, 4.9 and 5.1 respectively. It is implicitly given that the X chromosomes in these three sons are the same, and the amount of variation in these sons show that the copy number variation in this gene is not stagnant.

A similar line of thought can be applied to some of the other genes, implicating that the copy number can vary somewhat between generations. This is of course, assuming that the measurements are robust and not due to random fluctuations introduced method errors.

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. Assuming that there is a relation between copy number variation and meiotic drive, these genes are candidates for being involved in this process.

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 contrasting the copy number variation between the sexes. As a test, one of the most sexually grouped genes (E..46688) was mapped to the Y chromosome with no matches, which suggested that the sex-grouping is not due to methodological errors. 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, which is supported by the unequal sex ratios of each subspecies.

E..50351, E..22234, E..21637, E..22336, E48802 and XAGE5 might be good candidates of genes involved in meiotic drive processes. Unfortunately, expression data is missing on the three most widely copied genes (E..49971, E..42923, E.46688).

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

It would be appropriate to identify the ampliconic regions on each of the copy number variant genes. This can be done with dotplots. In order to investigate the cause of the amplification, it would make sense to identify SNPs on these regions, and investigate whether these SNPs are related to amplification.

The most ampliconic genes found in this study does not overlap with the genes with the most copy number-variant genes in Lucotte *et al.* 2018. This suggests that the ampliconic behaviour in all these genes has occurred after the Human-Chimpanzee speciation event.

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.

References

1. Skaletsky, H. *et al.* The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825–37, DOI: 10.1038/nature01722 (2003).

2. Jaenike, J. Sex chromosome meiotic drive. *Annu. Rev. Ecol. Syst.* **32**, 25–49, DOI: 10.1146/annurev.ecolsys.32.081501.113958 (2001). https://doi.org/10.1146/annurev.ecolsys.32.081501.113958.

3. 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.

4. 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.

5. 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.

6. 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).

7. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499**, DOI: 10.1038/nature12228 (2013).

8. Besenbacher, S., Hvilsom, C., Marquès-Bonet, T., Mailund, T. & Schierup, M. H. Direct estimation of mutations in great apes reveals significant recent human slowdown in the yearly mutation rate (2018).

9. Venn, O. *et al.* Strong male bias drives germline mutation in chimpanzees. *Science* **344**, 1272–1275, DOI: 10.1126/science.344.6189.1272 (2014). https://science.sciencemag.org/content/344/6189/1272.full.pdf.

Appendix

Subject	Sex	Species	Source
Julie-LWC21	female	Pan troglodytes ellioti	Prado-Martinez *et al.* 2013[7]
Akwaya-Jean	male	Pan troglodytes ellioti	Prado-Martinez *et al.* 2013[7]
Damian	male	Pan troglodytes ellioti	Prado-Martinez *et al.* 2013[7]
Koto	male	Pan troglodytes ellioti	Prado-Martinez *et al.* 2013[7]
Taweh	male	Pan troglodytes ellioti	Prado-Martinez *et al.* 2013[7]
Andromeda	female	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Harriet	female	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Kidongo	female	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Nakuu	female	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Bwambale	male	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Vincent	male	Pan troglodytes schweinfurthii	Prado-Martinez *et al.* 2013[7]
Clara	female	Pan troglodytes troglodytes	Prado-Martinez *et al.* 2013[7]
Doris	female	Pan troglodytes troglodytes	Prado-Martinez *et al.* 2013[7]
Julie-A959	female	Pan troglodytes troglodytes	Prado-Martinez *et al.* 2013[7]
Vaillant	male	Pan troglodytes troglodytes	Prado-Martinez *et al.* 2013[7]
Jimmie	female	Pan troglodytes verus	Prado-Martinez *et al.* 2013[7]
Bosco	male	Pan troglodytes verus	Prado-Martinez *et al.* 2013[7]
Clint	male	Pan troglodytes verus	Prado-Martinez *et al.* 2013[7]
Koby	male	Pan troglodytes verus	Prado-Martinez *et al.* 2013[7]
Donald	male	Pan troglodytes verus x troglodytes	Prado-Martinez *et al.* 2013[7]
Carolina	female	Pan troglodytes verus	Besenbacher *et al.* 2018[8]
Simliki	female	Pan troglodytes verus	Besenbacher *et al.* 2018[8]
Carl	male	Pan troglodytes verus	Besenbacher *et al.* 2018[8]
Frits	male	Pan troglodytes verus	Besenbacher *et al.* 2018[8]
Pearl	female	Pan troglodytes verus	Venn *et al.* 2014[9]
Marlies	female	Pan troglodytes verus	Venn *et al.* 2014[9]
Marco	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Dirk1	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Dennis	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Ruud	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Dylan	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Marlon	male	Pan troglodytes verus	Venn *et al.* 2014[9]
Pat	male	Pan troglodytes verus	Venn *et al.* 2014[9]

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

gene	length	min CN	max CN	median CN	SD	RSD	description
ENSPTRG00000049971	888	1.63	10.25	4.58	2.16	0.47	NA
ENSPTRG00000042923	882	0.74	8.25	3.40	1.90	0.56	NA
ENSPTRG00000046688	346	1.53	5.50	3.60	0.98	0.27	NA
ENSPTRG00000050351	1904	1.77	5.19	2.65	0.83	0.31	NA
ENSPTRG00000003827	640	0.21	3.44	1.50	0.79	0.53	protein phosphatase 1 regulatory subunit 14B
ENSPTRG00000050450	81688	0.53	3.33	1.57	0.79	0.50	NA
ENSPTRG00000046615	350	1.20	3.86	2.30	0.70	0.30	NA
ETDB	180	1.00	3.80	1.78	0.67	0.38	embryonic testis differentiation homolog B
ENSPTRG00000022234	5840	1.78	4.60	2.83	0.59	0.21	rhox homeobox family member 2
ENSPTRG00000023212	459	0.91	3.75	1.86	0.59	0.32	NA
ENSPTRG00000049942	14672	0.70	2.80	1.80	0.55	0.31	NA
ENSPTRG00000021637	20103	1.36	3.63	2.10	0.54	0.26	variable charge X-linked protein 3-like
ENSPTRG00000051860	74889	0.70	2.80	1.57	0.52	0.33	NA
OPN1LW & TEX28	NA	0.73	2.79	1.54	0.49	0.32	NA
ENSPTRG00000045311	642	0.56	2.40	1.50	0.48	0.32	family with sequence similarity 156 member B
ENSPTRG00000028324	52749	0.70	2.00	1.62	0.45	0.28	CD99 molecule (Xg blood group)
ENSPTRG00000048376	1070	1.14	2.67	1.67	0.42	0.25	NA
IL1RAPL2 & ENSPTRG00000052876	NA	1.21	3.13	1.80	0.41	0.23	NA
ENSPTRG00000049979	1693	1.30	3.00	1.78	0.39	0.22	heat shock transcription factor family X-linked member 3
ENSPTRG00000022336	945	1.33	3.00	2.00	0.38	0.19	SPANX A/D member 1
ENSPTRG00000049911	1202	1.83	3.50	2.29	0.34	0.15	NA
POLA1 & ENSPTRG00000041347	NA	1.15	2.38	1.67	0.32	0.19	NA
EDA & ENSPTRG00000047182	NA	1.17	2.38	1.67	0.29	0.17	NA
TCP11X2 & ENSPTRG00000048802	NA	1.77	3.00	2.00	0.28	0.14	NA
ENSPTRG00000049979 & ENSPTRG00000022357	NA	1.83	3.00	2.30	0.26	0.11	NA
TEX28	21367	1.50	2.67	1.79	0.25	0.14	testis expressed 28
OPN1LW	15677	1.27	2.33	1.59	0.24	0.15	opsin 1 long wave sensitive
SPANXN5	1250	1.50	2.50	1.83	0.23	0.13	SPANX family member N5
FGF13 & ENSPTRG00000016737	NA	1.25	2.20	1.71	0.22	0.13	NA
XAGE5	7192	1.83	2.50	2.20	0.16	0.07	X antigen family member 5
ENSPTRG00000048802	116637	1.65	2.25	2.00	0.13	0.06	NA

Table 2. Genes with highest SD of copy number. RSD is SD divided by median.