Transposable element expression with variation in sex chromosome number supports a toxic Y effect on human longevity

Jordan Teoli[1,2,3,4], Marie Fablet[2,5], Claire Bardel[1,2], Anamaria Necsulea[2], Audrey Labalme[6], Hervé Lejeune[7], Jean-François Lemaitre[2], François Gueyffier[2], Damien Sanlaville[6], Cristina Vieira[2]†*, Gabriel AB Marais[2,4,8,9]†*, Ingrid Plotton[1,3]†*

[1] Laboratoire de Biochimie et Biologie Moléculaire, Centre de Biologie et Pathologie Est, Hospices Civils de Lyon / Université Lyon 1; Bron, 69500, France.
[2] Laboratoire de Biométrie et Biologie Evolutive, UMR5558, CNRS / Université Lyon 1; Villeurbanne, F-69622, France.
[3] Institut Cellule Souche et Cerveau (SBRI), Unité INSERM 1208, Centre de Recherche INSERM; Bron, 69500, France.
[4] CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO Laboratório Associado, Campus de Vairão, Universidade do Porto; Vairão, 4485-661, Portugal.
[5] Institut universitaire de France.
[6] Service de Génétique, Hospices Civils de Lyon / Université Lyon 1; Bron, 69500, France.
[7] Service de médecine de la reproduction, Hôpital Femme-Mère-Enfant, Hospices Civils de Lyon; Bron, 69500, France.
[8] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto; Porto, 4099-002, Portugal.
[9] BIOPOLIS Program in Genomics, Biodiversity and Land Planning, CIBIO, Campus de Vairão; Vairão, 4485-661, Portugal.

†Equal contributions as senior authors

*Corresponding author
Correspondence: ingrid.plotton@chu-lyon.fr, gmarais@cibio.up.pt, cristina.vieira@univ-lyon1.fr

Abstract

Lifespan differences between sexes is a puzzling question in human and evolutionary biology. Sex chromosomes have been proposed to play a central role, and the Y male chromosome has been suspected of having a toxic genomic impact in this trait. As Y chromosomes are typically enriched in transposable elements (TE), this hypothesis suggests that in aged individuals, in whom TE are less repressed, an increase in somatic mutations and acceleration of the aging process will be observed. Using an unprecedented expression dataset from humans with atypical karyotypes, we show an increased TE expression related to the Y chromosome. These findings complement previous reports showing that the 47,XYY human population displays a shorter lifespan, and support the existence of a toxic Y effect on men lifespan.

Keywords: transposon, lifespan, gonosome, Klinefelter syndrome, transcriptomic, differential expression analysis

Introduction

Life expectancy at birth is on average five years longer in women than in men [1], a female-biased gap in longevity which is observed in almost all human populations [2] and which also explains why most supercentenarians are women [3]. The reasons for this consistent female-biased gap in longevity in humans remains a long-lasting conundrum. Both cultural and biological factors have been invoked [2,4]. For the latter, sex-specific production of hormones (e.g. androgens) as well as male-specific deleterious mitochondrial mutations are biological factors that - to some extent - contribute to the sex differences in survival prospects [5,6]. Recently, a new hypothesis has been proposed to explain this sex gap in longevity (reviewed in [7,8]). Differential survival among males and females are pervasive across the tree of life [9]. In both vertebrates and invertebrates, the heterogametic sex (male in XY systems, female in ZW ones) tend to die first, which suggests that sex chromosomes could contribute to establishing sex gaps in longevity [10,11]. As the heterogametic sex carries one type of each sex chromosome, it is therefore more frequently affected by recessive X-linked (or W-linked) deleterious genetic mutations (unguarded X hypothesis), but theory and empirical data suggest that this mechanism is probably not sufficient to explain the entire sex gap in longevity [12]. Moreover, Y (or W) chromosomes are often less recombinant, and repeated sequences tend to accumulate in these chromosomes. With age, epigenetic regulation mechanisms are less effective, and transposable elements (TE) can be reactivated. This was observed in aging fruit flies resulting in a surge in TE expression and an increase in the somatic mutation rate [13], higher in aged male than in aged female flies [14]. Experiments uncoupling phenotypic sex and sex chromosome content have shown that reduced longevity is caused by the presence of the Y, not by phenotypic sex [14]. This phenomenon, called the toxic Y effect, has been documented only in Drosophila species so far [15], and is still controversial [16].

The human Y chromosome is relatively small (about 57 Mb) compared to the X chromosome (about 156 Mb) but is an extremely repeat-rich chromosome compared to the X chromosome or autosomes [17,18]. If a toxic Y effect exists in humans, we should observe that - everything else being equal - more TE get reactivated in old 46,XY men than in old 46,XX women, as in Drosophila (hereafter referred to as prediction 1). Moreover, in humans, 47,XXY and 47,XYY individuals have strikingly different longevities [19,20]: 47,XYY is indeed associated with a 10-year-reduction in longevity while 47,XXY is associated with a 2-year-reduction only compared to 46,XY individuals. This is observed despite the extra-Y having smaller observable effects on the 47,XYY individual's biology and health than the extra-X in the 47,XXY karyotype that results in the Klinefelter syndrome (e.g. delays in first acquisitions, infertility, decreased bone mineral density ,…) [19,20]. If a toxic Y effect explains the aforementioned observations and contributes to establishing the human sex gap in longevity, TE expression should be associated with the number of Y chromosomes in a karyotype (hereafter referred to as prediction 2).

Results and Discussion

We tested prediction 1 using published blood RNA-seq data from the Genotype-Tissue Expression (GTEx) project. We compared TE expression in 46,XX female and 46,XY male individuals belonging to different age groups from 20 to 70 years old. We found that only a few TE families have expression patterns consistent with our prediction and that TE expression seems to show a peak of expression at 30-40 years old in males (see **Supplementary Materials - Analysis of TE expression in the GTEx dataset** and **Fig. S1 to S8, Tables S1 to S6, Data S1**). However, it was difficult to draw strong conclusions because of the intrinsic nature of the data from the GTEx database: the many donors in different conditions and the many different sequencing runs made the data quite noisy (potential confounding factors, no longitudinal follow-up of same individual, survival bias, etc…). Moreover, the rather low sequencing depth and read length are not ideal for studying TE expression (see **Materials and Methods**).

To test prediction 2, we generated RNA-seq data from blood samples of individuals with several gonosome compositions on the karyotypes. Blood samples were obtained from 25 individuals of different karyotype compositions (46,XX, 46,XY, 47,XXY and 47,XYY) (**Table S7**) from the Lyon university hospital. For each sample, we performed deep RNA-seq for a median of 174 million reads (min-max: 132 million – 206 million) *per* individual, producing an unprecedented dataset (**Table S7**). We first analyzed gene

expression (**Fig. S3, Data S1**) to validate our wet and dry lab protocols (see **Materials and Methods**) and to assess whether certain gene aging pathways (Gene Ontology (GO) - term analysis) were impacted by the karyotype. The number of significantly differentially expressed protein-coding genes varied according to the pair of karyotypes compared (**Fig. S9, Table S8**). Comparing 46,XX and 46,XY samples (typical male and female karyotypes) returned 0.10% and 0.09% of over-expressed and under-expressed protein-coding genes, respectively (**Table S8**). Previous work using much larger sample sizes yielded more sex-biased genes [21]; we were thus conservative here. We also found that most (13/20, 65%) male-biased protein-coding genes are Y-linked genes, and that most female-biased protein-coding genes are known X chromosome inactivation escapees (**Fig. S10, Tables S8 to S10, Data S2 and S3**), as expected from previous work [21]. It is indeed well-known that the extra-X chromosomes in 47,XXY or other karyotypes (e.g triple X) are inactivated; only the genes escaping inactivation exhibit expression changes [22,23]. Furthermore, we found *XIST* and *TSIX* genes to be expressed only in karyotypes containing two X chromosomes (47,XXY and 46,XX), which was also expected as these genes are implicated in the X inactivation occurring only in the presence of two or more X chromosomes [24–26]. Some autosomal protein-coding genes were also differentially expressed when comparing 46,XX and 46,XY samples (**Table S8**).

Using pairwise karyotype comparisons, we found that 47,XYY and 47,XXY samples are the ones with the highest number of differentially expressed protein-coding genes (~5-6%) compared to 46,XX samples (**Fig. S9, Table S8**). A GO-term analysis revealed that overexpressed genes in 47,XYY samples compared to 47,XXY samples were included in biological pathways (cell cycle, DNA damage/telomere stress induced senescence, senescence-associated secretory phenotype, cellular senescence, oxidative stress induced senescence) which may be related to a faster aging and cancer risk in 47,XYY individuals compared to 47,XXY (**Fig. S11, Data S4**).

Concerning TE expression (**Fig. S3, Data S1**), a principal component analysis (PCA) showed that males and females had different profiles of TE expression as indicated by the sex clustering in the PCA plot (**Fig. S12**), a pattern that we also found for gene expression (**Fig. S13**). Similarly to protein-coding genes, TE families exhibiting expression changes were more numerous in 47,XYY/46,XX and 47,XXY/46,XX karyotype comparisons (**Fig. S14, Table S11**): 46 differentially expressed TE families in 47,XYY and 58 in 47,XXY compared to 46,XX. Most differentially expressed TE families were found to be significantly over-expressed in 47,XYY (35 TE families), 47,XXY (51 TE families), and 46,XY (5 TE families) compared to 46,XX karyotype, while few were under-expressed (11, 7, and 1 TE families, respectively) (**Fig. S14, Table S11**). The differentially expressed TE families according to karyotype is shown in **Fig. 1-A and S15**. A similar pattern was found for a majority of differentially expressed TE families: a minimal expression in 46,XX, a maximal expression in 47,XYY, and an in-between expression in 46,XY and 47,XXY karyotypes, pointing to a contribution of the Y chromosome in establishing this pattern. This pattern was also observed when combining all the TEs (**Fig. 1-B**). Furthermore, when splitting individuals of a given karyotype in two age groups (lower and higher than the median for this group), we showed that TE expression tends to be higher later in life for 46,XY and 47,XYY individuals, which is in line with prediction 1, although the limited sample size prevents a firm conclusion (**Fig. S16**).

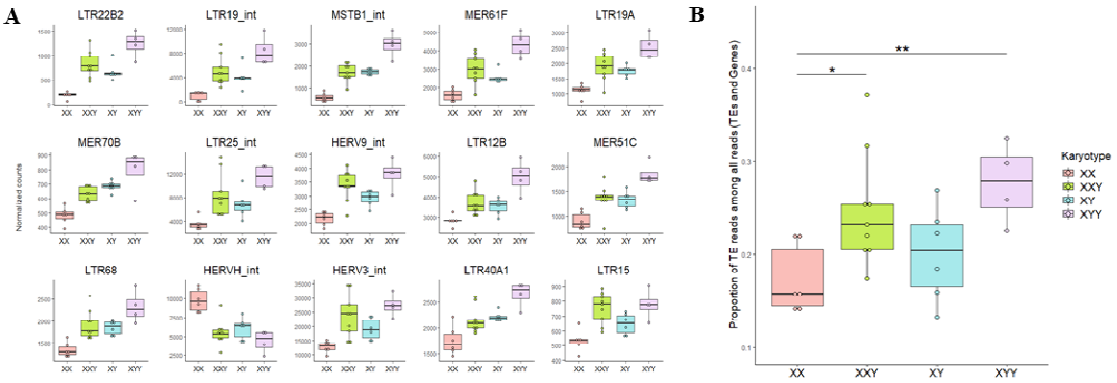

Figure 1. TE expression in individuals with normal and atypical sex chromosome content in our dataset. (A) Boxplots of the 15 most significantly differentially expressed TE families according to the karyotype using a likelihood-ratio test, after removing batch effect. (B) Boxplots of normalized counts combining all TE families after removing batch effect, regardless of age. (**A**) Boxplots for the other significantly differentially expressed TE families are in Supplemental material (**Fig. S15**). All of these 15 families had at least one copy on the Y chromosome and one copy on the X chromosome in the reference genome. (**B**) Raw p-values were less than 0.05 (*) for 47,XXY samples compared to 46,XX samples (P = 0.012), and less than 0.01 (**) for 47,XYY samples compared to 46,XY samples (P = 0.0095) and near 0.05 for 47,XYY samples compared to 46,XY samples (P = 0.067) using Wilcoxon test. Adjusted p-values using Holm method were no longer significant (P = 0.060, 0.057, and 0.267 respectively).

This pattern could be driven by the expression of TE insertions present in genes and in long non-coding RNA. We thus re-analyzed the data excluding insertions in exonic parts of all genes and in long non-coding RNA insertions. We also removed non-autonomous expression from introns of X and Y chromosomes (see **Materials and Methods and Fig. S17**). The results exhibited fewer significantly differentially expressed TE families, probably due to a reduction in statistical power, because of a reduction of counts per TE family. However, the pattern concerning the overall expression of TEs remained (**Fig. S18 to S21, Table S12**).

The analysis of our own data revealed strong changes in TE expression in 47,XYY individuals, and a clear trend towards a low expression in 46,XX individuals, high expression in 47,XYY, and intermediate expression in 46,XY and 47,XXY individuals for some TE families and for the global TE expression. Regardless of the age of the individuals, the results herein showed that TE global expression is higher in 47,XYY individuals compared to other karyotypes, although the comparison was significant only with the 46,XX individuals (**Fig. 1-B**). Interestingly, similar observations were made in Drosophila where Y-linked TE expression is stronger in individuals with an additional Y chromosome (i.e. XYY) compared to XY individuals, regardless of their age [14]. The present results suggest that a similar disruption of TE regulation is present in young individuals with a 47,XYY karyotype in humans. Furthermore, in Drosophila, Y-linked TE expression is stronger in aged flies compared to young flies for XY and XYY individuals but not for XXY individuals [14], which is in line with the trend we found in humans (**Fig. S16**).

The analyses carried out herein support both predictions of the toxic Y chromosome hypothesis, particularly when using the dataset that we have generated. We show an over-expression of TE related to the number of Y-chromosomes in the karyotype, and suggest that a toxic Y effect may contribute to the sex gap in longevity in humans, as shown in Drosophila ((*14*, *15*) but see [16]).

Sex gap in longevity is certainly a multifactorial trait and the toxic Y effect is just one of them. The present results indicate that relaxing TE control with aging or adding extra Y chromosomes will contribute to an increase in TE expression and potentially to an increase in somatic mutations. It is also important to bear in mind that an equilibrium is most certainly established in the interactions between chromosomes, and that adding Y chromosomes is disrupting it. The findings herein also suggest that future research in aging treatments needs to be sex-related and most probably TE-content related.

Materials and Methods

Biological material, RNA extraction and sequencing

GTEx dataset

The Genotype-Tissue Expression (GTEx) project regroups samples from a multitude of post-mortem organ donors and from a multitude of tissues. Samples were realized in PAXgene tissue fixation containers. After RNA extraction, libraries were prepared using an automated variant of the Illumina TruSeq RNA sample preparation protocol (Illumina Inc, California, USA) which was a non-stranded protocol using polyA selection. Then, libraries were sequenced in paired-end mode 76x2 on a HiSeq 2000 or 2500® (Illumina Inc, California, USA) (see https://gtexportal.org/home/methods and [27]).

To constitute the GTEx dataset, we extracted raw expression data (FASTQ containing Illumina reads) from the human expression atlas database GTEx via the dbGaP portal (dbGaP accession number phs000424.v8.p2) after authorization.

We focused on expression data obtained from blood, a tissue with one of the largest numbers of samples in the GTEx database. Thus, we considered only data of RNA-seq ("Assay.Type" variable) for which a file with sra extension was available ("DATASTORE.filetype" variable) and generated from RNA extracted from whole blood ("body_site" variable).

To reduce data heterogeneity and to get closer to the individuals we sampled in parallel to constitute our dataset (see below), we focused on: data collected in non-latino nor hispanic white subjects ("RACE" and "ETHNCTY" variables) and excluded data from subjects with persistent infection by the human immunodeficiency virus (HIV), the hepatitis C virus (HCV) or the hepatitis B virus (HBV) ("LBHCV1NT","LBHBHCVAB", "LBHIV1NT", "LBHIVAB", "LBHIVO", and "LBHBSAG" variables), or with dementia ("MHALZDMT" and "MHALZHMR" variables). Then, we made a restricted cohort ("no disease group") by also excluding data from subjects with current or antecedents of cancer ("MHCANCER5", "MHCANCERC", and "MHCANCERNM" variables), or cardiovascular diseases ("MHHRTATT", "MHHRTDIS", and "MHHRTDISB" variables), and a cohort including them ("group including cancer and cardiovascular diseases") to study the impact of current or antecedents of cancer or cardiovascular diseases in transposable element (TE) expression.

At this step, we removed samples identified by the GTEx analysis working group as suboptimal for use in analysis ("SMTORMVE" variable) and we kept only samples for which the material type experiment was described as RNA:Total RNA ("material type exp" variable). Then, we downloaded the remaining FASTQ files.

We generated an "age group" variable from the subjects' age by using five classes of 10-year ranges: [20,30],]30,40],]40,50],]50,60],]60,70].

Then, using PCA of gene expression and TE expression in the whole cohort and in the subgroups by sex or by age group, we identified a strong clustering of subjects who were organ donors OPO (Organ Procurement Organization), who were distinct from subjects who were not ("COHORT" variable) (**Fig. S1**). We also found that the "COHORT" variable had a statistical significant relation with the age group (Fisher test: $P < 0.001$) but not with the sex group (Fisher test: $P = 0.185$). Therefore, we considered the "COHORT" variable as acting as a confounding factor and we excluded the smallest group of subjects, i.e. subjects who were not organ donors OPO ($n = 24$). Finally, we obtained 80 files suitable for analysis in the no disease group (**Fig. S2**) and 102 in the group including cancer and cardiovascular diseases (see **Tables S1 and S4** for repartition of samples according to age and sex).

The median number of paired-reads (after trimming, see below) was about 44.6 million (Min: 27.0 – Max: 141.1 million) for the 80 FASTQ of the no disease group and was about 44.3 million (Min: 27.0 – Max: 141.1 million) for the 102 FASTQ in the group including cancer and cardiovascular diseases.

Dataset generated in this study: biological sample collection for 46,XX; 46,XY; 47,XXY; 47,XYY individuals

Twenty-five subjects were included: 13 had a sex chromosome aneuploidy (nine patients with a 47,XXY karyotype and four with a 47,XYY karyotype) and 12 had normal chromosomal formula (six subjects with 46,XX karyotype and six with 46,XY karyotype). Median age was 30.0 years old (inter-quartile interval: 25.0 – 35.0). Pairwise Wilcoxon test showed a significant difference in the age distribution only between 46,XX and 46,XXY ($P = 0.038$) which was no longer significant after Holm adjustment ($P = 0.230$) but likely due to a lack of statistical power. Therefore, we cannot exclude that age is a confounding factor in the relation between TE expression and karyotype, however it would be a conservatory bias for our predictions because 47,XXY (24.1 years old) and 47,XYY (29.5 years old) subjects have a lesser mean age than 46,XX (39.2 years old) and 46,XY (41.5 years old) subjects.

All subjects signed a consent form for genetic analysis and the study was approved by the ethics committee of the Lyon university hospital (number of the ethics committee: 22_385, number in the register of the *Commission nationale de l'informatique et des libertés*: 22_5385).

Blood samples were collected on a PAXgene® tube, kept 2 hours at room temperature, and then frozen at -80°C until extraction.

RNAs were thawed and extracted on a Maxwell® RSC system (Promega, Wisconsin, USA) with the Maxwell® RSC SimplyRNA Blood kit. Quantity (fluorometric concentration) and quality (RNA Integrity Number, RIN) of the extract were assessed by a fluorometric measurement (Quant-iT™ RNA Assay Kit, Broad Range from ThermoFisher Scientific, Massachusetts, USA) on a Spark® reader (Tecan, Männedorf, Switzerland) and by electrophoresis (TapeStation 4200® System from Agilent Technologies, California,

USA), respectively. Each sample was of acceptable RNA quantity (fluorometric concentration of nucleic acids between 26.9 and 109.3 ng/µl) and quality (RIN between 5.8 and 9.3).

Libraries were prepared using the Illumina TruSeq Stranded Total RNA Lp Gold kit (Illumina Inc, California, USA). Size homogeneity of the fragments was verified using the TapeStation 4200® System. Average length of fragments was between 405 and 498 bp (including adapters). Library concentrations were assessed by fluorometric measurement (Quant-iT™ 1X dsDNA HS Assay Kit from ThermoFisher) on a Spark® reader. Concentrations were between 19.3 and 133.0 nmol/L.

The sequencing was performed on a NovaSeq 6000® system (Illumina Inc, California, USA) in paired-end mode (2 x 151 cycles) having generated a median of 174 million (Min: 132 – Max: 206 million) pairs of reads per sample (after trimming, see below).

Demultiplexing was performed with bcl2fastq (v2.20) to obtain FASTQ files. The 25 subjects were divided into two sequencing runs (run 1: 14 subjects, run 2: 11 subjects) (**Table S7**).

Bioinformatic processing

Studying the TE expression is challenging due to their repeated nature. We used a methodology adapted to repeated sequences to map and estimate the expression level of TE.

From this point, FASTQ files from GTEx dataset or our dataset were processed in the same way, except that kallisto, TEcount and TEspeX tools were used in non-stranded mode for GTEx dataset and in stranded mode for our dataset when necessary.

See pipelines in **Fig. S3 and S17**.

Input files

FASTQ files were used after verification and improvement of quality with the FastQC® and Trimmomatic® softwares [28]. After trimming, read quality was good.

The whole human reference transcriptome (Human Release 32 GRCh38/hg38, gencode v32) was downloaded from the University of California Santa Cruz (UCSC) table browser (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=814613547_qUmlEGan4K2f39cHvb9AgrKdRVdK&clade=mammal&org=Human&db=hg38&hgta_group=genes&hgta_track=knownGene&hgta_table=0&hgta_regionType=genome&position=chr1%3A11%2C102%2C837-11%2C267%2C747&hgta_outputType=primaryTable&hgta_outFileName=) with output format = “sequence” (fasta file) and used for the differential gene expression analysis. The downloaded file contained 247541 transcript sequences (5’ and 3’ untranslated regions and coding DNA sequences of each gene were included, introns were discarded, and repeated regions were masked).

A GTF file corresponding to Release 32 GRCh38/hg38 of the human reference genome was downloaded in gencodegenes website https://www.gencodegenes.org/human/ (content = “comprehensive gene annotation”, regions = “CHR”). This file provided a correspondence between transcripts (227463 transcripts names) and genes (60609 genes, including 19947 protein coding-genes).

The human reference sequence file for repeats (including TE copy sequences) was downloaded on the UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=791366369_Zf0cNT7ykVM0ZQ0zErZRArSgvMEO&clade=mammal&org=Human&db=hg38&hgta_group=allTracks&hgta_track=rmsk&hgta_table=0&hgta_regionType=genome&position=chr1%3A11%2C102%2C837-11%2C267%2C747&hgta_outputType=primaryTable&hgta_outFileName=) with output format = “sequence” (fasta file) and RepeatMasker track from the “Dec. 2013 GRCh38/hg38” assembly of the human reference genome. This file contained sequences of 4886205 repeat copies, after excluding transfer RNAs, ribosomal RNAs, small RNAs, repeats U1 to U13, and satellites (microsatellites, GSATX, HSAT5). From this file, we built the “rosetta” file (**Data S1**) which provided a correspondence between the 4886205 repeat copy sequence names and the repeat family names (1270 repeat families). Since we did not have access to the genome of the participants, we could not determine the location of each repeat copy in each individual which may vary from one to another. Therefore, the rosetta file allowed us to regroup counts per repeat copy into counts per repeat family, independently of their location in the genome. Of note, two TE copies belong to the same family if at least 80% of their sequences on at least 80% of their lengths is identical [29].

Alignment and read counting

Using the transcript sequences from the whole human reference transcriptome, gene expression was quantified with Kallisto software (v0.46.1) [30]. The --rf-stranded option was added to deal with reverse stranded RNAseq data in our dataset only. Alignment to the reference transcriptome resulted in high mapping efficiency (**Table S7**).

Alignment and read counting on repeat families from the human repeat reference sequence file were realized with TEcount (v1.0.0) software from TEtools (https://github.com/l-modolo/TEtools). This tool uses Bowtie2 [31] as the aligner and we added the --nofw option to deal with reverse stranded RNAseq data in our dataset only. TEcount uses also the "rosetta" file to sum the read counts per repeat copy into read counts per repeat family [32].

Statistical analysis

Statistical analysis was performed using the R software (v 3.6.3). TheGTEx dataset and our dataset were computed separately using the same procedure described below.

We considered as gene or TE "expression" the number of counts aligned on each reference transcript or TE sequence and as "differential expression analysis" the statistical procedure to compare gene or TE expression between several conditions (i.e. sex, age group, or karyotype).

Data preparation

Differential expression (DE) analysis was performed separately for genes and for TEs. Two files were prepared for each purpose.

As a same gene can have several transcripts, the tximport R package (v1.14.2, Bioconductor) was used to sum the read counts per gene transcripts into read counts per gene using the GTF file mentioned above. Then, a file (txi file) containing the gene identifier (Ensembl nomenclature), the average gene length, and the estimated counts for all genes per sample was obtained.

For DE analysis on TE, we applied on the repeat family counts a normalization factor calculated only on the gene counts to avoid masking relevant biological information. Indeed, if all the repeat families in a sample were overexpressed, this relevant biological information would be erased by the normalization factor if this one was only calculated on repeat family counts [32]. To estimate the normalization factor, the gene identifiers and counts per gene were extracted from the txi file and merged with the TEcount file containing the repeat family names and counts per repeat family per sample. The length could not be included in the merged table since this variable was not available from the TEcount table (each repeat family is a collection of copies displaying a diversity of lengths, it is thus difficult to estimate the length of a TE family). This merged table was then imported in the DESeq2 tool (command DESeqDataSetFromMatrix). From this step, we removed counts on the following 24 snRNA, snpRNA, scRNA, or remaining satellites and rRNA repeat families: 7SLRNA, 7SK, LSAU, D20S16, REP522, SATR1, SATR2, ACRO1, ALR/Alpha, BSR/Beta, CER, 6kbHsap, TAR1, SST1, MSR1, SAR, GA-rich, G-rich, A-rich, HY1, HY3, HY4, HY5,5S. This allowed us to only consider the 1246 families corresponding to TE families out of the 1270 repeat families contained in the rosetta file.

For the DE analysis on genes, as the length was available for genes, a more reliable procedure was performed. The raw txi file was directly imported in DESeq2 (command DESeqDataSetFromTximport) and the average length of each gene could be used as a correction term (offset) in the DESeq2 statistical model described below to consider the length bias [33,34].

Differential expression analysis

The R package DESeq2 R (v1.26.0, Bioconductor) [33,34] was used to normalize the read counts to remove any depth sequencing bias or bias related to sequence length (when possible), and to fit a negative binomial model on the log2 of the normalized counts (dependent variable) according to two independent variables: sex (factor with 2 levels: female, male) and age group (factor with 5 levels: [20,30],]30,40],]40,50],]50,60],]60,70]) for the GTEx dataset, or karyotype (factor with 4 levels: XX, XY, XXY, XYY) and batch (factor with 2 levels: first batch, second batch) for our dataset. Indeed, we found a significant effect of the sequencing batch in our dataset that had to be taken into account in the statistical model (**Fig. S12 and S13**).

The statistical model provided variation of expression of a gene or a TE family between two conditions expressed as log2 fold-change (log2FC) (i.e. log base 2 of the ratio of predicted normalized counts between the two conditions).

Log2FC were compared to 0 with a Wald test. A likelihood-ratio test (LR test) was also performed to assess whether adding an independent variable in the model provided relevant information compared to the null model, i.e. to test whether gene or TE family expression was globally impacted by an independent variable having more than two classes (age group or karyotype variables). A 0.05 False-Discovery Rate (FDR) threshold value was used for significance and p-values were adjusted for multiple-testing using the Benjamini & Hochberg procedure [35].

The thresholds for the outlier and the low mean normalized read count filters were set as default. Graphics were generated after the data transformations proposed by the DESeq2 package authors: variance stabilizing transformation for PCA, adaptive shrinkage for MA plots, log2 transformation of the normalized counts for heatmaps [33,36].

The command removeBatchEffect (for unbalanced batches) from the limma R package (v3.46.0) was used before PCA representations, heatmap representations and before building boxplots of normalized counts according to the karyotype to deal with batch effect in our dataset.

When necessary, gene name, gene description, and gene localization were determined using the GTF file and the command gconvert from the R package gprofiler2 (v0.1.9) [37] from the Ensembl gene (ENSG) identifier.

Further analysis

GO term enrichment analysis:

A gene ontology (GO) analysis was performed on significantly differentially expressed genes (DE genes) in the six pairwise karyotype comparisons (47,XYY vs 46,XX, 47,XXY, 46,XX, 46,XY vs 46,XX, 47,XYY vs 46,XY, 47,XXY vs 46,XY, 47,XYY vs 47,XXY) after adjustment on the batch effect. The R package gprofiler2 (v0.1.9) was used [37] with the g:SCS algorithm for multiple testing correction. The significance threshold was set at 0.05.

Comparison of the global expression of TE:

We made another level of comparison of TE expression which assessed the global TE expression rather than TE family expression separately. To do that, we summed all the normalized counts of each TE family per individual and then we divided by the total number of normalized counts on genes and TE families per individual. The Wilcoxon test was used to compare global TE expression between groups. P-values were corrected for multiple testing using the Holm method [38].

Consensus-specific quantification of TE expression without TE insertions in exons and long non-coding RNA

To test the robustness of our results, we re-analyzed our dataset with another procedure which also permitted to avoid the spurious or non-autonomous expression of TE from exonic sequences and long non-coding RNA sequences. To do that, we used a tool named TEspeX (v2.0.1) [39], with "--strand reverse" option to deal with reverse stranded RNAseq data. We provided four input files to TEspeX. The first was a file containing the consensus TE sequences (1179 sequences) generated from the Dfam database using the instructions here https://github.com/fansalon/TEconsensus, keeping only "LINE" "SINE" "DNA" "LTR" "RC" "Other" "Retroposon" repeat classes (except for 7SLRNA). The second and third files containing cDNA and long non-coding RNA sequences (fasta files), respectively, were downloaded from https://www.gencodegenes.org/human/ (content = "Protein-coding transcript sequences" and "Long non-coding RNA transcript sequences", regions = "CHR", Release 43 GRCh38/hg38), as suggested by the creators of the tool. As we worked with total RNAseq data, we took advantage of the "--mask" TEspeX option to also remove spurious TE expression from introns of genes carried by the X or Y chromosome. We did not remove expression from intronic insertions of autosomal genes to meet our computer resources. For this purpose, we downloaded the intron sequences of genes carried by the X or Y chromosome on the UCSC table browser https://genome.ucsc.edu/cgi-bin/hgTables?hgsid=814613547_qUmlEGan4K2f39cHvb9AgrKdRVdK&clade=mammal&org=Human&db=hg38&hgta_group=genes&hgta_track=knownGene&hgta_table=0&hgta_regionType=genome&position=chr1%3A11%2C102%2C837-11%2C267%2C747&hgta_outputType=primaryTable&hgta_outFileName=

with output format = “sequence” (fasta file), track “GENCODE V43” from the “Dec. 2013 GRCh38/hg38” assembly of the human reference genome and without masking repeat sequences, then we removed duplicate intronic sequences using seqkit rmdup command from seqkit tool (v2.4.0).

TEspeX provided a file containing the number of counts per TE consensus sequence for each individual as output.

Appendices

Supplementary material including Supplementary text, Figures S1 to S21, Tables S1 to S12.

Data S1. Rosetta file

Data S2. List of all differentially expressed genes on X chromosome or Y chromosome according to karyotype using LR test in our dataset, first regardless the gene type and then for protein coding genes.

Data S3. List of all differentially expressed genes (regardless their type) in each pairwise karyotype comparisons after batch adjustment in our dataset

Data S4. GO analysis on our dataset

Acknowledgements

We thank Yasmine Zerdoumi and Maxime Vallée for their help in the data generation and data analysis, and Véréna Landel (DRS, Hospices Civils de Lyon) for reading the manuscript.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The GTEx data used for the analyses described in this manuscript were obtained from: dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000424.v8.p2.

Data, scripts, code, and supplementary information availability

Fastq files of our dataset were submitted to European Genome-phenome Archive (EGA ID of the study: EGAS00001007462).

Scripts and code were described in detail in the Methods section but can be supplied upon request.

Conflict of interest disclosure

The authors declare that they have no financial conflicts of interest in relation to the content of the article.

Funding

This study was funded by an internal collaborative grant of LBBE and by ANR (grant number ANR-20-CE02-0015).

References

1. World Health Organization. Life expectancy at birth (years). Available: https://www.who.int/data/gho/data/indicators/indicator-details/GHO/life-expectancy-at-birth-(years)

2. Rochelle TL, Yeung DKY, Bond MH, Li LMW. Predictors of the gender gap in life expectancy across 54 nations. Psychology, Health & Medicine. 2015;20: 129–138. doi:10.1080/13548506.2014.936884

3. Willcox BJ, Willcox DC, Ferrucci L. Secrets of Healthy Aging and Longevity From Exceptional Survivors Around the Globe: Lessons From Octogenarians to Supercentenarians. The Journals of Gerontology: Series A. 2008;63: 1181–1185. doi:10.1093/gerona/63.11.1181

4. Luy M. Causes of Male Excess Mortality: Insights from Cloistered Populations. Population and Development Review. 2003;29: 647–676.

5. Maklakov AA, Lummaa V. Evolution of sex differences in lifespan and aging: Causes and constraints: Prospects & Overviews. BioEssays. 2013;35: 717–724. doi:10.1002/bies.201300021

6. Milot E, Moreau C, Gagnon A, Cohen AA, Brais B, Labuda D. Mother's curse neutralizes natural selection against a human genetic disease over three centuries. Nature Ecology & Evolution. 2017;1: 1400–1406. doi:10.1038/s41559-017-0276-6

7. Marais G, Gaillard J-M, Vieira C, Plotton I, Sanlaville D, Gueyffier F, et al. Sex gap in aging and longevity: can sex chromosomes play a role? Biology of Sex Differences. 2018;9: 33. doi:10.1186/s13293-018-0181-y

8. Marais G, Lemaitre J-F, Vieira C. Y chromosome makes fruit flies die younger. PCI Evol Biol. 2020; 100105. doi:10.24072/pci.evolbiol.100105

9. Lemaître J-F, Ronget V, Tidière M, Allainé D, Berger V, Cohas A, et al. Sex differences in adult lifespan and aging rates of mortality across wild mammals. Proceedings of the National Academy of Sciences. 2020;117: 8546–8553. doi:10.1073/pnas.1911999117

10. Pipoly I, Bókony V, Kirkpatrick M, Donald PF, Székely T, Liker A. The genetic sex-determination system predicts adult sex ratios in tetrapods. Nature. 2015;527: 91–94. doi:10.1038/nature15380

11. Xirocostas ZA, Everingham SE, Moles AT. The sex with the reduced sex chromosome dies earlier: a comparison across the tree of life. Biol Lett. 2020;16: 20190867. doi:10.1098/rsbl.2019.0867

12. Connallon T, Beasley IJ, MDonough Y, Ruzicka F. How much does the unguarded X contribute to sex differences in life span? Evolution Letters. 2022;6: 319–329. doi:10.1002/evl3.292

13. Tsurumi A, Li W. Global heterochromatin loss: A unifying theory of aging? Epigenetics. 2012;7: 680–688. doi:10.4161/epi.20540

14. Brown EJ, Nguyen AH, Bachtrog D. The Y chromosome may contribute to sex-specific ageing in Drosophila. Nature Ecology & Evolution. 2020;4: 853–862. doi:10.1038/s41559-020-1179-5

15. Wei KH-C, Gibilisco L, Bachtrog D. Epigenetic conflict on a degenerating Y chromosome increases mutational burden in Drosophila males. Nat Commun. 2020;11: 5537. doi:10.1038/s41467-020-19134-9

16. Delanoue R, Clot C, Leray C, Pihl T, Hudry B. Y chromosome toxicity does not contribute to sex-specific differences in longevity. Nat Ecol Evol. 2023 [cited 18 Jul 2023]. doi:10.1038/s41559-023-02089-7

17. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003;423: 825–837. doi:10.1038/nature01722

18. Human Genome Assembly GRCh38.p14. Available: https://www.ncbi.nlm.nih.gov/grc/human/data

19. Bojesen A, Juul S, Birkebæk N, Gravholt CH. Increased Mortality in Klinefelter Syndrome. The Journal of Clinical Endocrinology & Metabolism. 2004;89: 3830–3834. doi:10.1210/jc.2004-0777

20. Stochholm K, Juul S, Gravholt CH. Diagnosis and mortality in 47,XYY persons: a registry study. Orphanet Journal of Rare Diseases. 2010;5: 15. doi:10.1186/1750-1172-5-15

21. Oliva M, Muñoz-Aguirre M, Kim-Hellmuth S, Wucher V, Gewirtz ADH, Cotter DJ, et al. The impact of sex on gene expression across human tissues. Science. 2020;369: eaba3066. doi:10.1126/science.aba3066

22. Monkhorst K, Jonkers I, Rentmeester E, Grosveld F, Gribnau J. X Inactivation Counting and Choice Is a Stochastic Process: Evidence for Involvement of an X-Linked Activator. Cell. 2008;132: 410–421. doi:10.1016/j.cell.2007.12.036

23. Zhang X, Hong D, Ma S, Ward T, Ho M, Pattni R, et al. Integrated functional genomic analyses of Klinefelter and Turner syndromes reveal global network effects of altered X chromosome dosage. Proc Natl Acad Sci USA. 2020;117: 4864–4873. doi:10.1073/pnas.1910003117

24. Loda A, Heard E. Xist RNA in action: Past, present, and future. Bartolomei MS, editor. PLOS Genetics. 2019;15: e1008333. doi:10.1371/journal.pgen.1008333

25. Hassan FM, Razik HAA, Wadie MS, Abdelfattah DS. XIST and RPS4Y1 long non-coding RNA transcriptome as sex biomarkers in different body fluids. Egyptian Journal of Forensic Sciences. 2019;9: 16. doi:10.1186/s41935-019-0122-x

26. Skakkebæk A, Nielsen MM, Trolle C, Vang S, Hornshøj H, Hedegaard J, et al. DNA hypermethylation and differential gene expression associated with Klinefelter syndrome. Scientific Reports. 2018;8: 13740. doi:10.1038/s41598-018-31780-0

27. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45: 580–585. doi:10.1038/ng.2653

28. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

29. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8: 973–982. doi:10.1038/nrg2165

30. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;34: 525–527. doi:10.1038/nbt.3519

31. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

32. Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Research. 2016;45: e17. doi:10.1093/nar/gkw953

33. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010;11: R106.

34. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15: 550. doi:10.1186/s13059-014-0550-8

35. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological). 1995;57: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

36. Stephens M. False discovery rates: a new deal. Biostatistics. 2016;18: 275–294. doi:10.1093/biostatistics/kxw041

37. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Research. 2019;47: W191–W198. doi:10.1093/nar/gkz369

38. Holm S. A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics. 1979;6: 65–70.

39. Ansaloni F, Gualandi N, Esposito M, Gustincich S, Sanges R. TEspeX: consensus-specific quantification of transposable element expression preventing biases from exonized fragments. Mathelier A, editor. Bioinformatics. 2022;38: 4430–4433. doi:10.1093/bioinformatics/btac526