BMC Bioinformatics

SOFTWARE Open Access

MADloy: robust detection of mosaic loss of chromosome Y from genotype-array-intensity data

Juan R. González[1,2,3*], Marcos López-Sánchez[4,5,6], Alejandro Cáceres[1,2], Pere Puig[3], Tonu Esko[7] and Luis A. Pérez-Jurado[4,5,6,8,9]

*Correspondence: juanr.gonzalez@isglobal.org
[1] Barcelona Institute for Global Health (ISGlobal), 08003 Barcelona, Spain
Full list of author information is available at the end of the article

Abstract

Background: Accurate protocols and methods to robustly detect the mosaic loss of chromosome Y (mLOY) are needed given its reported role in cancer, several age-related disorders and overall male mortality. Intensity SNP-array data have been used to infer mLOY status and to determine its prominent role in male disease. However, discrepancies of reported findings can be due to the uncertainty and variability of the methods used for mLOY detection and to the differences in the tissue-matrix used.

Results: We created a publicly available software tool called *MADloy* (Mosaic Alteration Detection for LOY) that incorporates existing methods and includes a new robust approach, allowing efficient calling in large studies and comparisons between methods. *MADloy* optimizes mLOY calling by correctly modeling the underlying reference population with no-mLOY status and incorporating B-deviation information. We observed improvements in the calling accuracy to previous methods, using experimentally validated samples, and an increment in the statistical power to detect associations with disease and mortality, using simulation studies and real dataset analyses. To understand discrepancies in mLOY detection across different tissues, we applied *MADloy* to detect the increment of mLOY cellularity in blood on 18 individuals after 3 years and to confirm that its detection in saliva was sub-optimal (41%). We additionally applied MADloy to detect the down-regulation genes in the chromosome Y in kidney and bladder tumors with mLOY, and to perform pathway analyses for the detection of mLOY in blood.

Conclusions: *MADloy* is a new software tool implemented in R for the easy and robust calling of mLOY status across different tissues aimed to facilitate its study in large epidemiological studies.

Keywords: Loss of chromosome Y, SNP array, Bioconductor

Background

Mosaic loss of the Y chromosome (mLOY) is the most commonly reported large structural somatic event [1]. That is the loss of the entire male chromosome that is observed in dividing somatic cells. Recent evidence suggests the important role of mLOY in numerous diseases, being a biological factor that contributes to overall male mortality [2, 3] and, therefore, is likely to play an important role in male-specific treatments of disease. In particular, mLOY in blood cells increases with age and is associated with smoking and with the risk of several age-related disorders, including hematological and non-hematological cancers, macular degeneration [4], Alzheimer's disease [5], major cardiovascular events [6] and suicidal behaviors [7]. Despite the accumulating evidence, the mechanisms that trigger mLOY and its clinical consequences are still poorly understood. While susceptibility loci and epigenetic marks for mLOY have been identified [8, 9], larger and more accurate association and functional studies are required; in particular, as some inconsistencies have been observed [10, 11]. Discrepancies of reported findings are likely due to the uncertainty and variability of the methods used for mLOY detection and to the differences in the tissue-matrix used to detect it. Therefore, current mLOY calling methods and protocols need to be improved and compared to confirm disease risks and mechanisms. Here, we propose *MADloy* a software tool that incorporates previous methods and implements a novel and robust approach. The software allows comparisons across all current analysis strategies and can be readily used on standard data formats such as PennCNV. In addition, we used the novel approach to compare the overlap of detecting mLOY on different tissue-matrices.

A prominent source of discrepancy for the associations between mLOY and male diseases can be due to the differences in the methods used to call mLOY status from SNP array intensity data [10]. The main signal used for mLOY calling is the log-R-Ratio (LRR) in chromosome Y which is a normalized measure of the total signal intensity for two alleles of the SNP. As a relative measure of the DNA content of a subject at a genomic locus for a group of individuals, men with mLOY are expected to show low SNP-LRR values across the 56-Mb male-specific region of chromosome Y, which excludes the homologous region between chromosomes X–Y, pseudoautosomal (PAR1 and PAR2) and X–Y transposed (XTR) [12]. Computing the mean LLR (mLRR-Y) in the region for each individual, Fosrberg et al. called mLOY status on those individuals with mLRR-Y lower than the 99% confidence interval of experimentally induced mLRR-Y variation of normal individuals [2]. As a threshold dependent method, this approach, here named *mLRR-Y*$_{thresh}$, is sensitive to the characterization of subjects with no-mLOY. Assuming that gains of chromosome Y (GOY) are rare, the positive side of the mLRR-Y distribution (centered at the peak) can be used to identify mLOY outliers by reflecting it to the negative mLRR-Y values to define the calling threshold. However, while GOY events are less frequent than LOY events, they are relevant in tumor tissues. The presence of a few of them is enough to affect the mLOY threshold. In addition, the method strongly assumes a symmetrical mLRR-Y distribution, which if not true can also affect the position of the threshold. These two uncontrolled effects can introduce misclassification in mLOY calling, reducing the power to find positive associations [13]. Alleviation for mLOY misclassification in the association analyses can be considered if mLRR-Y is taken as a quantitative continuous variable (*mLRR-Y*$_{quant}$) or as a proxy of mLOY cellularity (*mLRR-Y*$_{cellularity}$) [14]. However, these approaches are still limited by the fact that the underlying distribution is a clear mixture of individuals with gains, losses, or no changes in chromosome Y genetic content, a similar situation found in copy number variation calling [15, 16]. Therefore, improvements on the use of mLRR-Y to call mLOY status should include the robust identification of mLOY events as outliers from an mLRR-Y non-symmetric distribution. In addition, current methods have not used the B-allele frequency (BAF) signal in the homologous region between chromosomes X–Y as an independent signal to confirm findings. BAF is a normalized measure of the allelic intensity tatio of two SNP alleles, taking the value of 0.5 for heterozygous and 0 or 1 for homozygous individuals at the SNP. The deviation of BAF signal in heterozygous probes from its expected value, namely B-deviation, has been robustly used to detect different types of chromosomal events in mosaicism and compute their cellularity levels [17]. Taking these issues into account, we have therefore implemented a new mLOY detection method in *MADloy* that integrates robust outlier identification of mLRR-Y values with the use of B-deviation signal. We show the improved performance of the method against current approaches using simulations and real data sets. *MADloy* has been implemented as a Bioconductor package for comprehensive mLOY calling along with visualization functions.

Differences in mLOY detection between blood and buccal smear have also been proposed as a possible source of discrepancy between two large studies that estimated the mortality ratios associated with mLOY [10]. To address this issue, we studied 18 individuals with positive mLOY status detected with *MADloy* in blood at baseline and assessed the progression of mLOY cellularity in both blood and saliva in a follow-up visit at 3 years, to determine the extent to which the tissue matrices are comparable for mLOY detection. We finally illustrate the application of *MADloy* in the study of the transcriptional correlates of mLOY in blood and cancer. We analyzed data from the cancer genome atlas (TCGA) for kidney cancer as is has been associated with LOY [18] and for bladder cancer, because LOY appears to be frequent in urothelial bladder cancer [19].

Implementation

mLOY calling

The reference methods for calling mLOY status using genotype-array intensity are those described in [2] and [9]. Forsberg et al. [2] proposed to analyze the log R ratio (LRR) values of SNPs probes in the male-specific region of chromosome Y (mLRR-Y) in the 56-Mb region between PAR1 and PAR2 on chromosome Y (chrY:2,694,521-59,034,049; hg19/GRCh37). An important consideration is the XTR that is shared between X and Y chromosomes is also removed from the analysis because XTR can be affected by alterations in chromosome X, then redefining the mLRR-Y region: chrY 6,611,498-24,510,581; hg19/GRCh37.

Assuming that mLRR values follow a symmetrical distribution across subjects, individuals are LOY-scored based on the threshold that is defined as the lower limit of the 99% confidence interval of the experimentally induced mLRR-Y variation. Forsberg et al. proposed to model the symmetrical distribution by reflecting the positive values of mLRR-Y over its median (method $mLRR\text{-}Y_{thres}$). LRR is computed as the ratio between two intensities and therefore its distribution is likely to be skewed- In addition, the presence of large mLRR-Y values representing XYY gains cannot be fully discarded. Wright et al. [9], therefore, proposed to use mLRR-Y as a continuous variable to measure the degree of mLOY given by its cellularity content (method $mLRR\text{-}Y_{quant}$). However, it is clear that mLRR-Y is a multimodal distribution and ignoring this feature reduces the power in association studies, as it has been seen when copy number variant status is estimated using continuous intensities.

As an alternative method, we proposed to use a robust estimation of the threshold that determines gains and losses in chromosome Y as outliers of the mLRR distribution. This threshold is estimated by

$$median(mLRR) + 1.2\,IQR(mLRR).$$

It is a common statistical practice to use IQR for robust detection of outliers, including a wide range of different distributions, even for those where symmetry is not observed. As the frequency of mLOY is typically low and their mLRR values are extreme within to the reference population, it produces a distant small peak outlying the mLRR's distribution core. The IQR helps to dissect that small peak with a standard statistical procedure. Fosbergs's approach is less standard, assumes symmetry and can lose power to detect associations between mLOY and disease. We set the limit to 1.2 to allow for 5% false discovery calls in the case of normal distributions. Therefore, if the IQR dichotomization is applied to a population without mLOY, we expect a maximum of 5% false discovery rate (FDR). If the distribution is less dispersed than the normal, as typically observed, the FDR is expected to be further decreased. The proposed method is implemented in MADloy together with mLRR-Y_{thres} [2], mLRR-Y_{quant} [9] and mLRR-$Y_{cellularity}$ [14].

To reduce the number of false-positive calls, we also propose to combine the information obtained of mLRR-Y with the B-deviation values in X–Y homologous regions which include pseudoautosomal regions PAR1, PAR2, and XTR regions [12] whenever this data is available. The reason to use B-deviation in addition to LRR is that chromosome Y alterations also alter these values in the X–Y homologous regions, due to the allelic imbalance between chromosomes X and Y. We assume an expected BAF of heterozygous probes between 0.45 and 0.55 and by setting a B-deviation threshold of 0.05, we can identify when the BAF is altered. This information can be used as a quality control criterion by identifying those samples classified as LOY or GOY by a mLRR-Y method but with normal values of B-deviation. In addition, samples without alterations in the mLRR-Y values but with alterations in B-deviation classified can provide evidence of technical problems in the processing of the samples. We recommend visually inspecting discordant samples c and remove cases with clear contamination and other non-LOY events. To this end plots for chromosome X and Y are implemented in the MADloy.

LRR, mLRR-Y, BAF and B-deviation

mLOY calling is performed on data obtained in PennCNV format for each individual. The format contains SNP, chromosome, position, LRR, BAF and genotype information. LRR and BAF for Illumina arrays (see EGCUT dataset below) can be obtained using GenomeStudio tool (https://www.illumina.com/techniques/microarrays/array-data-analysis-experimental-design/genomestudio.html). For Affymetrix arrays (TCGA dataset), LRR and BAF values are obtained using Affymetrixc Power Tools with the Birdseed v2 algorithm (https://media.affymetrix.com/support/developer/powertools/changelog/index.html) and following the PennCNV-Affy method described in PennCNV webpage (https://penncnv.openbioinformatics.org/en/latest/user-guide/affy/).

For each individual, we computed the 5% trimmed-mean of LRR in autosomes to avoid regions having copy number alterations. The percentage was changed to 25% in the case of cancer samples, where numerous copy number alterations are expected. We then computed the normalized median of LRR-Y (mLRR-Y) using the trimmed-mean of LRR intensity in the autosomes to remove systematic biases across the array. Then, the B-deviation is computed for each sample by selecting the BAF values in the X–Y homologous region (PAR1, PAR2 and XTR) for non-homozygous probes. These probes were selected by setting an upper and lower BAF threshold of 0.8 and 0.2 for normal and variant homozygous genotypes. B-deviation is then calculated as the trimmed mean of the absolute difference between experimental BAF values and 0.5 (i.e. expected value) of the heterozygous probes.

Quality control

We performed quality control of the samples involved in the analyses by removing those samples with large LRR variability to avoid additional variability likely due to technical artifacts. We followed the Illumina manufacturer's recommendation of filtering samples with high variable LRR in autosomes (standard deviation of LRR > 0.28). This criterion was used for EGCUT dataset. For TCGA and NIA samples, that were genotyped using Affymetrix, we considered individuals with a normal variability of LRR those within 2 times the LRR standard deviation of autosomes.

MADloy software

We created *MADloy* (https://github.com/isglobal-brge/MADloy), an R package that automates mLOY detection for association studies implementing the methods previously described [20]. The package has been submitted to Bioconductor. The core functions include a pipeline to normalize data, perform quality control and summarize data from PennCNV format containing genome-wide information for LRR values. *MADloy* calls mLOY and gains of chromosome Y (XYY) using current methods. Visualization functions are also implemented to help inspect the LRR and BAF values of each sample. A vignette can be found at https://github.com/isglobal-brge/MADloy/tree/master/vignettes.

Experimental validation

Ten DNA samples belonging to EGCUT cohort with different proportions of mLOY detected by MADloy were validated using two Multiplex Ligation Probe-dependent Amplification (MLPA) panels, P070 (MRC-Holland Amsterdam, The Netherlands) covering all subtelomeric regions including the two pseudoautosomal regions (PAR1 and PAR2). Probes targeting the Y chromosome, and a custom-made panel with probes for *SRY* and several autosomal loci, were used to assess the copy number status of chromosome Y in relation to the control loci (autosomal and X chromosome). The MLPA reactions were performed as previously described [21] with some modifications for custom probe selection [17]. We used the relative peak height (RPH) method recommended by MRC-Holland. Non-mosaic losses and gains were expected at relative peak height between (0.5, 1.5) for the pseudoautosomal regions (normally disomic), and between (0, 2) for the Y-unique regions (normally monosomic).

Statistical analyses

Association studies between mLOY and age and cancer were performed in EGCUT and TCGA data using generalized linear models (Gaussian and binomial links, respectively). Association between cancer and mLOY in EGCUT data were adjusted by age by adding this variable in the model. The analysis of TCGA data included a random effect model for each individual since we compared normal vs tumor samples. In that case, the association with cancer did not require any adjustment for age given the paired nature of the data. Transcriptome data belonging to EGCUT data were obtained using HTA 2.0 microarray. Transcriptomic data for TCGA samples were obtained from RNA-sequencing available at RTCGA package. Both analyses were performed using *limma* package and the *voom* method was used to obtain continuous data from the RNAseq experiments of TCGA [22]. Models were adjusted for surrogate variables using *sva* package to control for experimental differences across samples [23]. Enrichment analyses were performed with *GOstats* Bioconductor library [24].

Simulation studies

We performed several simulations to determine the power of the association analyses considering mLOY status as continuous (Wright's method: mLRR-Y_{quant}), as categorical (Forsberg's: mLRR-Y_{thres} and MADloy's novel methods). We considered two main scenarios to assess the power to detect significant associations with continuous (age) or categorical (case/control) outcomes. Data were simulated using functions from the CNVassoc package which simulates gains and losses from continuous data (e.g. LRR), considered as a surrogate variable of mLOY [25]. LRR continuous data was generated using the mean and standard deviation of LRR in normal and LOY cases, and the standard deviation of the outcome using the values observed in EGCUT data. We aimed to simulate the real situation of analyzing the correlation of mLOY with case–control status or age. The effect size varied from 1 to 8, representing age changes for one unit change of mLOY. The sample size also varied (N = 200, 300, 500, 750, 1000, 1500). For categorical outcomes, the magnitude of the effect was measured as an odds ratio (OR). We simulated different scenarios with varying mean effects (beta = 1.5, 2, 2.5, 3, 3.5, 4) and ORs (1.5, 1.75, 2, 2.5, 3, 3.5, 4).

EGCUT data

SNP data of 682 individuals from three platforms: OmniX Human370CNV and Metabochip arrays were randomly selected from the Estonian Gene Expression Cohort (EGCUT, www.biobank.ee). Selected individuals were older than 18 years of age (mean age 51.4 ± 18.2 years). EGCUT comprises a large cohort of about 52,000 samples of the Estonian Genome Center Biobank, University of Tartu [26]. Data were genotyped using HumanCoreExome array and all the individuals included in the analysis had a genotyping success rate above 95%. Cryptic relatedness was tested with the PLINK v1.07 software. Only one of each detected relative pair (up to second cousins) was randomly chosen for the detection of genetic mosaicism. Sample mix-ups were corrected using MixupMapper [27]. All studies were performed following the ethical standards of the responsible committee on human experimentation, and with proper informed consent from all individuals tested. LRR and BAF were generated using GenomeStudio software.

TCGA data

We analyzed 346 paired solid tumor (n = 103), normal tissue (n = 126) and normal blood (n = 117) male samples of Kidney Renal Clear Cell Carcinoma (KIRC) and Bladder Urothelial Carcinoma (BLCA) belonging to The Cancer Genome Atlas (TCGA) project. Raw data obtained from the Affymetrix Genome-Wide Human SNP Array 6.0 chip were processed with Birdseed v2 algorithm. Clinical data were also downloaded to select male samples and for association analyses. RNAseq data was obtained from RTCGA.rnaseq Bioconductor package that provides counts of 20,532 annotated genes these two tumors.

Results

Improved detection of mLOY status

We developed *MADloy*, a bioinformatics tool available as a Bioconductor package, which implements accurate mLOY calling of SNP intensity data using mLRR-Y and B-deviation information across subjects, together with other state-of-the-art methods (https://www.github.com/isglobal-brge/MADloy). For each sample, MADloy first estimates the normalized mLRR-Y given by its ratio with the trimmed-mean of mLRR-Y values in the autosomes. The method uses a 5% trimmed-mean to avoid regions having copy number alterations. Calling using mLRR-Y information is given by the identification of unexpected negative values of the mLRR-Y distribution, centered by the expected chromosome Y ploidy for a man (− 0.45, one copy), that is lower than 1.2 times the distance of interquartile range. This distance contains ~ %95 of the area under a normal distribution. The interquartile range is robust in the presence of outliers and therefore is a suitable measure for simultaneous mLOY and GOY calling. The calling is complemented with B-deviation information in the PAR1/PAR2 regions to reduce the number of false positives of mLOY calls and to estimate their cellularity (see subsection "LRR, mLRR-Y, BAF and B-deviation" in the Implementation Section). We then examined whether this new calling method improved calling accuracy and the statistical power in association studies.

We first compared *MADloy* and *mLRR-Y*$_{thresh}$ calling methods on 682 males from the Estonian Gene Expression Cohort (EGCUT, www.biobank.ee). Thirty-eight individuals (5.6%) were discarded due to large mLRR-Y variability), see Methods. Figure 1a shows the ploidy-centered 5% trimmed mLRR-Y values for each subject with their respective calling status using *MADloy*. Figure 1b, c illustrate the expected pattern of mLRR-Y and BAF of a normal sample, where the mLRR-Y (blue line) is not far from the expected ploidy value that corresponds to one copy (− 0.45) (orange line) and the BAF along the PAR1 (vertical blue rectangle) and PAR2 (vertical yellow rectangle) regions is centered at 0.5 (right Y-scale). Figure 1b, c show the expected pattern of one sample with mLOY, for which mLRR-Y is lower than the ploidy value and a large BAF split is observed. The expected value of the B-deviation for heterozygous probes is 0.5 when no mLOY is present. The size of the split is proportional to the cellularity and is measured by B-deviation defined as the absolute difference between BAF values. Our results indicate that MADloy is able to detect mLOY with a cellular fraction larger than 5% having a very low sensitivity below this threshold.

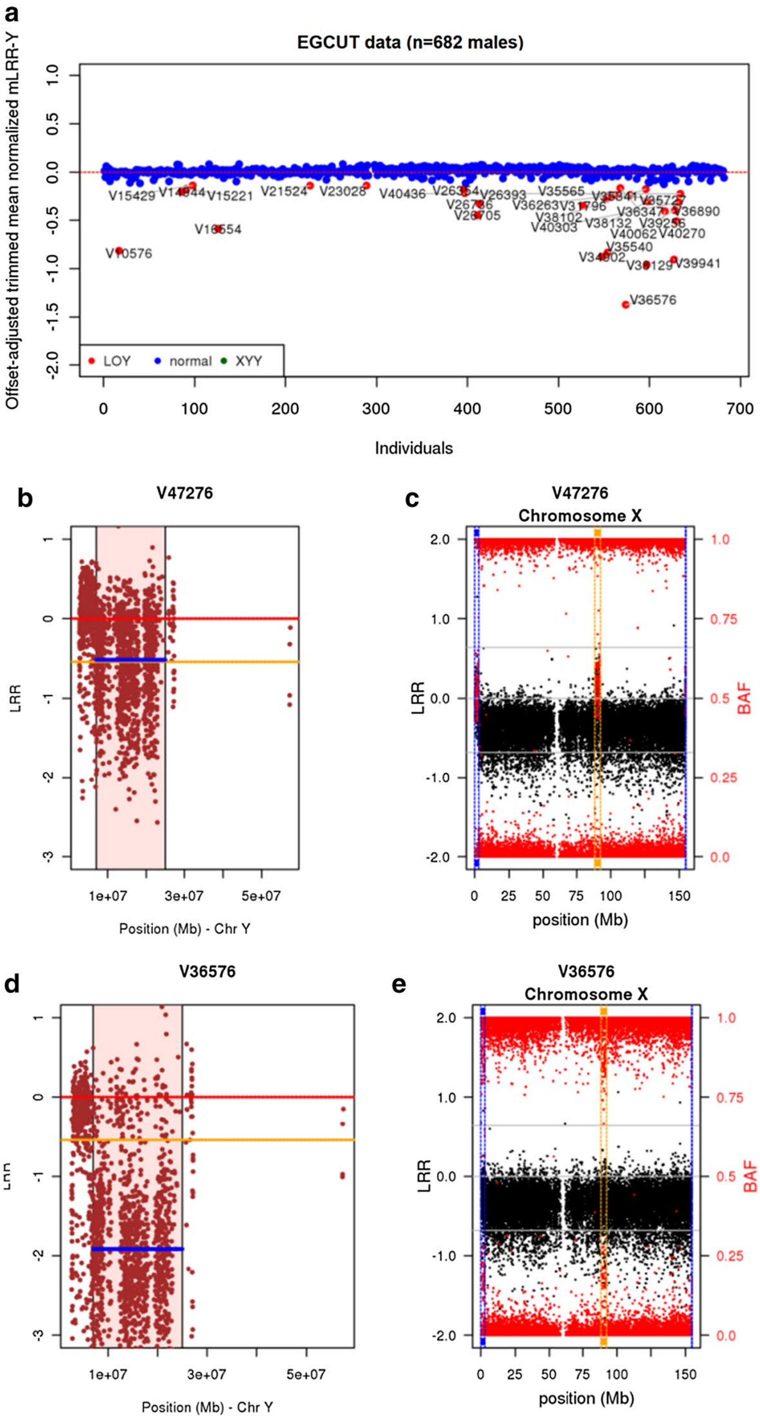

Fig. 1 *MADloy* detection in EGCUT data. **a** shows mLOY calling performed using mLRR-Y data by considering our proposed method to robustly detect outliers by considering the possibility of having GOY samples. **b** and **c** show the expected behavior of mLRR-Y and BAF data for a normal sample. **d** and **e** depict the expected values of mLRR-Y and BAF data of a mLOY sample. We observe a clear decrease of mLRR-Y (blue line) with respect to the expected value (yellow line) (**d**) and a split BAF in PAR1 (blue box) and XTR (yellow box) regions

Using *MADloy*, we detected a total of 30 (4.3%) individuals with X–Y allelic imbalance consistent with decreased Y chromosome dosage (Fig. 1a). Using the *mLRR-Y*$_{thresh}$ method, we called 56 samples with mLOY. A main reason for the difference is that the positive distribution of mLRR-Y is less variable than the negative part breaking the symmetry assumption and increasing the negative threshold at which mLOY is called by *mLRR-Y*$_{thresh.}$ As a result, more mLOY individuals are called. Additional file 1: Table S1 shows the discordant calls using *MADloy* and *mLRR-Y*$_{thresh}$. We investigated the different sources of discordance at the *mLRR-Y*$_{thresh}$ calling threshold. For instance, we observed that while V39233 has a lower mLRR-Y value than *mLRR-Y*$_{thresh}$, there is no BAF split in PAR1 and PAR2 regions (Additional file 1: Figure S1A–B). In addition, we observed that there are samples with low mLRR-Y due to other causes than mLOY. For instance, Additional file 1: Figure S1C–D shows a case with a BAF split (chromosome X) and LRR values (chromosome Y) consistent with the presence of an additional pair of XX chromosomes, in addition to the XY chromosomes of the individual. These additional X chromosomes are likely due to contamination of the sample with ~ 26% of DNA from another sample, caused by a pipetting error or less likely due to true chimerism.

To increase the experimental support of MADloy calling, we validated by MLPA, the positive mLOY status, as detected by MADloy, of 10 selected individuals from EGCUT. The results were fully concordant in all ten individuals. Figure 2a shows the LRR and BAF of probes across chromosomes Y and X for one individual. Figure 2b illustrates the validation of two mLOY cases (V32199 and V40611) with different cellularity (30% and 52%, respectively). We observed that for individuals with mLOY the relative peak heights (RPH) of Y probes were clearly reduced compared to the RPH of the X probe (Fig. 2b). We also observed a strong correlation with decreased expression of chromosome Y genes in samples from the same tissue which is also an indirect validation, while we are aware that being an indirect method a level misclassification could be expected.

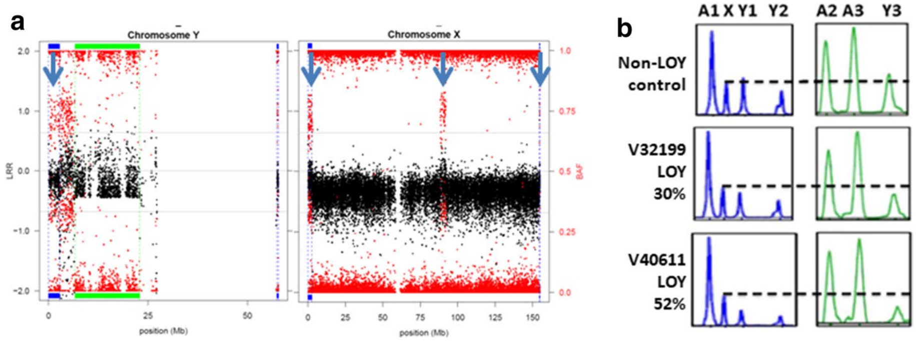

Fig. 2 Experimental mLOY validation. **a** LLR (black) and BAF (red) values of probes along chromosomes Y and X for an individual with positive mLOY status. Low values of LLR-Y are observed except for the PAR1 and PAR2 regions where a clear BAF split is observed. The BAF split is confirmed in the X chromosome. **b** MLPA validation of two individuals with mLOY with different cellularity content, compared with a normal individual (top). Two MLPA panels are shown for each individual, showing relative peak height (RPH) for probes in autosomes (A1, A2, A3) and in the X and Y chromosomes (Y1, Y2, Y3). Marking the peak for the chromosome X probe (dotted line), the figure shows a clear reduction in the RPH values of Y probes for the individuals with mLOY

To address the differences in mLOY detection across different tissues, we asked the extent to which mLOY status in blood could be detected in the buccal smear. 18 individuals from EGCUT, who were detected with mLOY using *MADloy* in blood, were randomly selected and re-contacted after 3 years to assess the progression of mLOY in blood and evaluate its detection by *MADloy* in buccal smear (Table 1). We observed that mLOY in blood persisted in time (Table 1). Detected from BAF signals, the estimated proportion of cells with mLOY increased in 16 cases (88%) while it decreased in 2 (12%). In saliva, we discarded one individual for the low quality of the sample and observed mLOY in only 7 individuals (41%). All cases of mLOY in saliva showed lower cellularity than mLOY in blood at followup. Using the same detection methods and procedures, we therefore confirmed that the detection power of mLOY in blood is substantially decreased if assessed in buccal smear.

mLOY improves the statistical power of association studies

Using a series of simulations, we first compared *MADloy* performance with mLRR-Y_{thresh} and mLRR-Y_{quant}. The simulations were performed using an independent simulator for copy number variation, CNVassoc assuming that the distribution of normal and mLOY cases, with the proportions observed for in the EGCUT study, followed those observed for CNVs. We studied the power to detect true associations between mLOY and a quantitative trait. mLOY calling was performed with MADloy and Fosberg's methods and associations were also assessed with mLRR-Y as a quantitative variable (Additional file 1: Figure S2). We found that *MADloy* had the largest statistical power in all

Table 1 Cellularity evolution of LOY detected in blood of 18 individuals after 3 years

Sample	Baseline detection		Follow-up at 3 years	
	Blood DNA	% cells	Blood DNA (% cells)	Saliva DNA (% cells)
V15429	Y loss	27	30	ND
V16554	Y loss	52	67	57
V16763	Y loss	41	50	ND
V19232	Y loss	41	53	ND
V22330	Y loss	52	46	ND
V23235	Y loss	37	48	27
V26736	Y loss	42	49	ND
V31014	Y loss	46	48	30
V31220	Y loss	29	39	17
V32199	Y loss	30	40	ND
V32632	Y loss	46	82	ND
V32752	Y loss	32	35	ND
V32850	Y loss	33	43	ND
V34568	Y loss	37	44	ND
V37288	Y loss	31	44	ND
V40611	Y loss	52	29	32
V44342	Y loss	35	37	20
V47558	Y loss	37	39	18

At follow-up mosaicism was also tested in saliva (*ND* not detectable) scenarios while mLRR-Y$_{\text{quant}}$ was the least powerful. In particular, we observed that an association of magnitude 2 per quantitative trait unit was detected with using *MADloy* with a power of 80% in 1,000 individuals while for *mLRR-Y*$_{thresh}$ a magnitude of >2.2 was required for the same power. We also observed similar results for dichotomous traits in case–control studies (Additional file 1: Figure S3). The differences in power were given by a lack of symmetry in the mLRR-Y distribution that resulted in misclassification by *mLRR-Y*$_{thresh}$. In addition, the simulations confirmed that treating mLRR-Y as quantitative when the underlying distribution is a mixture is a suboptimal approach.

We then compared in real setting the associations with age and cancer risk with mLOY called with MADloy and *mLRR-Y*$_{thresh}$ in EGCUT. In this case, we also aimed to evaluate the effect incorporating B-deviation signal in *MADloy* calling. We confirmed the association between mLOY and age. We found that mLOY called by *MADloy* with B-deviation showed the strongest and most significant association with age (beta $=21.21$, $P=2.0\times10^{-10}$). However, when mLOY status was called using mLRR-Y information only, either by *MADloy* or by *mLRR-Y*$_{thresh}$, the effect of the association was lower although these differences were not statistically significant (Fig. 3a). Note that the largest impact in the association was when considering *mLRR-Y*$_{thresh}$ method, suggesting again that the symmetry assumption is less accurate than the interquartile range to define the mLOY calling threshold. We observed analogous results for the association

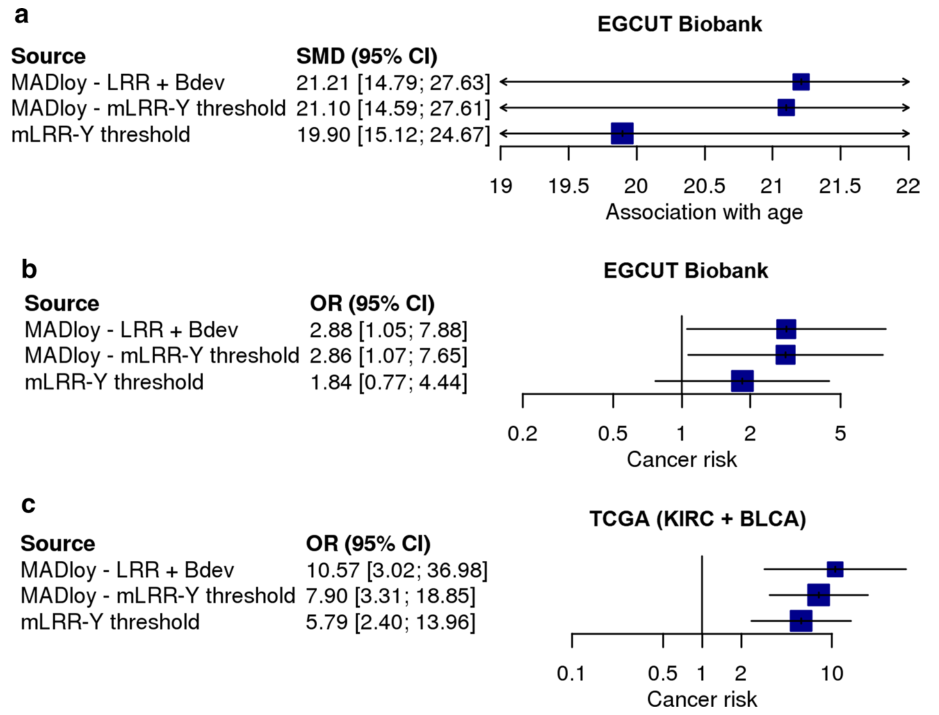

Fig. 3 Effect of mLOY misclassification. Effect estimates of three different association studies using three different calling methods. Panel **a**) corresponds to the analysis of a continuous phenotype (age) in EGCUT Biobank; panel **b**) is for a binary phenotype (cancer) in EGCUT Biobank; and panel **c**) for TCGA study. Two of them use only LRR information and two different threshold methods and one incorporates the information about B-deviation (Bdev). We observe as the effect is far from the null when using MADloy with both LRR and Bdev

between mLOY and cancer in EGCUT. For MADloy with and without B-deviation we observed similar significant increased risks of any tumor in models adjusted by age (OR 2.89, $P = 0.0390$, OR 2.86, $P = 0.0367$).However, for *mLRR-Y*$_{thresh}$, we did not find statistically significant differences (OR 1.84, $P = 0.1720$), confirming that the loss of power by methods that do not consider the possibility of having GOY nor the B-deviation can affect statistical significance (Fig. 3b).

We also assessed the different mLOY calling procedures in two cancer studies. We studied the differences in frequency of mLOY status between cancer and constitutional tissues of 346 samples from The Cancer Genome Atlas (TCGA) project. The samples belong to Clear Cell Carcinoma (KIRC) and Bladder Urothelial Carcinoma (BLCA) studies. The genotype data comprised 121 constitutional and 103 tumor samples. We first observed that *MADloy* detected *29* samples with mLOY (8.38%) while 34 were detected by *mLRR-Y*$_{thresh}$. Thus, *mLRR-Y*$_{thresh}$ shows a more liberal approach to mLOY detection, even under the presence of 2 GOY samples (0.6%) that are considered as part of the reference population. However, the GOY samples were called by *MADloy* showing the expected increase in mLRR-Y together with a clear BAF split in the PAR1 and PAR2 regions of chromosome X (Additional file 1: Figure S4).

We then tested the association with cancer status. We observed again that the strongest association was when mLOY was called with the full *MADloy* method (OR 10.6, $P = 4.65 \times 10^{-6}$). When mLOY was called with *mLRR-Y*$_{thresh}$, associations were also significant but their P values were higher in one order of magnitude ($P = 1.48 \times 10^{-5}$ and $P = 1.59 \times 10^{-5}$, respectively), in line with previous results and giving further evidence that more efficient use of the SNP intensity signals in mLOY calling can increment the power to detect associations with phenotypes (Fig. 3b). We also tested the association of tumor status with mLOY as a continuous variable (beta = − 2.32, $P = 9.2 \times 10^{-5}$), and with the cellularity of mLOY obtained by mLRR-Y$_{\text{cellularity}}$ (15) (beta = 0.98 $P = 1.59 \times 10^{-5}$). We, therefore, observed that while all results were consistent, the new method implemented in MADloy was the most powerful (e.g. showed the most significant p value).

Application of MADloy on finding transcriptomic signatures of mLOY

Differences in statistical power become crucial when studying the mechanisms underlying mLOY using genomic and transcriptomic data, as few misclassifications can lead to not detecting a relevant genome or transcriptome-wide association. We, therefore, applied *MADloy* to comparatively analyze transcriptomic data in cells from individuals with and without mLOY in blood in the EGCUT study and in cancer tissues in the TCGA study.

We searched for transcription correlates with mLOY in blood using the EGCUT study. We analyzed microarray expression data on 682 individuals (compared with age-matched controls with no mLOY) and observed 49 deregulated genes p value < 10^{-3}, five of which were significant with p value < 10^{-5}. These included *MED16*, *HNRNPKP1*, *TXNDC17*, *TMEM154*, and *CSF2RA*, the only gene in chromosome Y (PAR1). An enrichment analysis of the 49 selected genes using KEGG and GO databases showed several KEGG pathways that were significantly enriched in differentially expressed genes (Table 2). Meaningful pathways including DNA replication (p value = 0.004), mismatch repair (p value = 0.02) and homologous

Table 2 Enrichment analysis of blood transcriptomic data in EGCUT samples

Count	Size	FDR	OR enrich	Term	Genes
3	32	0.0039	10.76	DNA replication	MCM4, RPA1, RPA3
2	22	0.0204	10.18	Mismatch repair	RPA1, RPA3
3	64	0.0264	5.08	Antigen processing and presentation	CD4, HSP90AA1, NFYC
2	26	0.0280	8.47	Homologous recombination	RPA1, RPA3
3	71	0.0345	4.55	Hematopoietic cell lineage	CD4, CSF2RA, ITGA2B
3	81	0.0480	3.95	Ribosome	RPS15, RPS24, RPSA
1	5	0.0497	24.94	Vitamin B6 metabolism	PSAT1

Table includes KEGG categories that are over-enriched using differentially expressed genes at false discovery rate (FDR) lower than 0.05. Only terms with size ≥ 5 are shown

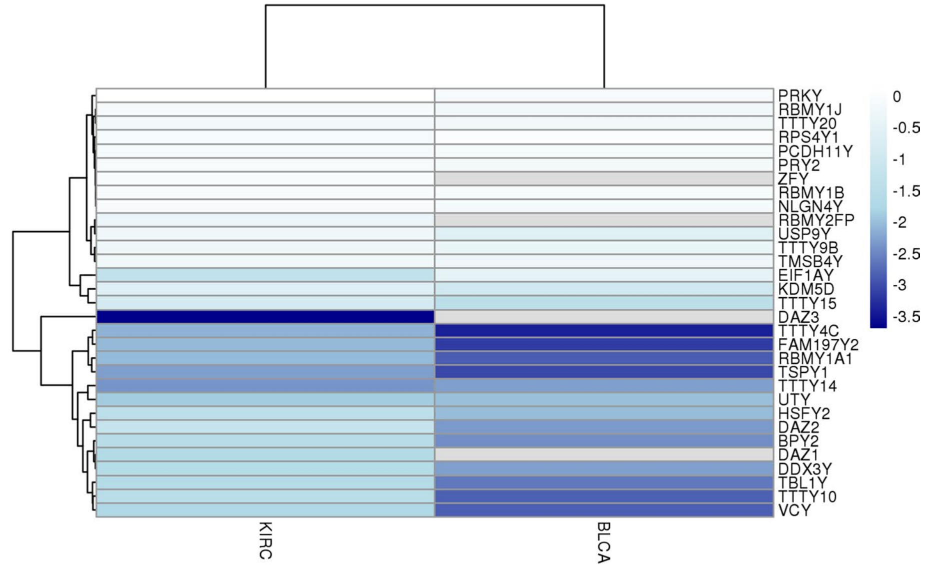

Fig. 4 Transcriptomic effects in tumor samples (TCGA). Downregulation of genes in chromosome Y when comparing samples with and without mLOY in KIRC and BLCA datasets of TCGA project

recombination (p value = 0.02) were found significant. Interestingly, only two genes located in the Y chromosome (*CSF2RA* and *EIF1AY)* were among the significantly deregulated gene in blood.

Transcriptome-wide association analysis of mLOY in TCGA was also performed with RNA-seq data in two different tumor datasets, the KIRC (Kidney renal clear cell carcinoma) and BLCA (Bladder urothelial carcinoma) studies. In both tumors, we found that 8 of the top-10 down-regulated genes by mLOY were in chromosome Y: *TTTY4C, UTY, TMSB4Y, USP9Y, ZFY, EIF1AY, RSP4Y1* and *TTTY15* (adjusted p values lower than 1.3×10^{-28}) (Fig. 4). This strong correlation with expression is fully supportive of the consistency of mLOY calling by *MADloy*, and indicates that mLOY in tumors leads to a detectable drop in transcription or extreme down-regulation of several genes across the chromosome Y.

Discussion

mLOY is emerging as an important marker of disease for aging men. Forthcoming studies will reveal the extent to which this biomarker can be used to design specific treatment to men. Here, we show that mLOY detection and quantification can be optimized with better analytical methods and by clearly defining the tissue matrix of interest. *MADloy* improves previous methods by using standard statistical procedures to call mLOY instead of more elaborated constructs that can reduce the power to detect associations. In addition, the incorporation of BAF-signal increases the accuracy of mLOY detection in samples with higher uncertainty.

Meanwhile, substantial differences in mLOY calling arise from the variety of in-house methods used. Here, we have implemented software in Bioconductor's framework that incorporates all the current methods described in the literature together with a new robust approach. We show that significant improvements in the detection accuracy of mLOY status from SNP array intensity data can be obtained by using a common matrix-tissue and optimization of the calling methods, by improving the calling threshold in mLRR-Y signal and accounting individuals with chromosome Y gains. Previous methods assumed symmetry in the mLRR-Y distribution and included individuals with gains in the reference population, affecting the threshold to detect mLOY status [2]. We observed that individuals with GOY are detectable, in particular in tumor tissues. Therefore, not accounting for asymmetry in the distribution can increase mLOY false positives, reducing the statistical power in the associations due to misclassification. Interestingly we also observed that treating mLRR-Y as a continuous variable did not substantially improve the associations. This is likely due to the multimodal nature of its distribution where individuals fall sharply into the normal or mLOY categories.

An additional gain in detection accuracy was observed by an independent analysis of the B-deviation signal [17]. While improvements were modest for the associations with age and cancer risk for mLOY in blood, we observed substantial increments in cancer tissues. The inclusion of this B-deviation as an additional signal in mLOY calling allowed us to detect instances where samples are likely contaminated and to confirm individuals with GOY. We have shown that samples classified as mLOY using mLRR-Y with discordant B-deviation values should be visually inspected to define contamination or other non-LOY events. Removing these cases from the analyses in a quality control phase will reduce misclassification errors and improve statistical power. In addition, greater accuracy of mLOY detection is needed if mLOY is to be considered as an individual marker of disease risk.

It has been suggested that discrepancies in associations between mLOY and disease risk or mortality could also be related to the different tissue sources of the DNA, despite significant correlation of mLOY calling in blood and buccal derived DNA from the same individuals [10]. We show here that the cellularity of mLOY in blood tends to increase with time, but the concordance with mLOY in buccal smear is only 41%, with lower cellularity in buccal samples in most cases. Therefore, different tissue sources could lead to differences in mortality ratios and other associations to mLOY, as previously reported [10]. On the other hand, the presence of mLOY across different tissues does suggest a common origin in each individual, consistent with a genetic predisposition that has also been documented by genome-wide association studies [8, 9].

Interestingly, the transcriptomic associations of mLOY are quite different in tumor samples from non-tumor blood samples. In kidney and bladder tumors, extreme downregulation of the entire Y is the main finding, a logical consequence of the decreased number of chromosome Y copies. However, in blood, only two Y chromosome genes are among the most deregulated ones, and the deregulation of autosomal genes mainly affects DNA replication, mismatch repair and homologous recombination pathways. This finding could be related to an increasingly oligoclonal leukocyte cell population in people with mLOY, which is consistent with the possibly compromised immune cell function of circulating leukocytes in people with mLOY as the risk factor for disease [2]. Mosaic gains and losses in chromosome Y can cause cancer by the downregulation of key genes in chromosome Y. It has been shown that overall decrease of chromosome Y transcript levels is a key element in the susceptibility to disease [28].

Conclusions

The mLOY is increasingly recognized as a male-specific risk factor for multiple diseases and, therefore, its interest in personalized treatment and risk management is likely to increase in the coming years. Specific disease and mechanistic studies require a robust estimation of the mLOY status of individuals. SNP intensity data is already available in large epidemiological studies where the effects of environmental conditions can also be investigated. The ability to call mLOY reliably using these data will increase the reproducibility of the findings. We show that *MADloy* is able to replicate the associations of mLOY in blood with cancer and aging while increasing statistical power obtained by current approaches. The method and the bioinformatics tool presented in this work are easy to use and scalable to large studies. Therefore, with *MADloy* we aim to facilitate the comparison between studies with a large number of individuals to better define the role of mLOY in complex diseases and its underlying mechanisms. *MADloy* allows the re-analysis of thousands of existing GWAS data in publicly available repositories.

Availability and requirements

Project name: MADloy.
Project home page: https://github.com/isglobal-brge/MADloy.
Operating system: Platform independent.
Programming language: R.
Other requirements: none.
License: GNU.
Any restrictions to use by non-academics: license needed.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12859-020-03768-z.

Additional file 1. Supplementary tables and figures.

Abbreviations

LOY: Loss of chromosome Y; GOY: Gains of chromosome Y; mLOY: Mosaic loss of chromosome Y; LRR: Log-R ratio; BAF: B-allele frequency; LRR-Y: Log-R ratio in chromosome Y; mLRR-Y: Median Log-R ratio in chromosome Y; PAR1:

Pseudoautosomal region 1; PAR2: Pseudoautosomal region 2; XTR: X–Y transposed region; EGCUT: The Estonian Biobank is a population-based biobank of the Estonian Genome Center of the University of Tartu; TCGA: The cancer genome atlas project; KIRC: Kidney renal clear cell carcinoma; BLCA: Bladder urothelial carcinoma; MLPA: Multiplex ligation probe-dependent amplification; mLRR-Y_{thresh}: Method for LOY detection proposed by Forsberg et al.; $mLRR_{quant}$: Method for LOY detection proposed by Wright et al.; $mLRR_{cellularity}$: Method for LOY detection proposed by Danielsson et al..

Acknowledgments
Not applicable.

Author contributions
JRG, ML, AC wrote the manuscript. LAPJ, PP and TE revised the manuscript. JRG, PP, AC proposed the statistical models. JRG and ML implemented the software. JRG and ML performed statistical analyses. TE provided data from the Estonia study. JRG and LAPJ supervised the project. All authors read and approved the final manuscript.

Funding
This research has received funding from Ministerio de Ciencia, Innovación y Universidades (MICIU), Agencia Estatal de Investigación (AEI) and Fondo Europeo de Desarrollo Regional, UE (RTI2018-100789-B-I00) also through the "Centro de Excelencia Severo Ochoa 2019–2023" Program (CEX2018-000806-S); and the Catalan Government through the CERCA Program. LAPJ lab is funded by the Catalan Department of Economy and Knowledge (SGR2014/1468, SGR2017/1974 and ICREA Acadèmia), and also acknowledges support from the Spanish Ministry of Economy and Competitiveness "Programa de Excelencia María de Maeztu" (MDM-2014-0370).

Availability of data and materials
The software described in the current study is available in the gitHub repository, https://github.com/isglobal-brge/MADloy**.** Data from the TCGA analyzed in the current study is available at the Bioconductor's repository through the RTCGA package (https://bioconductor.org/packages/release/bioc/html/RTCGA.html). Simulated and experimental data are included in this published article.

Ethics approval and consent to participate
The Estonian Biobank is a population-based biobank of the Estonian Genome Center of the University of Tartu (EGCUT). The project is conducted in accordance with the Estonian Gene Research Act (www.biobank.ee) and all subjects have been recruited randomly, on a voluntary basis by general practitioners and physicians in hospitals. All subjects provided written informed consent prior to participation and the approval for the study was granted by the Ethics Review Committee on Human Research at the University of Tartu. All other data used in the study is freely available and their consent to participate is available in their web page projects.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1] Barcelona Institute for Global Health (ISGlobal), 08003 Barcelona, Spain. [2] Centro de Investigación Biomédica en Red en Epidemiología Y Salud Pública (CIBERESP), Madrid, Spain. [3] Department of Mathematics, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain. [4] Genetics Unit, Universitat Pompeu Fabra, Barcelona, Spain. [5] Institut Hospital del Mar D'Investigacions Mediques (IMIM), Barcelona, Spain. [6] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain. [7] Estonian Genome Centre Science Centre, University of Tartu, Tartu, Estonia. [8] SA Clinical Genetics, Women's and Children's Hospital, Adelaide, Australia. [9] South Australian Health and Medical Research Institute, University of Adelaide, Adelaide, Australia.

Received: 30 October 2019 Accepted: 20 September 2020
Published online: 23 November 2020

References

1. Loftfield E, Zhou W, Graubard BI, Yeager M, Chanock SJ, Freedman ND, Machiela MJ. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. Sci Rep. 2018;8:1–10.
2. Forsberg LA, Rasi C, Malmqvist N, Davies H, Pasupulati S, Pakalapati G, Sandgren J, Diaz de Ståhl T, Zaghlool A, Giedraitis V, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat Genet. 2014;46:624–8.
3. Loftfield E, Zhou W, Graubard BI, Yeager M, Chanock SJ, Freedman ND, Machiela MJ. Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. Sci Rep. 2018;8:12316.
4. Grassmann F, Kiel C, den Hollander AI, Weeks DE, Lotery A, Cipriani V, Weber BHF, International Age-related Macular Degeneration Genomics Consortium (IAMDGC). Y chromosome mosaicism is associated with age-related macular degeneration. Eur J Hum Genet. 2019;27:36–41.
5. Dumanski JP, Lambert J-C, Rasi C, Giedraitis V, Davies H, Grenier-Boley B, Lindgren CM, Campion D, Dufouil C, Pasquier F, et al. Mosaic loss of chromosome Y in blood is associated with alzheimer disease. Am J Hum Genet. 2016;98:1208–19.

6. Haitjema S, Kofink D, van Setten J, van der Laan SW, Schoneveld AH, Eales J, Tomaszewski M, de Jager SCA, Pasterkamp G, Asselbergs FW, et al. Loss of Y chromosome in blood is associated with major cardiovascular events during follow-up in men after carotid endarterectomy. Circ Cardiovasc Genet. 2017;10:e001544.
7. Kimura A, Hishimoto A, Otsuka I, Okazaki S, Boku S, Horai T, Izumi T, Takahashi M, Ueno Y, Shirakawa O, et al. Loss of chromosome Y in blood, but not in brain, of suicide completers. PLoS ONE. 2018;13:e0190667.
8. Zhou W, Machiela MJ, Freedman ND, Rothman N, Malats N, Dagnall C, Caporaso N, Teras LT, Gaudet MM, Gapstur SM, et al. Mosaic loss of chromosome Y is associated with common variation near TCL1A. Nat Genet. 2016;48:563–8.
9. Wright DJ, Day FR, Kerrison ND, Zink F, Cardona A, Sulem P, Thompson DJ, Sigurjonsdottir S, Gudbjartsson DF, Helgason A, et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. Nat Genet. 2017;49:674–9.
10. Forsberg LA, Halvardson J, Rychlicka-Buniowska E, Danielsson M, Moghadam BT, Mattisson J, Rasi C, Davies H, Lind L, Giedraitis V, et al. Mosaic loss of chromosome Y in leukocytes matters. Nat Genet. 2019;51:4–7.
11. Zhou W, Machiela MJ, Freedman ND, Rothman N, Malats N, Dagnall C, Caporaso N, Teras LT, Gaudet MM, Gapstur SM, et al. Reply to 'Mosaic loss of chromosome Y in leukocytes matters.' Nat Genet. 2019;51:7–9.
12. Cotter DJ, Brotman SM, Wilson Sayres MA. Genetic diversity on the human X chromosome does not support a strict pseudoautosomal boundary. Genetics. 2016;203:485–92.
13. Wacholder S, Hartge P, Lubin JH, Dosemeci M. Non-differential misclassification and bias towards the null: a clarification. Occup Environ Med. 1995;52:557–8.
14. Danielsson M, Halvardson J, Davies H, Moghadam BT, Mattisson J, Rychlicka-Buniowska E, Jaszczyński J, Heintz J, Lannfelt L, Giedraitis V, et al. Intra-individual changes in the frequency of mosaic loss of chromosome Y over time estimated with a new method. 2019. https://doi.org/10.1101/631713.
15. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008;40:1245–52.
16. González JR, Subirana I, Escaramís G, Peraza S, Cáceres A, Estivill X, Armengol L. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. BMC Bioinform. 2009. https://doi.org/10.1186/1471-2105-10-172.
17. Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, Garcia-Closas M, Kogevinas M, Villa O, Hutchinson A, Earl J, Marenne G, et al. Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. Am J Hum Genet. 2010;87:129–38.
18. Arseneault M, Monlong J, Vasudev NS, Laskar RS, Safisamghabadi M, Harnden P, Egevad L, Nourbehesht N, Panichnantakul P, Holcatova I, et al. Loss of chromosome y leads to down regulation of KDM5D and KDM6C epigenetic modifiers in clear cell renal cell carcinoma. Rep.: Sci; 2017. p. 7.
19. Minner S, Kilgué A, Stahl P, Weikert S, Rink M, Dahlem R, Fisch M, Hppner W, Wagner W, Bokemeyer C, et al. Y chromosome loss is a frequent early event in urothelial bladder cancer. Pathology. 2010;42:356–9.
20. González JR, Cáceres A, Cáceres A. Omic Association Studies with R and Bioconductor. London: Chapman and Hall/CRC; 2019.
21. Van Os PGCE, Schouten JP. Multiplex ligation-dependent probe amplification (MLPA®) for the detection of copy number variation in genomic sequences. Methods Mol Biol. 2011;688:97–126.
22. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
23. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
24. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23:257–8.
25. Subirana I, Diaz-Uriarte R, Lucas G, Gonzalez JR. CNVassoc: Association analysis of CNV data using R. BMC Med Genomics. 2011. https://doi.org/10.1186/1755-8794-4-47.
26. Leitsalu L, Haller T, Esko T, Tammesoo M-L, Alavere H, Snieder H, Perola M, Ng PC, Mägi R, Milani L, et al. Cohort profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. Int J Epidemiol. 2015;44:1137–47.
27. Westra H-J, Jansen RC, Fehrmann RSN, te Meerman GJ, van Heel D, Wijmenga C, Franke L. MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. Bioinformatics. 2011;27:2104–11.
28. Cáceres A, Jene A, Esko T, Pérez-Jurado LA, González JR. Extreme down-regulation of chromosome Y and cancer risk in men. JNCI J Natl Cancer Inst. 2020. https://doi.org/10.1093/jnci/djz232.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.