**Article**

# Extreme down-regulation of chromosome Y and cancer risk in men

Alejandro Cáceres, PhD, Aina Jene, MsC, Tonu Esko, PhD, Luis A Pérez-Jurado, MD PhD FRACP, and Juan R González, PhD

**Author affiliations:** Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain (Cáceres, González). Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (Cáceres, González). Center for Genomics Regulation, Barcelona, Spain (Jene). Estonian Genome Centre Science Centre, University of Tartu, Tartu, Estonia (Esko). Genetics Unit, Universitat Pompeu Fabra, Institut Hospital del Mar d'Investigacions Mediques (IMIM) and Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona Spain (Pérez-Jurado). Women's and Children's Hospital, South Australian Health and Medical Research Institute & University of Adelaide, Adelaide, Australia (Pérez-Jurado). Department of Mathematics, Universitat Autònoma de Barcelona, Bellaterra, Spain (González)

**Correspondence to**: Alejandro Cáceres, PhD, Barcelona Institute for Global Health (ISGlobal), Doctor Aiguader 88, Barcelona 08003, Spain. email: alejandro.caceres@isglobal.org, Tel: +34 932147316 or Juan R. González, PhD, Barcelona Institute for Global Health (ISGlobal), Doctor Aiguader 88, Barcelona 08003, Spain. email: juanrgonzalez@isglobal.org Tel: +34 932147327

## Abstract

### Background

Understanding the biological differences between sexes in cancer is essential for personalized treatment and prevention. We hypothesized that the extreme down-regulation of chromosome Y gene expression (EDY) is a signature of cancer risk in men and the functional mediator of the reported association between the mosaic loss of chromosome Y (LOY) and cancer.

### Methods

We advanced a method to measure EDY from transcriptomic data. We studied EDY across 47 nondiseased tissues from the GTEx project (n = 371) and its association with cancer status across 12 cancer studies from the Cancer Genome Atlas (TCGA) (n = 1,774), and seven other studies (n = 7,562). Associations of EDY with cancer status and presence of loss-off function mutations in chromosome X were tested with logistic regression models, while a Fisher's test was used to assess genome-wide association of EDY with the proportion of copy number gains. All statistical tests were two-sided.

### Results

EDY was likely to occur in multiple nondiseased tissues ($P$<0.001) and statistically significantly associated with the EGFR tyrosine kinase inhibitor resistance pathway ($FDR$=0.028). EDY strongly associated with cancer risk in men (OR = 3.66, 95%CI = 1.58, 8.46, $P$ = 0.002), adjusted by LOY and age, and its variability was largely explained by several genes of the non-recombinant region (NRY) whose chromosome X homologs showed loss-of-function mutations that co-occurred with EDY during cancer (OR = 2.82, 95%CI =

1.32, 6.01, *P* = 0.007). EDY associated with a high proportion of *EGFR* amplifications (OR = 5.64, 95%CI = 3.70, 8.59, FDR<0.001) and *EGFR* overexpression, along with SRY hypomethylation and NRY hypermethylation, indicating alternative causes of EDY in cancer other than LOY. EDY associations were independently validated for different cancers and exposure to smoking, and its status was accurately predicted from individual methylation patterns.

**Conclusions**

EDY is a male-specific signature of cancer susceptibility that supports the escape from X-inactivation tumor suppressor hypothesis for genes that protect females with respect to males from cancer risk.

Men are more at risk and less likely to survive cancer than women *(1)*. Besides the different environments to which genders are exposed, sex-specific molecular processes are also important to explain sexual dimorphism in cancer. For instance, sex hormones are critically involved in cancer development and numerous loci in sex chromosomes are associated with cancer susceptibility *(2)*. In addition, the complete loss of chromosome Y (LOY) is a frequent event in tumor cells *(3-9)* and a specific risk factor for cancer in men when found in peripheral blood cells *(10-12)*. Given that LOY is the most common somatic mutation in men, there is a need to understand whether uncontrolled mitosis can lead to LOY or LOY, as an age-related condition *(13,14)*, can predispose to cancer *(15)*. A logical consequence of the presence of LOY in a tissue would be the reduction of the overall transcription output of Y across the affected tissue. As such, one should observe an association between the extreme down-regulation of chromosome Y (EDY) with cancer that, if stronger than LOY's, would indicate a directionality from LOY to cancer via EDY.

Studies have shown a strong association between aneuploidies in cancer and gene expression *(16)*. Consequently gene expression data have been used to identify the functional consequences of aneuploidies *(17)*. However, studies that measure the overall transcription output of an entire chromosome have not been reported. Therefore, we first proposed a method to measure EDY from transcriptomic data, obtained by either by RNA-sequencing or expression microarrays, and then confirmed the biological suitability of the measure by analyzing data from the Genotype Tissue-expression Project (GTEx) project over multiple nondiseased tissues. Using the Cancer Genome Altas (TCGA) and several microarray studies, we then studied the association between EDY and cancer risk in men. We also

investigated whether EDY status in tumors associated with differential methylation across Y and with copy number alterations in autosomes. We thus tested the hypothesis that the novel transcriptomic signature EDY is an important risk factor for cancer in men.

## Methods

### *Detection of EDY from transcriptome data*

We analyzed expression data from chromosome Y in the form of count data for RNA-sequencing and signal intensity for microarray experiments. For each individual, we measured the relative expression of the entire chromosome with respect to the autosomes. Having $N$ exons in chromosome Y, with $x_e$ read count for the $e$-th exon, we computed

$$y = \Sigma_{e=1,..N} \log_2(x_e + 1)/N$$

as a measure of the average expression of Y. Likewise, we obtained the mean expression in autosomes

$$a = \Sigma_{e=1,..M} \log_2(x_e + 1)/M$$

where $M$ is the number exons with count data in the autosomes. The relative amount of an individual's Y expression with respect to the individual's autosomes was then defined as

$$Ry = y - a.$$

We considered the extreme down-regulation for chromosome Y (EDY) as the extreme phenotype of Ry given by values lower than the 0.05 sample quantile, as it has been done for other extreme phenotypes *(18)*. The adequacy to treat EDY as a discontinuous extreme phenotype that is the consequence of LOY is supported by the observation that treating LOY itself as a continuous variable is sub-optimal *(19)*. In a study with *K* subjects, we then called individual *j* as having EDY if

$$\mathrm{Ry}_j < median(\mathrm{Ry}) - 1.2 \times IQR(\mathrm{Ry})$$

where IQR is the usual definition for the inter-quartile range of *Ry* values over subjects. The cutting threshold given by the expression above corresponds to the lower 5% of the data for different types of unimodal distributions. Given that the interquartile-range is robust for different distributions, in the case of array intensity data, we used similar definitions for EDY, computing the relative expression Ry from $x_e$ as the intensity value at probe *e*.

***Discovery and Validation Studies***

We studied the frequency of EDY in 47 nondiseased tissues using the version-6 RNA-sequencing data from the GTEx project (https://www.gtexportal.org/). Genome-wide SNP data was available for 298 males for whom we could determine their EDY status. We studied the association between EDY and cancer using the multiomic data for 28 TCGA cancer studies. We downloaded data from a total number of 10,642 samples, where 5,329 were from normal tissues and 5,313 were tumorous tissues. For validation, we downloaded from the ArrayEx-

press Archive (www.ebi.ac.uk/arrayexpress) a large expression matrix of 27,871 arrays with accession number E-MTAB-3732, the largest systematically annotated gene expression dataset of its kind. Gene expression data were searched in the GEO repository (www.ncbi.nlm.nih.gov/geo) for case-control studies of renal clear cell carcinoma and colorectal cancer. Their accession numbers are GSE36895 and GSE44076. We also downloaded transcriptomic data of two additional studies (GSE4573 and GSE5123) on lung squamous cell carcinoma with exposure to smoking. We downloaded normalized expression data and used female samples to check the lower limit of EDY detection in males. Probe annotation was made with Bioconductor biomaRt package.

We downloaded methylomic data from a case-control study on kidney cancer with accession number GSE61441. Given the strong pattern of methylation associated with EDY, we fitted an elastic-net model to build a subject-wise predictor of EDY from methylomic data, using glmnet and caret R packages. The model was trained in the 90% of TCGA cancer samples (n=1174) and validated in the other 10% of samples (n = 292). The model hyperparameters (mixing and smoothing) were estimated using 10-fold cross-validation. The list of CpGs and coefficients to build the predictive and an R function to get EDY prediction are available at: http://github.com/isglobalbrge/EDY. Further details are in the **Supplementary Methods**.

***Statistical analyses***

All analyses were performed using packages from Bioconductor version 3.8 and R version 3.5.2. Logistic regression models were fitted for testing the association between EDY with different outcomes, such as case-control status of the individuals or tumor status of biological samples of cancer patients. We used Bayesian regression models from the arm R package that gave consistent estimates for low frequencies of EDY cases. Random effects meta-

analyses were performed with the rma package where heterogeneity between studies was tested with a $\chi^2$ test. All models were adjusted by age and cancer type when available/needed. Main effect P-values were two sided and, if needed, corrected for multiple comparisons by false discovery rate (FDR). FDR and single-test P-values less than 0.05 were considered statistically significant. Processed data and the entire computer code, needed to completely reproduce our findings, have been made public in the figshare repository at: https://figshare.com/projects/Extreme_down-regulation_of_chromosome_Y_and_male_disease/58514.

## Results

### *EDY in nondiseased tissues*

We first studied EDY in nondiseased tissues analyzing RNA-sequencing data of 371 men across 47 tissues from the Genome Tissue Expression (GTEx) project. The average number of tissues per man was 12. We detected 140 subjects with EDY in at least one tissue (**Supplementary Figure 1**). There was large variability of EDY frequency between tissues (*mean* = 6.1%, *SD* = 3.7%) (**Supplementary Table 1**). We found high rates of individuals with EDY in more than one tissue and, therefore, hypothesized whether EDY was likely to appear in multiple tissues in a single individual, suggesting a genetic predisposition to it. Consequently, we first confirmed that individuals with EDY in one tissue were likely to show EDY in any other tissue (permutation test of tissue labels, $P$ <0.001). Then, due to the low power expected for the number of subjects, we performed enrichment analysis in genome-wide SNP associations for EDY status. Enrichment analyses were performed for a new variable $EDY_{>1\ tissue}$, defined as positive for individuals where EDY was found in more than one

tissue and negative otherwise. While no statistically significant associations were observed for $EDY_{>1\ tissue}$, we found that $EDY_{>2\ tissues}$ (N = 32) was statistically significantly enriched with SNPs in the EGFR tyrosine kinase inhibitor resistance pathway (*FDR* = 0.028). In this case, we also observed suggestive genome-wide associations mapping to susceptibility genes for basophil percentage of granulocytes (*GRIP1*) *(21)*; lung and gastric cancers and smoke-induced emphysema (*MMP12*) *(22-24)* and high and low-density cholesterol and triglycerides levels (*GPAM*) *(25)* (**Supplementary Figure 2**). Finally, in whole blood, the single genotyped tissue in GTEx where LOY could be called, only one of the three individuals with positive EDY was detected with LOY. Therefore, this novel signature EDY appears to be more common than LOY in nondisease individuals, can be identified across tissues and may have a genetic basis linked to several autosomal loci.

***EDY in 12 TCGA cancer studies***

We analyzed genomic and transcriptomic data of 12 cancer studies with normal and tumor samples of cancer patients from the TCGA project to establish whether EDY explained more cancer variability than LOY (**Supplementary Figure 3**). We called LOY from genotype data and EDY from transcriptomic data within each cancer study in all samples (normal/tumor) (**Table 1**). EDY was obtained with respect to the *Ry* distribution of samples with no loss and no gains in chromosome Y (**Figure 1**). As expected, the proportion of agreement between EDY and LOY status, comprising normal and tumor tissues, was high but varied across all 12 cancer studies (*mean* = 87%, SD = 6%). Comparing cancer to normal samples, we observed that the overall magnitude of the age-adjusted effect of EDY on cancer status (OR = 8.33, 95%CI = 3.30, 20.89, $P = 6.9\times10^{-6}$), remained statistically significant after adjusting by LOY, within each cancer study (OR = 3.66, 95%CI = 1.58, 8.46, *P* = 0.002). As all samples (normal/

tumor) are from cancer patients, there was no association between age and cancer status of the samples (OR = 0.99, 95%CI = 0.98, 1.00, *P* = 0.07). However, while we observed a statistically significant association between LOY and age (OR = 1.009, 95%CI = 1.003, 1.01, *P* = 0.001), we did not observe a statistically significant association between EDY and age (OR = 1.003, 95%CI = 0.99, 1.00, *P* = 0.2). Consistent with these findings, we observed that the association between EDY and cancer was robust under different age quartiles (Age 16-57: OR = 4.56, P <0.001; Age 57-64: OR = 5.03, , P <0.001; Age 64-71; OR = 3.76, P <0.001; Age 71-90: OR = 17.25, P =0.008).

Transcriptome-wide analyses in tumor samples revealed that the transcription levels of *DDX3Y*, *EIF1AY*, *KDM5D*, *RPS4Y1*, *UTY* and *ZFY* were statistically significantly down-regulated across all 12 cancers. Their joint down-regulation explained 89% of EDY's variability and 88% of LOY's variability (**Supplementary Figure 3**). Interestingly, these genes have four remarkable features. First, they are located in pairs in three distant regions of the non-recombinant region of Y (NRY) (Yp11.31: Mb 2.7-2.9 / Yq11.21: Mb 15.0-15.6 / Yq11.22: Mb 21.8-22.9), suggesting that they may share regulatory elements. Second, the genes encode proteins with important functions in cell cycle regulation: helicase (*DDX3Y*), translation initiation (*EIF1AY*), histone demethylation (*KDM5D* and *UTY/KDM6C*), transcriptional activation (*ZFY*) and ribosomal assembly (*RPS4Y1*). Third, these genes have homologs (*DDX3X*, *EIF1AX*, *KDM5C*, *KDM6A/UTX*) on the X chromosome that escape X-inactivation. And fourth male-biased loss of function (LoF) somatic mutations have been found in four of the X chromosome homologs of these genes across many cancers *(26)*. In line with this last feature, we observed that LoF mutations in the four X chromosome homologs co-occurred with LOY (OR = 3.59, 95%CI = 1.57, 8.18, P = 0.002) and EDY (OR = 2.82, , 95%CI = 1.32, 6.01, P = 0.007) during cancer.

Three further analyses in TCGA consistently showed that EDY provided a stronger signature of cancer than LOY. First, the meta-analysis between cancer status and the EDY derived from the NRY-gene signature was substantially more statistically significant than previous associations (OR = 8.14, 95%CI = 4.29, 15.40, P<0.001 ), remaining statistically significant after adjusting by LOY (OR = 3.61, 95%CI = 1.51, 8.63, *P* = 0.003). Second, Bayesian network analyses indicated that causal sequence -aging, LOY, EDY and cancer-, was more probable than -aging, cancer, LOY and EDY- (**Figure 2F**). Third, total EDY mediated 48.9% (95%CI = 25.3%, 66.0%) of the age-adjusted association between LOY and the cancer status of the samples. Overall, these observations on TCGA data support a possible cancer mechanism underlying LOY given by the simultaneous inactivation of NRY genes derived by EDY and their functional homologs on chromosome X by LoF mutations.

***Validation in independent studies***

Using data from independent transcriptomic studies, we performed numerous replication and consistency analyses (**Figure 3, Supplementary Table 2**). We first replicated the EDY association with colorectal (OR = 5.16, 95%CI = 1.30, 20.45, P = 0.01) and kidney cancer (OR = 20.09, 95%CI = 2.07, 195.11, P = 0.009) in two independent transcriptomic case/control studies, where tumor tissues were compared to normal tissues of cancer patients. In the kidney study, cancer was not associated with age (OR = 0.99, 95%CI = 0.93, 1.05, P = 0.8). EDY was not associated with age either (OR = 1.004, , 95%CI = 0.94, 1.06, P = 0.8) and, despite low numbers (12 cases and 17 controls), the association between EDY and cancer appeared to be consistent between two age strata (Age 35-59: OR = 14.9, 95%CI = 1.11, 201.05, P = 0.04, case/control = 5/8; Age 59-83: OR = 3.31, 95%CI = 0.80, 13.91, P = 0.09, case/control = 8/6). As LOY in blood is a risk factor for cancer, we then confirmed that

EDY in blood associated with cancer diagnosis in a large Estonian population sample (OR = 3.23, 95%CI = 1.24, 8.40, P = 0.01). We also aimed to determine the range of cancers associated with EDY, using a large collection of multi-disease expression arrays with 3,771 diseased tissues and 3,127 healthy-male tissues *(20)*. We observed strong positive associations for 8 cancer groups, negative associations for myeloma and other types of leukemia, and no association for lymphoma, neuroblastoma and prostate cancer (**Figure 3**). The range of cancers associated with EDY largely overlapped with cancer status of samples associated with LOY across all the 28 cancer studies from the TCGA. Interestingly, associations of LOY with leukemia and prostate cancers were statistically non-significant (**Supplementary Table 3**). Finally, in line with LOY's association with smoking *(27)*, we observed a statistically significant association of EDY in lung cancer with the number of cigarettes smoked per day (OR = 1.07, 95%CI = 1.02, 1.12, P = 0.003) in two studies (**Figure 3**), and confirmed the association with heavy smoking (>1 package/day, OR = 18.77, 95%CI = 1.02, 345.32, *P* = 0.04) in a third study.

***EDY association with copy number variants and methylation patterns***

To gain further insights on why EDY can be a stronger cancer signature than LOY, we studied whether EDY showed biological correlates in cancer that were independent of LOY. We analyzed the copy number variant (CNV) differences between EDY statuses in 3,034 tumors across all TCGA studies in windows of 1.25 Mb across the genome. Remarkably, we observed that, in individuals with no-LOY, EDY was strongly associated with higher proportion of copy number gains of *EGFR* (OR = 5.64, 95%CI = 3.70, 8.59, FDR<0.001) while, in individuals with LOY, EDY associated with lower proportion of copy number gains in regions containing *SOX4* (OR = 0.29, 95%CI = 0.17, 0.47, FDR<0.001), *NCOA2* and the short arm of

chromosome 12 (12p) (**Figure 4**). We also asked whether consistent differences in EDY were associated with methylation across chromosome Y, stratifying by LOY status and adjusting for tumor type. We thus observed a highly reproducible pattern of methylation-probe associations that was independent of LOY (**Figure 4, Supplementary Table 4**). We found statistically significant methylation changes in the NRY regions deregulated in EDY, the highest association being with hypomethylation surrounding *SRY* (cg04169747, OR = 0.96, 95%CI = 0.95, 0.97, FDR<0.001 and the highest changes being hypermethylation at *KDM5D* (cg15329860, 95%CI = 1.15, 1.29, FDR<0.001) *(28)* and *EIF1AY* (cg08820785, OR = 1.19, 95%CI = 1.12, 1.24, FDR<0.001). In line with the proportion of gains found in EDY, we observed statistically significant associations of *EGFR* and *SOX4* expression levels with reciprocal hypo and hyper-methylations of the same CpG sites (**Supplementary Tables 5-7**). In addition to the possible contribution of SRY hypomethylation, a gene hypermethylated after sex differentiation early in development *(29)*, our data reinforce the role of NRY genes in cancer sex bias *(26)*. Given the strong pattern of methylation associated with EDY, we used an elastic-net algorithm to build a subject-wise predictor of EDY from methylomic data, trained in the 90% of TCGA cancer samples and validated with 90.7% accuracy in the other 10%. In an independent methylomic study, we externally validated the association between the methylation-inferred EDY with kidney cancer (N = 92, OR = 45.4, 95%CI = 2.11, 977.32, P = 0.01) (**Figure 3**).

## Discussion

We have provided the first evidence of a path in males that leads from LOY and other genetic alterations, such as EGFR pathway activation, to cancer development through EDY. EDY is

a male-specific signature of cancer susceptibility that is strongly linked to LOY and chromosome Y methylation patterns, as well as to environmental exposures like smoking *(27)*. The high correlation between EDY and LOY confirmed that EDY is the most likely functional consequence of LOY, yet additional risk was observed for individuals with EDY and no LOY, suggesting that overall decrease of chromosome Y transcript levels is a key element in the susceptibility to disease. We found strong associations with cancer susceptibility, comparable to those of smoking. In the population-based EGCUT study of Estonian individuals, the frequency of EDY in blood in the general population (4.3%) and the fraction of individuals diagnosed with any type of cancer (9.4%) yielded an attributable risk of 16.2%, which is in range of the attributable risk to cancer due to smoking (*30*) but further studies are needed to refine these risk estimates.

In particular, our data provide additional evidence to support the escape from X-inactivation tumor suppressor hypothesis for genes that protect females with respect to males from cancer risk *(26)*, pointing to specific genes that accumulate male-biased mutations in cancer whose chromosome Y homologs on NRY define EDY. These genes (*DDX3Y*, *EIF1AY*, *KDM5D*, *RPS4Y1*, *UTY* and *ZFY)* regulate cell cycle through different mechanisms and behave as dosage-sensitive tumor suppressors. In addition to sex-biased LoF mutations on the gene copies of the X-chromosome, males would have a higher risk of first or second hits affecting the NRY copies, revealed by EDY derived from LOY and/or other genomic mechanisms associated with NRY hypermethylation. One of the main mechanisms of NRY hypermethylation seems to be related to *EGFR* gene dosage and polymorphisms in the pathway. *EGFR* codes for Epidermal Growth Factor Receptor, one of the four members of ErbB family of tyrosine kinase receptors whose catalytic activation leads to increase DNA methyltransferase activity resulting in increased global DNA methylation in some cancers

*(31,32)*. Our data also support a role for the EGFR pathway in the process of accumulated DNA methylation affecting the Y chromosome in male cancer progression.

Recent studies indicate that the risk factors of LOY include aging, smoking and air pollution (*10,27,33*). Given that they are also risk factors for cancer, they can confound the association between EDY and cancer. More detailed studies into the relationship of EDY to these and other cancer-related factors are needed to determine their role in EDY versus LOY susceptibility. Here, we observed that in tumor/healthy and cancer/control studies EDY was not associated with age, likewise cancer. In these studies, where age is matched, we observed that LOY, however, had a statistically significant association with age, suggesting neutral events deriving in LOY but not EDY. We additionally observed in the TCGA study that adjustment for smoking did not change the statistical significance of the association between EDY and cancer. While specific studies are needed to characterize these and other risk factors of EDY, our highly reliable predictions from methylation profiles indicate a strong role of environmental exposures. The methylation EDY-predictor is available at http://github.com/isglobal-brge/EDY, so its adequacy as a diagnostic or prognostic tool in different male cancers can be further tested in longitudinal studies.

## References

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA: a cancer journal for clinicians* 2019;69(1):7.

2. Clocchiatti A, Cora E, Zhang Y, et al. Sexual dimorphism in cancer. *Nat Rev Cancer.* 2016;16(5):330.

3. Bianchi NO. Y chromosome structural and functional changes in human malignant diseases. *Mutat Res Rev Mutat Res*. 2009;682(1):21–27.

4. Wright DJ, Day FR, Kerrison ND, et al. Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat Genet.* 2017;49(5):674.

5. Duijf PH, Schultz N, Benezra R. Cancer cells preferentially lose small chromosomes, *Int J Cancer.* 2013;132(10):2316.

6. Dürrbaum M, Storchová Z. Effects of aneuploidy on gene expression: implications for cancer. *FEBS J.* 2016;283(5):791.

7. Klatte T, Rao PN, De Martino M, et al. Cytogenetic profile predicts prognosis of patients with clear cell renal cell carcinoma. *J Clin Oncol.* 2009;27(5):746.

8. Nomdedeu M, Pereira A, Calvo X, et al., Clinical and biological significance of isolated y chromosome loss in myelodysplastic syndromes and chronic myelomonocytic leukemia. a report from the spanish mds group. *Leuk Res.* 2017;63:85.

9. Minner S, Kilgué A, Stahl P, et al. Y chromosome loss is a frequent early event in urothelial bladder cancer. *Pathology.* 2010;42(4):356.

10. Zhou W, Machiela MJ, Freedman ND, et al. Mosaic loss of chromosome Y is associated with common variation near tcl1a. *Nat Genet.* 2016;48(5);563.

11. Forsberg LA, Rasi C, Malmqvist N, et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. *Nat Genet*. 2014;46(6):624.

12. Rodríguez-Santiago B, Malats N, Rothman N, Armengol L, et al. Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am J Hum Genet.* 2010;87:129-138.

13. Holland AJ, Cleveland DW. Losing balance: the origin and impact of aneuploidy in cancer: ‘exploring aneuploidy: the significance of chromosomal imbalance’ review series. *EMBO Rep*. 2012;13(6):501.

14. Loftfield E, Zhou W, Graubard BI, et al. Predictors of mosaic chromosome y loss and associations with mortality in the UK Biobank. *Sci Rep*. 2018;8(1):12316.

15. Thompson D, Genovese G, Halvardson J, et al. Genetic predisposition to mosaic y chromosome loss in blood is associated with genomic instability in other tissues and susceptibility to non-haematological cancers. *bioRxiv*. 2019:514026.

16. Fehrmann RS, Karjalainen JM, Krajewska M, et al. Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat Genet.* 2015;47(2):115.

17. Carter SL, Eklund AC, Kohane IS, et al. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. *Nat Genet.* 2006;38(9):1043.

18. González JR, González-Carpio M, Hernández-Sáez R, et al. FTO risk haplotype among early onset and severe obesity cases in a population of western Spain. *Obesity.* 2012;20(4):909.

19. González JR, López-Sánchez M, Cáceres A, et al. A robust estimation of mosaic loss of chromosome Y from genotype-array-intensity data to improve disease risk associations and transcriptional effects. *BioRxiv*. 2019;https://doi.org/10.1101/764845.

20. Torrente A, Lukk M, Xue V, et al. Identification of cancer related genes using a comprehensive map of human gene expression. *PloS One*. 2016;11(6):e0157484.

21. Astle WJ, Elding H, Jiang T, et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell*. 2016;167(5):1415.

22. Su L, Zhou W, Asomaning K, et al. Christiani, Genotypes and haplotypes of matrix metalloproteinase 1, 3 and 12 genes and the risk of lung cancer. *Carcinogenesis.* 2005;27(5):1024.

23. Li Y, Sun DL, Duan YN, et al. Association of functional polymorphisms in MMPs genes with gastric cardias adenocarcinoma and esophageal squamous cell carcinoma in high incidence region of north China. *Mol Biol Rep*. 2010;37(1):197.

24. Churg A, Wang RD, Tai H, et al. Macrophage metalloelastase mediates acute cigarette smoke–induced inflammation via tumor necrosis factor-α release. *Am J Respir Crit Care Med*. 2003;167(8):1083.

25. Klarin D, Damrauer SM, Cho K, et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the million veteran program. *Nat Genet.* 2018;50(11):1514.

26. Dunford A, Weinstock DM, Savova V, et al. Tumor-suppressor genes that escape from x-inactivation contribute to cancer sex bias. *Nat Genet.* 2017;49:10.

27. Dumanski JP, Rasi C, Lönn M, et al. Smoking is associated with mosaic loss of chromosome Y. *Science.* 2015;347(6217):81.

28. Arseneault M, Monlong J, Vasudev NS, et al. Loss of chromosome y leads to down regulation of KDM5D and KDM6C epigenetic modifiers in clear cell renal cell carcinoma. *Sci Rep.* 2017;7:44876.

29. Nishino K, Hattori N, Tanaka S, Shiota K. DNA methylation-mediated control of SRY gene expression in mouse gonadal development. *J Biol Chem*. 2004;279:22306–22313..

30. Chyou PH, Nomura AM, Stemmermann GN. A prospective study of the attributable risk of cancer due to cigarette smoking. *Am. J. Public Health.* 1992;82(1):37.

31. Samudio-Ruiz SL, Hudson LG. Increased DNA methyltransferase activity and DNA methylation following epidermal growth factor stimulation in ovarian cancer cells. *Epigenetics*. 2012;7(3):216-224.

32. Bjaanæs MM, Fleischer T, Halvorsen AR, et al. Genome-wide DNA methylation analyses in lung adenocarcinomas: Association with EGFR, KRAS and TP53 mutation status, gene expression and prognosis. *Mol Oncol.* 2016;10(2):330–343.

33. Wong JY, Margolis HG, Machiela M, et al. Outdoor air pollution and mosaic loss of chromosome Y in older men from the Cardiovascular Health Study. *Environ. Int.* 2018;116:239.

**Funding**: This research has received funding from Ministerio de Ciencia, Innovación y Universidades de España and Fondo Europeo de Desarrollo, UE (RTI2018-100789-B-I00). 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).

**NOTES**

The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication. We would like to thank Francisco Real for his critical reading of the manuscript.

Conflicts of interest Disclosures: LAPJ is a founding partner and scientific advisor of qGenomics Laboratory. All other authors declare no conflict of interest.

**Table 1.** EDY and LOY status in 12 cancer studies of TCGA. EDY and LOY were estimated from transcriptomic and genomic data, respectively, obtained in both cancer and normal tissues*

| TCGA Cancer study | N | %EDY | %LOY | %Agreement EDY/LOY |
|---|---|---|---|---|
| Bladder Urothelial Carcinoma (BLCA) | 246 | 31.3 | 41.5 | 85.0 |
| Colon adenocarcinoma (COAD) | 108 | 31.5 | 55.6 | 74.1 |
| Esophageal carcinoma (ESCA) | 98 | 49.0 | 45.9 | 80.6 |
| Kidney Chromophobe (KICH) | 46 | 45.7 | 45.7 | 95.7 |
| Kidney renal clear cell carcinoma (KIRC) | 347 | 44.4 | 46.1 | 89.6 |
| Kidney renal papillary cell carcinoma (KIRP) | 165 | 77.0 | 77.6 | 95.8 |
| Liver hepatocellular carcinoma (LIHC) | 114 | 14.9 | 16.7 | 94.7 |
| Lung adenocarcinoma (LUAD) | 190 | 30.0 | 40.5 | 89.5 |
| Lung squamous cell carcinoma (LUSC) | 268 | 38.4 | 54.5 | 82.5 |
| Prostate adenocarcinoma (PRAD) | 413 | 11.1 | 7.7 | 91.8 |
| Rectum adenocarcinoma (READ) | 46 | 37.0 | 50.0 | 87.0 |
| Thyroid carcinoma (THCA) | 97 | 7.2 | 18.6 | 86.6 |

* EDY was computed with respect to samples with no gains or losses of chromosome Y. Theproportion of agreement between the measures was high but substantial differences were also observed.

**Figures and Table Legends**

**Figure 1. Relative chromosome Y expression (Ry) as a function of age for 12 cancer studies from the TCGA.** The figure shows tumor (black) and normal samples (blue). Samples with LOY, obtained from genotype intensity data, are shown in red triangles. Samples with EDY (green triangles) are those with low values of Ry, relative to samples with no chromosome Y losses or gains. While EDY and LOY status overlap, numerous individuals are observed with LOY but no EDY, particularly those with high values of Ry. Normal samples consistently have high Ry values across studies.

**Figure 2. Additive Bayesian network models for age, cancer, LOY and EDY for 12 cancer studies from the TCGA.** The left figure shows the driver model, where cancer depends on EDY, EDY on LOY and LOY on age. Maximum likelihood estimate and Bayes information criterion (BIC) are shown on top. The right figure shows the passenger model, where EDY depends on LOY, LOY on cancer and cancer on age. In the TCGA studies, the higher likelihood and lower BIC favor the driver model over the passenger model.

**Figure 3. EDY as a marker of cancer status of biological samples and individuals.** The figure shows the association between EDY and cancer status across different independent studies with publicly available data. (**A**) In the TCGA study (N = 1,774), the OR of EDY for tumor status of the biological samples of cancer patients was obtained from logistic-regression models adjusting by age for 12 different cancers. The overall estimate of the effect of EDY was computed by a random effects meta-analysis and its heterogeneity with a $\chi^2$ test. P-values are two sided. (**B**) The association between EDY and cancer status of individuals

was independently tested in colorectal (N = 142) and kidney (N = 29) cancer case-control studies (GSE66836, GSE36895) and in a population sample of 550 Estonian individuals (EGCUT). (**C**) A large transcriptomic dataset (N = 6,898) was used to assess EDY's association with multiple cancer diagnoses (E-MTAB-3732) and (**D**) two studies on lung squamous cell carcinoma (GSE5123, LUSC/TCGA, total N = 243) were used to test the association between EDY and cigarettes smoked per day.

**Figure 4. Association of EDY with chromosome Y methylation and genome-wide copy number variant proportion for individuals with and without LOY** (**A**) Number of cancer samples in all four LOY and EDY status across 12 cancer studies in TCGA (BLCA, COAD, ESCA, KICH, KIRC, KIRP, LIHC, LUAD, LUSC, PRAD, READ and THCA, described in Table 1). (**B**) Odds ratios of EDY for methylation sites across Y, stratified by LOY (LOY: vertical axis, no-LOY: horizontal axis). 52 statistically significant associations for both LOY statuses are colored in red. (**C**) Genome-wide differences in CNV proportion, positive (+) and negative (-), between No-EDY and EDY, and stratified by LOY (No-LOY: top, LOY: bottom).

# Supplementary Materials for

## Extreme down-regulation of chromosome Y and with cancer risk in men

Alejandro Cáceres, Aina Jene, Tonu Esko, Luis A Pérez-Jurado and Juan R González

Correspondence to: alejandro.caceres@isglobal.org, juanr.gonzalez@isglobal.org

### This PDF file includes:

Supplementary Methods

Figs. S1 to S5

Tables S1 to S7

# Supplementary Methods

## Detection of LOY from genotype data

We analyzed the summarized log R ratio (LRR) of allelic (microarray) intensities of all the SNPs in chromosome Y. The median of LRR in Y (mLRR-Y) per subject is as a measure of the subject's chromosome Y DNA content as has been used to call LOY status (11). We analyzed raw CEL files for TCGA microarray data that were processed to obtain the genome-wide LRR values in PenCNV format. LRR values in Y were obtained for the male-specific region between pseudoautosomal regions 1 and 2 (PAR1 and PAR2) and normalized with respect to the 25%-trimmed LRR mean for the autosomes to obtain mLRR-Y. Individual samples with large LRR variability (>3 SD) were removed from the analysis. LOY calling was performed for those individuals with low mLRR-Y using the MADloy R-package (https://github.com/isglobal-brge/MADloy) (19).

## GTEx data

We downloaded version-6 data from the GTEx project website. RNA-seq count data was obtained for 52 different tissues. Pair-ended RNA-seq was performed with Illumina HiSeq 2000 following the TrueSeq RNA protocol, see The GTEx Consortium (gtexportal.org). We analyzed exon-wise count data normalized in reads per kilobase per million mapped reads (RPKM). We were given access to download GTEx genotypes from dbGAP with accession number phs000424.v6.p1. Approximately 1.9 million SNPs were genotyped using whole blood samples with Illumina HumanOmni 2.5 M and 5M BeadChips. 47 male tissues were selected for EDY detection. Within each tissue, the males' relative Y expression ($Ry$) was compared to that of females, as the baseline noise given by the erroneous mapping of female reads to Y sequences.

We determined EDY status across 47 undiseased tissues in males of European ancestry (Figure S1). We detected 140 subjects, from a total of 371, with EDY in at least one tissue. There was a large variability of EDY frequency between tissues (max=20% in bladder, min=0% in stomach) (Table S1).To test whether EDY was likely to appear in multiple tissues, we first permuted the subject labels of EDY status 10,000 times within tissues, recounted the final number of subjects with EDY in any tissue and computed the probability of detecting 140 EDY cases or lower.

Genome-wide SNP data was available for 298 males for whom we could determine their EDY status. SNP data were filtered for minor allele frequency ($> 0.5\%$) and Hardy-Weinberg equilibrium. We used Bioconductor's snpStats package for SNP quality control and computed genome-wide principal components. Genome-wide associations between EDY and SNPs were tested in R with logistic regressions adjusting for age and principal components. For enrichment analyses, we extracted intra-genic SNPs and then pruned them by linkage disequilibrium ($R^2 < 0.2$). The significance level of a gene's association was computed by the combined P-value of its intra-genic SNPs, as implemented in the survcomp package. Enrichment analyses were performed for genes selected by nominal significance using the Bioconductor's clusterProfiler.

## TCGA data

We analyzed the multiomic data for 28 TCGA cancer studies. We used data from a total number of 10,642 samples, where 5,329 were from normal tissues and 5,313 were tumorous tissues. LOY status was inferred from Affymetrix Genome-Wide Human SNP 6.0 data processed with Birdseed v2 algorithm. LOY was called in 9,927 participants after quality control measures of LRR. We used Bioconductor packages RTCGA to download clinical and RNA-seq data. We estimated meta-analysis, mediation and Bayesian network models using the R packages metafor, mediation and abn. We analyzed exon-wise count data normalized in reads per kilobase per million mapped reads (RPKM), copy number variants (CNV) and methylation data to perform downstream analyses on male samples. Clinical, CNV and methylation data were obtained from the Bioconductor package RTCGA. For transcriptomic associations with EDY, we used Bioconductor's packages. We normalized count data with DESeq2, detected surrogate variables with SVA, corrected for batch effects with VOOM and fitted regression models with limma. Only probes with more than 15 counts in 25% of the samples were included. For CNV data we considered signal larger than log2(3/2) as gains and lower than -1 as copy number losses and tested the differences in the proportion of CNVs between EDY statuses within 1.25Mb windows across the genome, using an exact Fisher's test. For methylation data, we considered probes with more than 80% call rate and analyzed methylation percentage given by 100 times the methylation beta values. Logistic Bayesian regression models on EDY were fitted for all CPG probes in chromosome Y, adjusting by age and cancer type.

We counted the number of lost of function mutations within the genes DDX3X, EIF1AX, KDM5C and KDM6A/UTX for each individual within TCGA, as reported in the RTCGA.mutations package. We selected frameshift deletions and insertions, nonsense and splice site mutations within the genes. We tested the correlation between the presence of any of these mutations and EDY/LOY using logistic regression and adjusting by type of cancer and age.

## Validation Studies

Gene expression data were searched in the GEO repository (www.ncbi.nlm.nih.gov/geo), see main text. We also downloaded from ArrayExpress Archive (www.ebi.ac.uk/arrayexpress) a large expression matrix of 27,871 arrays with accession number E-MTAB-3732, the largest systematically annotated gene expression dataset of its kind. The data corresponds to a selection from a total 40,871 publicly available Affymetrix HG-U133Plus2 arrays, for which strict quality control and data normalization were applied (20). Samples annotations were available with tissue and disease status but not sex. Sex status was inferred from the clustering 3rd and 4th principal components of the probes within the X and Y chromosomes (Figure S4). We tested the accuracy of the inference using the study GSE12667 included in the E-MTAB-3732 collection and with reported sex. The study included 24 females and 21 males. When matching the reported sex with the inferred sex, we observed a 100% match. Female relative expression of Y (Ry) was further used as a noise measure to determine the quality of the sex inference (Figure S5). Diseased groups with less than 30 individuals were discarded, leaving at total 7,730 men.

We analyzed the gene expression levels in peripheral blood of individuals from the Estonian Gene Expression Cohort (EGCUT) (http://www.biobank.ee/). This cohort is composed of 1,074 randomly selected Estonian individuals ($37 \pm 16.6$ years; 50% females) from the $\sim$53,000 subjects in the Estonian Genome Center Biobank at the University of Tartu. Whole-genome gene-expression levels were obtained by Illumina HT12v3 arrays according to manufactures protocols. We analyzed transcriptomic data for 550 males, 52 of whom were diagnosed with cancer.

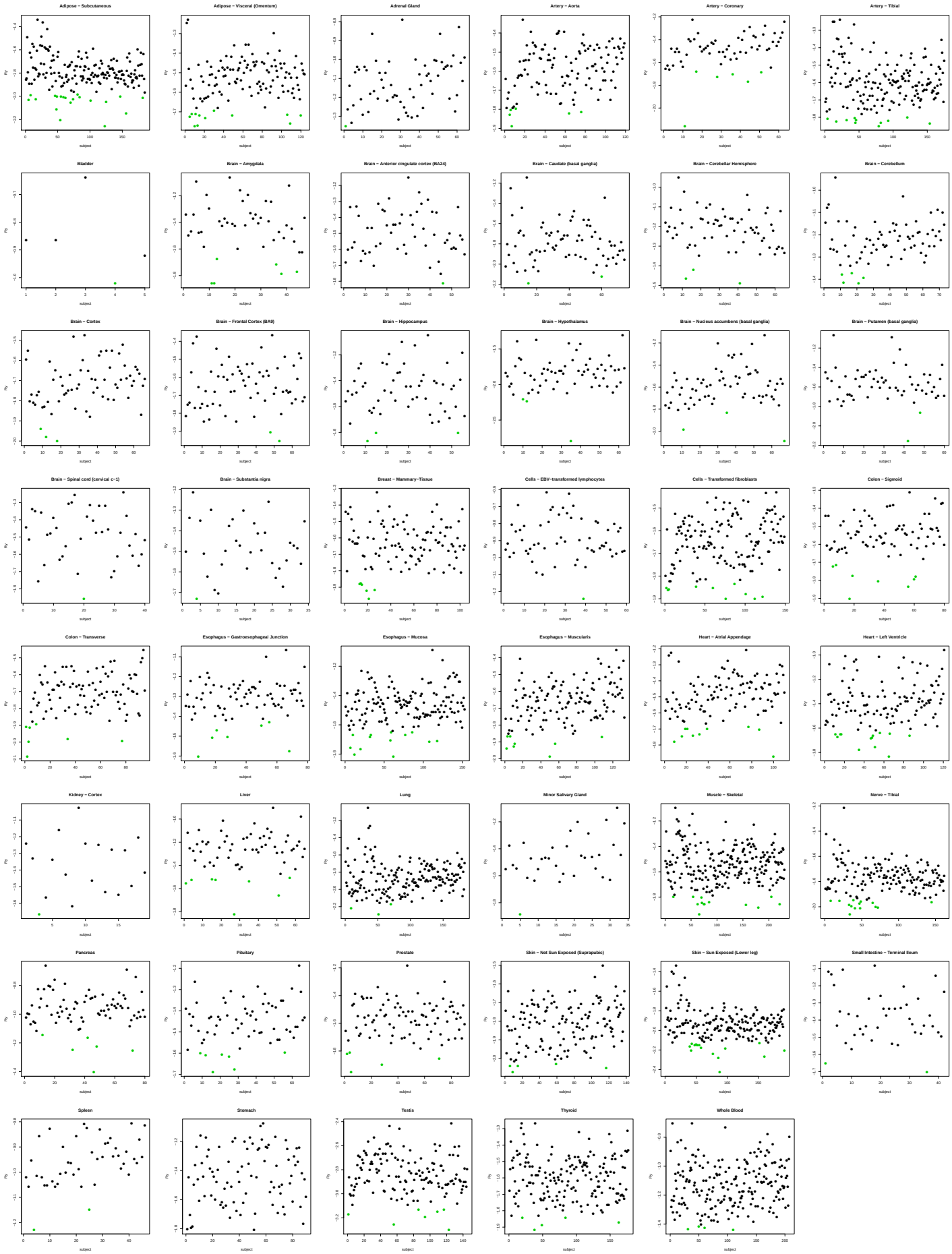



**Figure S1.** Relative expression of Y ($Ry$) for males across 47 human tissues from the GTEx project. EDY status of individuals is shown in green.

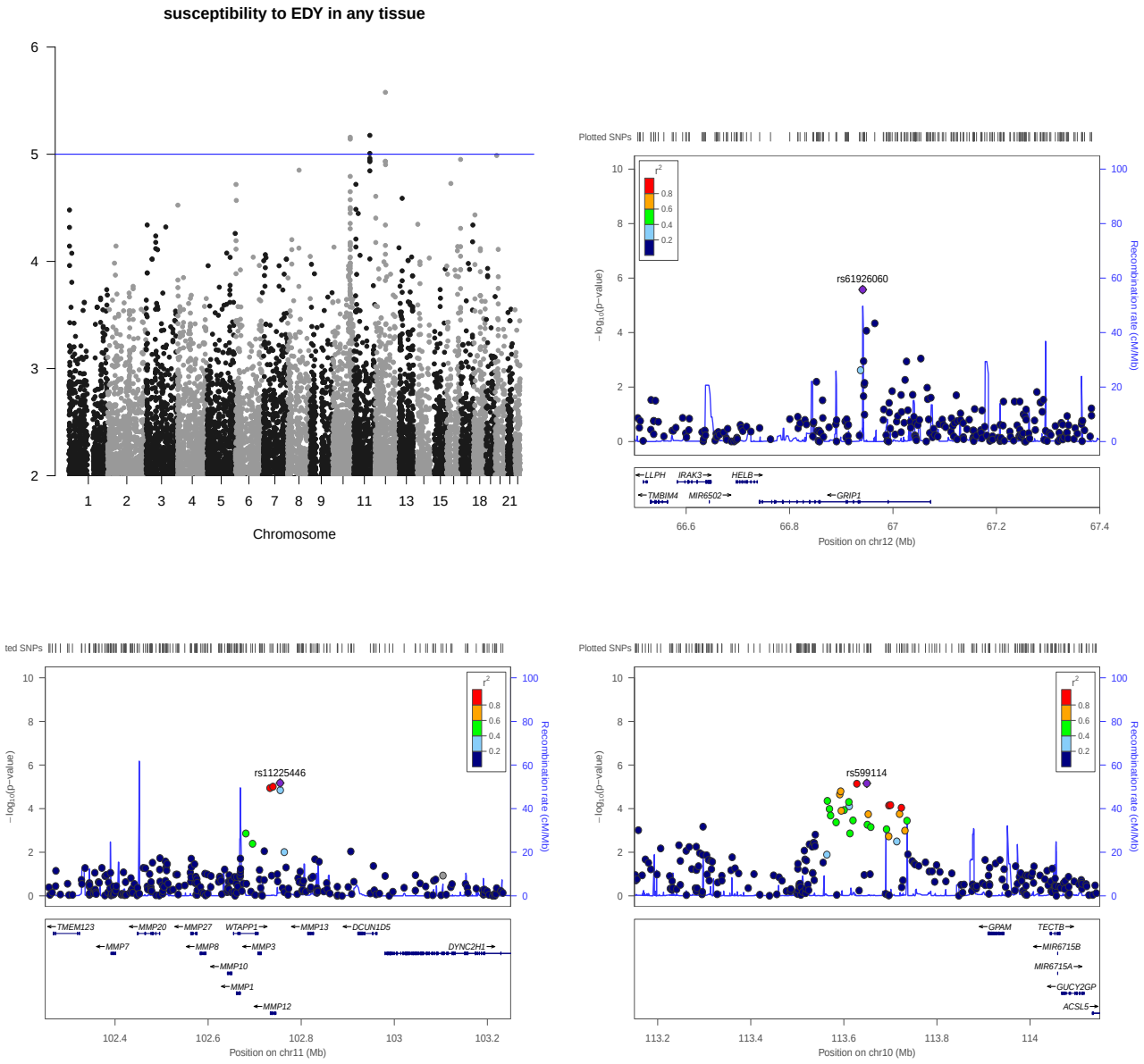



**Figure S2**. Genome-wide association analysis of EDY in more than 2 tissues of the 47 GTEx tissues.

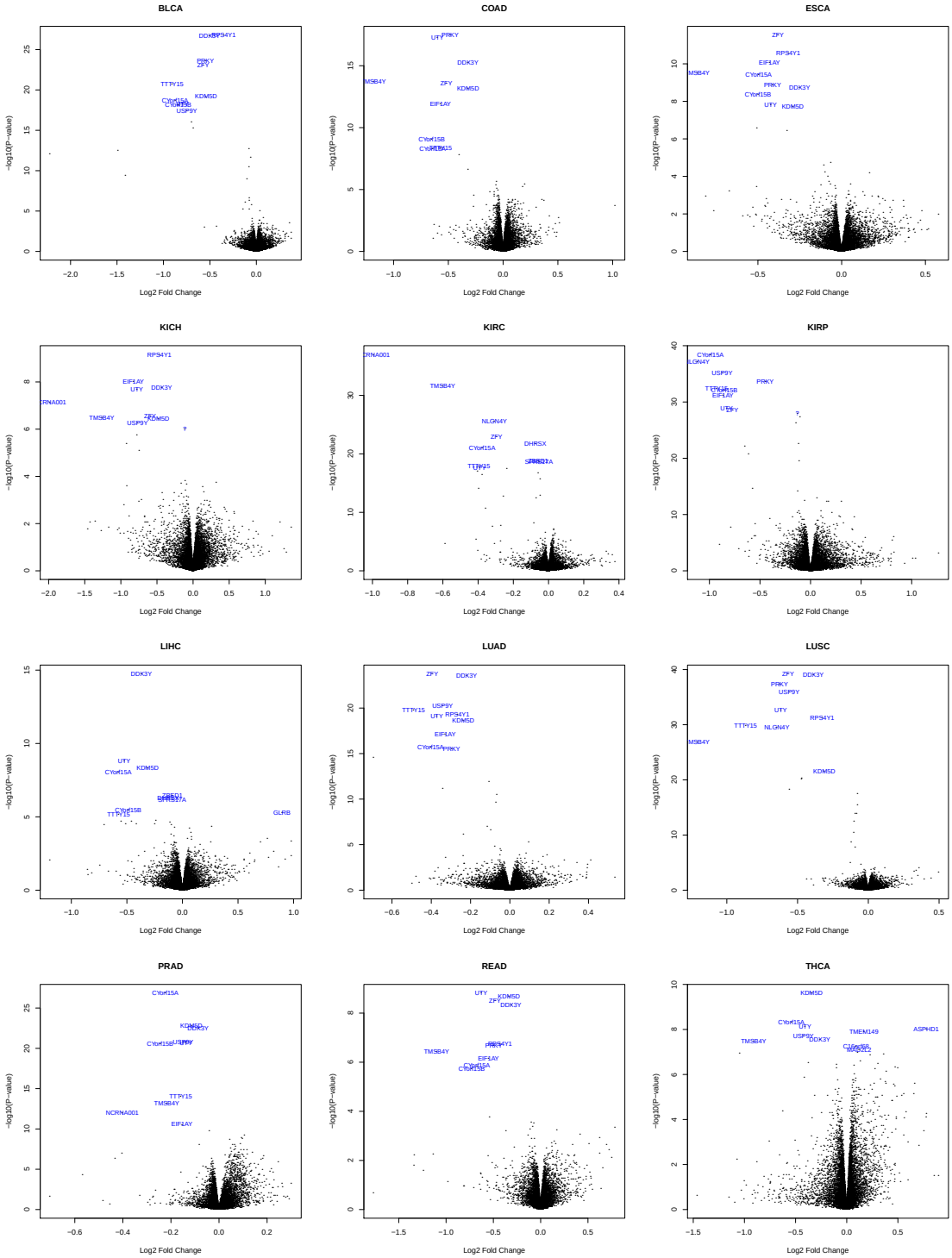



**Figure S3.** Volcano plots for differenential gene expression associated with EDY in 12 cancer studies from TCGA

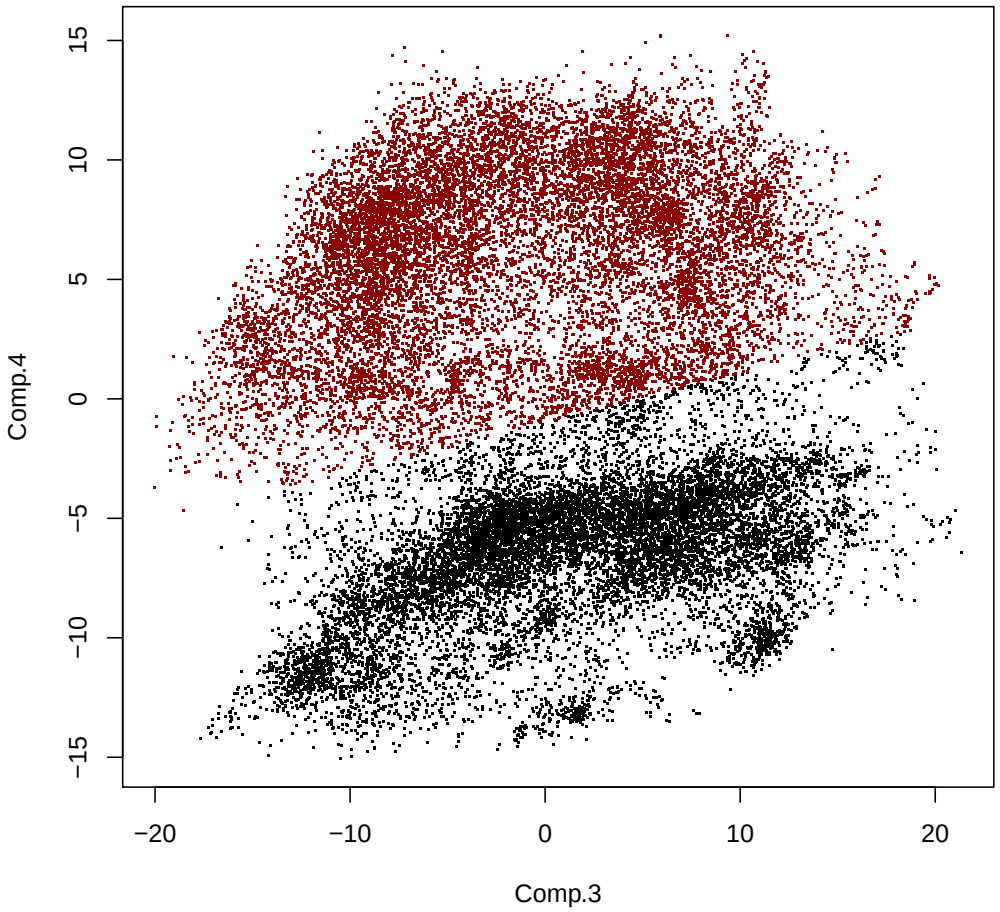



**Figure S4**. 3rd and 4th PCs of transcriptomic probes in X and Y across 27,887 arrays. Two clusters were identified (red and black) as trasncriptomic inferences on the sex of the individuals (males=black, females=red).

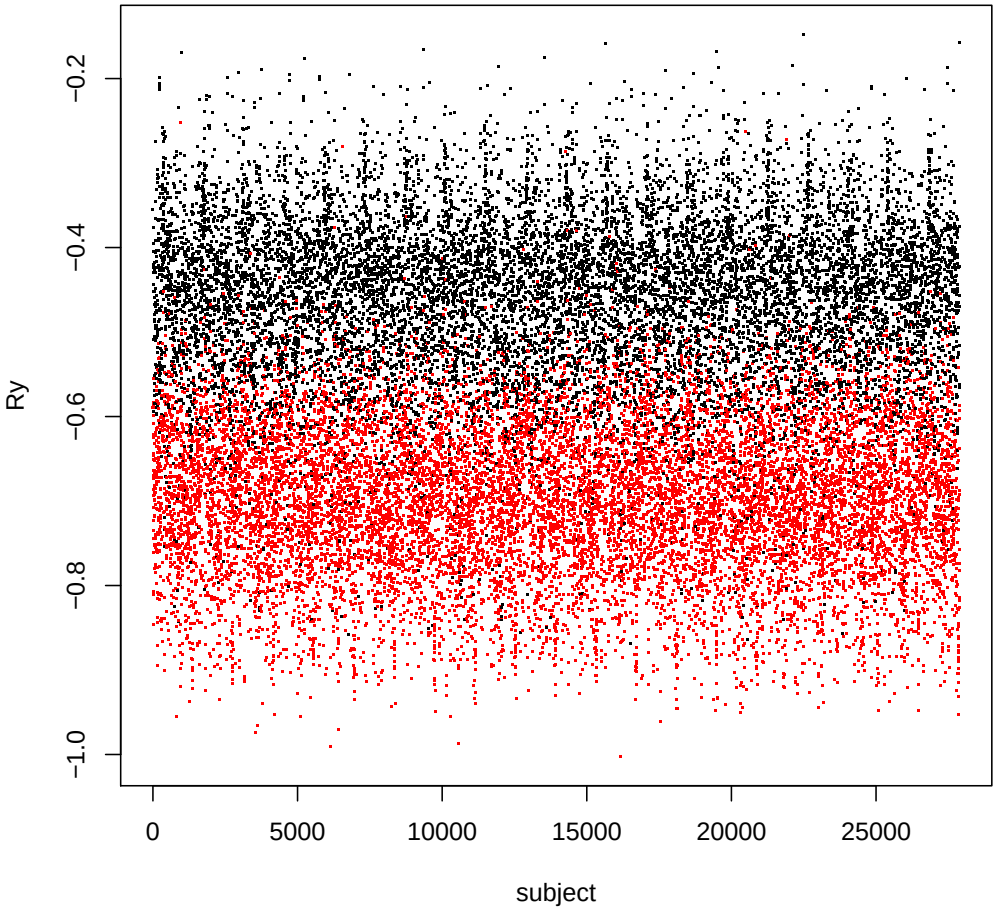



**Figure S5**. Relative expression of Y ($Ry$) for 27,887 arrays where each individual is colored by the sex inference given in Figure S7, confirming that inferred females have lower values of $Ry$ (noise) than males.

| tissue | EDY | % EDY | EDY in other tissues |
|---|---|---|---|
| Adipose - Subcutaneous | 20 | 10.81 | 13 (65 %) |
| Nerve - Tibial | 14 | 8.64 | 13 (92.9 %) |
| Heart - Left Ventricle | 13 | 10.74 | 11 (84.6 %) |
| Adipose - Visceral (Omentum) | 12 | 9.68 | 9 (75 %) |
| Artery - Tibial | 11 | 6.29 | 8 (72.7 %) |
| Esophagus - Mucosa | 13 | 8.50 | 8 (61.5 %) |
| Esophagus - Muscularis | 8 | 6.02 | 8 (100 %) |
| Heart - Atrial Appendage | 10 | 9.09 | 8 (80 %) |
| Muscle - Skeletal | 13 | 5.65 | 8 (61.5 %) |
| Skin - Sun Exposed (Lower leg) | 15 | 7.81 | 8 (53.3 %) |
| Liver | 8 | 12.31 | 7 (87.5 %) |
| Cells - Transformed fibroblasts | 10 | 6.71 | 6 (60 %) |
| Colon - Sigmoid | 8 | 10.00 | 6 (75 %) |
| Brain - Amygdala | 6 | 12.77 | 5 (83.3 %) |
| Breast - Mammary-Tissue | 6 | 5.77 | 5 (83.3 %) |
| Colon - Transverse | 7 | 7.37 | 5 (71.4 %) |
| Pancreas | 6 | 7.50 | 5 (83.3 %) |
| Pituitary | 7 | 10.45 | 5 (71.4 %) |
| Artery - Aorta | 6 | 5.04 | 4 (66.7 %) |
| Esophagus - Gastroesophageal Junction | 7 | 8.97 | 4 (57.1 %) |
| Prostate | 5 | 5.56 | 4 (80 %) |
| Skin - Not Sun Exposed (Suprapubic) | 5 | 3.62 | 4 (80 %) |
| Testis | 7 | 4.83 | 4 (57.1 %) |
| Thyroid | 5 | 2.82 | 4 (80 %) |
| Whole Blood | 4 | 1.92 | 4 (100 %) |
| Artery - Coronary | 6 | 9.52 | 3 (50 %) |
| Brain - Cerebellum | 5 | 6.85 | 3 (60 %) |
| Brain - Cortex | 3 | 4.55 | 3 (100 %) |
| Brain - Hippocampus | 3 | 5.36 | 3 (100 %) |
| Lung | 3 | 1.64 | 3 (100 %) |
| Brain - Cerebellar Hemisphere | 3 | 4.62 | 2 (66.7 %) |
| Brain - Hypothalamus | 3 | 4.76 | 2 (66.7 %) |
| Brain - Nucleus accumbens (basal ganglia) | 3 | 4.48 | 2 (66.7 %) |
| Small Intestine - Terminal Ileum | 2 | 4.76 | 2 (100 %) |
| Spleen | 2 | 4.35 | 2 (100 %) |
| Adrenal Gland | 1 | 1.56 | 1 (100 %) |
| Bladder | 1 | 20.00 | 1 (100 %) |
| Brain - Caudate (basal ganglia) | 2 | 2.70 | 1 (50 %) |
| Brain - Frontal Cortex (BA9) | 2 | 2.99 | 1 (50 %) |
| Brain - Putamen (basal ganglia) | 2 | 3.33 | 1 (50 %) |
| Brain - Spinal cord (cervical c-1) | 1 | 2.50 | 1 (100 %) |
| Cells - EBV-transformed lymphocytes | 1 | 1.69 | 1 (100 %) |
| Kidney - Cortex | 1 | 5.26 | 1 (100 %) |
| Minor Salivary Gland | 1 | 2.94 | 1 (100 %) |
| Brain - Anterior cingulate cortex (BA24) | 1 | 1.79 | 0 (0 %) |
| Brain - Substantia nigra | 1 | 2.94 | 0 (0 %) |
| Stomach | 0 | 0.00 | 0 (-) |

**Table S1**. EDY detection in GTEx

**Outcome: *tumor/healthy tissue of cancer patients***
**Study: TCGA**

| Biological sample | N (tumor/healthy) | % EDY |
|---|---|---|
| Bladder Urothelial Carcinoma (BLCA) | 204 (200/4) | 37.75 |
| Colon adenocarcinoma (COAD) | 98 (95/3) | 34.69 |
| Esophageal carcinoma (ESCA) | 83 (78/5) | 46.99 |
| Kidney Chromophobe (KICH) | 39 (27/12) | 53.85 |
| Kidney renal clear cell carcinoma (KIRC) | 324 (280/44) | 47.22 |
| Kidney renal papillary cell carcinoma (KIRP) | 158 (148/15) | 79.75 |
| Liver hepatocellular carcinoma (LIHC) | 97 (87/10) | 16.49 |
| Lung adenocarcinoma (LUAD) | 163 (148/15) | 34.97 |
| Lung squamous cell carcinoma (LUSC) | 230 (217/13) | 44.78 |
| Prostate adenocarcinoma (PRAD) | 373 (368/5) | 12.33 |
| Rectum adenocarcinoma (READ) | 37 (36/1) | 45.95 |
| Thyroid carcinoma (THCA) | 92 (85/7) | 7.61 |
| OVERALL | 1774 (1646/128) | 41.65 |

**Outcome: *case/control subject***
**Study: GSE66836**

| Disease status | N (subjects) | EDY(%) |
|---|---|---|
| Normal | 71 | 2.81 |
| Colorectal cancer | 71 | 15.49 |

**Study: GSE36895**

| | | |
|---|---|---|
| Normal | 12 | 2.81 |
| Kidney renal clear cell carcinoma | 17 | 15.49 |

**Study: EGCUT (*population based*)**

| | | |
|---|---|---|
| Normal (blood) | 498 | 8.74 |
| cancer diagnosis (blood) | 52 | 25 |

**Study: E-MTAB-3732 (*Multiple diseases*)**

| | | |
|---|---|---|
| Normal | 3112 | 3.6 |
| Cancer-colon | 283 | 35.3 |
| Cancer-gastric | 109 | 17.4 |
| Cancer-glioma | 157 | 24.8 |
| Cancer-hepatocellular carcinoma | 212 | 6.1 |
| Cancer-leukaemia | 840 | 0.2 |
| Cancer-lung | 291 | 19.2 |
| Cancer-lymphoma | 274 | 1.8 |
| Cancer-malanoma | 108 | 42.6 |
| Cancer-myeloma | 712 | 1 |
| Cancer-neuroblastoma | 51 | 2 |
| Cancer-pancreatic | 33 | 33.3 |
| Cancer-prostate | 214 | 0.9 |
| Cancer-undifferentiated sarcoma | 34 | 23.5 |

**Outcome: *EDY in tumors of LUSC patients***
**Study: GSE4573**

| EDY | N | > 1 package/day ( % ) |
|---|---|---|
| no | 77 | 57.14 |
| yes | 5 | 100 |

**Study: GSE5123**

| EDY | N | Mean (cigarrets/day) |
|---|---|---|
| no | 33 | 3.09 |
| yes | 4 | 5.57 |

**Table S2**. List of studies analyzed with their EDY detection frequency.

| Cancer study | N | % LOY | OR | P |
|---|---|---|---|---|
| Adrenocortical carcinoma (ACC) | 53 | 32.10 | 167.66 | 3.1e-03** |
| Bladder Urothelial Carcinoma (BLCA) | 470 | 30.00 | 3.63 | 8.1e-09*** |
| Breast invasive carcinoma (BRCA) | 14 | 0.00 | 1.00 | 1.0e+00 |
| Cholangiocarcinoma (CHOL) | 28 | 25.00 | 6.37 | 5.4e-02 |
| Colon adenocarcinoma (COAD) | 213 | 36.60 | 6.97 | 8.7e-09*** |
| Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC) | 28 | 28.60 | 8.58 | 2.6e-02* |
| Esophageal carcinoma (ESCA) | 185 | 24.90 | 62.55 | 1.7e-06*** |
| Glioblastoma multiforme (GBM) | 570 | 15.10 | 0.96 | 8.5e-01 |
| Head and Neck squamous cell carcinoma (HNSC) | 630 | 29.00 | 10.66 | 4.5e-24*** |
| Kidney Chromophobe (KICH) | 68 | 30.90 | 33.82 | 9.9e-05*** |
| Kidney renal clear cell carcinoma (KIRC) | 597 | 31.30 | 9.76 | 2.6e-23*** |
| Kidney renal papillary cell carcinoma (KIRP) | 302 | 47.70 | 38.14 | 4.8e-26*** |
| Acute Myeloid Leukemia (LAML) | 121 | 24.80 | 0.91 | 8.3e-01 |
| Brain Lower Grade Glioma (LGG) | 520 | 5.80 | 3.72 | 3.3e-03** |
| Liver hepatocellular carcinoma (LIHC) | 201 | 12.40 | 3.85 | 6.6e-03** |
| Lung adenocarcinoma (LUAD) | 344 | 28.50 | 5.19 | 3.3e-09*** |
| Lung squamous cell carcinoma (LUSC) | 495 | 34.90 | 10.81 | 1.0e-22*** |
| Pancreatic adenocarcinoma (PAAD) | 179 | 34.60 | 7.11 | 9.8e-08*** |
| Pheochromocytoma and Paraganglioma (PCPG) | 139 | 1.40 | 5.39 | 2.7e-01 |
| Prostate adenocarcinoma (PRAD) | 801 | 9.40 | 0.65 | 7.5e-02 |
| Rectum adenocarcinoma (READ) | 90 | 32.20 | 7.93 | 1.2e-04*** |
| Sarcoma (SARC) | 223 | 9.00 | 15.09 | 1.4e-03** |
| Skin Cutaneous Melanoma (SKCM) | 550 | 21.60 | 8.71 | 2.0e-14*** |
| Stomach adenocarcinoma (STAD) | 336 | 34.20 | 7.69 | 2.1e-13*** |
| Testicular Germ Cell Tumors (TGCT) | 237 | 36.70 | 40.47 | 1.4e-14*** |
| Thyroid carcinoma (THCA) | 173 | 13.30 | 2.87 | 3.0e-02* |
| Thymoma (THYM) | 103 | 21.40 | 0.77 | 5.9e-01 |
| Uveal Melanoma (UVM) | 60 | 28.30 | 0.86 | 7.8e-01 |
| OVERALL EFFECT | 7730 | 23.90 | 5.78 | 1.13e-12*** |

**Table S3**. LOY detection in 28 cancer studies from the TCGA project. The total number of samples in each study is given (N) together with the percentage of individuals with LOY (% LOY). Odds ratios (OR) and P-values, corresponding to the associations between LOY and cancer status of the samples, are also shown.

| cpg | Gene | OR (strat NoLOY) | P | OR (strat LOY) | P | combined P |
|---|---|---|---|---|---|---|
| cg04169747 | SRY | 0.96 | 2.67e-06 | 0.92 | 6.05e-28 | 1.24e-31 |
| cg27636129 | SRY | 0.97 | 1.84e-04 | 0.92 | 3.46e-28 | 4.63e-30 |
| cg09595415 | SRY | 0.98 | 1.51e-03 | 0.92 | 6.05e-28 | 6.40e-29 |
| cg00479827 | NCRNA00185 | 0.93 | 3.12e-06 | 0.89 | 1.40e-22 | 2.79e-26 |
| cg02107461 | | 1.02 | 2.01e-02 | 1.08 | 8.37e-25 | 1.01e-24 |
| cg23834181 | | 1.02 | 1.30e-03 | 1.08 | 3.15e-23 | 2.43e-24 |
| cg18163559 | | 1.03 | 7.08e-03 | 1.08 | 1.06e-21 | 4.07e-22 |
| cg27443332 | NLGN4Y | 0.98 | 2.30e-03 | 0.94 | 3.88e-21 | 4.83e-22 |
| cg13805219 | | 1.03 | 3.54e-03 | 1.07 | 3.88e-20 | 7.05e-21 |
| cg04691144 | NLGN4Y | 0.98 | 2.98e-02 | 0.94 | 1.08e-19 | 1.55e-19 |
| cg02233183 | NLGN4Y | 0.98 | 2.93e-02 | 0.92 | 1.96e-19 | 2.74e-19 |
| cg11898347 | SRY | 0.97 | 3.91e-05 | 0.95 | 6.89e-16 | 1.24e-18 |
| cg09230658 | | 1.03 | 2.08e-03 | 1.12 | 1.79e-17 | 1.71e-18 |
| cg14303457 | | 1.05 | 1.18e-04 | 1.08 | 4.37e-16 | 2.34e-18 |
| cg01463110 | | 1.03 | 4.11e-03 | 1.06 | 4.09e-17 | 7.42e-18 |
| cg01900066 | EIF1AY | 0.96 | 2.02e-02 | 0.92 | 3.10e-17 | 2.69e-17 |
| cg27049643 | KDM5D | 1.06 | 8.50e-03 | 1.15 | 2.49e-16 | 8.81e-17 |
| cg04021548 | NLGN4Y | 1.02 | 4.76e-02 | 1.08 | 5.19e-17 | 1.03e-16 |
| cg08528516 | DDX3Y | 0.94 | 4.46e-02 | 0.86 | 1.09e-16 | 1.99e-16 |
| cg18188392 | | 0.98 | 1.58e-02 | 0.95 | 1.08e-15 | 6.76e-16 |
| cg26517491 | KDM5D | 1.09 | 1.33e-02 | 1.18 | 3.81e-15 | 1.96e-15 |
| cg03767353 | | 0.96 | 5.75e-04 | 0.92 | 1.55e-13 | 3.39e-15 |
| cg15563434 | | 0.95 | 2.13e-05 | 0.94 | 9.50e-12 | 7.50e-15 |
| cg00063477 | EIF1AY | 0.95 | 9.22e-04 | 0.90 | 3.21e-13 | 1.09e-14 |
| cg13654344 | SRY | 0.97 | 4.58e-04 | 0.95 | 7.94e-13 | 1.33e-14 |
| cg15027426 | BCORL2 | 0.97 | 5.68e-04 | 0.95 | 3.63e-12 | 7.18e-14 |
| cg07939587 | | 0.97 | 2.48e-03 | 0.93 | 1.13e-12 | 9.68e-14 |
| cg15329860 | KDM5D | 1.05 | 4.06e-02 | 1.22 | 4.69e-13 | 6.20e-13 |
| cg08820785 | EIF1AY | 1.03 | 4.29e-02 | 1.19 | 9.96e-12 | 1.26e-11 |
| cg27214488 | NLGN4Y | 0.98 | 2.16e-02 | 0.95 | 2.98e-11 | 1.87e-11 |
| cg14463736 | TMSB4Y | 1.02 | 1.67e-02 | 1.06 | 2.25e-09 | 9.39e-10 |
| cg15794778 | BCORL2 | 0.98 | 5.66e-03 | 0.96 | 8.54e-08 | 1.08e-08 |
| cg04303809 | | 0.97 | 2.89e-02 | 0.96 | 5.31e-08 | 3.27e-08 |
| cg14720093 | | 0.98 | 3.93e-02 | 0.95 | 5.86e-08 | 4.80e-08 |
| cg13845521 | TTTY14 | 1.02 | 1.96e-03 | 0.97 | 2.35e-06 | 9.30e-08 |
| cg14210405 | | 0.99 | 3.88e-02 | 0.97 | 1.65e-07 | 1.27e-07 |
| cg02582450 | | 0.98 | 3.27e-02 | 0.96 | 3.02e-07 | 1.92e-07 |
| cg09748856 | NLGN4Y | 0.96 | 4.28e-02 | 0.96 | 3.93e-07 | 3.18e-07 |
| cg00876332 | BCORL2 | 0.98 | 5.33e-03 | 0.97 | 7.62e-06 | 7.31e-07 |
| cg03359666 | | 1.02 | 7.23e-03 | 1.03 | 1.83e-05 | 2.23e-06 |
| cg10172760 | EIF1AY | 1.02 | 2.53e-02 | 1.05 | 6.18e-06 | 2.61e-06 |
| cg10338539 | SRY | 0.97 | 2.16e-03 | 0.97 | 7.64e-05 | 2.74e-06 |
| cg20106158 | | 0.97 | 6.63e-03 | 1.04 | 1.34e-04 | 1.33e-05 |
| cg06060201 | LOC401630;LOC401629 | 1.04 | 1.46e-03 | 1.02 | 3.65e-03 | 7.01e-05 |
| cg25012987 | LOC401629;LOC401630 | 1.04 | 2.98e-03 | 1.03 | 1.92e-03 | 7.49e-05 |
| cg16292375 | LOC401629;LOC401630 | 1.03 | 8.79e-03 | 1.02 | 7.57e-04 | 8.60e-05 |
| cg01141334 | | 0.98 | 4.46e-02 | 1.02 | 4.42e-04 | 2.33e-04 |
| cg10593480 | EIF1AY | 1.03 | 8.72e-03 | 1.04 | 2.33e-03 | 2.40e-04 |
| cg14170959 | ZFY | 1.02 | 2.87e-02 | 0.98 | 8.12e-04 | 2.72e-04 |
| cg00639218 | | 1.02 | 3.81e-03 | 1.02 | 1.07e-02 | 4.53e-04 |
| cg04964672 | | 0.97 | 1.76e-02 | 0.97 | 5.67e-03 | 1.02e-03 |
| cg16894943 | LOC401630;LOC401629 | 1.02 | 1.61e-02 | 1.01 | 1.62e-02 | 2.42e-03 |

**Table S4**. Significant methylation differences associated with EDY status, stratified by LOY.

| cpg | genenms | Beta | Adjusted P |
|---|---|---|---|
| cg10691859 | NCRNA00185 | 1.08 | 1.36e-07 |
| cg01463110 | | 1.73 | 2.07e-07 |
| cg01900066 | EIF1AY | 0.94 | 1.08e-06 |
| cg05618150 | PRKY | -1.62 | 1.09e-06 |
| cg15662272 | KDM5D | -0.72 | 1.22e-06 |
| cg08528516 | DDX3Y | 0.69 | 4.95e-06 |
| cg14442616 | DDX3Y | 0.71 | 7.45e-06 |
| cg00063477 | EIF1AY | 0.78 | 8.11e-06 |
| cg27049643 | KDM5D | -0.63 | 9.89e-06 |
| cg01086462 | | 0.68 | 1.08e-05 |
| cg15329860 | KDM5D | -0.53 | 1.81e-05 |
| cg25815185 | KDM5D | -0.45 | 4.38e-05 |
| cg04169747 | SRY | -1.44 | 4.40e-05 |
| cg26517491 | KDM5D | -0.51 | 4.80e-05 |
| cg02129146 | | 1.04 | 6.32e-05 |
| cg27254225 | | 0.62 | 6.67e-05 |
| cg27433982 | ZFY | -0.45 | 1.20e-04 |
| cg05128824 | DDX3Y | 0.59 | 1.38e-04 |

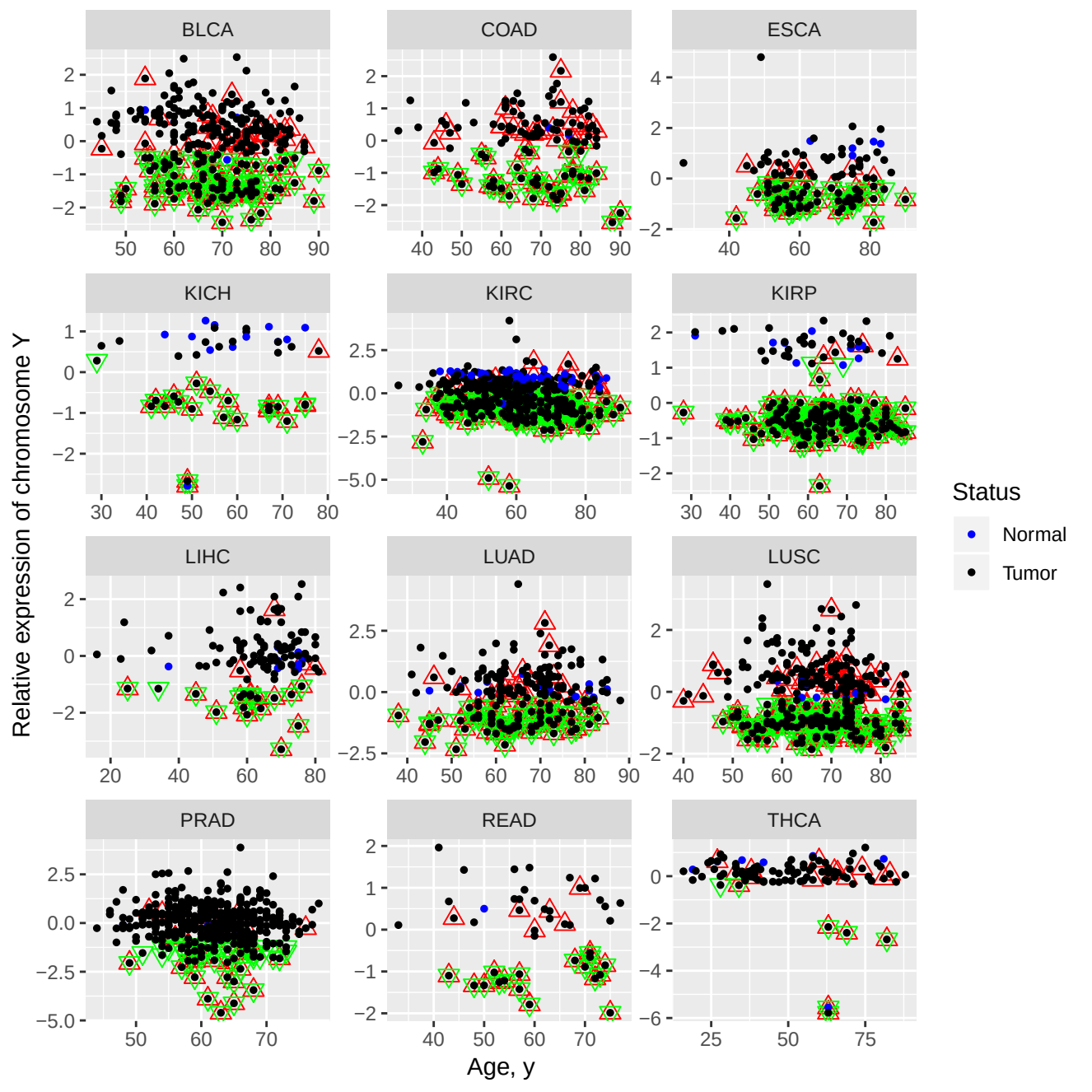
**Table S5**. Significant effects of *EGFR* on methylation probes across Y, adjusted by EDY, LOY, age and cancer type. *EGFR* expression associates with four genes (*DDX3Y*, *EIF1AY*, *KDM5D* and *ZFY* ) of the six-gene transcriptomic signature of EDY.

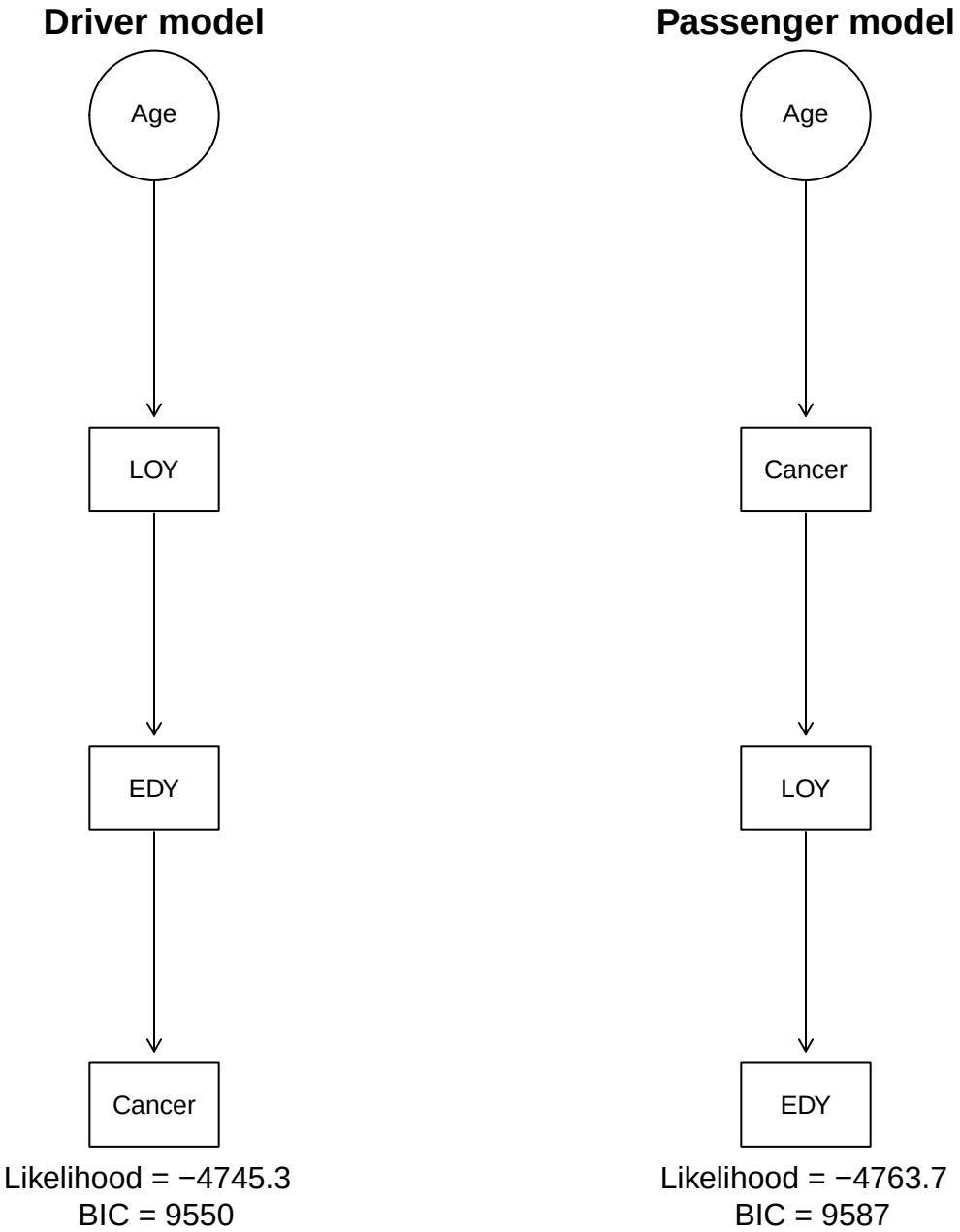
| cpg | genenms | Beta | P |
|---|---|---|---|
| cg27636129 | SRY | 3.71 | 3.37e-11 |
| cg04169747 | SRY | 3.34 | 3.69e-10 |
| cg09595415 | SRY | 3.49 | 5.03e-10 |
| cg11898347 | SRY | 3.61 | 4.65e-09 |
| cg10338539 | SRY | 2.51 | 4.73e-07 |
| cg17816615 | DDX3Y | -1.78 | 2.58e-06 |
| cg03601053 | DDX3Y | -1.18 | 2.73e-06 |
| cg13654344 | SRY | 2.35 | 3.64e-06 |
| cg04576441 | | -1.70 | 4.27e-06 |
| cg15027426 | BCORL2 | 1.99 | 1.74e-05 |
| cg01463110 | | -2.10 | 3.16e-05 |
| cg20106158 | | 1.21 | 1.30e-04 |
| cg00639218 | | -2.20 | 1.95e-04 |

**Table S6**. Significant effects of *SOX4* on methylation probes across Y, adjusted by EDY, LOY, age and cancer type.

| cpg | genenms | Beta | P |
|---|---|---|---|
| cg18077436 | DDX3Y | 0.94 | 1.96e-05 |
| cg15662272 | KDM5D | -0.55 | 1.01e-04 |

**Table S7**. Significant effects of *NCOA2* on methylation probes across Y, adjusted by EDY, LOY, age and cancer type.


BLCA
COAD
ESCA
KICH
KIRC
KIRP
LIHC
LUAD
LUSC
PRAD
READ
THCA
Relative expression of chromosome Y
Age, y
Status
Normal
Tumor

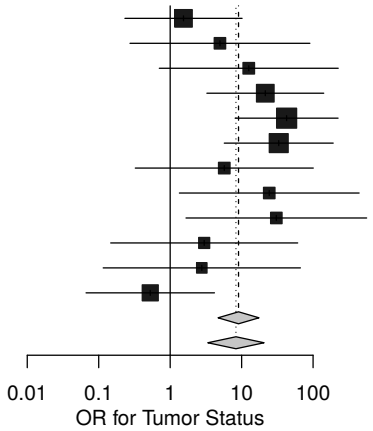

Driver model
Age
LOY
EDY
Cancer
Likelihood = −4745.3
BIC = 9550
Passenger model
Age
Cancer
LOY
EDY
Likelihood = −4763.7
BIC = 9587

A

**Tumor vs. normal tissue EDY**

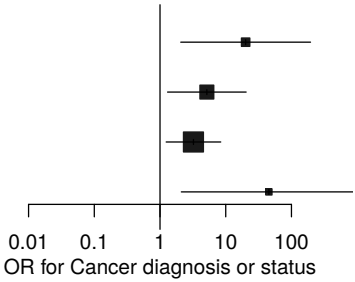
| **Meta−analysis** | **OR (95% CI)** |
|---|---|
| Bladder (BLCA) | 1.54 (0.23 to 10.15) |
| Colon (COAD) | 4.97 (0.27 to 90.17) |
| Esophageal (ESCA) | 12.59 (0.70 to 224.88) |
| Kidney (KICH) | 21.44 (3.26 to 140.99) |
| Kidney CCC (KIRC) | 42.62 (8.11 to 223.89) |
| Kidney PCC (KIRP) | 33.06 (5.72 to 191.25) |
| Hepatocellular (LIHC) | 5.72 (0.32 to 100.68) |
| Lung (LUAD) | 24.34 (1.34 to 440.82) |
| Lung squamous (LUSC) | 30.59 (1.66 to 562.18) |
| Prostate (PRAD) | 2.99 (0.15 to 60.94) |
| Rectum (READ) | 2.77 (0.12 to 66.49) |
| Thyroid (THCA) | 0.53 (0.07 to 4.17) |
| Total (fixed effect) | 9.05 (4.68 to 17.52) |
| Total (random effects) | 8.33 (3.36 to 20.62) |

Heterogeneity: $\chi^2_{11} = 19.38$ $P = .05$




B

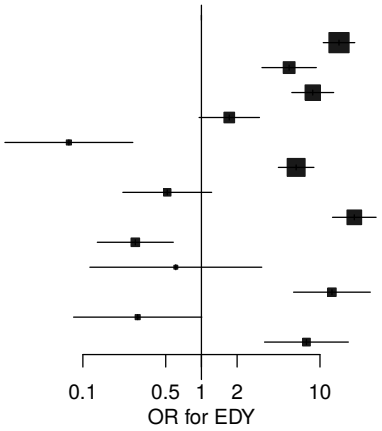
| **Effect of EDY on cancer** | **OR (95% CI)** |
|---|---|
| Cancer patients | |
| GSE36895: Kidney CCC (tumor vs. normal tissue EDY) | 20.09 (2.07 to 195.11) |
| Case−control | |
| GSE44076: Colorectal (tumor vs. normal tissue EDY) | 5.16 (1.30 to 20.45) |
| Population Based | |
| EGCUT: Any cancer (difference of EDY in blood) | 3.23 (1.24 to 8.40) |
| Methylation inferred EDY | |
| GSE61441: Kidney CCC (tumor vs. normal tissue EDY) | 45.40 (2.11 to 977.32) |




C

**Tumor vs. normal tissue EDY**

| **Multiple−regression** | **OR (95% CI)** |
|---|---|
| Normal−Ref | 1.00 |
| Colon | 14.47 (10.65 to 19.66) |
| Gastric | 5.49 ( 3.24 to 9.32) |
| Glioma | 8.69 ( 5.79 to 13.05) |
| Hepatocellular | 1.72 ( 0.96 to 3.09) |
| Leukaemia | 0.08 ( 0.02 to 0.26) |
| Lung | 6.30 ( 4.46 to 8.89) |
| Lymphoma | 0.51 ( 0.22 to 1.22) |
| Melanoma | 19.48 (12.75 to 29.75) |
| Myeloma | 0.28 ( 0.13 to 0.58) |
| Neuroblastoma | 0.61 ( 0.11 to 3.23) |
| Pancreatic | 12.62 ( 5.99 to 26.59) |
| Prostate | 0.29 ( 0.08 to 1.01) |
| Sarcoma (undifferentiated) | 7.69 ( 3.41 to 17.35) |

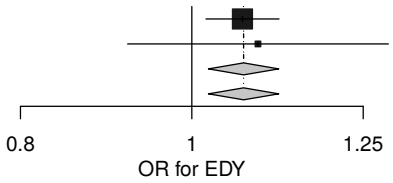


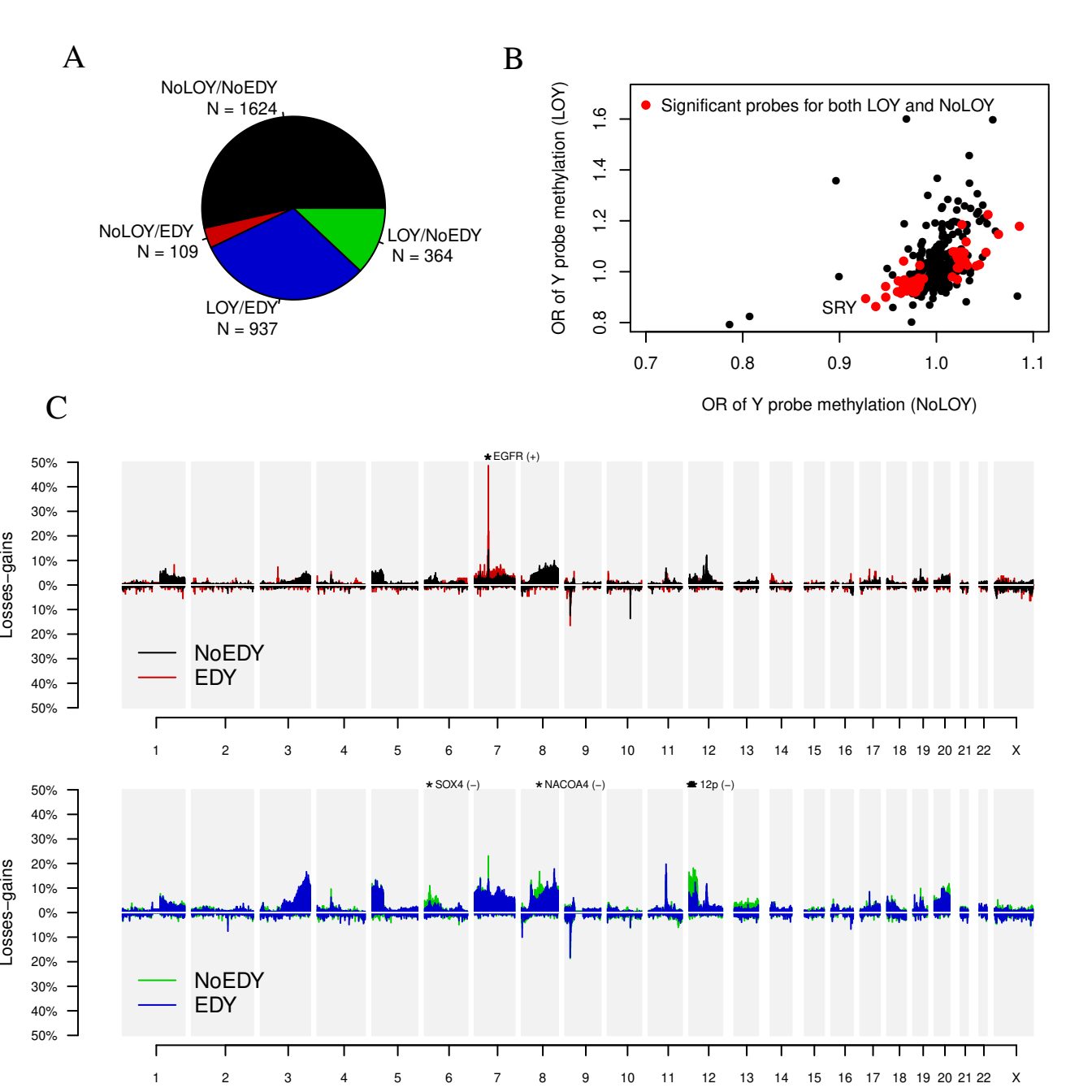

D

**Smoking cigarette/day (tumor tissue)**

| **Meta−analysis** | **OR (95% CI)** |
|---|---|
| GSE5123: Lung squamous cell carcinoma | 1.07 (1.02 to 1.12) |
| LUSC (TCGA): Lung squamous cell carcinoma | 1.09 (0.92 to 1.29) |
| Total (fixed effect) | 1.07 (1.02 to 1.12) |
| Total (random effects) | 1.07 (1.02 to 1.12) |

Heterogeneity: $\chi^2_1 = 0.05$ $P = .82$




A
NoLOY/NoEDY
N = 1624
NoLOY/EDY
N = 109
LOY/NoEDY
N = 364
LOY/EDY
N = 937
B
Significant probes for both LOY and NoLOY
SRY
OR of Y probe methylation (LOY)
OR of Y probe methylation (NoLOY)
0.8
1.0
1.2
1.4
1.6
0.7
0.8
0.9
1.0
1.1
C
EGFR (+)
SOX4 (−)
NACOA4 (−)
12p (−)
Losses−gains
NoEDY
EDY
50%
40%
30%
20%
10%
0%
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X