universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

"Pan-cancer analysis of chromosome Y loss"

verfasst von / submitted by

Christoph Kreitzer

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Biol. Dr. Ulrich Technau

Contents

Abstract

The last decade provided a deluge of tumor sequencing data; however, the interrogation of chromosome Y in genome-wide tumor analysis was almost universally neglected. Thus, its impact on tumorigenesis and progression remains largely elusive.

Here, we conducted a pan-cancer analysis of chromosome Y loss (LOY) across >13,000 patients spanning over 45 cancer types. We validated our findings using orthogonal data resources and further established correlations with clinicopathological and genomic factors. Furthermore, the prognostic value of LOY in various malignancies is elucidated.

Our study confirmed the suitability of MSK-IMPACT sequencing data to study chromosome Y aberrations in both normal and tumor specimen, respectively. We observed LOY in 34.9% of male specimen; however, with marked differences across and within cancer types. LOY was positively correlated with chromosomal instability, expressed through the fraction of the genome altered. Moreover, somatic mutations in *KDM5C*, *KDM6A*, and *CRLF2* were associated with LOY in various tumor lineages. Lastly, we provide evidence of the prognostic value of LOY in prostate adenocarcinoma, where overall survival is increased in patients retaining the Y chromosome.

We conclude that LOY is common and largely associated with p53-mediated genomic instability. Hence, for most tumor types, it depicts a byproduct of CIN. However, for prostate adenocarcinomas, the Y chromosome is of clinical significance, presumably through epigenetic factors such as *KDM5D*. Taken together, our study provides a comprehensive catalog of LOY estimates, confirms many of the recent findings, and further warrants a resource for continuative investigations.

Keywords: Next-generation sequencing, Copy-number alterations, Y chromosome, Tumor evolution, Tumor suppressor genes, Prostate cancer

1. Introduction

Sex-unspecific cancer incidence and mortality rates among female and male patients are disproportionate (Edgren *et al.*, 2012; Siegel, Miller and Jemal, 2017; Li *et al.*, 2018). A recent report demonstrated a male cancer mortality preference in the United States in 32 out of 36 cancer types between 1977 and 2006 (Cook *et al.*, 2011). Reasons for these sex differences include different physiology (Rubin *et al.*, 2020), the effect of sex hormones (Lopes-Ramos, Quackenbush and DeMeo, 2020) and environmental exposure (Scarselli *et al.*, 2018) among others. Most obvious, however, the intrinsic genetic sex differences caused by the dichotomous sex chromosomes. The human Y chromosome, the male-specific sex chromosome, evolved from an ancestral autosome hundreds of millions of years ago (Jobling and Tyler-Smith, 2003; Guo *et al.*, 2020). Despite its roles for testis determination and spermatogenesis it plays essential roles for male viability. It acquired a sex-determining function, and subsequent series of inversions suppressed crossing over with the X chromosome (Rice, 1996; Hughes *et al.*, 2010; Bachtrog, 2013). Over millions of years, genetic decay left only three percent of its ancestral genes that survived on the Y chromosome (Bachtrog, 2013; Bellott *et al.*, 2014). Hence, it was long considered a genomic wasteland.

Structurally, the Y chromosome depicts an acrocentric chromosome composed of two pseudoautosomal regions (PARs), a short arm (Yp), and a longer arm (Yq) (Sauter *et al.*, 1995). While the PARs undergo genetic recombination during meiotic crossover, the male-specific segment (MSY) is haploid and comprised of approximately 23 Mb of euchromatic and 40 Mb of heterochromatic sequences (Figure 1, Skaletsky et al., 2003). The heterochromatic block contains no protein-coding genes, differs in length, and is mainly composed of long, homogenous tandem repeats (Skaletsky *et al.*, 2003). Conversely, the euchromatin depicts a mosaic of three distinct classes: X-transposed, X-degenerate, and ampliconic (Figure 1).

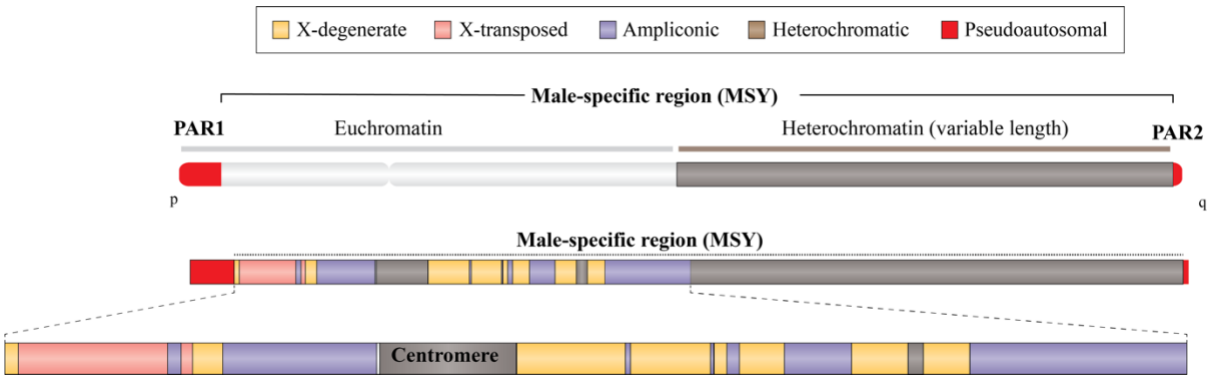


Figure 1: The structure of the human Y-chromosome. The structure of the Y-chromosome was largely adopted from Skaletsky et al., 2003. Note, that designates DNA classes such as X-degenerate, X-transposed, and ampliconic are not scaled.

The MSY genes evolved during about 300 million years after beginning of X-Y differentiation (Lahn and Page, 1999). Several MSY genes are ubiquitously expressed (Godfrey *et al.*, 2020), and function as regulators of gene expression and serve protein stability as maintaining the ancestral dosage of homologous X-Y pairs (Bellott *et al.*, 2014). Moreover, several MSY genes are evolutionary conserved and depict high nucleotide similarities with its functional X-linked homolog. For example, the ubiquitously transcribed tetratricopeptide repeat-containing, Y-linked, UTY gene is 80.2% identical to *UTX*, which resides on the X-chromosome. Other MSY genes that lack X homologs have been amplified and belong to five multicopy gene families BPY2, CDY, DAZ, PRY and XKRY (Godfrey *et al.*, 2020) which show greater than 99% nucleotide sequence similarity (Skaletsky *et al.*, 2003). Little is known about the role of specific Y-linked genes in cancer, yet anecdotal evidence suggests relevance to disease biology (Qi *et al.*, 2022). In prostate tumors for example, expression of the Y-linked histone demethylase KDM5D is associated with response to chemotherapy (Komura *et al.*, 2016; Plch, Hrabeta and Eckschlager, 2019). A different study reported that entire or partial deletions of the male-specific Y chromosome, including the histone H3 lysine 4 (H3K4) demethylase JARID1D (*KDM5D*), represses gene expression programs associated with cell invasiveness and suppresses the invasion of prostate cancer cells in vitro and in vivo (Li *et al.*, 2016). Another reports shows that *KDM5D* loss through LOY increases viability of renal cancer cell lines (Arseneault *et al.*, 2017). On the other hand, there are also reports implicating a oncogenic role of chromosome Y. Oram et al. provide evidence that *TSPY* potentiates cell proliferation and tumorigenesis by promoting cell cycle progression in various cell lines (Oram *et al.*, 2006). Mosaic loss of chromosome Y (mLOY) describes a chromosome Y aneuploidy acquired during lifetime and it is the most common post-zygotic variant described in normal hematopoietic cells (Forsberg, Gisselsson and Dumanski, 2017; Danielsson *et al.*, 2020). mLOY has been

noticed for more than half a century (Jacobs *et al.*, 1963), but functional implications were occult until recently. Recent studies postulated that the frequency of mLOY varies depending on the population and age group studied (Thompson *et al.*, 2019; Danielsson *et al.*, 2020) and further that mLOY is associated with increased risk for all-cause mortality, various forms of cancer and Alzheimer's disease, as well as other common human diseases (Forsberg *et al.*, 2014; Danielsson *et al.*, 2020; Guo *et al.*, 2020).

Mosaic loss of chromosome Y (mLOY) must not be confused with chromosome Y loss (LOY) in tumorigenesis. While mLOY describes chromosome alterations in adult males' normal/healthy blood cells, LOY describes somatic chromosomal alterations in malignant tumor tissue. LOY has previously been described in several solid tumors, including esophageal carcinoma, pancreatic, urothelial bladder, colorectal and prostate cancer (Hunter *et al.*, 1993; Wallrapp *et al.*, 2001; Minner *et al.*, 2010; Agahozo *et al.*, 2020). Furthermore, one report show great inter-tumor heterogeneity among LOY rates in renal cell carcinomas (Kovacs *et al.*, 1991). While papillary renal cell carcinomas exhibited 85% LOY, only 30% of non-papillary RCC exhibited loss of chromosome Y (Kovacs *et al.*, 1991). Importantly, while recent literature provides abundant evidence for vast differences of LOY estimates among different cancer types, it's worth mentioning that there are no two concordant data sources reporting the same estimates among the same tumor lineages. For example, while Qi and colleagues (Qi *et al.*, 2022) found LOY in 2.6% of prostate adenocarcinomas, Priestly et al. (Priestley *et al.*, 2019) reported on 18.7%. These discrepancies may arise from various source; some of which include (i) sample size of cohort studied, (ii) tissue site samples (i.e., primary, or metastatic tumor tissue), (iii) analysis method deployed and/or (iv) the consideration of mLOY. The detection of LOY can broadly be split into two areas. The usage of wet-lab techniques such as tissue microarrays (TMA) that are constructed to perform immunohistochemistry and fluorescent in situ hybridization (FISH) or the application of dry-lab methods such as next-generation sequencing (NGS) technologies. While molecular cytogenetics approaches are generally labor-intensive, they circumvent inherited obstacles associated with LOY detection from sequencing data. Particularly, the haploid nature of the Y chromosome, the homology with the X chromosome, and the repetitive sequence as a result from many gene expansions.

To date, there are only two large-scale cancer studies that investigate LOY (Priestley *et al.*, 2019; Qi *et al.*, 2022). However, both studies lack (i) a concise discrimination between mosaic loss of chromosome Y (mLOY) and LOY, which may result in erroneous estimates, (ii) a comprehensive cohort including primary and metastatic tissue samples, (iii) LOY estimates for only a small number of cancer types (i.e., not including rare malignancies) and (iv) a precise

elaboration whether chromosome Y is purposefully targeted (i.e., deleted or retained) in tumor tissue and hence provide a tumor suppressive or promoting function. A precise elaboration on these factors, in turn would further fuel the usage of chromosome Y as a prognostic biomarker that potentially may serve therapeutic opportunities.

We here present the investigation of chromosome Y loss of the largest -to our knowledge– cancer cohort. This cohort comprise >13,000 male patients and spans over 45 different tumor types with varying number of histological subtypes. Moreover, we demonstrate the eligibility of targeted sequencing panels to assess mLOY in peripheral blood and LOY in tumor tissue, respectively. Lastly, we provide evidence of the clinical significance associated with LOY and further highlight tumor suppressive functions associated with chromosome Y.

2. Material and Methods

2.1 Study cohort and prospective sequencing

The study cohort consisted of 89,735 tumor samples that were sequenced at Memorial Sloan Kettering Cancer Center until July 12, 2022. Ethical approval for the study was obtained from the Memorial Sloan Kettering Cancer Center Institutional Review Board (IRB), and written informed consent was obtained from all patients for tumor sequencing and access to their medical records, including detailed demographic, pathology, and treatment information. Tumor profiling was performed using Memorial Sloan Kettering Integrated Molecular Profiling of Actionable Cancer Targets (MSK-IMPACT) clinical sequencing assay. This assay utilized a hybridization capture-based, next-generation sequencing platform (Cheng et al., 2015). Briefly, DNA fragments were captured and sequenced as paired-end 100 - bp reads on an Illumina HiSeq 2500 instrument. The sequencing data were then aligned to the human GRCh37/hg19 reference genome using BWA-mem alignment algorithm. Further information have been published elsewhere (Cheng *et al.*, 2015; Zehir *et al.*, 2017). Tumor types were classified using the ONCOTREE classification system (http://www.cbioportal.org/oncotree/). For colorectal cancer, we further stratified the samples into microsatellite stable (MSS) and microsatellite instable (MSI) tumors. This classification was based on the presence of oncogenic POLE mutation (as defined via OncoKB, Chakravarty *et al.*, 2017) or MSIsensor score > 10, Niu *et al.*, 2014). We excluded any lymphoid and myeloid tumor that were analyzed using the MSK-ACCESS129 or HEME-400/486 assays (Ptashkin *et al.*, 2022). Additionally, we excluded female sample, samples from patients younger than 18 years, and samples with unknown primary cancer site. In cases where multiple tumor samples were available for a patient, we selected one based on criteria such as tumor purity, sequence coverage, and the most recent gene panel used. Furthermore, we excluded samples that did not meet quality metrics for overall copy-number fit, as defined in facets-preview (https://github.com/taylor-lab/facets-preview). Finally, we excluded 942 samples that showed indications of mosaic loss of the Y-chromosome (see below).

2.2 Detection of chromosome Y mosaicism (mLOY)

To detect chromosome Y aneuploidies, specifically mosaic loss of chromosome Y (mLOY) in ‘normal/healthy’ soma, we utilized peripheral blood tissue (Forsberg *et al.*, 2014, 2019; Forsberg, Gisselsson and Dumanski, 2017). This analysis focused on peripheral blood samples

obtained from 14,080 male individuals with no history of hematological malignancies at the time of sampling. These blood samples were collected concurrently with tumor sample acquisition and underwent MSK-IMPACT panel testing if the participants provided consent to protocol IRB #12-245 (Cheng *et al.*, 2015). We employed the MADSEQ R-package, which enabled us to assess the sequencing read-depth and GC content information for each targeted region (http://ykong2.github.io/MADSEQ/). The specific targeted regions varied based on the gene panel used, ranging from 6,623 tiling intervals for MSK-IMPACT 341 to 9,488 for MSK-IMPACT 505. Among these tiling intervals, 13 baits were designed specifically to enrich genomic DNA from chromosome Y (Supplementary table 2). To account for variations in sequence coverage, we performed GC content correction using a loess regression (http://ykong2.github.io/MADSEQ/). To calculate the relative DNA concentration of individual chromosomes, we compared the median read depth of targeted chromosome to that of all autosomes (chromosomes 1-22). To mitigate batch effects, we applied local regression median from a kernel density estimation, as described in previous studies (Forsberg *et al.*, 2014; Danielsson *et al.*, 2020). These batch effects represent deviations from the expected DNA ratios (1 for autosomes and 0.5 for allosomes). The corrected values were then multiplied by two to approximate the ploidy.

Subsequently, we determined the lower 2.5th percentile with the surrounding 95% confidence interval using the jmuOutlier package (https://www.rdocumentation.org/packages/jmuOutlier/versions/2.2) for each sequencing gene-panel employed. Accordingly, we derived four independent ploidy cut-offs (Supplementary figure 1B). Samples with an extrapolated ploidy smaller than respective ploidy cut- offs were considered as mLOY and further excluded from analysis. Association studies were performed using linear regression models from the base R-package (R Core Team 2021).

2.2 Determination of sequence coverage and mapping qualities

To determine the sequence coverage and mapping qualities of genetic loci on chromosome Y, we obtained gene annotations from Ensembl (GRCh38.p3) using the R/Bioconductor package biomaRt (Durinck *et al.*, 2009; accession: 07/01/2022). To ensure compatibility with the GRCh37/hg19 reference genome coordinates, we utilized the liftover tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) to transfer the annotations. DNA elements without distinct descriptions (HGNC) were excluded. This process resulted in annotations for a total of 429 transcriptional and non-transcriptional active elements. We used Rsamtools (Morgan *et al.*, 2022) to assess sequence coverage and mapping qualities at given genetic loci

(n=429). Specifically, we utilized the scanBam and countBam modules with the following flags: isUnmappedQuery = FALSE, isNotPassingQualityControls = FALSE, isSecondaryAlignment = FALSE, isDuplicate = FALSE. By applying the snp-pileup function with default parameters (https://github.com/mskcc/htstools), we obtained read-depth data, also known as count-matrices, from tumor/normal pairs. Single nucleotide polymorphism (SNP) locations were retrieved from the human single nucleotide polymorphism database (dbSNP) build 137.

2.3 Allele-specific DNA copy number and chromosome Y loss analysis

We determined total, allele-specific and integer DNA copy number genome-wide as well as tumor purity and ploidy using a modified version of Facets (v.0.5.6, Shen and Seshan, 2016). The modified version, called FacetsY (https://github.com/BastienNguyen/facetsY), was specifically designed to incorporate sequencing reads from the Y-chromosome. To ensure data quality, sequencing reads with err.thresh = 10 and del.thresh = 10 were discarded. Moreover, we only considered chromosome Y loci that exhibited sufficient coverage and reasonable mapping quality (Supplementary figure 2A, B) for segmentation. Reads mapping to either of the pseudoautosomal regions (PAR1 or PAR2) and the heterochromatic region of chromosome Y were excluded from the analysis. This resulted in a total aggregate length of 20.4 Mb (range = 2,654 – 27,800 kbp), representing approximately 34% of the Y-chromosome (Figure 1), which was used for copy-number determination.

Each matched tumor/normal pair underwent a two-pass analysis. In the first pass, we estimated tumor purity and ploidy (cval = 100), while in the second pass, we focused on detecting focal events with higher sensitivity (cval = 50). The remaining parameters were set to default. Integer copy numbers and the associated cellular fraction estimates for each segment were obtained using the expectation-maximization (EM) algorithm (Shen and Seshan, 2016). Tumors were classified as having undergone whole-genome doubling (WGD) if more than 50% of their autosomal genome exhibited a major copy number of two or greater (Bielski, Donoghue, *et al.*, 2018; Bielski, Zehir, *et al.*, 2018). The quality of FacetsY fits was assessed using predefined criteria described in facets-preview (https://github.com/taylor-lab/facets-preview).

For the determination of discrete chromosome Y copy numbers, we applied the following logic (Qi *et al.*, 2022). The expected maximum Y wild-type (WT) copy number was defined as round(ploidy/2) if round(ploidy) is even or round(ploidy/2+1) if round (ploidy) is odd. The minimum Y WT copy number (min.Y.WT) was defined as round(ploidy/2) if round(ploidy) is even or round(ploidy/2-1) if round(ploidy) is odd (Qi *et al.*, 2022). We classified a segment as

‘gained’ if the total copy number (tcn) was greater than max.Y.WT and as ‘lost’ if tcn was less than min.Y.WT. A segment of the Y-chromosome was considered ‘WT’ if min.Y.WT < tcn < max.Y.WT (Qi *et al.*, 2022). Additional configurations, such as partial gains or losses, relative losses and gain/loss are outlined in Figure 4A.

2.4 Validation cohort: Exome re-sequencing

To validate our findings, we assembled a validation cohort comprising 769 tumor samples for which both MSK-IMPACT and whole-exome recapture sequencing (WES-recapture) were available (Supplementary table 4). It is worth noting that WES offers a higher sensitivity for detecting DNA copy number alterations compared to targeted sequencing of known cancer genes (Jonsson *et al.*, 2019). All tumor samples included in the analysis were recaptures of existing sequencing libraries. Details regarding the analysis procedures for WES-recapture samples can be found elsewhere (Jonsson *et al.*, 2019).

The validation cohort encompassed various cancer types, including prostate adenocarcinoma (n = 96; 12.5%), bladder urothelial carcinoma (n = 77; 10%), renal clear cell carcinoma (n = 60; 7.8%) and more than ten other cancer types (Supplementary table 4).

Similar to the main cohort, we applied FacetsY to the validation cohort with the same methodology described above, except for adjusting the cval for the purity-run to 500 and 200 for the high-sensitivity run. All other parameters were kept at their default values. To assess copy-number alterations, we extracted individual copy-number log ratios (CnLR) and calculated the median when more than ten independent loci were available for CnLR determination.

To evaluate the concordance between MSK-IMPACT and WES-recapture samples in terms of CnLR and tumor purity, we performed correlation analysis using Pearson’s method implemented in the stats package (version 3.6.2) of the R programming langue (R Core Team (2017).

2.5 LOY detection in TCGA and RNA-Seq analysis

To investigate the presence of LOY in an independent cohort of kidney renal clear cell carcinoma (n = 446, Creighton *et al.*, 2013), we conducted DNA copy number alteration analysis using the same methodology as for the WES-recapture samples, as outlined above. Following the exclusion of female samples and the application of the quality metrics described previously (https://github.com/taylor-lab/facets-preview), we identified 290 male tumor samples suitable for LOY detection and subsequent RNA-seq analysis.

For the RNA-seq analysis, we obtained expression levels (log_2) of *VHL*, *KDM5C*, *KDM5D*, *DDX3Y* and *UTY* genes using the cBioPortal for Cancer Genomics (http://www.cbioportal.org, Cerami *et al.*, 2012). Respective expression levels per gene were then stratified based on the chromosome Y status of the samples. To assess differences in gene expression levels between wild-type and LOY samples, we performed the Mann-Whitney U test, with a significance level of $P < 0.05$ considered as statistically significant.

2.6 Ancestry

Recently, global ancestral contributions and admixture of continental populations were quantitatively inferred using genetic markers captured by MSK-IMPACT (Arora *et al.*, 2022). We employed this dataset (n=11,294) and established correlations with LOY estimates.

2.7 GISTIC analysis for recurrent focal alteration detection

To identify recurrent focal alterations on the Y chromosome, we employed GISTIC2.0 (Mermel *et al.*, 2011) on a subset of our cohort consisting of n = 329 tumor samples that exhibited imbalances on either chromosome Y arm. This analysis was conducted using the GenePattern platform (https://www.genepattern.org) with parameter --rx 0 to include sex chromosomes, while leaving the remaining parameters at their default values. From the GISTIC2.0 output files, we extracted the gain and loss peaks, along with their corresponding genomic boundaries, G-scores, and adjusted p-values. These extracted results were then utilized for visualization and downstream analysis. To ensure statistical significance, we focused on peaks with adjusted p-values < 0.05.

2.8 Fraction genome altered, microsatellite instability and tumor mutational burden

To assess the extent of genomic alterations, we calculated the fraction of genome altered (FGA) for each sample. FGA represents the percentage of the genome exhibiting absolute log2 copy ratios great than 0.2. The log2 copy-number ratios were derived following the methodology described in Cheng et al. (Cheng *et al.*, 2015). Microsatellite instability (MSI) status was determined for all tumor samples using MSIsensor, a validated tool for assessing microsatellite instability (Niu *et al.*, 2014). MSIsensor scores indicate the proportion of unstable microsatellites among the tested microsatellites (Niu *et al.*, 2014). Tumors with an MSIsensor score greater than or equal to ten were classified as MSI tumors (Supplementary figure 9B). Due to the limited number of MSI tumors in some tumor types, our analysis focused on bladder (n = 814; 15 MSI), esophagogastric (n = 683; 44 MSI), prostate (n = 1614; 36 MSI), small bowel (n = 54; 11 MSI), and hypermutated colorectal cancers (n = 172; 158 MSI). The ‘n’ value

represents the total samples size of respective tumor type, and the number attached indicates the number of MSI tumors.

Tumor mutational burden (TMB) was calculated for each sample as the total number of nonsynonymous mutations divided by the number of bases sequenced (Vokes *et al.*, 2019).

2.9 Average sequencing depth of the Y-chromosome

We inferred the average sequencing depth of the Y-chromosome using the files generated by the MADSEQ package (Cheng et al., 2015; Zehir et al., 2017) as described in 'Detection of chromosome Y mosaicism'. Briefly, the MADSEQ package utilizes a set of user-defined input probe intervals, specifically the gene intervals included in the respective gene-panels. The average sequencing depth across these regions was calculated. Additionally, MADSEQ performs a GC-bias correction using loess smoothing to account for any potential biases introduced by GC content. This correction helps ensure accurate estimation of the sequencing depth. To assess the impact of tumor sample quality on the average sequencing depth, we categorized samples based on their -FacetsY- inferred purity. Tumor samples with inferred purities lower than 0.3 were classified as low-quality samples, while those with purities greater than 0.8 were considered high purity samples. To evaluate the differences in average sequencing depth between the two groups, the Mann-Whitney U test was employed.

2.10 Somatic mutations

Somatic mutations (substitutions and small insertions and deletions), gene-level focal copy number alterations, and structural rearrangements were detected with a clinically validated pipeline as previously described (Cheng *et al.*, 2015; Zehir *et al.*, 2017). Somatic alterations were classified as oncogenic or likely oncogenic using OncoKB (Chakravarty *et al.*, 2017).

2.11 Multivariable regression models

To investigate the genetic associations with LOY, we employed a multivariable logistic regression model to determine the probability of LOY occurrence. Our model incorporated various factors, including somatic mutations, SCNAs (somatic copy number alterations), and fusions detected in at least 3% of the entire cohort withing the 505 genes sequenced. These factors were encoded as binary predictor variables. In addition to the aforementioned genetic factors, we incorporated other relevant variables into our model, including cancer subtypes, TP53 mutational status (yes/no), FGA, purity, ploidy, and MSI-type (MSS or MSI). To assess multicollinearity resulting from correlated predictor variables, we employed variance inflation factors (VIF) (Bielski, Zehir, *et al.*, 2018). Variables with a VIF greater than 4 were excluded

from the model to ensure reliable estimates. Furthermore, we performed adjusted p-value calculations using the Benjamini-Hochberg method to account for multiple hypothesis testing. Associations with a false discovery rate (FDR) less than 0.01 were deemed statistically significant.

2.12 Survival analysis

We conducted both univariate and multivariate survival analyses to assess the impact of various factors on overall survival. We employed Cox proportional hazard regression models for these analyses, which allowed us to examine the association between predictor variables and survival outcomes while accounting for other covariates (Bielski, Zehir, *et al.*, 2018; Jonsson *et al.*, 2019). For the visualization of survival outcomes, Kaplan-Meier survival curves were generated using the survival package in R (Therneau *et al.*, 2023).

In our analysis, overall survival was measured in months and calculated as the difference between the data of the initial procedure, when prospective sequencing was performed, and the data of the last follow-up. We utilized the Wald statistic to obtain p-values for the univariate survival analyses. For the multivariate analyses, the likelihood ratio test (LRT) statistic was employed to assess the significance of the predictor variables while considering the effects of other covariates.

By employing these survival analysis techniques, we aimed to investigate the impact of clinicopathological and genomic factors on overall survival and identify potential prognostic factors in our study population.

2.13 Statistical testing and data visualization

All statistical analyses were performed using R statistical software (v3.5.2) with the following packages: survival (v2.43-3), ggplot2 (v3.1.0), survminer (v0.4.3), forestplot (v.1.7.2), factorial2x2 (v0.2.0), cowplot (v1.1.0) and tidyverse (v1.2.1). All tests of statistical significance were two-sided. P-values were corrected for multiple comparisons using the Benjamini–Hochberg False Discovery Rate (FDR) method or the Bonferroni method, as noted. All source codes for the present analyses are available at https://github.com/chris-kreitzer/Y_chromosome_loss/tree/main.

3. Results

3.1 Overview of the study cohort

In this study, we included a total of 13,138 male tumor specimen. These specimens underwent prospective matched tumor and normal genomic profiling using one of the following gene panels: MSK-IMPACT341, 410, 468 or 505 (Cheng *et al.*, 2015) (Figure 2A). We excluded hematologic malignancies, including myeloid and lymphoid tumors (Ptashkin *et al.*, 2022) as well as pediatric patients below the age of 18 at the time of sequencing, cancers of unknown primary origin, and female tumor samples (n = 60,161). We further discarded tumor samples without available matched normal, and those with insufficient quality (Methods).

The final cohort consisted of 8,514 (64.8%) primary tumor samples and 4,624 (35.2%) metastatic tumor samples, encompassing 47 different tumor types with varying number of histological subtypes (Figure 2B and Supplementary table 1, http://www.cbioportal.org/oncotree/). Among the primary tumors, 1,186 (13.6%) were derived from microsatellite stable (MSS) colorectal patients, followed by 1,004 (12.6%) prostate cancer patients and 949 (10.9%) non-small cell lung cancer patients. The majority of sequenced samples obtained from metastatic sites originated from the liver (n = 1909; 40.8%), lymph nodes (n = 1706; 36.4%) and lung (n = 988; 21.1%). The median age at sequencing was 65 years, ranging from a median of 26 years for desmoplastic small-round-cell tumor (DSRCT) to a median of 74 years for male patients with urethral urothelial carcinoma (UCU).

Lastly, we excluded 942 samples with indication of chromosome Y aneuploidies in blood cells, known as mosaic loss of the Y chromosome (mLOY) (Forsberg *et al.*, 2014; Danielsson *et al.*, 2020). mLOY is the most common somatic genetic aberration in blood cells (Thompson *et al.*, 2019; Guo *et al.*, 2020), and it is crucial to exclude these samples from further analysis. Otherwise, tumor Y copies can be 'gained' relative to the control; or if the fraction of somatic LOY in the control (i.e., blood) and tumor are approximately equal, no difference will be detected (Qi *et al.*, 2022). We assessed mLOY in peripheral blood tissue (Methods) and found a bi-modal distribution of extrapolated chromosome Y ploidies across all samples studied. (Supplementary figure 1A). We then stratified chromosome Y ploidies according to respective gene panel deployed and identified four independent ploidy cut-offs indicating some level of mLOY (Supplementary figure 1B). For instance, blood samples with chromosome Y ploidy < 0.78 that were sequenced using the IM6 gene panel (targeting 468 genes, Cheng *et al.*, 2015) were considered to have lost the chromosome Y in at least a major fraction of cells, and

therefore excluded from further analysis (Methods). Furthermore, we observed a decline in chromosome Y ploidies with increasing age, a phenomenon repeatedly shown in the past (Danielsson *et al.*, 2020). Patients with mLOY were significantly older than those showing an intact Y-chromosome in peripheral blood tissue (72 years vs. 64 years, $P < 2e^{-16}$, Mann- Whitney U test, Figure 2C).

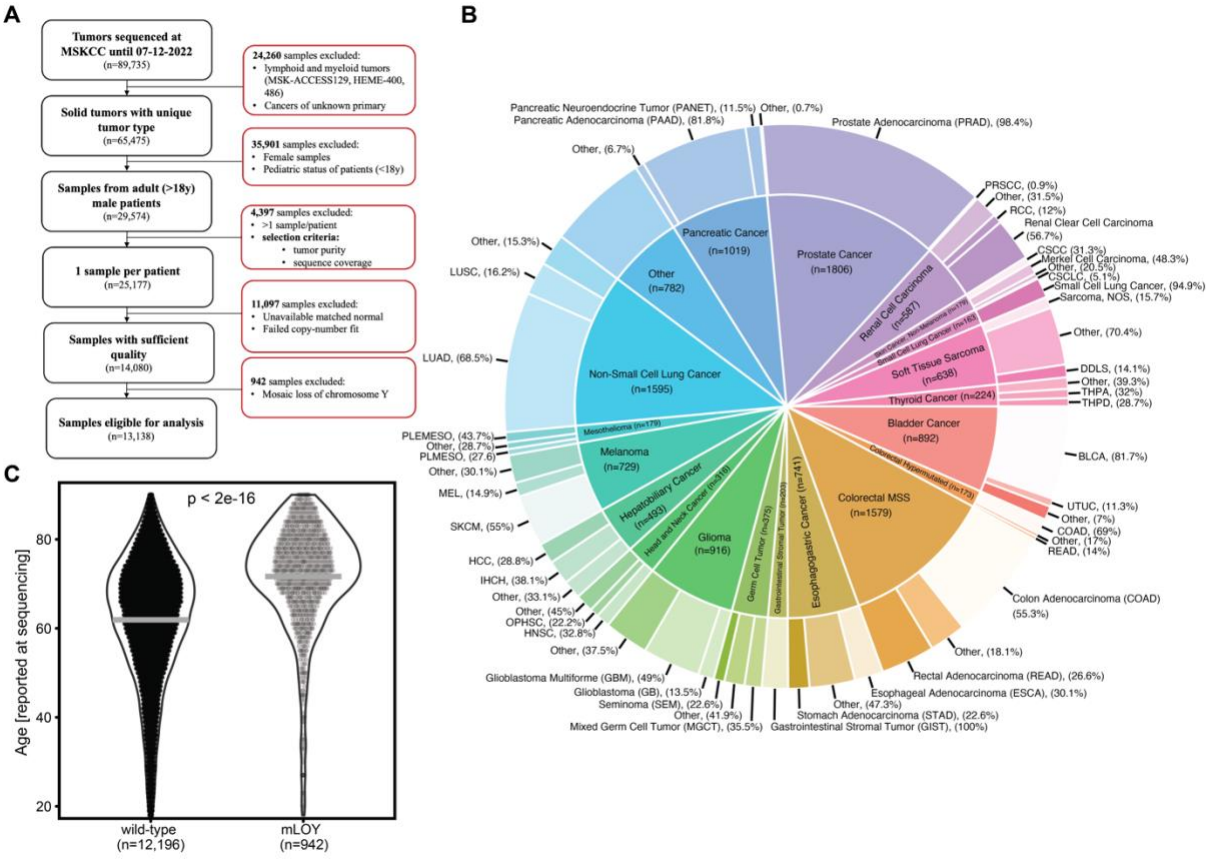


Figure 2: Study cohort overview and assessment of mosaic loss of the Y chromosome (mLOY). A) Consort flow diagram illustrating the selection criteria for the analysis cohort. B) Donut plot showcasing the cancer types and their respective histological subtypes included in the present study. Supplementary table 1 provides abbreviations for the histological subtypes (i.e., oncotree code) along with their corresponding sample sizes. C) Association between age and mosaic loss of the Y-chromosome. Patients' age (y-axis in years) is plotted against binary mLOY assignments (Methods). The median age in years is indicated by the grey horizontal line. Statistical comparison was performed using the Mann-Whitney U test. A P-value < 0.05 was considered statistically significant.

3.2 MSK-IMPACT is suitable for chromosome Y loss determination

Accurate sequence read-depth distribution in matched tumor/normal samples is of uttermost importance for copy number profiling (Krumm *et al.*, 2012; Shen and Seshan, 2016). Consequently, we investigated individual sequencing files to assess potential biased coverage data, considering the repetitive nature of the Y chromosome (Skaletsky *et al.*, 2003, Figure 1). Despite the inclusion of only 13 baits (Supplementary table *2*) in the MSK-IMPACT sequencing assays to enrich DNA from chromosome Y, we observed uneven sequence coverage

at 63 different loci (14.7%, Methods). Among these, only 17 genetic elements (including 11 protein-coding genes; Supplementary table 3) were covered in more than 75% of the analyzed sequencing files (Supplementary figure 2A), indicating sporadic coverage by off-target reads (Mangul *et al.*, 2021). Additionally, while consistent coverage was observed for 15 genetic elements, the pseudogene *CCNQP2* and poly(ADP-ribose) polymerase family member 4 pseudogene 1, PARP4P1 were not captured using either MSK-IMPACT gene panel IM3 or IM5 (Supplementary figure 2B). Next, we assessed the average mapping qualities (BWA-MAPQ) of aligned sequencing reads. Three genes, namely *FAM197Y1*, *CHEK2P1* and *PCDH11Y* showed an average BWA-MAPQ < 22 (Supplementary figure 2C), indicating ambiguous alignments and were therefore flagged for exclusion when determining tumor copy-number profiles. For example, the low mapping qualities of *PCDH11Y* may be attributed to its close nucleotide similarity (98%) with its X-homologue, *PCDH11X* (Blanco *et al.*, 2000). Similarly, *CHEK2P1* is part of the centromere, which is known to contain numerous tandem repeats (Giunta and Funabiki, 2017).

Before applying FacetsY to determine copy-number alterations and chromosome Y loss on the pan-cancer cohort (Figure 2A, B and Methods), we assembled a validation cohort of matched whole-exome re-sequenced (WES- recapture) samples (n = 769, Supplementary table 4) to gain insight into the validity of our approach. Notably, WES-recapture samples exhibited significantly greater per-base coverage, number of loci and feature coverage at the Y- chromosome (Supplementary figure 2D). We used the median copy-number log ratio (CnLR; Methods) to compare LOY calls between the two cohorts (Figure 3A) and found a strong correlation (Pearson's r: 0.94, $P < 2e^{-16}$, Figure 3B). This suggests that even with a reduced representation of chromosome Y (i.e., MSK-IMPACT sequenced samples), accurate tumor copy number calls are achievable. Furthermore, purity estimates of matched MSK- IMPACT and WES-recapture samples exhibited a strong correlation (Pearson's r: 0.95, $P < 2e^{-16}$, Figure 3C), further supporting our assumption of correct copy-number fits in both cohorts. Overall, reasonable correlations were established, underpinning the utilization of MSK-IMPACT samples to determine chromosome Y loss. Although a few samples deviated from the expected results (i.e., deviation of | median CnLR | > 1 between the two cohorts, Figure 3B), these deviations can largely be explained by the method employed (discussed in Supplementary figure 3).

Lastly, we determined the accuracy of FacetsY by analyzing the kidney renal clear cell carcinoma (KIRC, n = 210) cohort of the TCGA project (Creighton *et al.*, 2013). Specifically, we examined whether LOY calls correlated with mRNA expression levels of chromosome Y

genes. We identified n = 92 male primary KIRC tumors (43.8%) that exhibited LOY. As expected, genes *KDM5D*, *DDX3Y*, and *UTY* showed significantly decreased mRNA expression in LOY samples compared to wild-type samples (Figure 3D, $P < 0.05$, Mann- Whitney U test). Furthermore, we conducted the same analysis for two independent autosomal genes, *KDM5C* residing on the X-chromosome and *VHL* (von Hippel-Lindau tumor suppressor), and found no significant difference in expression levels based on the chromosome Y status (Figure 3C, $P > 0.05$, Mann-Whitney U test).

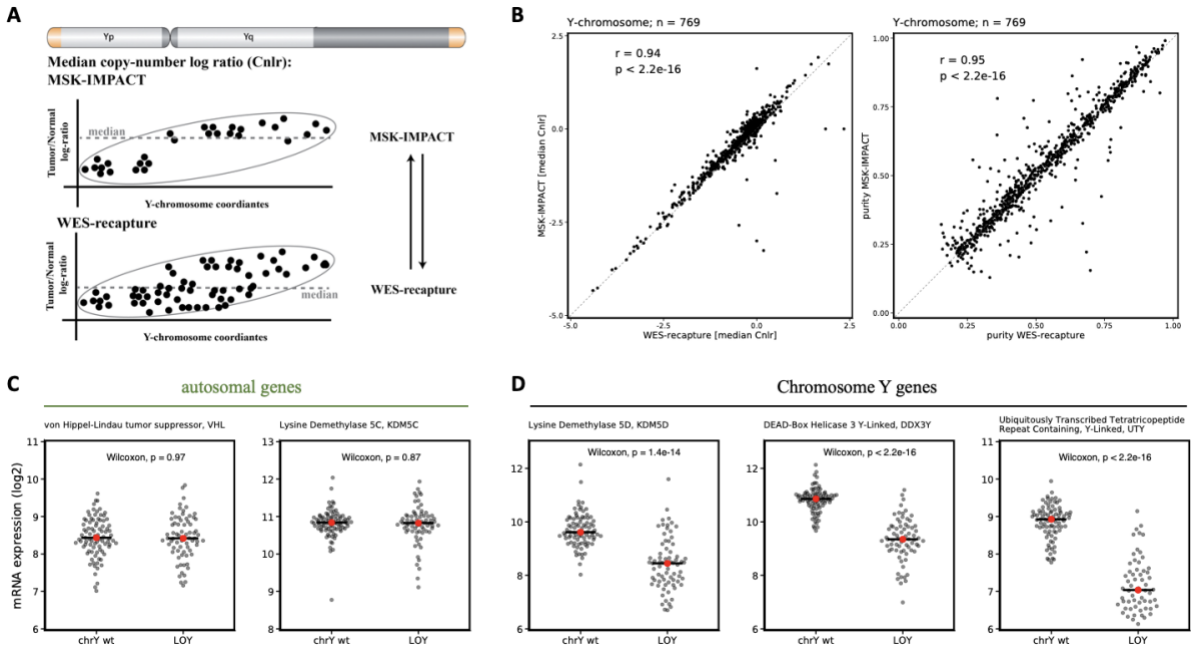


Figure 3: MSK-IMPACT depicts a reliable data source to study LOY in tumor samples. A) The cartoon depicts the approach used to compare LOY calls between MSK-IMPACT and WES-recapture (validation cohort). (Top plot): It shows the PAR-regions (orange), the centromere, the heteromeric block (dark grey) and regions with sufficient mapping quality (light grey) used for copy-number determination (range: 2,654-27,800 kbp, hg19/GRCh37). (Middle/bottom plot): The plot represents individual tumor/normal ratios (black dots). Note that these are randomly placed and not scaled to the outline of the Y-chromosome. B) The left plot shows the median copy-number log ratios (CnLR) across the matched MSK-IMPACT (y-axis) and WES-recapture samples (x-axis). The right plot highlights the purity estimates. Each black dot represents a tumor sample. Correlation coefficients were determined using Pearson's method. C) mRNA expression levels (log_2(value+1)) of TCGA, KIRC (n=290) samples are plotted for selected autosomal genes while D) shows mRNA expression levels for chromosome-Y genes. Samples are stratified by the copy-number state of the Y-chromosome. The red dot (+horizontal black line) represents the median mRNA expression value. LOY = loss of Y; wt = wild-type.

3.3 Loss of chromosome Y varies across tumor types

A total of 13,138 male tumor samples were eligible for studying chromosome Y loss in tumor tissues (Figure 2A). We observed a wide range of Y copy number events (Figure 4A, Methods), with the most frequent event being complete loss of the Y-chromosome (32.6%, Figure 4B). Relative loss of the Y chromosome (rLOY), which refers to the loss of a chromosome copy while maintaining an overall ploidy >2, occurred in 1.3% of male tumors, while partial losses

were observed in only 131 samples (1%; Figure 4A). Overall, 4738 male tumors (34.9%) exhibited either complete, relative, or partial loss(es) of the Y-chromosome. In contrast, chromosomal gains accounted for 21% of samples investigated, with single whole-chromosome duplications being the predominant type (19.9%). The rate of LOY varied significantly among different cancer types, with esophagogastric tumors being the most affected (64.5%), while glioma samples showed a much lower rate of 10.5% (Figure 4B). We also observed variations in LOY rates among molecularly distinct subtypes of diseases. For example, while 41.9% of all microsatellite stable (MSS) colorectal cancer exhibited LOY, the fraction was only 11.9% in hypermutated (i.e., microsatellite instable or oncogenic POLE mutation) colorectal cancers. We further analyzed LOY rates in individual cancer types and sampling sites. Overall, there was no significant difference in LOY rates between primary and metastatic tumors (Figure 4B, P = 0.83, Mann-Whitney U test). However, several cancer types showed distinct variations. For instance, ampullary tumors had a LOY rate of 50% in primary samples, which increased to 73% in metastatic sites. In contrast, small bowel tumors showed a LOY rate of 46% in primary samples, but only 14% in metastatic (Figure 4B) samples. We also examined whether LOY rates were influenced by age. In contrast to age-related mLOY (Figure 2C, Danielsson *et al.*, 2020), which showed a steady decline with age, we observed a relative plateau in LOY rates in tumors samples. Approximately 20% of patients younger than 30 years and 35% of patients older than 50 years exhibited LOY (Supplementary figure 4A), suggesting that chromosome Y loss in tumor samples is age independent. Recently, Arora, K., et al. demonstrated that the global ancestral contribution and admixture of continental populations can be quantitatively inferred using markers captured by the MSK-IMACT clinical panel (Arora *et al.*, 2022). Leveraging this data, we employed it to model ancestry calls along with LOY rates. Interestingly, we found that LOY rates were highest in East Asian populations (34.3%) and Ashkenazi Jewish-European populations (34.1%), whereas Native American populations (NAM) exhibited a lower rate of complete loss of the Y chromosome at 25% (Supplementary figure 4B). Furthermore, a recent study investigated LOY rates in primary male tumors as part of the TCGA project (Qi *et al.*, 2022). We established good agreement between the LOY rates reported in TCGA samples and our present estimates (Figure 4C, $P = 1.56e^{-08}$, r = 0.95, Pearson's product moment correlation) further supporting the validity of our results.

Lastly, we found that, besides tumor types, LOY rates markedly differed among histologically distinct subtypes of disease (Supplementary figure 5). For instance, papillary renal cell carcinomas (PRCC) exhibited a high LOY rate of 86.6%, whereas renal clear cell carcinoma

(CCRCC) experienced a lower rate of 45.9% LOY (Figure 4D, P = $2.1e^{-08}$, Proportion test). Additionally, papillary thyroid tumors showed minimal evidence of LOY (5.6%) compared to Hurthle cell thyroid cancers, which exhibited a higher LOY rate of 54.5% (Figure 4E, P = $6.1e^{-07}$, Proportion test). This observation aligns with the notion that papillary thyroid tumors are primarily driven by oncogenes and have a relatively stable genome (Agrawal *et al.*, 2014), whereas Hurthle cell thyroid cancers display extensive chromosomal aberrations (Supplementary figure 6).

In summary, like other patterns of chromosome gains and losses, the rates of different Y chromosome events vary across cancer types and further exhibit notable differences within histological subtypes. Due to its frequency and the potential impact of losing an entire chromosome without a backup copy, we focused our further analyses on complete LOY.

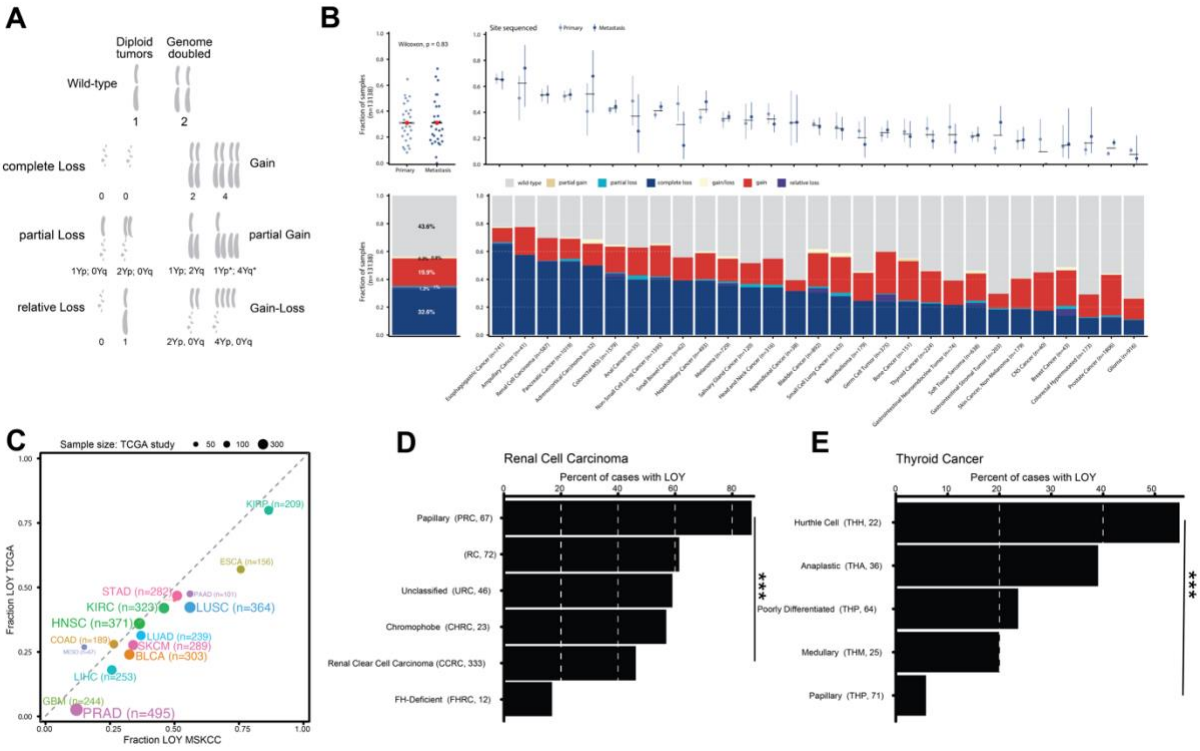


Figure 4: LOY rates across tumor types. A) The figure illustrates various copy-number configurations for the Y-chromosome, considering whether a sample underwent whole-genome duplication (WGD) or not. The depicted cartoon provides examples of different scenarios. For instance, partial loss occurring on the Yq arm indicates the absence of functional DNA sequence before or after the WGD event (0Yq). Relative losses are observed only in WGD+ tumors. The cartoon also demonstrates partial gains on the Yq arm (4Yq). The copy-numberr configurations were determined based on the most parsimonious explanation (Van Loo et al., 2010; Bielski, Donoghue, et al., 2018). For example: in cases where a 2Yp;4Yq configurations is observed in WGD+ samples, it is assumed that a partial gain occurred before the WGD event. B) The bottom barchart shows the fraction of LOY across the whole cohort (stacked bar chart on the left) and across cancer types (x-axis, stacked bar chart on the right). Cancer types were sorted based fraction of complete LOY. Top, shows the distribution of LOY rates across the sampling site (left, whole cohort) and on the right for individual cancer types. Red dots and horizontal lines indicate the mean across both groups, respectively. C) The jitter plot showcases the LOY rates across matched tumor types found in both the TCGA-cohort (y-axis, (Qi et al., 2022) and MSKCC-cohort (x-axis). Only cancer types present in both studies are shown. The P-value, obtained through Pearson's product-moment correlation, indicates the level of correlation between the LOY rates from the two cohorts. TCGA tumor type abbreviations can be found in Supplementary table 5. D) The plot displays LOY rates (x-axis) across different histologically distinct subtypes (y-axis) of renal cell carcinomas. E) Similarly, the plot shows LOY rates (x-axis) across different histologically distinct subtypes (y-axis) of thyroid tumors. The ONCOTREE codes and sample sizes are indicated in parathesis. The significance level is denoted by "***" to indicate $P < 0.001$.

3.4 Focal alteration signals are rarely seen on the Y chromosome

A total stretch of 25.3 Mb, covering approximately 63% of the Yp (range: 2,7-10,5 Mb) and 30% of the Yq arm (range: 14-28 Mb), was used for copy-number segmentation. Among the tumor samples, the majority (97.5%) exhibited whole-chromosome events, meaning that a single discrete copy-number solution was identified (e.g., loss, gain or wild-type). However, in the remaining 329 samples (2.5%), several chromosomal-arm imbalances were detected (Figure 5A). The most common imbalances were partial losses (39.2%), followed by gain/losses (36.8%) and partial gains (10.0%, Figure 5B). We further investigated the pattern

of chromosomal-arm instabilities. For 262 samples, we found one chromosome arm gained while the opposite arm was lost (i.e., gain/loss). Notably, the Yp arm was more prone to gains compared to the Yq arm (Figure 5C). Additionally, a majority of losses occurred on the Yq arm, demonstrating a significant association ($P < 0.001$, Proportion test). Indeed, several cases with focal deletions spanning a region of 1.3 Mb (21.6-22.9 Mb) were observed, involving protein coding genes *KDM5D*, *EIF1AY* and *RPS4Y2* (Supplementary figure 7). Partial gains predominantly occurred on the Yp arm (66.6%), whereas partial losses affected 81.5% of the Yq arm. These findings indicate a non-random association (Figure 5C, P value = $2.2e^{-16}$, Proportion test) between gains affecting the Yp arm and losses predominantly occurring on the Yq arm. Furthermore, several tumor types exhibited a elevated number of tumors with imbalanced Y chromosomes (Supplementary figure 8).

To identify putative recurrent somatic copy-number alterations (SCNAs) on the Y chromosome and assess their statistical significance, we utilized GISITC2.0 (Mermel *et al.*, 2011), Methods), on our segmentation data. No significant peaks of deletions or amplifications on the Y chromosome were detected ($Q < 0.05$), suggesting that focal and recurrent alterations are rare (Figure 5D). However, we did observe several strong amplification peaks on chromosome 1, 11 and 12 encompassing known oncogenes (Bailey *et al.*, 2018). Notably, cyclin D1, (*CCND1*) showed recurrent amplifications, consistent with its role as an oncogene. Indeed, Casimiro and colleagues reported that high expression of *CCND1* correlate with increased chromosomal instability (CIN) (Casimiro *et al.*, 2012). Similarly, *CDKN2A*, a tumor suppressor gene, was recurrently deleted, further supporting the notion that chromosomal-arm imbalances observed on the Y chromosome may be a consequence of general CIN rather than negative selection.

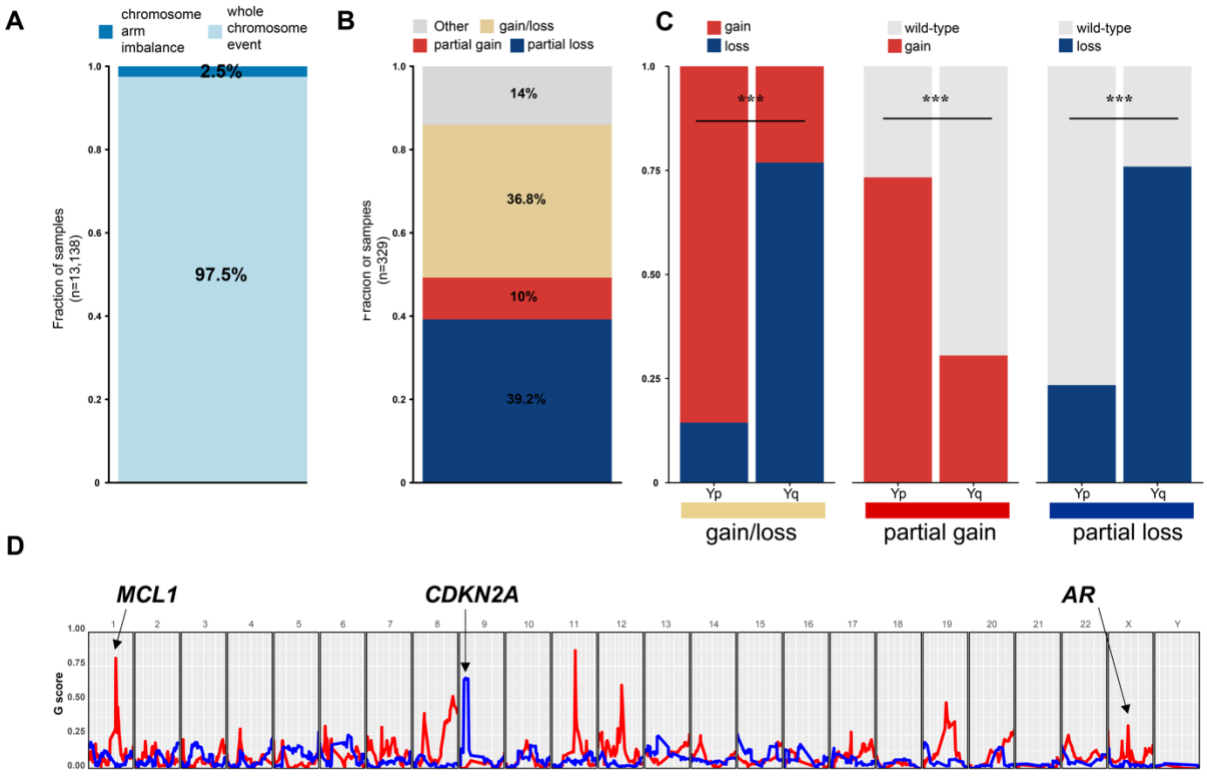


Figure 5: Chromosomal-arm imbalances observed on the Y-chromosome. A) The fraction of male tumors with either whole-chromosome events (light blue) or chromosome-arm imbalances (dark blue) called by FacetsY is shown. B) The bar chart illustrates the relative contribution of individual chromosomal-arm imbalances. The category "Others" (grey) represents samples where more than one segment with the same classification was called. For example, a sample may have experienced two independent gains, but the overall effect is the same. C) The first plot in this section displays the relative fraction of gains and losses mapped to either arm for samples with gain/loss imbalances. The second plot shows the relative fraction of gained segments, and the third plot shows the relative fraction of lost segments mapping to either chromosome-arm. Significance levels are indicated with "***" ($P < 0.001$). D) The line chart presents GISTIC (Mermel et al., 2011) G-scores for each chromosome (top) of samples with chromosome Y imbalances. Deletion signals are depicted by blue curves while amplification signals are represented by red curves. The chart highlights two genes as candidate oncogenes (*MCL1*, *AR*) and one tumor suppressor gene (*CDKN2A*).

3.5 LOY is common in aneuploid tumors

Given the rate and variability of LOY across cancer types, we sought to determine whether an association between various genomic and clinicopathological features and LOY existed. We ran a multivariable logistic regression model and found that LOY was significantly associated with the fraction of genome altered (FGA) (Figure 6A, $P = 4.8e^{-153}$); a result expected, given that FGA expresses the percentage of genome that has been affected by copy number gains or losses (Chakraborty *et al.*, 2020). We further explored whether the loss of chromosome Y, as a small and gene-poor chromosome, was a result of general genomic instability and subsequent aneuploidy. We observed a moderate correlation (rho = 0.6, P = 0.0007, Spearman's rank correlation) between the number of autosomal chromosome arms lost and the fraction of LOY (Figure 6B). This suggests that LOY follows the rates of general aneuploidy. Furthermore, this

pattern persisted even when considering correlations with other measures of aneuploidy (Supplementary figure 9A). However, while renal cell carcinomas exhibited elevated LOY rates compared to their overall aneuploidies scores, hypermutated colorectal and germ cell tumors showed disproportionately high autosomal aberrations relative to their LOY rates, indicating the need for further investigations. The multivariable model also predicted that age at the time of sequencing increased the odds of a tumor losing the Y chromosome by a factor of 0.004 ($P = 1.1e10^{-3}$, Wald test). Although the effect size was minor, it was significant, potentially stemming from the steady increase in LOY rates in patients aged 30-50 years (Supplementary figure 4A). Ploidy was another variable significantly associated with LOY (Figure 6A), in line with previous studies demonstrating a correlation between ploidy, aneuploidy scores, and whole-genome doublings (WGD, Taylor *et al.*, 2018). Additionally, consistent with unstable genomes, tumors that had undergone genome doubling were more likely to exhibit various copy number changes on chromosome Y (Figure 6C). Specifically, complete, relative, and partial losses were significantly elevated in WGD+ tumors.

Furthermore, we utilized MSIsensor scores (Niu *et al.*, 2014) that enables a discrimination between MSI and MSS tumors (Supplementary figure 9B) and found that the fraction of LOY was lower in MSI tumors (Figure 6D), although the sample size was relatively small, limiting the exploratory power of the analysis (Methods). However, MSI tumors showed decreased odds of experiencing LOY which hold statistical significance ($P = 0.017$, Wald test).

Interestingly, tumor purity (i.e., the fraction of normal cell contamination) emerged as another significant contributing variable (Figure 6A, $P = 4.8e10^{-47}$, Wald test). Tumors with low purity were more likely to exhibit LOY, as indicated by a moderate negative correlation (Figure 6E, rho = - 0.49, Spearman's rank correlation). We tested the hypothesis that this phenomenon could be explained by sequencing coverage and whether biased tumor-to-normal sequencing ratios could account for the higher incidence of LOY in low purity samples. However, when comparing the average sequencing depth across low purity (< 0.3) and high purity (> 0.8) samples we didn't determine any significant difference (Figure 6F, $P = 0.56$, Mann-Whitney U test). This observation poses a paradox that merits further investigations. Taken together, the results presented indicate that LOY may be a byproduct of the cells' aberrant genomes, potentially caused by CIN.

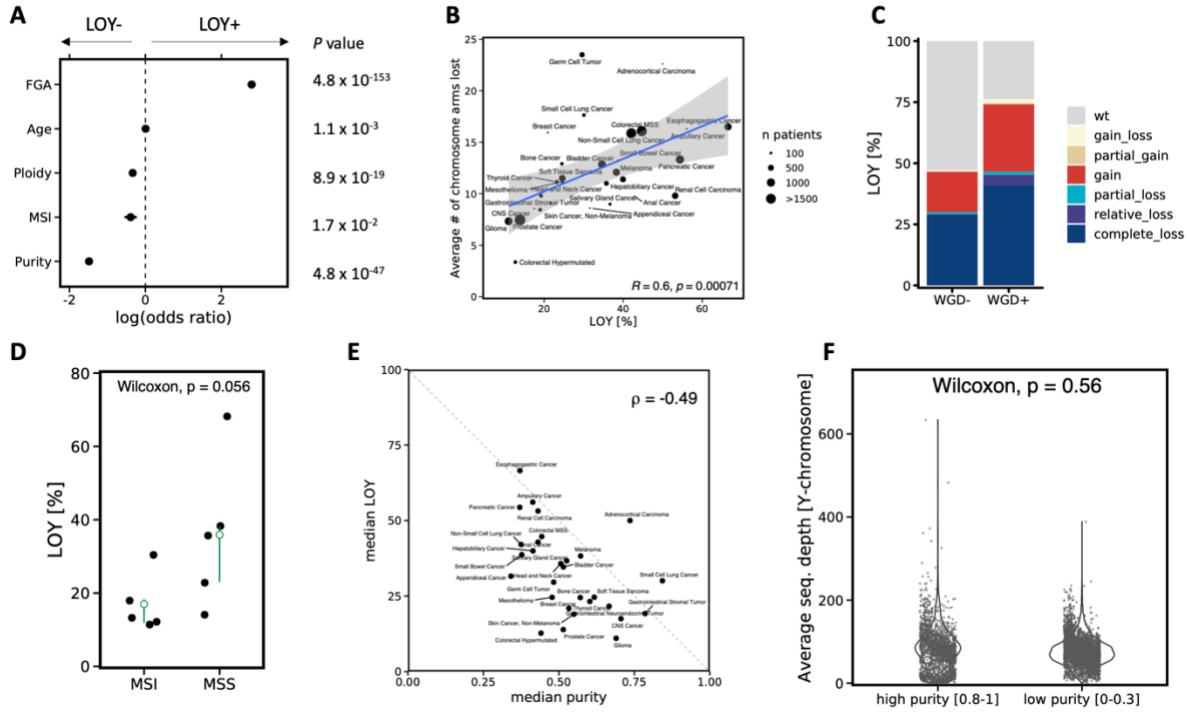


Figure 6: Genome correlates with LOY. A) Multivariable logistic regression model: The figure shows the statistically significant associations between different variables and LOY. The y-axis represents the nominal P-values (< 0.001) associated with each variable. The error bars indicate the log odds ratio plus or minus two times the standard error. B) This panel illustrates the correlation between the rate of LOY (x-axis) and the average number of autosomal chromosome arms lost (y-axis). The blue line represents the trendline generated by the smoothing method 'lm', and the shaded area represents the 95% predicted confidence interval. C) The graph shows the fraction of chromosome Y alterations in tumors that underwent either no whole-genome doubling (WGD-) or whole-genome doubling (WGD+). D) This plot focuses on five specific cancer types (bladder, esophagogastric, prostate, small-bowel, and colorectal cancer) with sufficient sample size in both microsatellite instable (MSI) and microsatellite stable (MSS) groups. It compares the fraction of LOY (y-axis) between MSI and MSS tumors, with green circles representing the median and horizontal lines denoting the lower and upper quartiles. E) This correlation plot displays the average LOY rate and median purity for individual cancer types analyzed. The correlation coefficient, calculated using the Pearson product-moment correlation, is indicated. F) The violin plot compares the average sequencing depth (y-axis) between high and low purity tumor samples (x-axis). Statistical significance was evaluated using the Mann-Whitney U test.

3.6 Association of LOY with point mutations

Next, we sought to identify associations between LOY and somatic mutations in cancer driver genes. Tumor mutational burden (TMB), which refers to the total number of mutations found in cancer cells (Vokes *et al.*, 2019), did not differ significantly between wild-type and LOY tumors (Supplementary figure 10A). However, the most significantly associated driver in the entire cohort was *TP53*, with point mutations or deep deletions present in 60.6% of the tumour experiencing LOY (Figure 7A). The majority of truncating mutations were represented by p53 C-terminal R213* mutants (87%), while missense mutations were predominantly found at residue R273 (67%), a well-known mutational hotspot site (Shirole *et al.*, 2016). Interestingly, tumor samples experiencing LOY while expressing wild-type TP53 proteins (n = 1,859 samples), frequently exhibited deep deletions in *CDKN2A* (22%), missense mutations of *KRAS*

(16%) and damaging mutations in *ARID1A* (1.8%, Supplementary figure 11). It is worth noting that the frequencies of these alterations varied significantly across cancer types, reflecting their idiosyncratic biology (Zehir *et al.*, 2017). For example, while *CDKN2A* deletions were observed in all cancer types investigated, *KRAS* missense mutations were rarely seen in renal cell carcinomas (1.4%; 4 oncogenic mutations out of 279 samples). Similarly, *APC* mutations were predominantly found in colorectal cancers (>80% damaging mutations), while being absent or rare in most other caner types (Supplementary figure 10B).

To uncover recurrent genetic lesions associated with LOY, we employed an approach recently suggested by Canisius et al. (Canisius, Martens and Wessels, 2016) and Milosevic et al. (Milosevic *et al.*, 2012). As expected, *TP53* mutations showed the strongest association with LOY at the pan-cancer level (Supplementary figure 10A, $P = 6.18e^{-141}$, Fisher's exact test). Additionally, *CDKN2A* ($P = 2.38e^{-26}$, Fisher's exact test) and *KRAS* mutations ($P = 1.39e^{-24}$, Fisher's exact test) exhibited significant co-occurrence with LOY across multiple cancer types, although these associations diminished when stratifying for individual cancer types (Supplementary figure 10B). At the cancer type level, we observed positive associations between LOY and *CRLF2* deletions in glioma ($P = 3.14e^{-37}$, Fisher's exact test), *KDM5C* alterations in renal cell carcinoma ($P = 3.3e^{-13}$, Fisher's exact test), and *KDM6A* in bladder cancer (Supplementary figure 10B, $P = 1.8e^{-11}$, Fisher's exact test).

Considering the numerous clinicopathological variables that could potentially confound LOY predictions (Figure 6A), we performed additional analyses using a multivariable logistic regression model adjusted for cancer types, TP53 status (i.e., mutated or wild-type), and several other parameters (Methods). Even after adjustment, the model confirmed an association between CRLF2 deletions and LOY in glioma and bladder cancers (Figure 7B, adjusted p- value = $2.2e^{-11}$). Moreover, the lysine demethylase 5C (*KDM5C*) significantly co-occurred with LOY in renal cell carcinoma and *KDM6A* alterations were associated with LOY in bladder cancer (adjusted p-value = $1.7e^{-07}$). Several other recurrent alterations were independently associated with LOY (adjusted p-value < 0.1, Supplementary table 6). For instance, we observed a significant association between LOY and *PTPRD*, a receptor tyrosine phosphatase, functioning as TSG gene in glioma (Veeriah *et al.*, 2009), confirming a putative link between genomic instability and LOY (Figure 6B). Curiously, the adjusted model didn't confirm an association between LOY and somatic mutations in *CDKN2A*, *NFE2L2* and *LRP1B* – genes previously speculated to be associated with LOY within and across cancer types (Qi *et al.*, 2022).

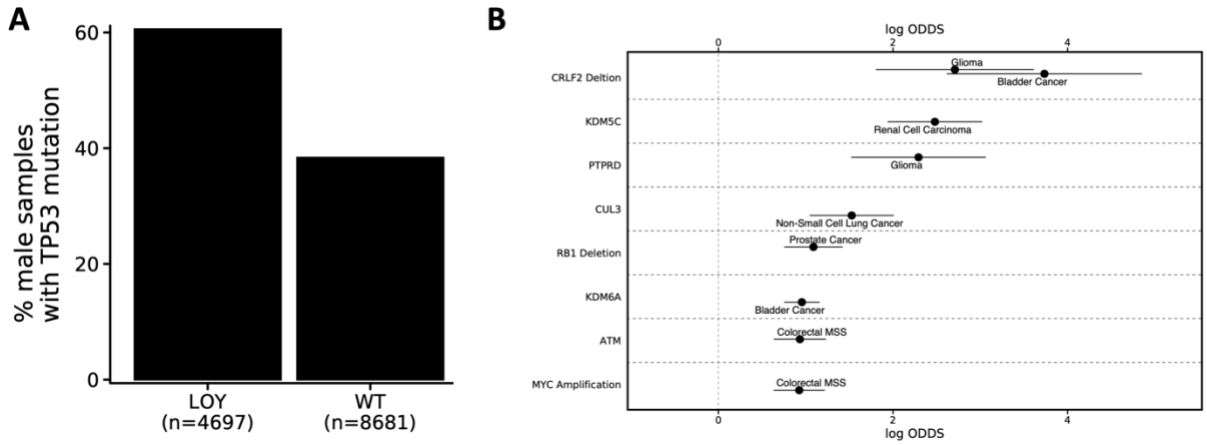


Figure 7: Somatic mutations and LOY. A) Fraction of male tumors with *TP53* damaging mutation in wild-type and LOY cases. B) Statistically significant associations (y-axis) with LOY across individual cancer types assessed by a multivariate regression model. The error bars indicate the log odds ratio (x-axis) plus or minus two times the standard error (horizontal black lines).

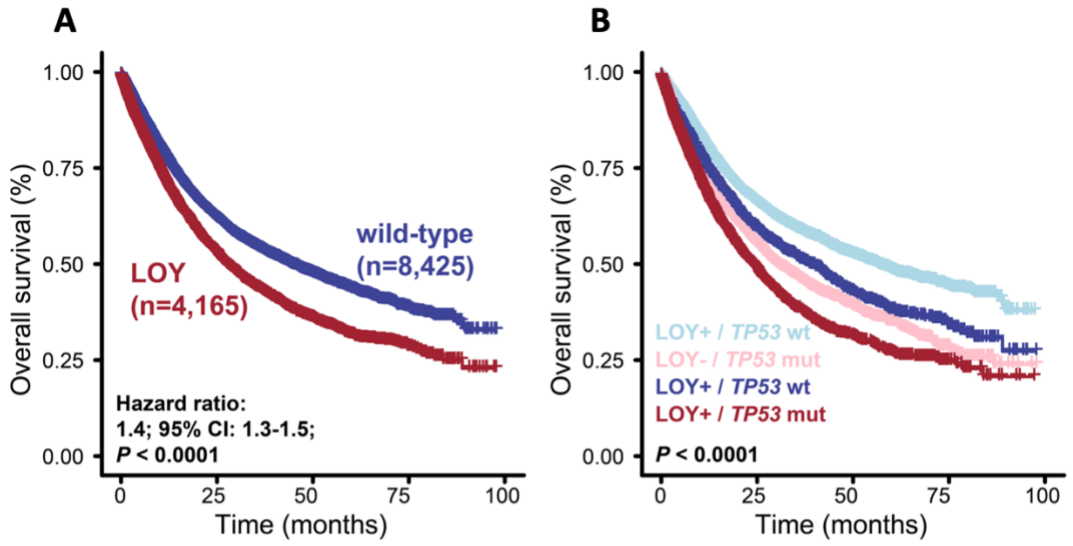
3.7 Chromosome Y loss depicts an independent prognostic factor in selected cancer types

We examined the overall survival (OS) of male samples from the time of sequencing based on the status of chromosome Y (complete loss vs. wild-type). The median OS for male patients with LOY was 27.6 (95% CI: 25.8-29.8) compared to 44.2 months (95% CI: 41.7-47.1) for intact Y chromosomes. This finding indicates that LOY is associated with poorer overall survival (Figure 8A, HR = 1.35; 95% CI: 1.28-1.43; $P < 0.001$, Wald test). Despite the influence of tumor type in a pan-cancer analysis, LOY remained significantly associated with decreased OS even after adjusting for cancer type and TP53 mutation status (Figure 8B, HR = 1.17; 95% CI: 1.11-1.25; $P < 0.001$; Wald test). At the individual tumor type level, LOY was significantly associated with poor OS ($P < 0.05$, Wald test) in gastrointestinal neuroendocrine, pancreatic, non-small cell lung, melanoma, bone, and prostate cancer. However, after adjusting for *TP53* mutational status, only melanoma, pancreatic, non-small cell lung, and prostate cancer showed lower OS in LOY cases (Supplementary figure 12B).

We then focused on prostate cancer, as it exhibited the highest hazard ratio (HR) associated with LOY, and a recent publication indicated that tumors deficient in the lysine-specific demethylase 5D (KDM5D) exhibited aggressive phenotypes (Komura *et al.*, 2016). In prostate cancer samples, 11.7% showed complete LOY, while a small fraction of samples showed gain/loss or partial loss configurations (Figure 5A, B). Interestingly, the majority losses (39/45, 86.7%) mapped to the Yq arm encompassing the *KDM5D* locus, suggesting a directed, focal deletion event (Supplementary figure 13A). We also observed a significant enrichment of oncogenic alterations in *TP53*, *AR*, *FOXA1*, *MYC*, *ATM* and *RB1* in prostate tumors with LOY

(Supplementary figure 13B). However, after controlling for *TP53* mutation status, FGA, purity, and other confounding factors in a multivariate logistic regression model (Methods), only *RB1* deletions, *ATM* mutations, and *SPOP* mutations remained significantly enriched in LOY cases (Supplementary figure 13C, $P < 0.05$, Wald test).

Using these genes in a multivariate survival model, while also correcting for tumor site, MSI-status, histological subgroup, age and TP53 status, we found further evidence that LOY is an independent predictor of survival (Figure 8C, HR: 1.37; 95% CI: 1.18-1.60; $P < 0.001$; Wald test). This result suggests that chromosome Y may have tumor-suppressive functions in prostate cancer, as recently suggested by various studies (Blair *et al.*, 2011; Harmeyer *et al.*, 2017; Tricarico *et al.*, 2020). Moreover, the analysis evidently highlighted the importance of stratifying tumor types by histological subgroups as prostate neuroendocrine carcinomas showed a significantly elevated HR compared to prostate adenocarcinomas (Figure 8C).



C

Variable	Level	N	Hazard ratio	HR (95% CI)	P-value
Chromosome Y	intact	1302		Reference	
	loss	200		1.37 (1.18, 1.60)	**<0.001**
Sample Site	Primary	807		Reference	
	Metastasis	695		1.05 (0.94, 1.17)	0.36
MSI-Status	Stable	1468		Reference	
	Instable	34		0.94 (0.66, 1.33)	0.71
Cancer Type Detailed	Prostate Adenocarcinoma (PRAD)	1479		Reference	
	Prostate Neuroendocrine Carcinoma (PRNE)	8		4.22 (2.05, 8.68)	**<0.001**
	Prostate Small Cell Carcinoma (PRSCC)	14		1.08 (0.61, 1.91)	0.78
	Prostate Squamous Cell Carcinoma (PRSC)	1		0.95 (0.13, 6.75)	0.96
Age		1502		1.01 (1.00, 1.01)	**0.01**
***TP53*-Status**	wild type	1017		Reference	
	mutated	485		1.22 (1.09, 1.37)	**<0.001**
TMPRSS2/ETS	wild type	1046		Reference	
	mutated	456		1.02 (0.91, 1.15)	0.68
***RB1*-Status**	wild type	1396		Reference	
	mutated	106		1.28 (1.03, 1.60)	**0.03**

Figure 8: LOY depicts an independent prognostic factor for overall survival in selected cancer types. A) Kaplan-Meier survival curve for male tumors with LOY compared to wild-type tumors. B) Kaplan-Meier survival curve for a multivariate cox-regression model, adjusted for TP53 mutational status. C) Multivariate model for OS for prostate cancer patients. Square boxes correspond to the hazard ratio (HR, x-axis) and the ranges indicate the 95% confidence interval. Numbers for respective categories are provided in column “N”. P-values were obtained from Wald-test.

4 Discussion and Conclusion

The current study aimed to investigate chromosome Y loss in a large cohort of over 13,000 male tumor specimens, representing the largest assembled cohort to date. Our analyses provide valuable insights into the dynamics of chromosome Y loss across more than 45 different tumor types. Using the MSK-IMPACT panel sequencing assays, we successfully determined both mosaic loss of chromosome Y (mLOY) and chromosome Y loss in tumor tissue (LOY).

Overall, we observed 32.6% of complete LOY which is in great agreement with recent estimates (Priestley et al., 2019; Qi et al., 2022). However, we note that we observed notable variations in LOY rates among different tumor types, highlighting the significance of distinguishing molecularly distinct cancer types for understanding disease progression Moreover, we established a significant association between chromosomal instability, quantified as the fraction of the genome altered (FGA) and LOY. This finding suggests that LOY may be a consequence of general chromosomal instability in male tumors. Additionally, we found a significant enrichment of *TP53* mutations in cases of LOY. This observation further supports our hypothesis that LOY is linked to overall chromosomal instability, as *TP53* is known to play a crucial role in maintaining genome stability (Eischen, 2016; Bernard et al., 2020; Feroz and Sheikh, 2020).

Furthermore, our analysis revealed that LOY may serve as a prognostic factor, as patients who retained their Y chromosome demonstrated significantly improved overall survival compared to those who experienced its loss. Notably, in prostate cancer patients we found that LOY is an independent predictor of survival, even after adjusting survival models for *TP53* mutational status, histological subtypes, and other putative confounding factors.

Overall, our study contributes to a better understanding of the role of LOY in male tumors and underscores the importance of considering chromosomal instability and *TP53* mutations in the context of LOY. Further research is warranted to uncover the precise mechanisms driving LOY to further elucidate the unique vulnerabilities that could be therapeutically exploited.

4.1 MSK-IMPACT for mLOY and LOY detection

The MSK-IMPACT sequencing assay, as described by Cheng et al. (Cheng et al., 2015) is a targeted sequencing approach that allows for the assessment of the mutational landscape of specific cancer genes. In our study, we questioned whether the limited representation of the genome in this assay would be sufficient for evaluating the status of chromosome Y in available

samples. Our findings demonstrate that the MSK-IMPACT sequenced samples can indeed be used for detecting both mosaic loss of chromosome Y (mLOY) and LOY in tumor samples.

Although these two approaches may sound similar, they represent two fundamentally different concepts. Firstly, age-related somatic loss of chromosome Y (mLOY) is a common occurrence in normal hematopoietic cells and has been associated with an increased risk of all-cause mortality (Forsberg et al., 2014, 2019). Although we didn't aim to establish associations between mLOY and clinicopathological factors, we conducted this analysis as part of the subsequent copy-number profiling in tumor samples. For instance, in tumor samples, chromosome Y copies can be 'gained' relative to the control (i.e., blood cells), or if the fraction of somatic LOY is approximately equal between the control and tumor no difference will be detected (Qi et al., 2022). With the exclusion of mLOY samples, we assume to decrease the amount of false-positive LOY calls in tumor samples.

Overall, we found that 942 samples (7.2%) exhibited some degree of mLOY, which aligns with recent estimates from Forsberg et al. who reported an 8.2% prevalence in a different cohort (Forsberg et al., 2014). Additionally, our data clearly supported the notion of a decline in chromosome Y ploidy with increasing age, consistent with previous reports.

Importantly, we witnessed vastly different mLOY estimates depending on which analysis strategy deployed. In particular, a uniform treatment of all MSK-IMPACT sequenced samples (i.e., across all gene-panels deployed) led to a conservative ploidy cut-off where signs of mLOY would be undetected in the majority of samples. Rather, we determined a unique ploidy cut-off for each targeted gene panel respectively, that allowed precise mLOY callings.

Furthermore, we opted for a binary assignment (yes/no) rather than a continuous determination of the fraction of cells experiencing mLOY (Danielsson et al., 2020). While this approach may be conservative, it was not the primary purpose of the current study to precisely identify mLOY dynamics from peripheral blood cells.

Subsequently, we asked whether samples that underwent MSK-IMPACT and whole-exome recapture (WES-recapture) sequencing provide us with the same CNA-profiles. Specifically, since the sequence representation of MSK-IMPACT sequenced samples is significantly lower compared to WES-recapture samples. Indeed, despite the lower sequence representation in the targeted gene panel sequencing, we confirmed that overall ploidy, purity, and copy-number alterations profiles were in excellent agreement between the two methods.

Furthermore, we validated our method, FacetsY, by applying it to primary tumors of renal cell carcinomas (KIRC) as part of the TCGA project (Creighton et al., 2013). We obtained consistent estimates of LOY in KIRC as previously published (Qi et al., 2022). Additionally,

we independently verified LOY calls with the lack of mRNA expression of genes located on the Y chromosome, further validating the reliability of our methodology. It is important to note, however, that we didn't investigate whether chromosome Y genes are dosage dependent or not; meaning that gene expression levels are proportional to the tumor copy number estimated. This aspect will be the focus of future research endeavors.

4.2 LOY across various tumor types

Overall, we found a LOY rate of 32.6% in the samples investigated, which aligns well with recent estimates (Priestley et al., 2019; Qi et al., 2022). Moreover, we observed strong correlations between the LOY rates in various tumor types in a recent study on male tumors from the TCGA project and the estimates from our study (Qi et al., 2022). However, we did notice a slight trend towards higher LOY rates in our cohort. We attribute this difference to the inclusion of mixed samples obtained from primary and metastatic sites, as well as the challenges in directly comparing tumor histologies. However, our estimates differ significantly from those obtained through wet-lab methods. For instance, Lukeis et al. demonstrated that 70% of NSCLC samples exhibited LOY using karyotyping and cytogenetic techniques, which is in stark contrast to our estimates of 37% (Lukeis et al., 1990). On the contrary, while Kovacs et al. determined a LOY rate of 85% in papillary renal cell carcinomas (PRCC) (Kovacs et al., 1991), we found that 86.6% of PRCC lost their Y chromosome. We believe that deviations from wet-lab techniques are mainly attributable to low samples sizes, as cytogenetic approach were labor-intensive.

Our study provides valuable insights into LOY dynamics across different cancer types, although it is important to consider the distinct lineages within each type. For example, we observed vast differences between LOY estimates across various histological subtypes of renal cell carcinomas and thyroid cancers. While papillary thyroid cancers only show 5.6% of complete LOY, Hurthle cell thyroid cancers exhibited LOY in over 54% of samples investigated.

Besides LOY, we observed chromosome Y gains in 19.9% of the samples analyzed, and a minor fraction of samples showing both a gain and a loss event on the Y chromosome at the same time (i.e., gain/loss). Although we didn't explicitly focus on chromosomal gains, we saw that samples that experienced imbalanced chromosome Y arms are enriched in bladder, prostate, and non-small cell lung cancer patients. We saw that whenever a gain and loss event happened at the same time, there is a significant trend towards gains happing on the Yp arm while the Yq arm seems to be predominantly deleted. Interestingly, there are several tumor suppressor genes,

such as *KDM5D* or *EIF1AY*, residing on the Yq arm, which may explain the elevated rate of chromosomal losses at this region. However, we note that GISTIC didn't identify any significantly recurrent focal deletions signals from those samples, hence tumor-suppressive functions of the Y chromosome remain a speculation. Furthermore, we note that this analysis is impacted by the low number of samples analyzed (n = 329).

4.3 LOY dynamics and it's prognostic value

We found that the fraction genome altered is the single most predictive variable for LOY. However, germ cell and renal cell carcinomas exhibit disproportionately high aneuploidy or LOY rates, deviating from the predictions of the model. It is known that germ cell tumors are prone to high levels of aneuploidy while having fewer somatic mutations (Shen et al., 2018), but it is unclear why the LOY rates do not follow this trend. It is possible that the Y chromosome serves important functions in younger patients and hence a loss would result in detrimental cellular behavior. On the other hand, renal cell carcinomas show disproportionally low aneuploidy compared to LOY rates, which requires further investigations. However, we speculate that the loss of chromosome Y is context, and cancer type dependent. That means, that some genes from the Y chromosome may fulfil tumor-suppressive functions in renal cell carcinomas and that the elimination of chromosome Y is a logical consequence for 'successful' tumor progression. This aspect merits further research efforts and depicts a major limitation to our study.

Besides FGA, we also found that purity has a significant confounding effect on LOY prediction. The reason behind this is currently unknown, but it does not appear to be related to sequencing coverage. Rather, we believe that purity isn't related to LOY per se, more that this association derived from tumor types that contribute most to the pan-cancer cohort and are known to have low purities in general (e.g., prostate, non-small cell lung cancer).

Lastly, we have demonstrated that LOY has prognostic value in prostate cancer, as male patients with LOY experience significantly reduced overall survival. This result even persists when including potential confounding factors in survival analysis, which further strengthens our finding.

In summary, LOY appears to be an independent predictive biomarker in prostate cancers. Its detection through standard clinical assays is relatively straightforward and hence would greatly contribute to improved genetic reports in future. However, there are still large knowledge gaps. For instance, the Y chromosome usually lacks a backup copy, meaning that LOY results in the loss of several unique and ubiquitously expressed genes, such as *KDM5D*, *KDM6A*, and

RPS4Y1, which may have important implications for cell fitness. However, we don't know yet whether some of those functions can be substituted by X-linked homologous genes, or not. The investigation of this aspect as well as further functional studies are needed to complement the existing literature on Y-linked tumor suppressors in tumor types with frequent LOY.

5 Supplementary Information

5.1 Supplementary tables

Supplementary table 1: Cancer type summary of the study cohort

Supplementary table 2: Baits (tiling intervals) deployed for chromosome Y DNA enrichment for MSK-IMPACT gene panels 341, 410, 468 and 505

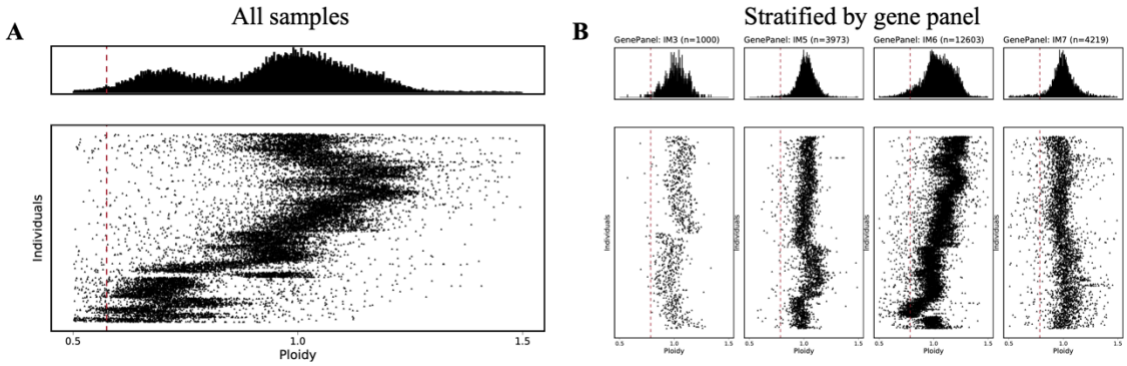
Supplementary table 3: Chromosome Y loci that were covered in >75% of sequencing files investigated

Supplementary table 4: Whole-exome recapture sequenced validation cohort (n = 769)

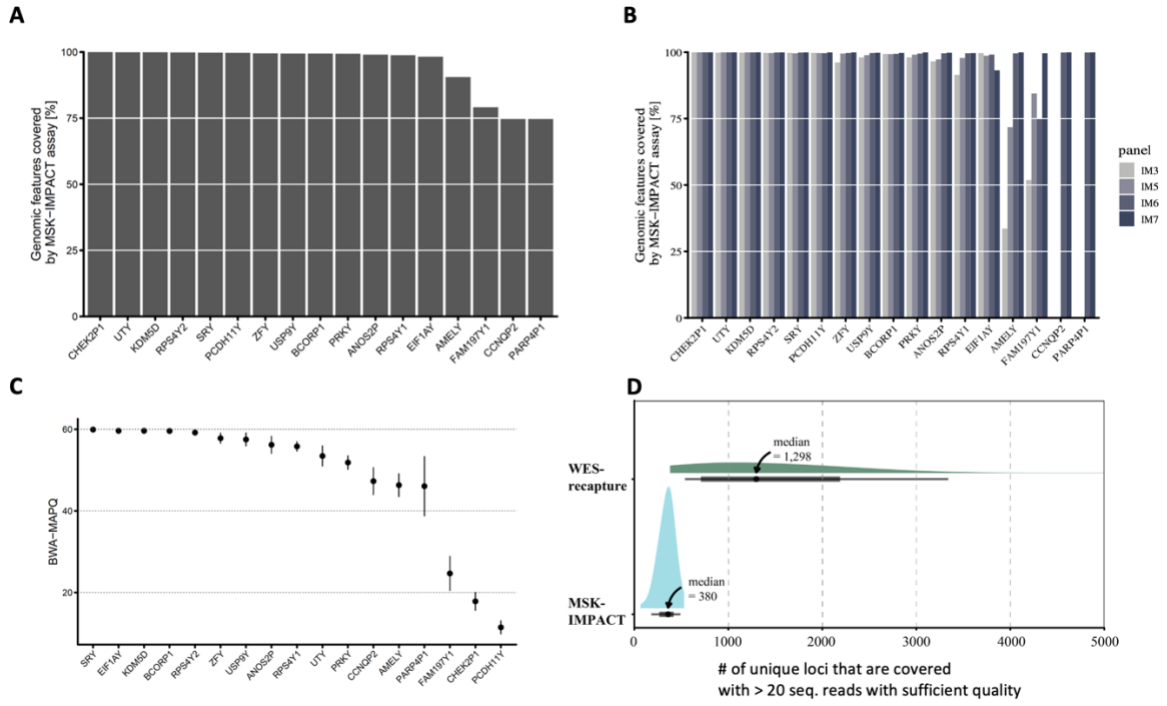
Supplementary table 5: TCGA cancer-study abbreviations used in Figure Figure 4C

Supplementary table 6: Results of the multivariable logistic regression model showing the association of mutations with LOY

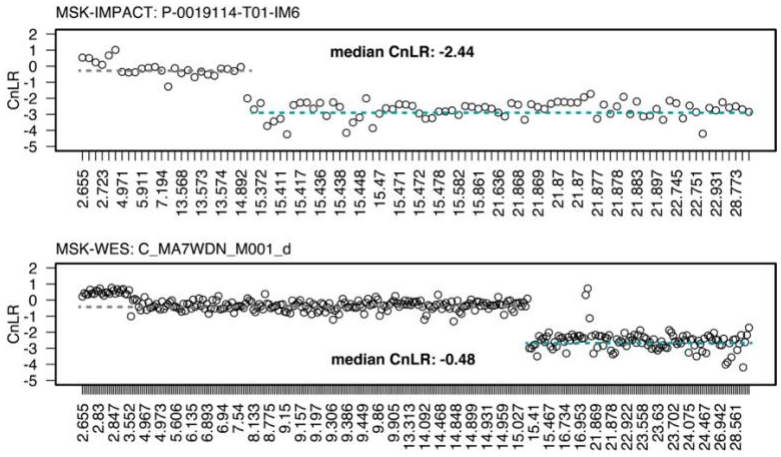
5.2 Supplementary figures



Supplementary figure 1: Mosaic loss of chromosome Y detection using MSK-IMPACT targeted gene panel sequencing. A) Extrapolated chromosome Y ploidies (x-axis) from normal blood cells are plotted for each individual sample with a histogram (top) showing the ploidy-distribution across the whole cohort. The red dotted line indicates the ploidy threshold (0.57) that was deployed as cut-off to determine mLOY (Methods). B) Extrapolated chromosome Y ploidies stratified by gene-panel deployed. Red lines indicate the 95% percentile cut-off deployed to flag samples with signs of mLOY. The following cut-offs were used: IM3 = 0.84; IM5 = 0.89; IM6 = 0.78; IM7 = 0.78.

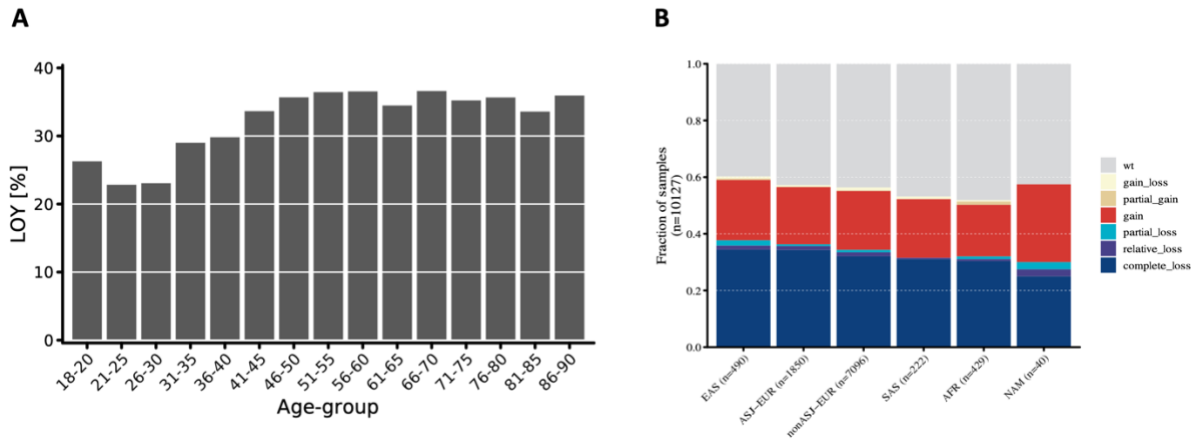


Supplementary figure 2: Detailed representation of chromosome Y using MSK-IMPACT sequencing assays. A) Genetic elements (x-axis) that are covered in >75% of samples (n=21,369) analyzed in the present study. B) Sequence coverage of chromosome Y elements depend on which sequencing assay was deployed. Gene panel IM3 (grey bars) depicts the initial release targeting 341 genes, whereas IM7 (purple bars) targets 505 genes. C) Average mapping qualities (BWA-MAPQ) of retained genetic elements are drawn. Dots depict the mean, while corresponding error bars indicate the standard deviation. D) WES-recapture samples (top, green density plot) show on average more unique loci that are covered with >20 sequencing reads with sufficient quality than MSK-IMPACT sequenced samples (bottom, blue density plot).



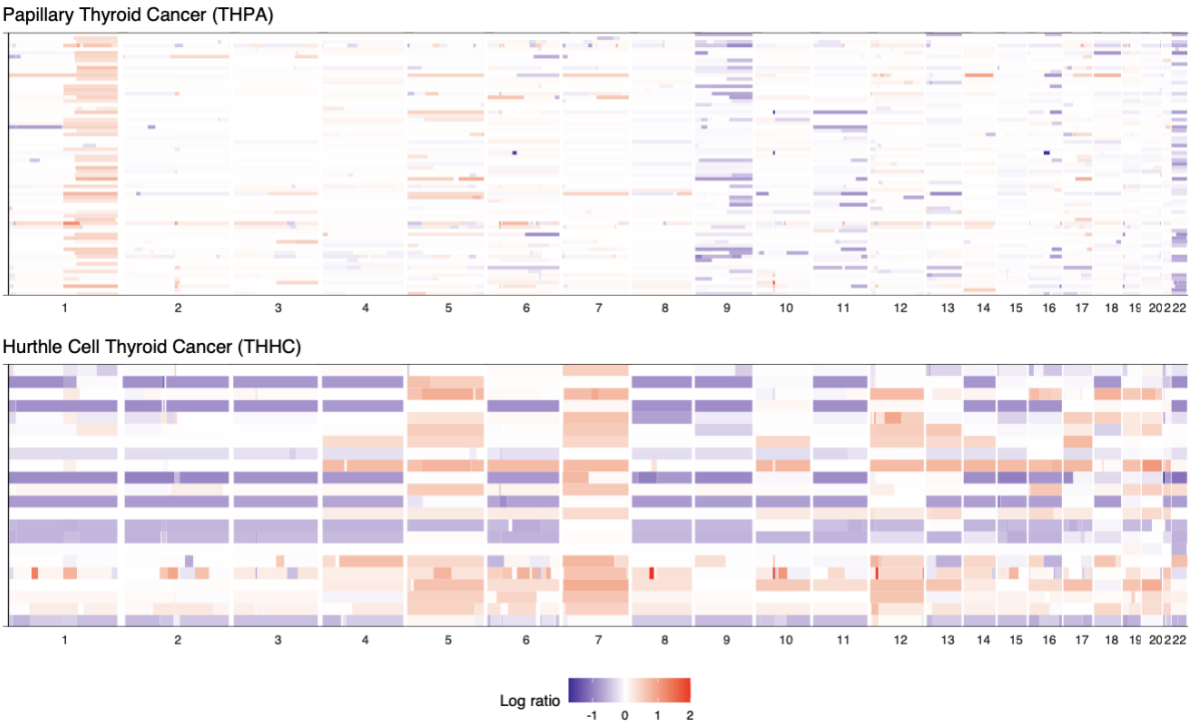
Supplementary figure 3: Exemplary tumor sample (P-0019114-T01-IM6) that highlights the reason for the disagreement between the CnLR estimates obtained from MSK-IOMPACT and WES-recapture sequencing experiments. A) The plot shows individual CnLR estimates (y-axis) across the Y-chromosome (x-axis) for the MSK-IMPACT sequenced sample. B) The matched WES-recaptured sample is depicted with CnLR estimates on the y-axis, while the respective Y-chromosome coordinates are shown on the x-axis. Both plots include grey and turquoise lines representing local averages.

It is important to note that since the median across the entire Y-chromosome was used to approximate the copy-number, this metric is highly dependent on the number of individual observations. In this example, two regions with different copy-number states are evident. The grey line indicates a region of length 12.65 Mb (range: 2.655 – 15.372 Mb) with a median CnLR of - 0.238 for the MSK-IMPACT sequenced sample. Similarly, for the WES-recaptured sample, a region of 12.76 (range: 2.655-15.41 Mb) with a median CnLR of -0.36 was obtained. The second segment clearly indicates a loss in both sequencing strategies. However, due to the higher genomic loci content in WES-recaptured samples and the central tendency assessed with the median, different central estimates were observed despite making the same observations. This helps to address the issue of outliers observed in Figure 3B.

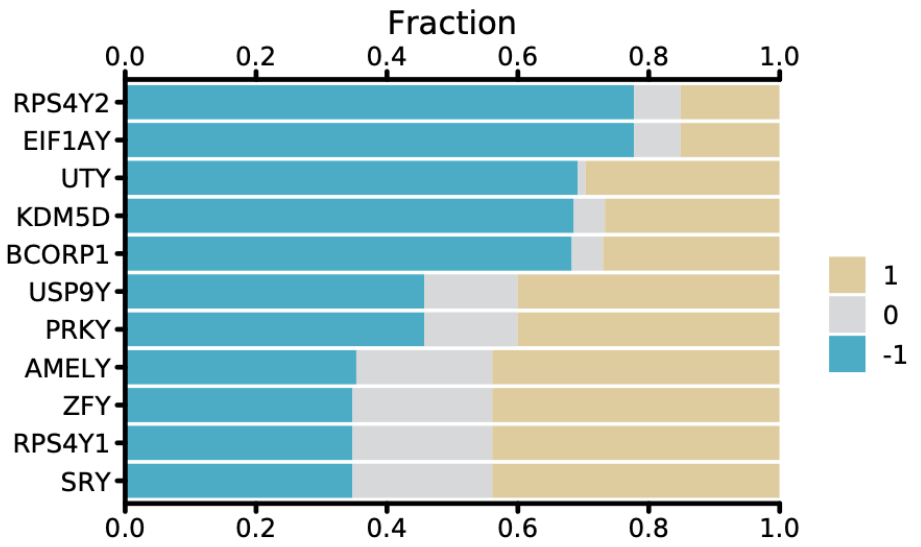


Supplementary figure 4: LOY rates across age groups and ancestry calls. A) LOY rates (y-axis) are binned by age-groups (x-axis) in 5y intervals. Note that the youngest study participants were 18y. B) The relative Y-copy number configurations (y-axis) are plotted against individuals with different ancestry estimates (x-axis). EAS = East Asian; ASJ-EUR = Ashkenazi Jewish (ASJ)-European; EUR = European; SAS = South Asian; AFR = African; NAM = Native American.

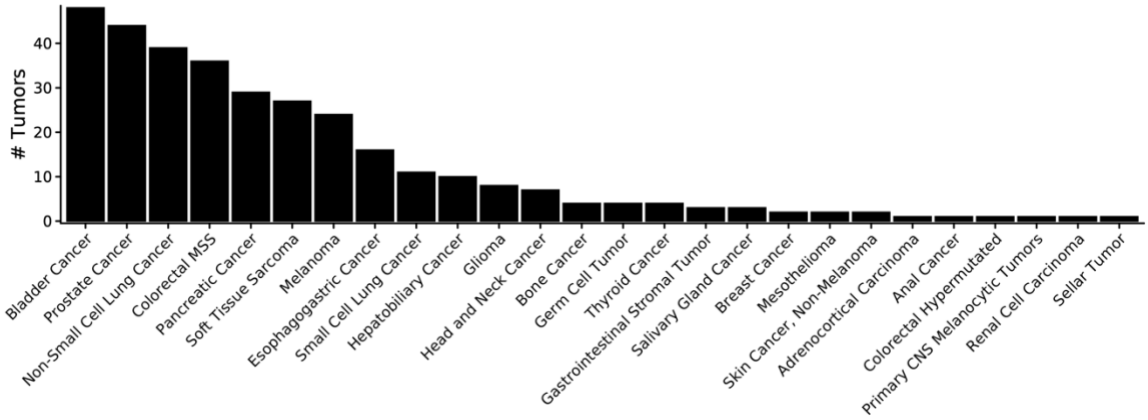
Supplementary figure 5: LOY rates across histological subgroups in cancer types investigated in the present study.



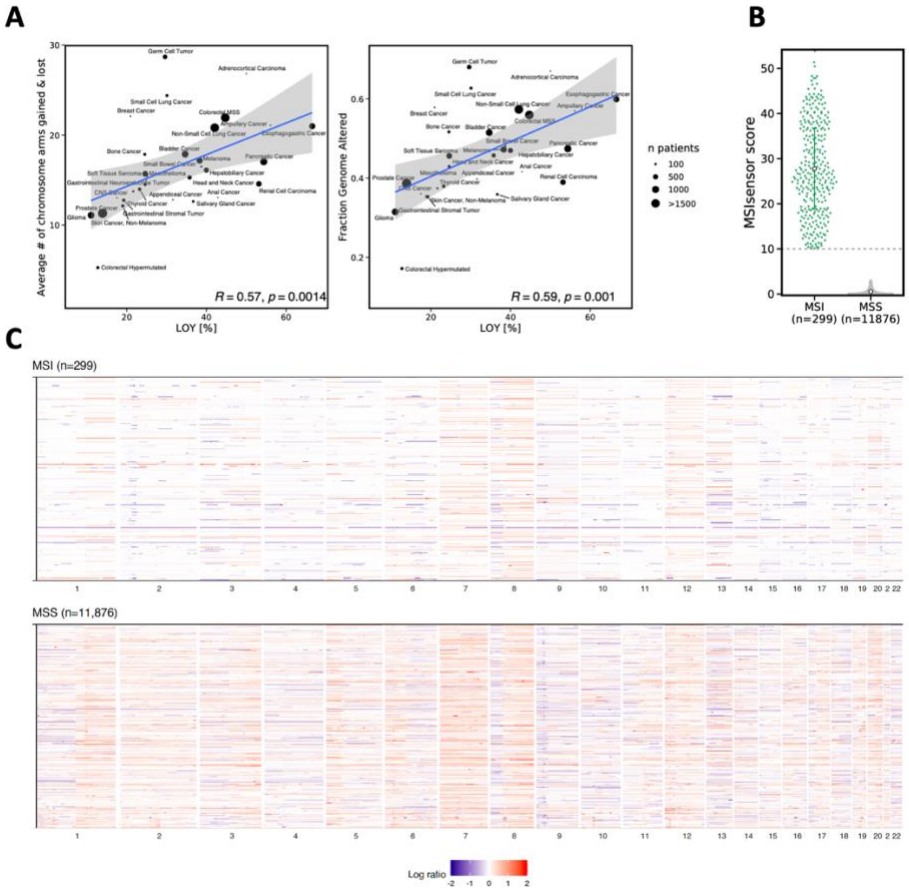
Supplementary figure 6: Somatic copy-number segmentation plot for A) papillary and B) Hurthle cell thyroid cancers. Segmentation means are capped to -2 and +2, respectively.



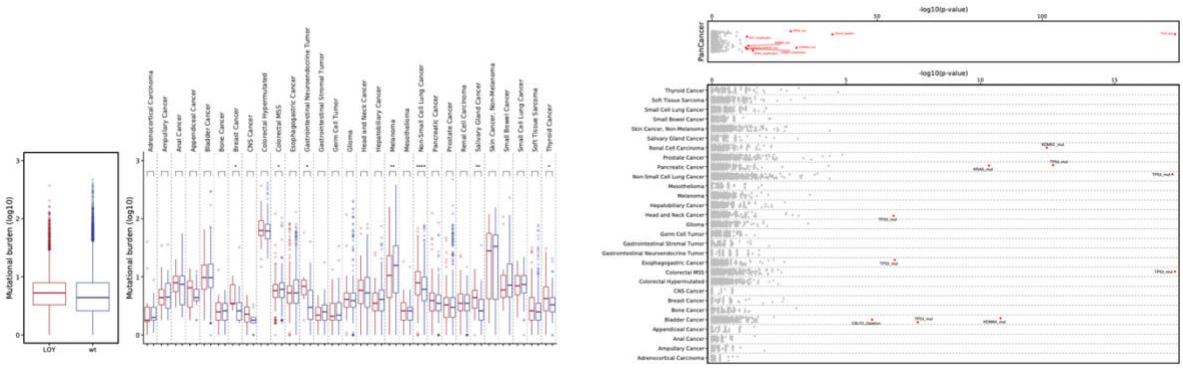
Supplementary figure 7: Frequency of chromosome Y gene deletions (-1) and gains (1) seen in 329 samples with chromosomal-arm imbalances.



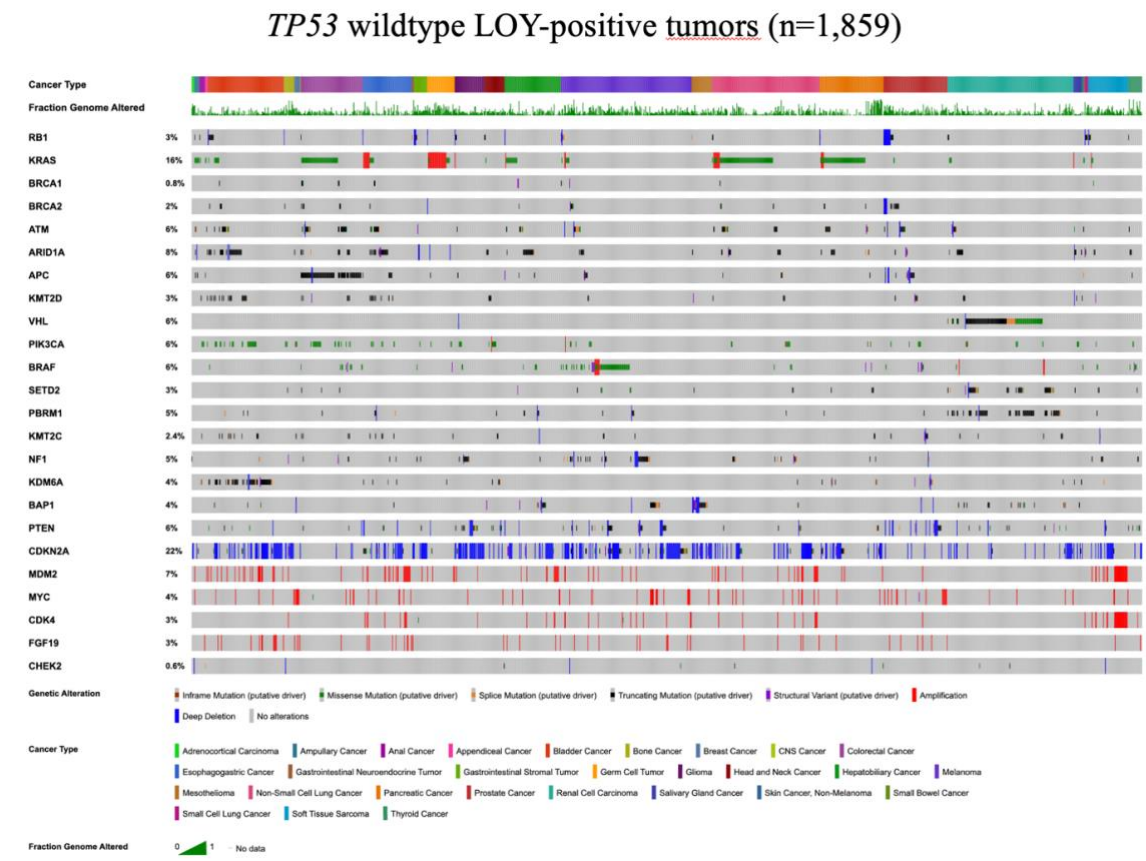
Supplementary figure 8: Absolute number of male tumors (y-axis) with chromosome Y imbalanced (Methods) across different tumor types (x-axis)



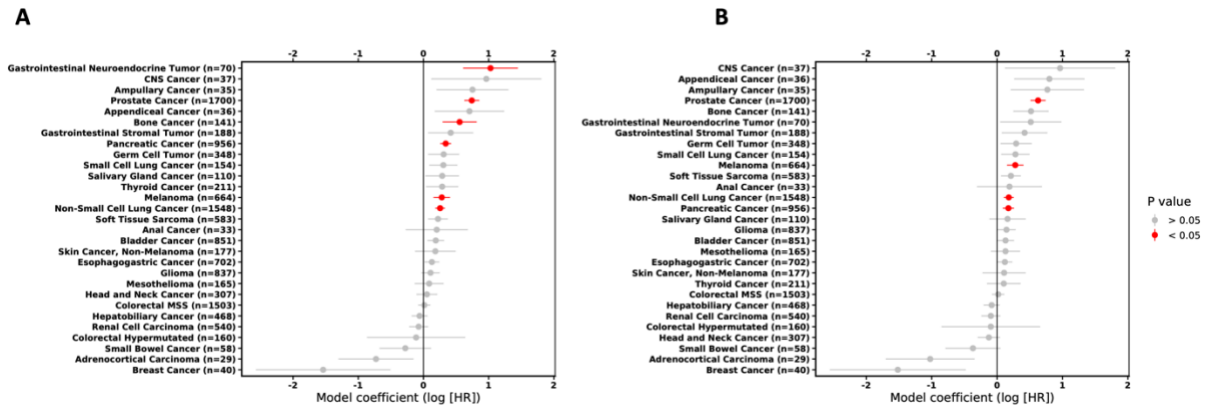
Supplementary figure 9: LOY is common in aneuploid tumors. A) The average number of chromosome arms that are lost and/or gained (y-axis) are plotted against the fraction of LOY (x-axis), while the second plot shows the fraction of genome altered (FGA) in percent against the LOY rate in various cancer types. The blue line derived from a linear modeling fit with a subsequent smoothing showing the 95% CI in grey. R and p are derived from correlation analysis using Spearman's method. B) MSIsensor scores (Niu et al., 2014) were used to distinguish microsatellite stable (MSS) and microsatellite instable (MSI) tumors (x-axis). Tumor samples showing MSIsensor scores >10 were considered as MSI. C) Genome-wide (chromosome numbers underneath the panels) CNA-segmentation patterns in MSI (top) and MSS (bottom) tumors. Note that bluish bars indicated losses (i.e., log ratios <0) while reddish segments indicate gains with log ratios >0.



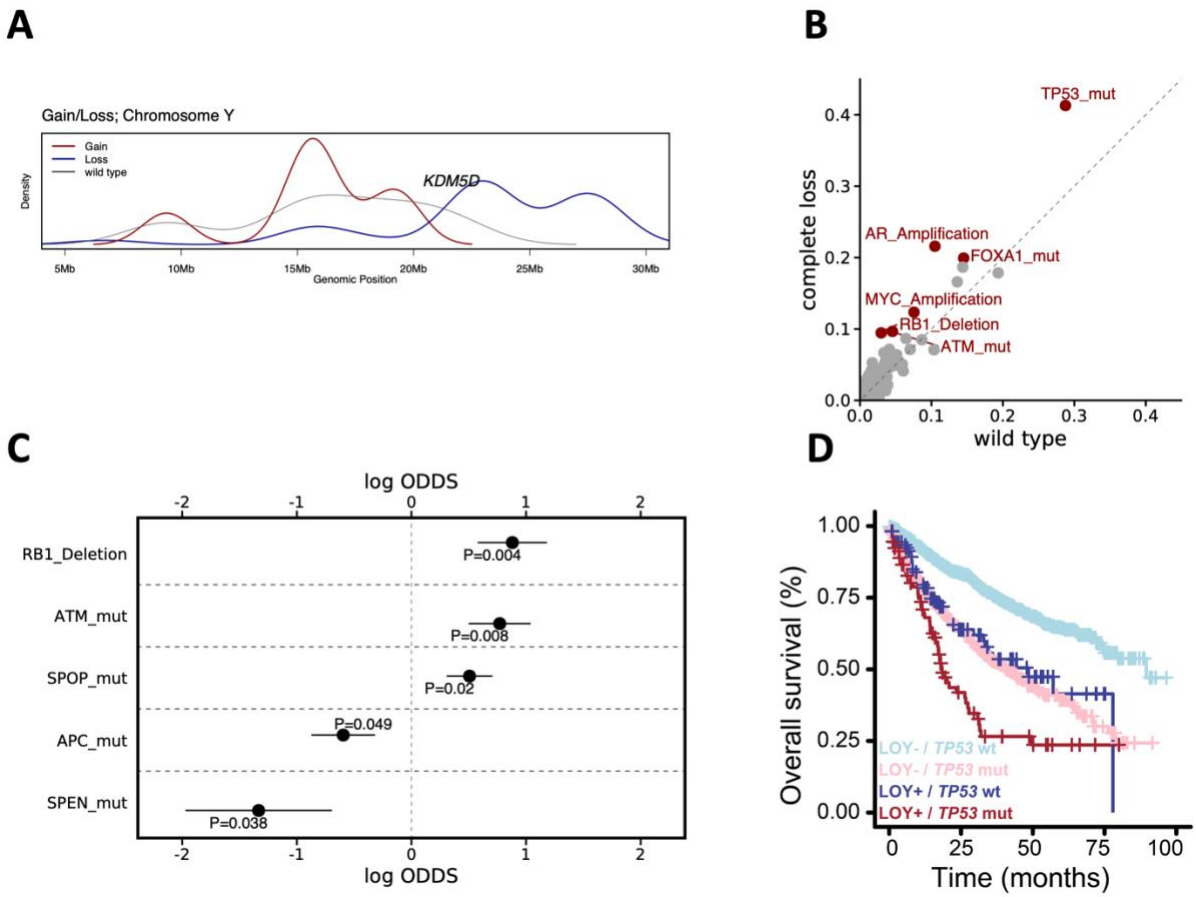
Supplementary figure 10: Association of point mutations and LOY. A) (left) Mutational burden, expressed as the logarithm of the total number of mutations is shown for LOY tumors (red) and wild-type (blue) tumors. (right). TMB stratified by LOY status and tumor types; *P < 0.05, **P<0.01, ***P<0.001, ****P<0.0001 by Mann-Whitney U test. B) Dot chart showing significantly enriched mutations (red dots) in various tumor lineages. Top plots indicate a pan-cancer overview. Note that all mutations, either point- or structural or CNA mutations were considered (see Methods).



Supplementary figure 11: Alteration frequencies in LOY positive and TP53 wildtype tumor samples



Supplementary figure 12: A) Univariate and B) multivariate survival analysis (controlling for TP53 damaging mutations) across cancer types analyzed. Tumor types are sorted by the log hazard ratio (HR). Dots highlight HR (log) estimates with corresponding 95% confidence interval. Red dots indicate cancer types where p-values are < 0.05 (Wald test).



Supplementary figure 13: Mutations and overall survival in prostate cancer patients. A) Density of breakpoints of imbalanced chromosome Y samples, show greatest loss-density around the *KDM5D* locus. B) Somatic mutations associated with LOY (Y-axis) compared to wild-type cases (x-axis). Red dots indicate genes where the alteration difference between LOY and wild-type cases is > 5%. C) Multivariate logistic regression model, showing genes (y-axis) that are enriched in LOY cases (log ODDS > 0) or depleted (log ODDS < 0). D) Cox-proportional hazard model adjusted for TP53 mutational status.

6 List of Figures

7 Deutsche Zusammenfassung

Die letzten zehn Jahre haben eine Flut von Tumor-Sequenzierungsdata geliefert. Dennoch wurde die Untersuchung des Chromosoms Y in genomweiten Analysen nahezu einheitlich vernachlässigt. Daher ist die spezifische Rolle des Y-Chromosoms in Bezug auf Tumorentstehung und -progression weitgehend unklar.

In der vorliegenden Studie haben wir eine Analyse von Chromosom-Y-Aberrationen bei über 13,000 Patienten in über 45 verschiedenen Krebsarten durchgeführt. Wir haben unsere Ergebnisse mit orthogonalen Datenquellen validiert und Korrelationen mit klinisch-pathologischen und genomischen Faktoren festgestellt. Darüber hinaus wird der prognostische Wert von LOY in verschiedenen Malignitäten beleuchtet.

Unsere Studie bestätigte die Eignung der MSK-IMPACT Sequenzierungsmethode und Daten zur Untersuchung von Chromosom-Y-Aberrationen sowohl in normalen als auch in Tumorproben. Wir haben im Durchschnitt 34.9% LOY in den männlichen Tumorproben detektiert, jedoch mit deutlichen Unterschieden zwischen und innerhalb von verschiedenen Krebsarten. Wir haben eine positive Korrelation von LOY mit chromosomaler Instabilität, ausgedrückt durch den Anteil des veränderten Genoms, festgestellt. Darüber hinaus wurden somatische Mutationen in *KDM5C*, *KDM6A* und *CRLF2* mit LOY in verschiedenen Tumorlinien assoziiert. Letztlich konnten wir zeigen, dass LOY einen prognostischen Wert bei Prostata Adenokarzinomen birgt, da die Gesamtüberlebensdauer bei Patienten mit Y-Chromosomen länger ist.

Wir kommen zu dem Schluss, dass LOY häufig vorkommt und weitgehend mit p53- vermittelter genomischer Instabilität assoziiert ist. Daher stellt es für die meisten Tumorarten ein Nebenprodukt der chromosomalen Instabilität dar. Bei Prostata Adenokarzinomen hat das Y-Chromosom jedoch eine klinische Bedeutung, vermutlich durch epigenetische Faktoren wie *KDM5D*. Zusammenfassend liefert unsere Studie einen umfassenden Katalog von LOY-Schätzungen, bestätigt viele der jüngsten Erkenntnisse und bietet eine Grundlage für weiterführende Untersuchungen.

8 Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor Nikolaus Schultz. Niki is a brilliant and genuine researcher, and I am very thankful for the opportunity to join his outstanding research facility. He was incredibly supportive during our meetings, and always surprised and stimulated me with new thoughts, ideas, and approaches. Apart from research, he also guided parts of my personal development – which was momentous, especially during the last few years of Corona. Thank you, Niki, that my personal needs have equal value than research! I would also like to thank my colleagues Subhi, Henry, Bastien, Walid and many others for always taking time listening to my crazy thoughts, for fruitful conversations and discussions and random banter. In addition, I would like to thank my internal supervisor Dr. Ulrich Technau, who agreed on supervising this work internally, so that this research could actually be realized. Furthermore, the financial aid that supported me, is gratefully acknowledged.

Finally, I want to enunciate my strongest appreciation and gratitude to my friends for relentlessly supporting me in any possible way and for advice whenever I had to make decisions. Your unceasing encouragement is invaluable!

Thank you all of you – You are all part of the person who I am…

9 References

Agahozo, M.C. *et al.* (2020) 'Loss of Y-Chromosome during Male Breast Carcinogenesis', *Cancers*, 12(3), p. 631. Available at: https://doi.org/10.3390/cancers12030631.

Agrawal, N. *et al.* (2014) 'Integrated Genomic Characterization of Papillary Thyroid Carcinoma', *Cell*, 159(3), pp. 676–690. Available at: https://doi.org/10.1016/j.cell.2014.09.050.

Arora, K. *et al.* (2022) 'Genetic Ancestry Correlates with Somatic Differences in a Real-World Clinical Cancer Sequencing Cohort', *Cancer discovery*, 12(11), pp. 2552–2565. Available at: https://doi.org/10.1158/2159-8290.CD-22-0312.

Arseneault, M. *et al.* (2017) 'Loss of chromosome Y leads to down regulation of KDM5D and KDM6C epigenetic modifiers in clear cell renal cell carcinoma', *Scientific Reports*, 7(1), p. 44876. Available at: https://doi.org/10.1038/srep44876.

Bachtrog, D. (2013) 'Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration', *Nature Reviews Genetics*, 14(2), pp. 113–124. Available at: https://doi.org/10.1038/nrg3366.

Bailey, M.H. *et al.* (2018) 'Comprehensive Characterization of Cancer Driver Genes and Mutations', *Cell*, 173(2), pp. 371-385.e18. Available at: https://doi.org/10.1016/j.cell.2018.02.060.

Barbari, S.R. and Shcherbakova, P.V. (2017) 'Replicative DNA polymerase defects in human cancers: Consequences, mechanisms, and implications for therapy', *DNA Repair*, 56, pp. 16–25. Available at: https://doi.org/10.1016/j.dnarep.2017.06.003.

Bellott, D.W. *et al.* (2014) 'Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators', *Nature*, 508(7497), pp. 494–499. Available at: https://doi.org/10.1038/nature13206.

Bernard, E. *et al.* (2020) 'Implications of TP53 allelic state for genome stability, clinical presentation and outcomes in myelodysplastic syndromes', *Nature Medicine*, 26(10), pp. 1549–1556. Available at: https://doi.org/10.1038/s41591-020-1008-z.

Bielski, C.M., Zehir, A., *et al.* (2018) 'Genome doubling shapes the evolution and prognosis of advanced cancers', *Nature Genetics*, 50(8), pp. 1189–1195. Available at: https://doi.org/10.1038/s41588-018-0165-1.

Bielski, C.M., Donoghue, M.T.A., *et al.* (2018) 'Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer', *Cancer Cell*, 34(5), pp. 852-862.e4. Available at: https://doi.org/10.1016/j.ccell.2018.10.003.

Blair, L.P. *et al.* (2011) 'Epigenetic Regulation by Lysine Demethylase 5 (KDM5) Enzymes in Cancer', *Cancers*, 3(1), pp. 1383–1404. Available at: https://doi.org/10.3390/cancers3011383.

Blanco, P. *et al.* (2000) 'Conservation of PCDHX in mammals; expression of human X/Y genes predominantly in brain', *Mammalian Genome*, 11(10), pp. 906–914. Available at: https://doi.org/10.1007/s003350010177.

Canisius, S., Martens, J.W.M. and Wessels, L.F.A. (2016) 'A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence', *Genome Biology*, 17(1), p. 261. Available at: https://doi.org/10.1186/s13059-016-1114-x.

Casimiro, M.C. *et al.* (2012) 'ChIP sequencing of cyclin D1 reveals a transcriptional role in chromosomal instability in mice', *The Journal of Clinical Investigation*, 122(3), pp. 833–843. Available at: https://doi.org/10.1172/JCI60256.

Cerami, E. *et al.* (2012) 'The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data', *Cancer Discovery*, 2(5), pp. 401–404. Available at: https://doi.org/10.1158/2159-8290.CD-12-0095.

Chakraborty, G. *et al.* (2020) 'Fraction genome altered (FGA) to regulate both cell autonomous and non-cell autonomous functions in prostate cancer and its effect on prostate cancer aggressiveness.', *Journal of Clinical Oncology*, 38(6_suppl), pp. 347–347. Available at: https://doi.org/10.1200/JCO.2020.38.6_suppl.347.

Chakravarty, D. *et al.* (2017) 'OncoKB: A Precision Oncology Knowledge Base', *JCO Precision Oncology*, (1), pp. 1–16. Available at: https://doi.org/10.1200/PO.17.00011.

Cheng, D.T. *et al.* (2015) 'Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology', *The Journal of Molecular Diagnostics*, 17(3), pp. 251–264. Available at: https://doi.org/10.1016/j.jmoldx.2014.12.006.

Cook, M.B. *et al.* (2011) 'Sex Disparities in Cancer Mortality and Survival', *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 20(8), pp. 1629–1637. Available at: https://doi.org/10.1158/1055-9965.EPI-11-0246.

Creighton, C.J. *et al.* (2013) 'Comprehensive molecular characterization of clear cell renal cell carcinoma', *Nature*, 499(7456), pp. 43–49. Available at: https://doi.org/10.1038/nature12222.

Danielsson, M. *et al.* (2020) 'Longitudinal changes in the frequency of mosaic chromosome Y loss in peripheral blood cells of aging men varies profoundly between individuals', *European Journal of Human Genetics*, 28(3), pp. 349–357. Available at: https://doi.org/10.1038/s41431-019-0533-z.

Durinck, S. *et al.* (2009) 'Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt', *Nature Protocols*, 4(8), pp. 1184–1191. Available at: https://doi.org/10.1038/nprot.2009.97.

Edgren, G. *et al.* (2012) 'Enigmatic sex disparities in cancer incidence', *European Journal of Epidemiology*, 27(3), pp. 187–196. Available at: https://doi.org/10.1007/s10654-011-9647-5.

Eischen, C.M. (2016) 'Genome Stability Requires p53', *Cold Spring Harbor Perspectives in Medicine*, 6(6), p. a026096. Available at: https://doi.org/10.1101/cshperspect.a026096.

Feroz, W. and Sheikh, A.M.A. (2020) 'Exploring the multiple roles of guardian of the genome: P53', *Egyptian Journal of Medical Human Genetics*, 21(1), p. 49. Available at: https://doi.org/10.1186/s43042-020-00089-x.

Forsberg, L.A. *et al.* (2014) 'Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer', *Nature Genetics*, 46(6), pp. 624–628. Available at: https://doi.org/10.1038/ng.2966.

Forsberg, L.A. *et al.* (2019) 'Mosaic loss of chromosome Y in leukocytes matters', *Nature Genetics*, 51(1), pp. 4–7. Available at: https://doi.org/10.1038/s41588-018-0267-9.

Forsberg, L.A., Gisselsson, D. and Dumanski, J.P. (2017) 'Mosaicism in health and disease - clones picking up speed', *Nature Reviews. Genetics*, 18(2), pp. 128–142. Available at: https://doi.org/10.1038/nrg.2016.145.

Giunta, S. and Funabiki, H. (2017) 'Integrity of the human centromere DNA repeats is protected by CENP-A, CENP-C, and CENP-T', *Proceedings of the National Academy of Sciences*, 114(8), pp. 1928–1933. Available at: https://doi.org/10.1073/pnas.1615133114.

Godfrey, A.K. *et al.* (2020) 'Quantitative analysis of Y-Chromosome gene expression across 36 human tissues', *Genome Research*, 30(6), pp. 860–873. Available at: https://doi.org/10.1101/gr.261248.120.

Guo, X. *et al.* (2020) 'Mosaic loss of human Y chromosome: what, how and why', *Human Genetics*, 139(4), pp. 421–446. Available at: https://doi.org/10.1007/s00439-020-02114-w.

Harmeyer, K.M. *et al.* (2017) 'JARID1 Histone Demethylases: Emerging Targets in Cancer', *Trends in cancer*, 3(10), pp. 713–725. Available at: https://doi.org/10.1016/j.trecan.2017.08.004.

Hughes, J.F. *et al.* (2010) 'Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content', *Nature*, 463(7280), pp. 536–539. Available at: https://doi.org/10.1038/nature08700.

Hunter, S. *et al.* (1993) 'Y chromosome loss in esophageal carcinoma: an in situ hybridization study', *Genes, Chromosomes & Cancer*, 8(3), pp. 172–177. Available at: https://doi.org/10.1002/gcc.2870080306.

Jacobs, P.A. *et al.* (1963) 'Change of human chromosome count distribution with age: evidence for a sex differences', *Nature*, 197, pp. 1080–1081. Available at: https://doi.org/10.1038/1971080a0.

Jobling, M.A. and Tyler-Smith, C. (2003) 'The human Y chromosome: an evolutionary marker comes of age', *Nature Reviews Genetics*, 4(8), pp. 598–612. Available at: https://doi.org/10.1038/nrg1124.

Jonsson, P. *et al.* (2019) 'Tumour lineage shapes BRCA-mediated phenotypes', *Nature*, 571(7766), pp. 576–579. Available at: https://doi.org/10.1038/s41586-019-1382-1.

Komura, K. *et al.* (2016) 'Resistance to docetaxel in prostate cancer is associated with androgen receptor activation and loss of KDM5D expression', *Proceedings of the National Academy of Sciences of the United States of America*, 113(22), pp. 6259–6264. Available at: https://doi.org/10.1073/pnas.1600420113.

Kovacs, G. *et al.* (1991) 'Cytogenetics of papillary renal cell tumors', *Genes, Chromosomes and Cancer*, 3(4), pp. 249–255. Available at: https://doi.org/10.1002/gcc.2870030403.

Krumm, N. *et al.* (2012) 'Copy number variation detection and genotyping from exome sequence data', *Genome Research*, 22(8), pp. 1525–1532. Available at: https://doi.org/10.1101/gr.138115.112.

Lahn, B.T. and Page, D.C. (1999) 'Four Evolutionary Strata on the Human X Chromosome', *Science*, 286(5441), pp. 964–967. Available at: https://doi.org/10.1126/science.286.5441.964.

Li, C.H. *et al.* (2018) 'Sex Differences in Cancer Driver Genes and Biomarkers', *Cancer Research*, 78(19), pp. 5527–5537. Available at: https://doi.org/10.1158/0008-5472.CAN-18-0362.

Li, N. *et al.* (2016) 'JARID1D Is a Suppressor and Prognostic Marker of Prostate Cancer Invasion and Metastasis', *Cancer Research*, 76(4), pp. 831–843. Available at: https://doi.org/10.1158/0008-5472.CAN-15-0906.

Lopes-Ramos, C.M., Quackenbush, J. and DeMeo, D.L. (2020) 'Genome-Wide Sex and Gender Differences in Cancer', *Frontiers in Oncology*, 10, p. 597788. Available at: https://doi.org/10.3389/fonc.2020.597788.

Lukeis, R. *et al.* (1990) 'Cytogenetics of non-small cell lung cancer: Analysis of consistent non-random abnormalities', *Genes, Chromosomes and Cancer*, 2(2), pp. 116–124. Available at: https://doi.org/10.1002/gcc.2870020207.

Mangul, S. *et al.* (2021) 'Seeing beyond the target: Leveraging off-target reads in targeted clinical tumor sequencing to identify prognostic biomarkers'. bioRxiv, p. 2021.05.28.446240. Available at: https://doi.org/10.1101/2021.05.28.446240.

Mermel, C.H. *et al.* (2011) 'GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers', *Genome Biology*, 12(4), p. R41. Available at: https://doi.org/10.1186/gb-2011-12-4-r41.

Milosevic, J.D. *et al.* (2012) 'Clinical significance of genetic aberrations in secondary acute myeloid leukemia', *American Journal of Hematology*, 87(11), pp. 1010–1016. Available at: https://doi.org/10.1002/ajh.23309.

Minner, S. *et al.* (2010) 'Y chromosome loss is a frequent early event in urothelial bladder cancer', *Pathology*, 42(4), pp. 356–359. Available at: https://doi.org/10.3109/00313021003767298.

Morgan, M. *et al.* (2022) 'Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import'. Bioconductor version: Release (3.15). Available at: https://doi.org/10.18129/B9.bioc.Rsamtools.

Niu, B. *et al.* (2014) 'MSIsensor: microsatellite instability detection using paired tumor-normal sequence data', *Bioinformatics*, 30(7), pp. 1015–1016. Available at: https://doi.org/10.1093/bioinformatics/btt755.

Oram, S.W. *et al.* (2006) 'TSPY potentiates cell proliferation and tumorigenesis by promoting cell cycle progression in HeLa and NIH3T3 cells', *BMC cancer*, 6, p. 154. Available at: https://doi.org/10.1186/1471-2407-6-154.

Plch, J., Hrabeta, J. and Eckschlager, T. (2019) 'KDM5 demethylases and their role in cancer cell chemoresistance', *International Journal of Cancer*, 144(2), pp. 221–231. Available at: https://doi.org/10.1002/ijc.31881.

Priestley, P. *et al.* (2019) 'Pan-cancer whole-genome analyses of metastatic solid tumours', *Nature*, 575(7781), pp. 210–216. Available at: https://doi.org/10.1038/s41586-019-1689-y.

Ptashkin, R.N. *et al.* (2022) 'Enhanced clinical assessment of hematologic malignancies through routine paired tumor:normal sequencing'. medRxiv, p. 2022.10.03.22280675. Available at: https://doi.org/10.1101/2022.10.03.22280675.

Qi, M. *et al.* (2022) 'Loss of chromosome Y in primary tumors'. bioRxiv, p. 2022.08.22.504831. Available at: https://doi.org/10.1101/2022.08.22.504831.

Rice, W.R. (1996) ‘Evolution of the Y Sex Chromosome in Animals: Y chromosomes evolve through the degeneration of autosomes’, *BioScience*, 46(5), pp. 331–343. Available at: https://doi.org/10.2307/1312947.

Rubin, J.B. *et al.* (2020) ‘Sex differences in cancer mechanisms’, *Biology of Sex Differences*, 11, p. 17. Available at: https://doi.org/10.1186/s13293-020-00291-x.

Sauter, G. *et al.* (1995) ‘Y chromosome loss detected by FISH in bladder cancer’, *Cancer Genetics and Cytogenetics*, 82(2), pp. 163–169. Available at: https://doi.org/10.1016/0165-4608(95)00030-S.

Scarselli, A. *et al.* (2018) ‘Gender differences in occupational exposure to carcinogens among Italian workers’, *BMC Public Health*, 18(1), p. 413. Available at: https://doi.org/10.1186/s12889-018-5332-x.

Shen, H. *et al.* (2018) ‘Integrated Molecular Characterization of Testicular Germ Cell Tumors’, *Cell Reports*, 23(11), pp. 3392–3406. Available at: https://doi.org/10.1016/j.celrep.2018.05.039.

Shen, R. and Seshan, V.E. (2016) ‘FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing’, *Nucleic Acids Research*, 44(16), p. e131. Available at: https://doi.org/10.1093/nar/gkw520.

Shinbrot, E. *et al.* (2014) ‘Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication’, *Genome Research*, 24(11), pp. 1740–1750. Available at: https://doi.org/10.1101/gr.174789.114.

Shirole, N.H. *et al.* (2016) ‘TP53 exon-6 truncating mutations produce separation of function isoforms with pro-tumorigenic functions’, *eLife*. Edited by J.M. Espinosa, 5, p. e17929. Available at: https://doi.org/10.7554/eLife.17929.

Siegel, R.L., Miller, K.D. and Jemal, A. (2017) ‘Cancer Statistics, 2017’, *CA: a cancer journal for clinicians*, 67(1), pp. 7–30. Available at: https://doi.org/10.3322/caac.21387.

Skaletsky, H. *et al.* (2003) ‘The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes’, *Nature*, 423(6942), pp. 825–837. Available at: https://doi.org/10.1038/nature01722.

Taylor, A.M. *et al.* (2018) ‘Genomic and Functional Approaches to Understanding Cancer Aneuploidy’, *Cancer Cell*, 33(4), pp. 676-689.e3. Available at: https://doi.org/10.1016/j.ccell.2018.03.007.

Therneau, T.M. *et al.* (2023) ‘survival: Survival Analysis’. Available at: https://cran.r-project.org/web/packages/survival/index.html (Accessed: 18 June 2023).

Thompson, D.J. *et al.* (2019) 'Genetic predisposition to mosaic Y chromosome loss in blood', *Nature*, 575(7784), pp. 652–657. Available at: https://doi.org/10.1038/s41586-019-1765-3.

Tricarico, R. *et al.* (2020) 'X- and Y-linked chromatin-modifying genes as regulators of sex-specific cancer incidence and prognosis', *Clinical cancer research : an official journal of the American Association for Cancer Research*, 26(21), pp. 5567–5578. Available at: https://doi.org/10.1158/1078-0432.CCR-20-1741.

Van Loo, P. *et al.* (2010) 'Allele-specific copy number analysis of tumors', *Proceedings of the National Academy of Sciences*, 107(39), pp. 16910–16915. Available at: https://doi.org/10.1073/pnas.1009843107.

Veeriah, S. *et al.* (2009) 'The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers', *Proceedings of the National Academy of Sciences*, 106(23), pp. 9435–9440. Available at: https://doi.org/10.1073/pnas.0900571106.

Vokes, N.I. *et al.* (2019) 'Harmonization of Tumor Mutational Burden Quantification and Association With Response to Immune Checkpoint Blockade in Non–Small-Cell Lung Cancer', *JCO Precision Oncology*, (3), pp. 1–12. Available at: https://doi.org/10.1200/PO.19.00171.

Wallrapp, C. *et al.* (2001) 'Loss of the Y chromosome is a frequent chromosomal imbalance in pancreatic cancer and allows differentiation to chronic pancreatitis', *International Journal of Cancer*, 91(3), pp. 340–344. Available at: https://doi.org/10.1002/1097-0215(200002)9999:9999<::aid-ijc1014>3.0.co;2-u.

Yaeger, R. *et al.* (2018) 'Clinical Sequencing Defines the Genomic Landscape of Metastatic Colorectal Cancer', *Cancer Cell*, 33(1), pp. 125-136.e3. Available at: https://doi.org/10.1016/j.ccell.2017.12.004.

Zehir, A. *et al.* (2017) 'Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients', *Nature Medicine*, 23(6), pp. 703–713. Available at: https://doi.org/10.1038/nm.4333.