Mosaic loss of chromosome Y is associated with common variation near TCL1A

Weiyin Zhou1,2,17, Mitchell J Machiela1,17, Neal D Freedman1,17, Nathaniel Rothman1, Nuria Malats3, Casey Dagnall1,2, Neil Caporaso1, Lauren T Teras4, Mia M Gaudet4, Susan M Gapstur4, Victoria L Stevens4, Kevin B Jacobs2,5, Joshua Sampson1, Demetrious Albanes1, Stephanie Weinstein1, Jarmo Virtamo6, Sonja Berndt1, Robert N Hoover1, Amanda Black1, Debra Silverman1, Jonine Figueroa1, Montserrat Garcia-Closas1,7, Francisco X Real3,8, Julie Earl3, Gaele Marenne3, Benjamin Rodriguez-Santiago8–10, Margaret Karagas11, Alison Johnson12, Molly Schwenn13, Xifeng Wu14, Jian Gu14, Yuanqing Ye14, Amy Hutchinson1,2, Margaret Tucker1, Luis A Perez-Jurado8,9,15, Michael Dean1,16, Meredith Yeager1,2,18 & Stephen J Chanock1,18

Mosaic loss of chromosome Y (mLOY) leading to gonosomal XY/XY commonly occurs during aging, particularly in smokers. We investigated whether mLOY was associated with non-hematological cancer in three prospective cohorts (8,679 cancer cases and 5,110 cancer-free controls) and genetic susceptibility to mLOY. Overall, mLOY was observed in 7% of men, and its prevalence increased with age (per-year odds ratio (OR) = 1.13, 95% confidence interval (CI) = 1.12–1.15; P < 2 × 10−16), reaching 18.7% among men over 80 years old. mLOY was associated with current smoking (OR = 2.35, 95% CI = 1.82–3.03; P = 5.55 × 10−11), but the association weakened with years after cessation. mLOY was not consistently associated with overall or specific cancer risk (for example, bladder, lung or prostate cancer) nor with cancer survival after diagnosis (multivariate-adjusted hazard ratio = 0.87, 95% CI = 0.73–1.04; P = 0.12). In a genome-wide association study, we observed the first example of a common susceptibility locus for genetic mosaicism, specifically mLOY, which maps to TCL1A at 14q32.13, marked by rs2887399 (OR = 1.55, 95% CI = 1.36–1.78; P = 1.37 × 10−10).

mLOY refers to loss of the Y chromosome in a subset of cells while the remainder of cells retain the normal chromosome. For more than four decades, it has been noted that a fraction of healthy men lose all or some portion of the Y chromosome over the course of their lifetime1. Moreover, several studies have reported mLOY in males of advanced age, suggesting that mLOY is associated with aging and increasing hypodiploidy1–4. Other studies have suggested that mLOY is associated with specific hematological disorders, including acute myelogenous leukemia (AML), myelodysplastic syndrome (MDS) and preleukemias5–9.

SNP genotyping arrays have become an important tool for discovering common variants that contribute to human diseases10. Two widely used applications of this technology are genome-wide association studies (GWAS) and copy number variant (CNV) analyses for large-scale mosaic autosomal aberrations11–18. SNP microarray data have also been used to investigate mosaicism on the sex chromosomes19,20. Other studies have evaluated next-generation sequencing data to detect mosaic mutations at the base-pair level21–23. Although each of these studies of mosaic CNVs, mosaic uniparental disomies, mosaic single-nucleotide variants (SNVs) and mLOY ascertain different aspects of the biological process of clonal expansion, taken together, these studies suggest that the frequency of mosaic events—both large and small—increases with age. This trend could reflect either a deterioration in the capacity to maintain a stable genome or, alternatively, a decline in stem cell diversity22–25.

We investigated the association between mLOY and age at DNA collection, smoking status, DNA source (derived from blood or buccal material), inferred ancestry, genetic susceptibility to mLOY, non-hematological cancer risk and cancer-specific survival in subjects from three prospective cohorts: the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC), the Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial (PLCO) and the Cancer Prevention

1Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), US National Institutes of Health (NIH), Bethesda, Maryland, USA. 2Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Leidos Biomedical Research, Inc., Frederick, Maryland, USA. 3Cancer Genomics Research Center (CNIO), Madrid, Spain. 4Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA. 5Bioinformatics, LLC, Gaithersburg, Maryland, USA. 6Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland. 7Division of Genetics and Epidemiology, Institute for Cancer Research, London, UK. 8Departamento de Ciencias Experimentales i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain. 9Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain. 10Quantitative Genomic Medicine Laboratory, qGenomics, Barcelona, Spain. 11Biostatistics and Epidemiology Section, Dartmouth Medical School, Lebanon, New Hampshire, USA. 12Vermont Cancer Registry, Burlington, Vermont, USA. 13Maine Cancer Registry, Augusta, Maine, USA. 14Department of Epidemiology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. 15Hospital del Mar Research Institute (IMIM), Barcelona, Spain. 16Laboratory of Experimental Immunology, Center for Cancer Research, NCI-Frederick, Frederick, Maryland, USA. 17These authors contributed equally to this work. 18These authors jointly directed this work. Correspondence should be addressed to M.Y. (yeagerm@mail.nih.gov) or S.J.C. (chanocks@mail.nih.gov).

Received 20 July 2015; accepted 16 March 2016; published online 11 April 2016; doi:10.1038/ng.3545
The fraction of men with mLOY across 5-year age groups for all subjects (n = 13,729). The fraction of men with chromosome Y loss is calculated as the number of men in an age group with chromosome Y loss divided by the total number of men in that age group. Error bars represent 95% Jeffreys confidence intervals around the proportion estimates. (a) mLOY is associated with older age at DNA collection, with frequencies of 1.18% in individuals less than 60 years of age and 18.71% in those 80 years or older (OR per year of age = 1.13, 95% CI = 1.12–1.15; P = 2 × 10⁻¹⁶). The χ² test for trend among the six age groups shows that there is significant evidence that the fraction of men with chromosome Y loss increases with age (P = 2.2 × 10⁻¹⁶). (b) Scatterplot for age versus fraction of men with chromosome Y loss. There is an overall increasing trend for the fraction of men with chromosome Y loss until age 80 years. After age 80 years, the trend becomes unstable, reflecting the limited number of subjects in this age group in our study (Supplementary Table 4). All statistical tests were two-sided.

Study II (CPS-II) (Supplementary Table 1). SNP microarray data generated using Illumina Infinium arrays for GWAS that had sufficient coverage of the Y chromosome (HumanHap610K, Hap1M, Omni1M and Omni2.5M arrays) were used to detect mLOY in DNA isolated from blood or buccal cells for 8,679 males with non-hematological cancer and 5,110 cancer-free adult male controls. The hybridization data from all subjects were examined for deviations from the expected log₂-transformed intensity ratio (Online Methods) representing evidence of loss or gain of the male-specific region of chromosome Y (MSY; chr. Y: 6,671,498–22,919,969, hg18/Build 36) (Supplementary Fig. 1a). We observed 970 men (7.03%) with detectable mLOY, and the estimated fraction of affected cells was 22.7–73.4% (Online Methods and Supplementary Fig. 1c), which differs from the fraction for autosomes (7–95% mosaicism)⁴,¹¹. No significant difference in mLOY frequency was observed when adjusting for DNA source, either blood or buccal (P = 0.33), or by genotyping array (P = 0.14; Supplementary Table 2).

We validated the accuracy of our mLOY detection method by qPCR analysis of markers for 15 genes distributed across the MSY (Supplementary Fig. 1b)²⁶. We selected 124 subjects from the cohorts with probable mLOY for validation by qPCR assays. The concordance rate between the SNP microarrays and qPCR was 87.9% (Supplementary Table 3), with no differences observed by array type. Among the 13,789 males scanned, we found evidence of chromosome Y gain for 133 males (0.96%). For the males with suspected chromosome Y gain, we had DNA available from 69 individuals for qPCR validation. The concordance rate was only 49% for mosaic gains (Supplementary Table 3), with validation corresponding to substantial log R ratio (LRR) deviation from baseline. After removing 34 men with validated chromosome Y gain, we excluded 26 participants with similarly large positive deviations from the baseline after manual review. The resulting data set included 8,632 cancer cases and 5,097 controls for analysis of chromosome Y loss. Of the 8,632 cases, 5,545 had their blood or buccal cells collected at least 1 year before cancer diagnosis. The majority of these participants (n = 5,369) were diagnosed with bladder, lung or prostate cancer.

The most significant association for mLOY was with increasing age: in 13,729 men, the frequency of detectable mLOY increased from 1.18% for men under 60 years of age to 18.71% for men over 80 years of age. The proportion of individuals with mLOY increased for every age stratum up to 80 years of age (P < 2.2 × 10⁻¹⁶); for older participants, the estimates became less stable because of the small numbers of subjects (Fig. 1a,b and Supplementary Table 4). Adjusting for smoking status (current smoking and, for former smokers, years since cessation), ancestry, source of DNA and contributing study, we observed evidence for an association between mLOY and age (OR per year of age = 1.13, 95% CI = 1.12–1.15; P < 2.00 × 10⁻¹⁶) (Table 1 and Supplementary Table 5). This association with age was of greater magnitude than that observed previously for large-scale structural

### Table 1 Predictors of mLOY

| Age (ATBC + PLCO + CPS-II) | n with Y loss | n normal | Total | Proportion with mLOY (%) | OR (95% CI) | P value |
|----------------------------|--------------|----------|-------|--------------------------|-------------|---------|
| < 20                       | 970          | 12,759   | 13,729| 7.07                     | 1.13 (1.12–1.15) | <2.00 × 10⁻¹⁶|
| Smoking status (PLCO + CPS-II) | | | | | | |
| Never-smoker               | 209          | 3,199    | 3,408 | 6.13 | Reference |
| Former smoker              | 446          | 4,964    | 5,410 | 8.24 | 1.33 (1.12–1.57) | 0.001   |
| Current smoker             | 100          | 934      | 1,034 | 9.67 | 2.35 (1.82–3.03) | 5.55 × 10⁻¹¹|
| Years since cessation (PLCO + CPS-II) | | | | | | |
| 1–4                        | 38           | 332      | 370   | 10.27 | 2.15 (1.49–3.10) | 3.83 × 10⁻⁵|
| 5–10                       | 64           | 567      | 631   | 10.14 | 1.92 (1.43–2.58) | 1.26 × 10⁻⁵|
| 11–20                      | 107          | 1,167    | 1,274 | 8.40 | 1.49 (1.17–1.90) | 0.001   |
| >20                        | 231          | 2,809    | 3,040 | 7.60 | 1.10 (0.91–1.34) | 0.333   |
| Cigarettes per day in current smokers (ATBC + PLCO + CPS-II) | | | | | | |
| 1–10                       | 61           | 727      | 788   | 7.74 | Reference |
| 11–20                      | 154          | 2,201    | 2,355 | 6.54 | 1.04 (0.76–1.43) | 0.800   |
| 21–30                      | 66           | 1,206    | 1,272 | 5.19 | 0.92 (0.63–1.33) | 0.648   |
| >30                        | 26           | 429      | 455   | 5.71 | 0.95 (0.58–1.55) | 0.835   |

Age refers to the age at DNA collection, and the odds ratio estimate is for a 1-year increase in age. For the association between age and mLOY, the analysis was adjusted for smoking status (current smoking and, for former smokers, the number of years since cessation), estimated admixture proportion, DNA source and contributing study. For the association between mLOY and the number of years since cessation and the number of cigarettes smoked per day, the analyses were adjusted for age, estimated admixture proportion, DNA source and contributing study.
Figure 2  mLOY and smoking analysis. (a) Proportion of males with mLOY across strata for 5-year age group and smoking status for subjects from the PLCO and CPS-II studies (n = 9,859). Error bars represent 95% Jeffery’s confidence intervals around the proportion estimates. Currently smoking men ≥75 years old have 13.90-fold increased odds of having mLOY as compared to non-smoking men <65 years old (95% CI = 6.60–29.26; P = 4.13 × 10−12). (b) Association of current smoking and years since cessation with mLOY in adjusted logistic regression models (n = 8,825). The dashed red line represents an OR of 1.0 for the referent never-smokers. Error bars represent Wald 95% confidence intervals around the odds ratio estimates. All statistical tests were two-sided.

autosomal mosaic events (OR = 1.05, 95% CI = 1.04–1.07)15 and for mosaic SNVs (OR = 1.08, 95% CI = 1.07–1.09)22.

The frequency of individuals exhibiting age-related mLOY was nearly tenfold greater than that observed for autosomal mosaicism (0.7–2.0%)14–16. A detectable mosaic autosomal abnormality (>2 Mb in length) was observed in 130 (0.95%) men16. This rate is substantially lower than what was observed for mLOY in the same individuals (7.07%), indicating that mLOY is the most frequent large-scale chromosomal somatic event. Of the 970 males in our study with mLOY, 18 (1.86%) also had evidence of detectable autosomal mosaicism (Supplementary Tables 6 and 7), suggesting that men with mLOY are more likely to harbor large autosomal mosaic events than men without mLOY (OR = 2.13, 95% CI = 1.22–3.55; P = 0.005).

We investigated the association with smoking status in 9,859 subjects (1,034 current smokers, 5,410 former smokers, 3,408 never-smokers and 7 individuals of unknown status) from the PLCO and CPS-II studies, excluding participants from ATBC, which recruited only current smokers. The frequency of mLOY was higher in ever-smokers (65.4%) than in never-smokers (34.6%), with the highest frequency among current smokers. Relative to never-smokers who were less than 65 years old, current smokers over 75 years old had increased risk of mLOY (OR = 13.9, 95% CI = 6.60–29.26; P = 4.13 × 10−12) (Fig. 2a and Supplementary Table 8). Adjusting for age, ancestry, DNA source and study, we observed an association of mLOY with current smoking (OR = 2.35, 95% CI = 1.82–3.03; P = 5.55 × 10−11) and former smoking (OR = 1.33, 95% CI = 1.12–1.57; P = 0.001) (Table 1 and Supplementary Table 9). This result is consistent with a previous report that men who smoke are at greater risk of mLOY, with observed odds ratios for current smoking versus non-smoking that ranged from 2.4 (95% CI = 1.6–3.6) to 4.3 (95% CI = 2.8–6.7)19 for the three studies included. Mosaic SNVs have also been associated with smoking with comparable odds ratio estimates (OR = 2.2)23. It is notable that previous studies have not reported an association between smoking and large-scale (>2-Mb) autosomal mosaicism14,15.

Because of differences in mLOY between current and former smokers, the risk of mLOY progressively declined with increasing number of years after cessation, with OR = 2.15 within 1–4 years of quitting (95% CI = 1.49–3.10; P = 3.83 × 10−5), OR = 1.92 within 5–10 years of quitting (95% CI = 1.43–2.58; P = 1.26 × 10−5) and OR = 1.49 within 11–20 years of quitting (95% CI = 1.17–1.90; P = 0.001) relative to never smoking. By 20 years after cessation, there was no evidence for association (OR = 1.10, 95% CI = 0.91–1.34, P = 0.33) (Fig. 2b, Table 1 and Supplementary Table 10). Our findings suggest that smoking has a long-lasting impact on mLOY, influencing its frequency perhaps more than a decade after quitting, but the association wanes with long-term cessation. Among 4,904 current smokers, we observed no association of mLOY with smoking intensity, measured by cigarettes per day (Table 1 and Supplementary Table 11).

It has been proposed that autosomal mosaicism may be associated with risk for certain solid tumors, but a connection has not been definitely established14–16. Moreover, others have suggested that mLOY is associated with cancer risk overall20. We investigated the frequency of mLOY in blood or buccal DNA and solid tumor risk in 5,545 subjects with cancer from whom DNA was collected at least 1 year before cancer diagnosis and in 5,097 cancer-free individuals (Supplementary Table 12). Overall, mLOY was slightly more common in men who went on to develop cancer (6.67%) than in cancer-free controls (5.49%) (unadjusted OR = 1.23, 95% CI = 1.04–1.45; P = 0.012; multivariate-adjusted OR = 1.19, 95% CI = 1.00–1.42; P = 0.047) (Table 2, Supplementary Fig. 2 and Supplementary Table 13). Furthermore, we investigated a possible association between cancer risk and mLOY for cancer types with sufficient sample size, including bladder, lung and prostate cancers in adjusted analyses (using continuous age, smoking status (current smoking and, for former smokers, years since quitting), ancestry, source of DNA and study). In the cohort studies with DNA collected 1 year or more before cancer diagnosis, we observed a possible association between mLOY and risk of bladder cancer (OR = 1.47, 95% CI = 1.09–1.99; P = 0.011) and risk of prostate cancer (OR = 1.35, 95% CI = 1.04–1.74; P = 0.024) but no evidence for a relationship with lung cancer (OR = 0.90, 95% CI = 0.69–1.18, P = 0.45) (Table 2). This latter point is striking given that lung cancer is more strongly associated with smoking than either bladder or prostate cancer. We examined cases diagnosed at or before

Table 2 Association between mLOY and incident cancer overall and stratified by date of DNA collection

| Cancer Type | DNA 1 year before cancer diagnosis | DNA at and after cancer diagnosis | All |
|-------------|----------------------------------|----------------------------------|-----|
|             | n | OR (95% CI) | P | n | OR (95% CI) | P | n | OR (95% CI) | P |
| Bladder     | 731 | 1.47 (1.09–1.99) | 0.011 | 558 | 2.01 (1.47–2.75) | 1.40 × 10−5 | 1,289 | 1.69 (1.33–2.13) | 1.26 × 10−5 |
| Lung        | 1,908 | 0.90 (0.69–1.18) | 0.450 | 381 | 0.81 (0.48–1.38) | 0.44 | 2,289 | 0.90 (0.70–1.17) | 0.44 |
| Prostate    | 2,730 | 1.35 (1.04–1.74) | 0.024 | 2,093 | 1.53 (1.17–2.01) | 1.84 × 10−3 | 4,823 | 1.43 (1.19–1.73) | 1.40 × 10−4 |
| Combined    | 5,545 | 1.19 (1.00–1.42) | 0.047 | 3,087 | 1.52 (1.23–1.87) | 8.40 × 10−5 | 8,632 | 1.31 (1.13–1.53) | 4.79 × 10−4 |

The analysis used mLOY as the predictor variable and cancer type as the response variable and was adjusted for continuous age, smoking status (current smoking and, for former smokers, the number of years since cessation), estimated admixture proportion, DNA source and contributing study. A total of 5,097 cancer-free controls were used as a reference. Other cancers in addition to bladder, lung and prostate cancers were included when performing the combined cancer analysis.

NATURE GENETICS VOLUME 48 | NUMBER 5 | MAY 2016

© 2016 Nature America, Inc. All rights reserved.
biospecimen sampling. For bladder and prostate cancers but not lung cancer, when examining DNA obtained contemporaneously with cancer diagnosis, we observed somewhat higher risk estimates for individuals with mLOY that could reflect effects of treatment modalities (chemotherapy, surgery and/or radiation therapy). We additionally examined the possible association between mLOY and bladder cancer risk in three case–control studies (total of 2,062 cases and 2,064 controls) but found no association (OR = 1.17, 95% CI = 0.93–1.48; P = 0.18; Supplementary Table 14). Together, these results provide limited support for the hypothesis that mLOY is a strong risk factor for common solid tumors, and larger studies will be needed to investigate further given the current estimated effect sizes.

A recent report suggested that mLOY was associated with all-cause and cancer-related mortality in a cohort of 982 participants who were free from cancer at the study baseline. As cancer mortality reflects both developing cancer and dying from it, we examined whether mLOY might be associated with subsequent overall and cancer-specific mortality in our cancer cases, restricting our analysis to cases with DNA collected at least 1 year before diagnosis and available follow-up (n = 5,340). We observed little evidence for an association with either end point, whether in Kaplan–Meier survival curves (Fig. 3a,b) or in unadjusted or multivariate-adjusted Cox proportional hazard models (Supplementary Table 15). After adjusting for age at diagnosis, smoking status (current smokers and, for former smokers, the number of years since cessation), smoking intensity in pack-years, body mass index (BMI) and contributing study, the hazard ratio for mortality from all causes was 0.89 (95% CI = 0.76–1.04; P = 0.15) and the hazard ratio for mortality from cancer was 0.87 (95% CI = 0.73–1.04; P = 0.12). Similar findings were observed for bladder, lung and prostate cancers separately (Fig. 3c–e and Supplementary Table 15).

We conducted a GWAS to identify regions associated with risk for mLOY in the three cohorts, analyzed separately and in a meta-analysis. We adjusted for smoking status (ever- versus never-smoker) and principal components significantly associated with mLOY in each cohort (P < 0.05; reported previously to account for subtle differences in population substructure in participants of European background). The analysis included 895 men with detected mLOY and 11,474 men with no detected mLOY. The P-value distribution from the combined meta-analysis had an inflation factor (λ) of 1.015, as depicted in the quantile–quantile plot (Supplementary Fig. 3). We observed a significant association with SNP rs2887399 at 14q32.13 (OR = 1.55, 95% CI = 1.36–1.78; P = 3.17 × 10−10) (Fig. 4 and Supplementary Fig. 4). The major risk allele (G) has a frequency of 0.77 in the CEU (European-ancestry) population. The effect estimates were consistent across the three studies, and there was no evidence for heterogeneity (P = 0.86; Supplementary Fig. 5). The relationship remained robust when not adjusting for smoking status (OR = 1.57, 95% CI = 1.36–1.80; P = 6.46 × 10−11). The rs2887399 variant maps to the 5′ end of the TCL1A gene (encoding T cell leukemia/lymphoma 1A), which functions as a co-activator of the cell survival kinase AKT and has been implicated in T cell and B cell hematological malignancies, mainly because recurrent chromosomal rearrangements bring TCL1A in close proximity to the T cell antigen receptor gene. To our knowledge, the rs2887399 variant is the first common variant associated with a clonal expansion phenotype beyond the known association of a JAK2 haplotype with mosaicism for the common JAK2 variant encoding p.Val617Phe (refs. 29–32). This genetic finding linking germline variation to somatic mosaicism could lead to an understanding of how clonal hematopoiesis relates to many chronic diseases; it is plausible that the susceptibility haplotype could contribute to actual loss of the Y chromosome or that it could be permissive for clonal expansion. Further work is needed to fine-map the region and investigate its biological underpinnings of the development of mLOY.

In summary, mLOY is the most common large-scale detectable mosaic chromosomal event in males, and it has a striking association with aging and cigarette smoking, which is attenuated by years after cessation. We observed limited evidence for mLOY as a strong risk factor for three common cancer types in men, and we did not observe an association with survival after cancer diagnosis. Contrary to previous evidence suggesting that men with mLOY have a substantially higher likelihood of dying from cancer, our study provides little evidence for this hypothesis. Together, these data suggest that age and smoking have a substantial effect on the development of mLOY.
but that there is insufficient evidence to conclude that mLOY is a major risk factor for non-hematological cancer in men. Lastly, our GWAS identified a locus at 14q32.13 associated with mLOY, which may provide insight into the biological basis of mLOY in relation to smoking and aging in men.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. A complete list of the mLOY events identified in this study is provided as the Supplementary Data Set.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

This project has been funded in whole or in part with federal funds from the National Cancer Institute, US National Institutes of Health, under contract HHSN261200800001E. The content of this publication does not necessarily reflect the views or policies of the US Department of Health and Human Services nor does mention of trade names, commercial products or organizations imply endorsement by the US government.

AUTHOR CONTRIBUTIONS

M.Y. and S.J.C. conceived the analysis. W.Z., M.J.M., N.D.F., M.D., M.Y. and S.J.C. designed the study. W.Z., M.J.M., N.D.F., N.R., N.C., M.D., M.Y. and S.J.C. interpreted the primary results. W.Z., M.J.M., N.D.F., K.B.J., F.X.R., B.R.-S., L.A.P.-J., M.D. and M.Y. developed the study methods. W.Z., M.J.M., N.D.F., C.D. and K.B.J. analyzed the data. W.Z., M.J.M. and A.H. were responsible for production and analysis of the genotype data. W.Z., M.J.M., N.D.F. and J.S. performed statistical analysis. W.Z., M.J.M., N.D.F., M.D., M.Y. and S.J.C. drafted the manuscript. M.T., R.N.H. and S.J.C. provided vital programmatic and institutional support. N.M., L.T.T., M.M.G., S.M.G., V.L.S., D.A., S.W., J.V., S.B., A.D.S., J.F., M.G.-C., F.X.R., J.E., G.M., B.R.-S., M.K., M.S., X.W., J.G., Y.Y. and L.A.P.-J. contributed data or samples. All authors contributed critical feedback, review and approval of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pierre, R.V. & Haagland, H.C. Age-associated aneuploidy: loss of Y chromosome from human bone marrow cells with aging. Cancer 30, 889–894 (1972).
2. Wong, A.K. et al. Loss of the Y chromosome: an age-related or clonal phenomenon in acute myelogenous leukaemia/myelodysplastic syndrome? Arch. Pathol. Lab. Med. 132, 1329–1332 (2008).
3. Jacobs, P.A., Brunton, M., Court Brown, W.M., Doll, R. & Goldstein, H. Change of Y chromosome loss and aging: in situ hybridization studies on human interphase nuclei. Am. J. Hum. Genet. 57, 1143–1150 (1995).
4. Jacobs, P.A., Koschorz, B., Berntehler, U., Grimm, T. & Schmid, M. Sex chromosome loss and aging: an age-related or clonal phenomenon? Nat. Genet. 197, 1080–1081 (1963).
5.abelovich, D., Yehuda, O., Ben-Neriah, S. & Or, R. Loss of Y chromosome. An age-related event or a cytogenetic marker of a malignant clone? Cancer Genet. Cytogenet. 76, 70–71 (1994).
6. Herens, C. et al. Loss of the Y chromosome in bone marrow cells: results on 1907 consecutive cases of leukaemia and preleukaemia. Clin. Lab. Haematol. 21, 17–20 (1999).
7. Wiktor, A. et al. Clinical significance of Y chromosome loss in hematologic disease. Genes Chromosom. Cancer 27, 11–16 (2000).
8. Wong, A.K. et al. Loss of the Y chromosome: an age-related or clonal phenomenon in acute myelogenous leukaemia/myelodysplastic syndrome? Arch. Pathol. Lab. Med. 132, 1329–1332 (2008).
9. Zhang, L.J., Shin, E.S., Yu, Z.X. & Li, S.B. Molecular genetic evidence of Y chromosome loss in male patients with hematological disorders. Chin. Med. J. (Engl.) 120, 2002–2005 (2007).
10. Chian, S. Cancer biology: genome-wide association studies. in World Cancer Research Report 2014 (eds. Stewart, B. & Wild, C.P.) 193–202 (International Agency for Research on Cancer, 2014).
11. Conrad, D.F. et al. Origins and functional impact of copy number variation in the human genome. Nature 464, 704–712 (2010).
12. Gonzalez, J.R. et al. A fast and accurate method to detect allelic genomic imbalances underlying mosaic reanamorphisms using SNP array data. BMC Bioinformatics 12, 166 (2011).
13. Ihsara, A. et al. Population analysis of large copy number variants and hotspots of human genetic disease. Am. J. Hum. Genet. 84, 148–161 (2009).
14. Jacobs, K.B. et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat. Genet. 44, 651–658 (2012).
15. Laurie, C.C. et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat. Genet. 44, 642–650 (2012).
16. Machiela, M.J. et al. Characterization of large structural genetic mosaicism in human autosomes. Am. J. Hum. Genet. 96, 487–497 (2015).
17. McCarron, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. Nat. Genet. 39 (suppl. 7), S37–S42 (2007).
18. Pfeffer, D.A. et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res. 16, 1136–1148 (2006).
19. Durmanski, J.P. et al. Smoking is associated with mosaic loss of chromosome Y. Science 347, 81–83 (2015).
20. Forsberg, L.A. et al. Mosaic loss of chromosome Y in peripheral blood is associated with shorter survival and higher risk of cancer. Nat. Genet. 46, 624–628 (2014).
21. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. 20, 1472–1478 (2014).
22. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. 371, 2488–2498 (2014).
23. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. 371, 2477–2487 (2014).
24. Fernández, L.C., Torres, M. & Real, F.X. Somatic mosaicism: on the road to cancer. Nat. Rev. Cancer 16, 43–55 (2016).
25. Machiela, M.J. & Chanock, S.J. Detectable clonal mosaicism in the human genome. Semin. Hematol. 50, 348–359 (2013).
26. Skaletsky, H. et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature 423, 825–837 (2003).
27. Laine, J., Künstle, G., Obata, T., Sha, M. & Noguchi, M. The protooncogene TCL1 is an Akt kinase coactivator. Mol. Cell 6, 395–407 (2000).
28. Virgilio, L. et al. Deregulated expression of TCL1 causes T cell leukemia in mice. Proc. Natl. Acad. Sci. USA 95, 3885–3889 (1998).
29. Olcaydu, D. et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. Nat. Genet. 41, 450–454 (2009).
30. Olcaydu, D. et al. The ‘GGCC’ haplotype of JAK2 confers susceptibility to JAK2 exon 12 mutation-positive polycythemia vera. Leukemia 23, 1924–1926 (2009).
31. Jones, A.V. et al. JAK2 haplotype is a major risk factor for the development of myeloproliferative neoplasms. Nat. Genet. 41, 446–449 (2009).
32. Kilpivaara, O. et al. A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. Nat. Genet. 41, 455–459 (2009).
33. Machiela, M.J. & Chanock, S.J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 31, 3555–3557 (2015).
ONLINE METHODS

Study overview. The analyzed data consist of 13,789 males drawn from cancer GWAS conducted in three prospective cohorts. The mean age at DNA collection was 67.26 years for all participants. The studies were approved by the institutional ethics committee of each participating hospital and the Institutional Review Board (IRB) of the National Cancer Institute. Written informed consent was obtained from all individuals.

DNA was extracted from peripheral circulating leukocytes (70.38%) and buccal samples (29.62%) for men drawn from the three cohorts (Supplementary Table 1). Genomic DNA was screened and analyzed at the National Cancer Institute according to the standard sample handling process of the Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics (DCEG). AmpFISTR Identifier assays confirmed that each sample was concordant with reported sex and removed samples with evidence of contamination. Genotyping was carried out on one of four Illumina Infinium SNP arrays (Human Hap610K, Hap1M, Omni1M and Omni2.5M arrays), each of which has an adequate number of probes specific to the Y chromosome. Additional scanning on other microarray chips was not used because of the inadequate probe coverage for the Y chromosome. Of all the participants, 94.99% were detected as having ≥80% European ancestry and 2.39% were detected as having ≥80% African ancestry. Case–control studies for bladder cancer were also included and were drawn from two studies carried out in Spain and the New England Bladder Cancer Study, previously scanned with the Human Hap1M and Hap610K chips34, respectively. The MD Anderson bladder cancer study was initially scanned on a chip with inadequate probe coverage for the Y chromosome35, and we therefore detected mLOY in this study using the qPCR assay described below (Supplementary Table 14).

Analysis of log2-transformed intensity ratio and B allele frequency generation. Sample intensity files (two files per sample, for the red and green channels) were loaded into Illumina GenomeStudio software. The intensity data were normalized using the Illumina five-step self-normalization procedures, which used information contained on the array itself to convert raw x and y (allele A and allele B) signal intensities to normalized values. The LRR and B allele frequency (BAF) values for each assay were exported from GenomeStudio software using the Genotype Final Report (GFR) format.

Test region for detecting chromosome Y abnormality. We extensively examined loci across the Y chromosome for the four commercial chip types with adequate chromosome Y coverage and used the MST from 6,671, 498–22,919,969 (hg18/Build 36) as the test region for detecting chromosome Y mosaicism, as this region provides relatively stable signal intensity and is outside of a majority of the regions containing genes with multiple copies26 (Supplementary Fig. 1a).

Chromosome Y mosaicism detection method. Chromosome Y mosaicism was detected using LRR, which is the normalized measure of total signal intensity and provides data on relative copy number. Subjects were examined for deviations from the expected log2-transformed intensity ratio for evidence of loss of the MSY. A minimum mean threshold of LRR ≤0.15 was used to identify a chromosome Y loss event. A minimum mean threshold of LRR ≥0.15 was used to define a chromosome Y gain event. Samples with mean LRR values falling below or above these thresholds were called as mosaic chromosome Y losses and gains, respectively. To minimize false discovery of chromosome Y mosaicism, the ratio of the mean LRR to the standard deviation for the test region was calculated. A minimum threshold for the ratio was set to 0.25 to filter out samples with excessive noise in their LRR values. For potential mLOY, each chromosome Y plot was manually reviewed and suspect events were further excluded in subsequent analyses.

qPCR validation. qPCR was used to evaluate the ratio of chromosome Y signal to the signal for an autosomal single-copy gene. Fifteen qPCR gene assays spanning the p and q arms of the Y chromosome (Supplementary Fig. 1b) were run in duplex with RPPH1 (RNase P) as the reference gene, known to be single copy26 (Supplementary Table 16).

For each sample, 5 ng of DNA, as determined by Quant-iT PicoGreen dsDNA quantification (Life Technologies), was transferred to a LightCycler-compatible 384-well plate (Roche) and dried down. An internal standard curve (comprising a serial dilution, with seven target ratios, of pooled male genomic DNA samples with no detectable chromosome Y loss, with a pool of female genomic DNA samples) and assay control samples (three target ratios; prepared similarly to the standard curve) were applied to the assay plates to guide analysis and indicate the overall quality of assay performance. All experimental and control samples were assayed in triplicate on each plate.

qPCR was performed using a 5-µl reaction volume consisting of 2.5 µl of LightCycler 480 Probes Master Mix (Roche), 2.0 µl of MBG Water, 0.25 µl of 20× TaqMan Copy Number Reference Assay, human RNase P (Life Technologies) and 0.25 µl of 20× TaqMan Copy Number Assay specific to the marker of interest (Life Technologies). Thermal cycling was performed on a LightCycler 480 (Roche) with PCR conditions consisting of holding at 95 °C for 5 min, denaturation at 95 °C for 15 s and annealing at 60 °C for 30 s, with fluorescence data collection, for 45 cycles.

LightCycler software (release 1.5.0) was used for initial analysis of raw data. Using absolute quantification analysis with the second-derivative maximum method and high-confidence detection algorithm, single target sequences were quantified and values were expressed as a ratio (target/reference) based on the internal standard curve of known ratios. The ratios of the 15 assays were averaged to yield an overall chromosome Y signal ratio.

Estimation of the proportion of cells with chromosome Y loss. A quadratic regression model was used to fit the average qPCR ratio and mean LRR data pairs, with mean LRR as the predictor variable (x) and average qPCR ratio as the response variable (y), to create a predictive polynomial equation \( y = a_0 + a_1 \times x + a_2 \times x^2 \), where \( a_0, a_1 \) and \( a_2 \) represent coefficient estimates. Only the data points from subjects having consensus event calls between qPCR and chip data for chromosome Y loss and normal, with coefficient of variation (CV) ≤10% from qPCR data, were used to generate such relationships (n = 98 subjects). For qPCR, the standard curve used simulates known ratios for the Y chromosome ranging from 10% to 100% (corresponding to 90% and no loss, respectively). This is achieved by diluting a pool of male samples with no mLOY with a pool of female samples to simulate these percentages of loss. Data for chromosome Y gain were not included when building the predictive model, as the qPCR ratio for the estimated amount of gain was not precise because the data were extrapolated outside the experimental copy number range of 0.1 to 1 defined by the standard curve. For each mean LRR, the corresponding copy number ratio can be predicted by inserting the mean LRR into the quadratic equation. The percentage of cells with chromosome Y loss equals 1 minus the average chromosome Y signal ratio (Supplementary Fig. 1c); for example, a mean LRR of 0.15 corresponds to a frequency of chromosome Y loss of 22.7%. We also performed this analysis for the case–control studies and found similar results (data not shown).

Logistic regression analysis. All of the logistic regression models were generated in R using the glm function with quasi-binomial family and logit as a link function. To determine the relationship between individuals with chromosome Y loss and their age at DNA collection, smoking behavior, DNA source, ancestry and study cohort, we fit several models that regressed the presence of chromosome Y loss for each individual on relevant covariates in a logistic model. To determine the relationship between individuals with chromosome Y loss and cancer diagnosis, we fit several models that regressed the occurrence of a cancer diagnosis for each individual on relevant covariates in a logistic model. The following covariate terms were defined for each individual: (i) age of DNA (a continuous measure of age at DNA collection); (ii-a) smoking status (a categorical variable with three levels: current smoker, former smoker and never-smoker (reference group)); (ii-b) number of years since cessation of smoking (a categorical variable with six levels: current smoker, 1–4 years, 5–10 years, 11–20 years and >20 years since quitting smoking, and never-smoker (reference group)); (iii) DNA source (a categorical variable with two levels: individuals who contributed DNA derived from a buccal sample and from a blood sample (reference group)); (iv) East Asian ancestry (a continuous measure of admixture estimation); (v) African ancestry (a continuous measure of...
admixture estimation); and (vi) study (a categorical variable with three levels: ATBC, PLCO and CPS-II (reference group)).

Survival analysis. Cox proportional hazard models were generated in R using the Survival package. Using age as the time scale, the start year was the year of cancer diagnosis and the end year was the year of death or censorship. For total survival analysis, the event indicator equals 1 for individuals who died during the study follow-up and 0 for those who did not. For cancer survival analysis, the event indicator equals 1 for individuals with a cancer-related cause of death during the period of the study and 0 for those who were alive at the end of the study or had a non-cancer-related cause of death. We observed similar associations when we started the follow-up time at the year of DNA collection. All case subjects used in the analysis had DNA collected at least 1 year before a cancer diagnosis. The following covariate terms were defined for each individual: (i) age of diagnosis (a continuous measure of age at diagnosis); (ii) BMI (a continuous measure of body mass index); (iii) pack-years (a categorical variable with five levels: >60, 40–60, 20–40 and <20 pack-years, and never-smoker (reference group)); and (iv) mLOY (a categorical variable with two levels: 1 for individuals identified as having chromosome Y loss and 0 for individuals without chromosome Y loss. All other covariates were defined as in the logistic regression models.

GWAS analysis. A GWAS was conducted adjusting for smoking status (ever- versus never-smoker) and principal components that were significantly associated with mLOY in each cohort. The combined meta-analysis of men from the three cohorts consisted of 928 men with detectable mLOY and 12,118 men without evidence of mLOY (Supplementary Fig. 5). To ensure robust association, a further analysis was carried out that did not adjust for smoking status.

Other analyses. Frequency plots were generated using R. Confidence intervals on frequencies are reported using 95% confidence bounds from the Jeffery interval method and were generated in R using the binom package. The 95% confidence interval unadjusted analysis of count data and frequencies was performed using Fisher’s exact test for contingency tables, as implemented in the R software package. The GLU software package was used to estimate admixture coefficients for each subject.

34. Rothman, N. et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nat. Genet. 42, 978–984 (2010).
35. Wu, X. et al. Genetic variation in the prostate stem cell antigen gene PSCA confers susceptibility to urinary bladder cancer. Nat. Genet. 41, 991–995 (2009).