

ARTICLE

Received 26 Sep 2015 | Accepted 6 May 2016 | Published 13 Jun 2016

DOI: 10.1038/ncomms11843

OPEN

# Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome

Mitchell J. Machiela *et al.*<sup>#</sup>

To investigate large structural clonal mosaicism of chromosome X, we analysed the SNP microarray intensity data of 38,303 women from cancer genome-wide association studies (20,878 cases and 17,425 controls) and detected 124 mosaic X events > 2 Mb in 97 (0.25%) women. Here we show rates for X-chromosome mosaicism are four times higher than mean autosomal rates; X mosaic events more often include the entire chromosome and participants with X events more likely harbour autosomal mosaic events. X mosaicism frequency increases with age (0.11% in 50-year olds; 0.45% in 75-year olds), as reported for Y and autosomes. Methylation array analyses of 33 women with X mosaicism indicate events preferentially involve the inactive X chromosome. Our results provide further evidence that the sex chromosomes undergo mosaic events more frequently than autosomes, which could have implications for understanding the underlying mechanisms of mosaic events and their possible contribution to risk for chronic diseases.

---

Correspondence and requests for materials should be addressed to S.J.C. (email: chanocks@mail.nih.gov)

<sup>#</sup>A full list of authors and their affiliations appears at the end of the paper.

Genetic mosaicism is classically defined as the coexistence of clonal cellular populations harbouring two or more distinct genotypes<sup>1</sup>. To date, detectable mosaicism has been reported in apparently healthy individuals as well as in patients with rare diseases, such as neurofibromatosis type II (NF2), trisomy 21, naevus sebaceous and Proteus syndrome<sup>2–9</sup>. Emerging data from consortia of genome-wide association studies (GWAS)<sup>3,5,6,10–12</sup> have demonstrated large autosomal mosaicism (events >2 Mb in size) in DNA collected from peripheral leukocytes and buccal epithelium. These studies suggest that autosomal mosaicism is associated with aging, hematologic cancer risk, and possibly ancestry and male sex. Whereas autosomal mosaicism is detectable in <2% of older individuals, recent studies indicate that large mosaic events may be far more common for the Y chromosome, and in particular among older men who smoke cigarettes<sup>13–15</sup>.

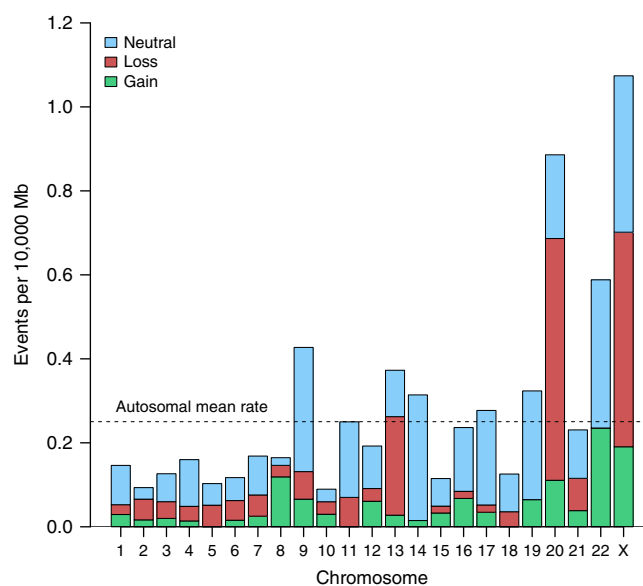
The functional consequences of detectable chromosomal mosaicism remain to be fully determined. A number of groups have reported detectable genetic mosaicism of single-nucleotide mutations in the general population, particularly in genes implicated in hematopoietic disorders such as leukaemias and lymphomas<sup>2,4,16</sup>. Point-mutation events could reflect early, preleukemic clones and separately could increase risk for cardiovascular events<sup>4</sup>. Moreover, many reports have shown phenotypic consequences of chromosomal mosaicism that vary by genomic location of the event, developmental timing, tissue type involved and percentage of cells affected<sup>7–9</sup>. In prospective cohort studies, it has been possible to detect large mosaic structural events in blood samples of individuals who eventually develop chronic leukaemia, as early as 14 years before diagnosis, suggesting detection of a subset of events that eventually become manifest as part of the molecular profile of leukaemia<sup>3,5,17</sup>.

To date, reports have not systematically addressed the frequency and characteristics of X chromosomal mosaicism. The X chromosome is unique among the human chromosomes in that normal women carry two copies and normal men carry one. To compensate for dosage differences between sexes, one copy of the female X chromosome is rendered transcriptionally inactive in a process called X inactivation<sup>18</sup>. In humans, the inactive X-chromosome (Xi) is randomly chosen early in development. Once established, X inactivation is generally irreversible and stably maintained through mitotic divisions. Established mechanisms for maintaining X inactivation include expression of the non-coding *XIST* RNA, chromatin modifications, changes in nuclear scaffold proteins, and DNA methylation<sup>19–23</sup>. Sequence data from cancer genomes suggest that the X chromosome, particularly the female Xi, has a higher somatic mutation load of point mutations than the autosomes<sup>24</sup>. It has been postulated that the observed higher load of somatic point mutations could be directly related to the timing of Xi replication, which occurs late and is faster than either the active X-chromosome (Xa) or the autosomes<sup>25–27</sup>. Although these and other data suggest that X-chromosome mosaicism may be detectable at a prevalence higher than that observed on the autosomes<sup>28–30</sup>, little is known about its frequency in the population or basic characteristics of the distribution and types of gains, losses and acquired loss of heterozygosity.

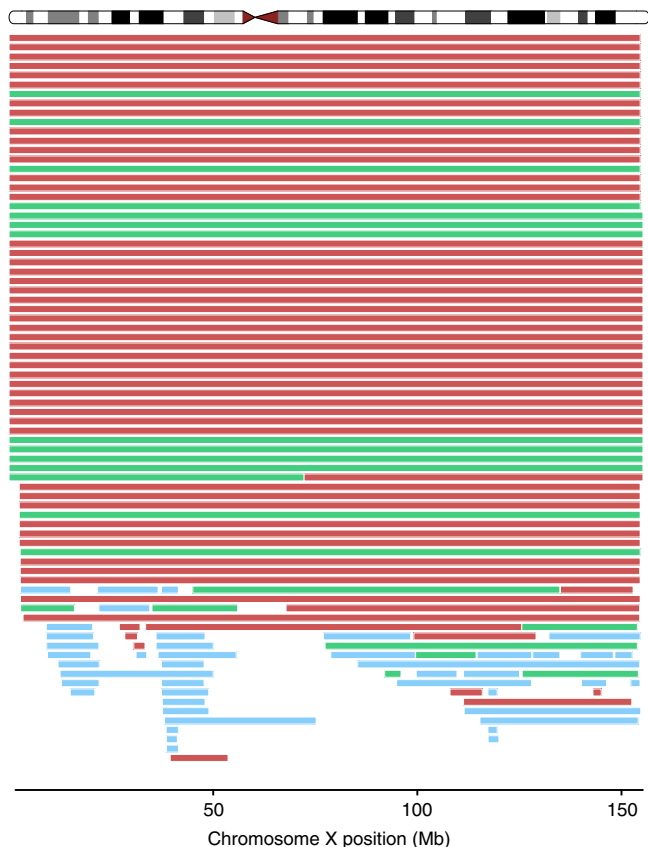
In this report, we investigate the frequency of large-scale chromosome X mosaicism (>2 Mb) in blood or buccal samples from 38,303 women. We observe an overall frequency of X mosaicism of ~0.25%, roughly four times the mean autosomal rate. The frequency of X mosaicism increases with increasing age, but is not associated with non-haematologic cancer risk. Further investigations by methylation analyses suggest the inactive X chromosome is preferentially gained or lost in X mosaic events.

## Results

**Detected chromosome X events.** Using a segmentation algorithm, we conducted a systematic scan of large structural detectable mosaicism on the X chromosomes of 38,303 women (20,878 cancer cases and 17,425 cancer-free controls), who had been previously examined for autosomal mosaicism<sup>3,11,12</sup>. In total, 124 mosaic events greater than 2 Mb in size were detected on the X chromosomes of 97 of the 38,303 women who were scanned (0.25%, Supplementary Table 1, Supplementary Table 2); all detected cases of trisomy X and XO (Turner's syndrome) were removed from subsequent analyses ( $n=5$ ). Of the 97 women with detected X events, 15 (15%) had more than one event detected on their X chromosome, with one woman having as many as five events. The base-pair adjusted rate of mosaic X events was 1.07 events per 10,000 Mb, over fourfold higher than the mean 0.25 events per 10,000 Mb rate observed across the autosomes<sup>12</sup> ( $P$  value =  $1.32 \times 10^{-5}$ , Fig. 1). Significantly elevated rates were observed for the X chromosome in comparison with all autosomes except for chromosome 20 (chr20 = 0.89, chrX = 1.07 events per 10,000 MB;  $P$  value = 0.29). The 124 mosaic X events consisted of 59 mosaic losses, 43 mosaic copy-neutral events and 22 mosaic gains (Fig. 2, Supplementary Fig. 1). These events mostly included the whole chromosome, with a fraction (37%) mapping to the interstitial region (Table 1). Few events were found at either the centromeric or telomeric ends. Most whole-X-chromosome events were mosaic losses. Interstitial events were primarily mosaic copy-neutral loss of heterozygosity, which have been less extensively documented in the cytogenetic literature on chromosome X (Supplementary Table 3). Two notable clusters of interstitial mosaic copy-neutral events are centered at approximately 26 and 49 Mb (NCBI36/hg18, Fig. 2). While X-chromosome mosaic events were more common than autosomal events, the mean proportion of cells with X-chromosome mosaicism tended to be lower than the mosaic proportion with autosomal events overall ( $X=0.299$ , autosomes = 0.359,  $P$  value = 0.01, Supplementary Fig. 2), however, this association was not observed in cancer-free individuals ( $P$  value = 0.10). Women with an X-chromosome mosaic event had a significantly higher likelihood of harbouring an autosomal event relative to women without detectable X



**Figure 1 | Adjusted mean rate of events by chromosome.** A comparison of detected mosaic events >2 Mb in size in the autosomes to the X chromosome (X events = 124, Autosomal events = 430).

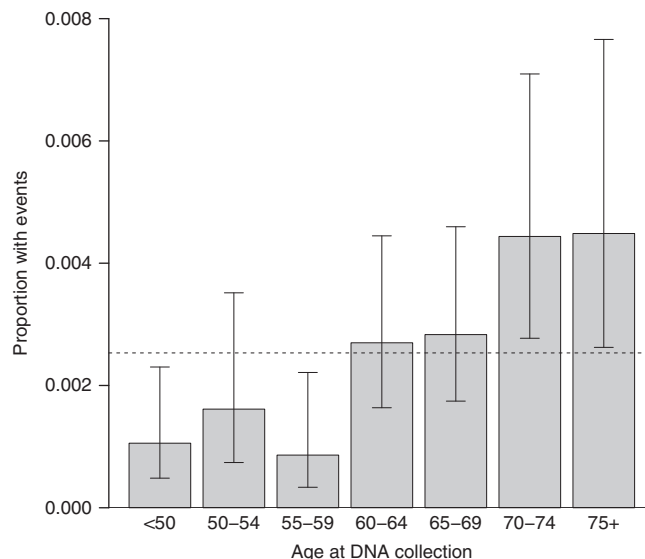


**Figure 2 | Detected mosaic events on the X chromosome.** Mosaic losses ( $N=59$ ) are in red, mosaic gains ( $N=22$ ) are green, and mosaic copy-neutral events ( $N=43$ ) are in blue.

mosaicism (unadjusted odds ratio ( $OR_{unadj}$ ) = 16.7, 95% confidence interval (CI) = 8.3–33.6,  $P$  value =  $2.5 \times 10^{-15}$ ), even after adjusting for age (adjusted odds ratio ( $OR_{adj}$ ) = 15.6, 95% CI = 7.3–33.0,  $P$  value =  $8.6 \times 10^{-13}$ ).

**Validation by qPCR.** Detected X mosaic events were experimentally validated using a set of 12 quantitative PCR assays (qPCR) across chromosome X. Specifically, we estimated copy-number ratios for 26 events across 25 females with single-nucleotide polymorphism (SNP) microarray-detected X mosaicism with a range of mosaic proportions from 6 to 88%. In the 18 mosaic samples with events that spanned the entire X chromosome, the concordance rate was 100% for gains and 80% for losses (Supplementary Table 5). An inspection of the discordant copy-loss samples called as copy-neutral events revealed qPCR copy-number values near the calling threshold, or samples with low mosaic proportions. For detected mosaic events spanning only a portion of the X chromosome, four of the eight (50%) showed evidence for mosaic copy-number changes by qPCR, although only 25% were concordant in copy-number state with qPCR (Supplementary Table 5), suggesting the limited subsets of qPCR probes that spanned events may have been insufficient to adequately call copy-number states.

**X mosaicism in men.** We also examined X-chromosome mosaicism in men. Although we identified 187 men with suggestive evidence of X-chromosome mosaicism (from 43,735 scanned participants), results from qPCR validation in 39 men with available DNA were poor (15% concordance). Calling X-chromosome mosaicism is inherently more challenging in men

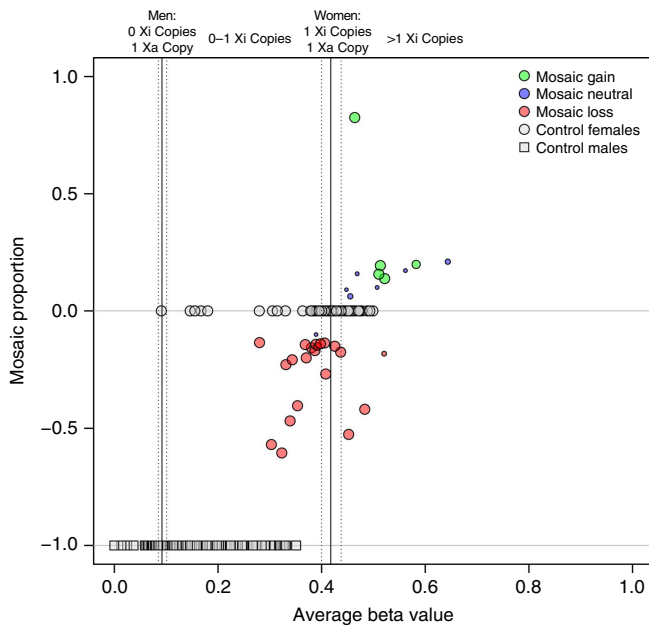


**Figure 3 | Unadjusted age relationship with X mosaicism.** Dashed line represents the mean overall proportion with mosaic X events across all age groups and error bars represent 95% Wilson confidence intervals ( $N=31,982$ ).

as their possession of a single X-chromosome precludes analysis with the B-allele frequency (BAF). Although certainly of interest, further refinement of the calling algorithm is required before we can reliably call detectable X mosaicism in men. All subsequent analyses of X mosaicism reported herein are restricted to women.

**X mosaicism associations.** Detectable X mosaicism increases with age, with more events in older women than in younger women. The estimated frequency of X mosaicism was 0.11% in women under 50 years of age and 0.45% in women 75 years or older (Fig. 3). Multivariate analyses adjusted for ancestry, cancer status and study found a statistically significant association with an OR of 1.04 per 1-year increase in age (95% CI = 1.01–1.06,  $P$  value = 0.005), with a 20-year increase in age resulting in over twice the odds of acquiring a mosaic event on the X chromosome. Altogether with prior evidence from autosomes and the Y chromosome<sup>12,13</sup>, our data suggest that each human chromosome is susceptible to age-related structural deterioration related to clonal mosaicism, but at distinct rates. Y mosaic events are more common than X events, and X events are more common than those in autosomes. These frequencies may reflect intrinsic differences in the mechanisms by which each type of chromosome is replicated or protected against age-related DNA damage<sup>26</sup>.

Comparable to what we reported for the autosomes (in over 127,000 individuals scanned), we found little to no evidence for an overall association between X mosaicism and non-haematologic cancer ( $P$  value = 0.19)<sup>3,5,12</sup>. An analysis by cancer site found at most a marginally significant association between X mosaicism and lung cancer risk (OR = 1.89, 95% CI = 1.02–3.50,  $P$  value = 0.042; 26 lung cancer cases with mosaicism). However, we had only a limited sample size, were unable to adequately adjust for the major lung cancer risk factor, cigarette smoking and we did not consider multiple comparisons across cancer types. We did not detect an association between X mosaicism and ancestry (three continental populations: European, African and East Asian,  $P$  value = 0.40) that was detected in prior autosomal mosaicism analyses<sup>12</sup>. In addition, we did not find evidence for an association between X mosaicism and smoking for a subset of women with available smoking



**Figure 4 | Chromosome X methylation beta values by estimated mosaic proportion.** Average beta values (range: 0.0–1.0) indicate amount of methylation at a genomic locus where low values indicate hypomethylation and high values indicate hypermethylation. X methylation beta values are plotted for promoter probes spanning a mosaic X event and points are sized for the number of probes. The estimated mosaic proportion (positive versus negative) is determined from average probe beta z-scores. Control men ( $N=1,665$ ) and women ( $N=136$ ) are shown as light grey squares and circles. Mosaic females ( $N=48$ ) are plotted as green, blue and red circles for mosaic gains, copy-neutral events and losses, respectively. Solid and dashed black lines are median and interquartile range for control men (left) and women (right) beta values.

information ( $N=19,197$ , ever smoker versus never smoker  $P$  value = 0.54). Interestingly, an association was found between DNA source and X mosaicism in which X mosaicism was more frequent in buccal cells as compared with leukocytes (OR = 3.50, 95% CI = 1.45–8.46,  $P$  value = 0.005). Larger studies are required to confirm these findings.

**Methylation analysis.** To investigate the molecular basis of X-chromosome mosaicism, we used Illumina Human-Methylation450 microarray data for a subset of mosaic females with sufficient DNA to determine whether mosaic events are preferential for either the Xa or Xi. Established sex-specific differences in chromosome X promoter methylation<sup>31,32</sup> provide an opportunity to determine whether the pattern of large structural mosaic events parallels what has previously been reported for analyses of somatic mutations in cancer, namely, events more likely occurring in the inactive X-chromosome<sup>24</sup>. After we completed a rigorous quality control process for methylation microarray data in a control population of 1,665 men and 136 women, probes in gene promoter sites on the X chromosome were extracted and filtered to focus analyses on a reference set of probes that were differentially methylated between men and women, as these are the locations that are inactivated on Xi (Supplementary Fig. 3)<sup>31,32</sup>. Methylation beta values for the resulting set of 1,888 probes were evaluated for differences from normal expected values in women (beta values greater than expected suggest mosaic gain of Xi and less than expected suggest mosaic loss of Xi) (Fig. 4). Of the 21 women

**Table 1 | Chromosomal arm location of detected mosaic autosomal and X events.**

	Autosomes	X chromosome
Interstitial	148 (34.4%)	46 (37.1%)
Spans centromere	17 (4.0%)	2 (1.6%)
Telomeric p	95 (22.1%)	4 (3.2%)
Telomeric q	148 (34.4%)	12 (9.7%)
Whole	22 (5.1%)	60 (48.4%)
	430 (100%)	124 (100%)

with mosaic losses, 16 had evidence for a loss of the Xi chromosome. Similarly, all 5 women with mosaic gains had evidence suggesting a mosaic gain of Xi. For mosaic copy-neutral events, 6 women showed evidence for a loss of a portion of the Xa and a replacement with Xi and one woman showed evidence for a loss of a portion of Xi and a replacement with the Xa. Our combined data for mosaic gains and losses suggest that Xi is preferentially involved in mosaic copy-number changes, with Xi more commonly altered in mosaic losses and preferentially gained for mosaic gains ( $P$  value = 0.002). Mosaic events on the X chromosome that do not follow this trend, particularly the five mosaic losses with evidence for a loss of the Xa, could represent normal variation, perhaps due to different DNA extraction techniques, noise in the methylation assay or statistical outliers. Alternatively, chromosome X events could occur early in female development, perhaps at a time that precedes X-inactivation, and thus X-inactivation could only occur in cells with more than one X chromosome.

## Discussion

Our analysis using SNP microarray intensities identified detectable mosaic events on the female X chromosome that occur at higher frequencies than mosaic events on the autosomes. We observed evidence that individual women with mosaic events of the X chromosome are also more likely to have mosaic events of the autosomes. Furthermore, X mosaic events are more likely to involve the inactive X chromosome than the active X chromosome, and thus might be phenotypically neutral. As with autosomal and Y mosaicism, X mosaicism increases with age.

For decades, it has been apparent that an appreciable fraction of paediatric developmental disorders are directly attributable to a spectrum of mosaic events (for example, from point mutations to large structural alterations) that can also influence clinical course<sup>9,33–35</sup>. Our data indicate that substantial numbers of adults also possess mosaic chromosomes in blood and buccal cells, suggesting the genome undergoes somatic alterations that either are generated later due to less efficient protective mechanisms or were perhaps tolerated from early age and subsequently expanded due to less efficient mechanisms for retaining genomic stability.

A limitation of our analysis is the low level of validation for partial chromosome copy-neutral events. Because of both the smaller event size and the need for log R ratio (LRR) baseline correction, our array-based detection algorithm together with qPCR-based validation yielded a low level of concordance. Further work is needed to improve the calling algorithms, which could also be accelerated by the analysis of larger samples sizes, ultimately leading to more precise measurement of mosaic X-chromosomal events.

It is striking that the frequency of large megabase mosaicism is higher in the inactive X as well as the Y chromosome compared with the autosomes. This higher frequency of mosaicism on sex chromosomes could be a reflection of less cell selection because

the inactive X is transcriptionally inactive while the Y chromosome has the smallest number of genes. Future studies are needed to understand the mechanisms responsible for the generation and selection of these mosaic alterations in sex and autosomal chromosomes, which occur at different frequencies. In turn, insights into the underlying mechanisms as well as the differences in frequencies of large structural genetic mosaicism should provide an important foundation for understanding their contribution to health and chronic diseases<sup>6,36,37</sup>.

## Methods

**Study population.** The data set was drawn from cancer GWAS of solid tumours performed at the National Cancer Institute Division of Cancer Epidemiology and Genetics and the Cancer Genome Research Laboratory. In total, peripheral leukocyte or buccal epithelial DNA was available for 20,878 solid tumour cancer cases and 17,425 cancer-free controls. DNA was genotyped on one or more commercially available Illumina Infinium Human SNP array (Hap300, Hap240, Hap550, Hap610, Hap660, Hap 1, Omni Express, Omni 1, Omni 2.5 and Omni 5). Quality control procedures were applied after genotyping and samples were clustered in batches to optimize accuracy and minimize batch effects. All GWAS studies were reviewed by the Institutional Review Board of the National Cancer Institute and those of the participating study centers. Informed consent was received for each study participant before study enrollment.

**Detection algorithm.** BAF and LRR are two metrics used to detect mosaic events. BAF is a measure of allelic imbalance and used to quantify deviation of an individual's SNP genotype from expected AA, AB and BB genotype clusters. Contiguous runs of heterozygous SNPs with BAF values that deviate from the expected value of 0.5 are evidence for mosaicism. The LRR value of an individual's SNP is a proxy for copy number. LRR values are the  $\log_2$  of the ratio of observed SNP intensity value to expected intensity value. LRR values greater than expected baseline LRR suggest copy gain and less than expected baseline LRR suggest copy loss. The expected baseline LRR was calculated from women within each clustering group based on the ratio of males and females in the original genotyping cluster group. All BAF and LRR values were calculated using methods described<sup>38</sup> and renormalized as outlined previously<sup>3</sup>.

For female participants, BAF and LRR values were systematically scanned across the X chromosome. Chromosomes were segmented for mosaic events using circular binary segmentation (CBS) on BAF values with the BAF segmentation package<sup>39</sup>. Segments <2 Mb in size were filtered out to control the false-positive rate. Gaussian mixture models were fit to BAF bands to assign event type given the best-fitting model (2–4 Gaussian components). Event copy-number state was assigned based on LRR values with baselines adjusted for the number of men present within original genotyping cluster groups. For whole-chromosome mosaic X events, LRR deviations of 0.01 and  $-0.01$  were used to classify events as gain and losses, respectively. For mosaic X events encompassing only a portion of the X chromosome, we chose a more conservative threshold of 0.05 and  $-0.05$  for gains and losses due to greater LRR variation due to the reduced number of X probes that spanned the events. Mosaic proportions were estimated using deviation from the expected BAF given the LRR defined copy-number state. Further details are outlined in our prior work on autosomal mosaicism<sup>3</sup>.

**Quantitative PCR.** qPCR assays were selected to determine copy-number status of 12 regions spanning the X chromosome by normalizing to an autosomal gene, RNase P, which is present in two copies in a diploid genome (Supplementary Table 4). One additional assay was run to validate the presence of the Y chromosome. According to Quant-iT PicoGreen dsDNA quantitation (Life Technologies, Grand Island, NY), 5 ng of sample DNA were transferred into LightCycler-compatible 384-well plates (Roche, Indianapolis, IN) in triplicate and dried down. Two internal standard curves were run separately in each plate, pooled gDNA samples of males and pooled gDNA samples of females, both with no detectable X chromosome loss/gain, and serially diluted to 6 concentrations. qPCR was performed using 5  $\mu$ l reaction volumes consisting of: 2.5  $\mu$ l of LightCycler 480 Probes Master Mix (Roche, Indianapolis, IN), 2.0  $\mu$ l of MBG Water, 0.25  $\mu$ l of 20  $\times$  TaqMan Copy Number Reference Assay, RNase P (Life Technologies, Grand Island, NY), and 0.25  $\mu$ l of specific 20  $\times$  TaqMan Copy Number Assay (Life Technologies, Grand Island, NY). Thermal cycling was performed on a LightCycler 480 (Roche) where PCR conditions consisted of: 95 °C hold for 5 min, denature at 95 °C for 15 s, anneal at 60 °C for 30 s, with fluorescence data collection over 45 cycles. All experimental and control samples were assayed in triplicate on each plate, separately for all 12 individual target assays.

The LightCycler software (Release 1.5.0) was used for initial analysis of the raw data, utilizing the absolute quantification analysis with the second derivative maximum method and high-confidence detection algorithm, to yield a crossing threshold (Ct) for all replicates. The Ct for each assay was used to interpolate concentration of target and reference sequences using the standard curves. The ratio of target to reference was multiplied by 2 to determine the diploid amount of

X chromosome in that region. The ratios of the 12 assays were then averaged to yield an overall X-chromosome signal ratio. Seventy-five normal copy-number controls were used to estimate normal probe ratio means and s.d. A value of 3 s.d. above the normal mean ratio was used as the threshold to call gains and a value of 3 s.d. below the normal mean ratio was the threshold for calling losses.

**Methylation arrays.** After Quant-iT PicoGreen dsDNA quantitation (Life Technologies, Grand Island, NY), 1,000 ng of sample DNA were treated with sodium bisulfite using the EZ-96 DNA Methylation MagPrep Kit (Zymo Research, Irvine, CA) to convert unmethylated cytosine residues to uracils (detected as thymidines), leaving 5'-methylcytosines residues unaffected. Bisulfite-treated samples were denatured, neutralized and then whole-genome amplified, isothermally, to increase the amount of DNA template. The amplified product was enzymatically fragmented, precipitated and resuspended in hybridization buffer. Samples were hybridized overnight on Infinium HumanMethylation450 BeadChips (Illumina Inc., San Diego, CA), which allowed fragmented DNA to anneal to locus-specific 50mers (covalently linked to one of over 500,000 bead types). Single-base extension of oligonucleotides on the BeadChip, using the captured DNA as template, incorporated tagged nucleotides on the BeadChip, which were subsequently fluorophore labelled during staining. BeadChips were scanned by an Illumina iScan at two wavelengths to create image and intensity files. An internal control, a DNA sample from a lymphoblastoid cell line NA07057 (Coriell Cell Repositories, Camden, NJ), was utilized to confirm the efficiency of bisulfite conversion and subsequent methylation analysis.

Methylation beta values are indicators of site-specific methylation with a theoretical range from 0 to 1, where low values indicate hypomethylation and high values indicate hypermethylation. Raw beta intensity values were extracted for probes in promoter sites on the X chromosome and further filtered to include only probes that are differentially methylated between women (Xa/Xi) and men (Xa). A control sample of available men ( $N=1,665$ ) and women ( $N=136$ ) was used to determine expected beta value means and s.d. Using the RnBeads R library, promoter probes were selected that had mean beta values between 0.35 and 0.5 and s.d. <0.09 in women and mean beta values <0.15 and s.d. <0.05 in men (Supplementary Fig. 3). This left a total of 1,888 differentially methylated probes that spanned 212 promoter sites across the X chromosome for analysis. For each mosaic female, mean beta values and z-scores were calculated for all differentially methylated promoter probes that spanned detected mosaic X events in an effort to determine changes in methylation profiles and thus phase mosaic events to the Xa or Xi chromosomes. The mosaic proportions were calculated from SNP microarray per cent mosaicism values. Only X events spanning 5 or more promoter regions were used for the analysis.

**Statistical analysis.** All statistical analyses were performed on a 64 bit Windows build of R 3.0.1 "Good Sport". Multivariate analyses used logistic regression models (glm procedure) with X mosaicism as the dependent variable and adjusted for age of DNA collection, study indicator variables, cancer status (case = 1, control = 0), and genetically inferred ancestry (%European, %African and %Asian) unless otherwise specified. Inferred ancestry proportions were estimated for each individual using reference populations from the HapMap project<sup>40</sup> with the GLU software package (<https://code.google.com/p/glu-genetics/>) using the struct.admix module. Confidence intervals for plots are Wilson intervals. All reported *P* values are two-sided.

**Data availability.** Original study data has been posted in dbGaP (<http://www.ncbi.nlm.nih.gov/gap>) under accession numbers phs000093.v2.p2, phs000336.v1.p1, phs000351.v1.p1, phs000361.v1.p1, phs000652.v1.p1, phs000716.v1.p1, phs000734.v1.p1, phs000396.v1.p1, phs000147.v2.p1, phs000346.v2.p1, phs000863.v1.p1 and phs000206.v5.p3. Data on called event features, location and individual characteristics are available in Supplementary Table 1. Methylation array beta values for events are presented in Supplementary Table 6 and raw data is posted in dbGaP under accession number phs001112.v1.p1. The methylation data has been deposited in dbGaP under accession code phs001112.v1.p1

## References

- Strachan, T., Read, A. P. & Strachan, T. *Human Molecular Genetics* xxv 781 (Garland Science, 2011).
- Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
- Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
- Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
- Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
- Bonnefond, A. *et al.* Association between large detectable clonal mosaicism and type 2 diabetes with vascular complications. *Nat. Genet.* **45**, 1040–1043 (2013).
- Youssoufian, H. & Pyeritz, R. E. Mechanisms and consequences of somatic mosaicism in humans. *Nat. Rev. Genet.* **3**, 748–758 (2002).

8. Machiela, M. J. & Chanock, S. J. Detectable clonal mosaicism in the human genome. *Semin. Hematol.* **50**, 348–359 (2013).
9. Biesecker, L. G. & Spinner, N. B. A genomic view of mosaicism and human disease. *Nat. Rev. Genet.* **14**, 307–320 (2013).
10. Forsberg, L. A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).
11. Rodriguez-Santiago, B. *et al.* Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am. J. Hum. Genet.* **87**, 129–138 (2010).
12. Machiela, M. J. *et al.* Characterization of large structural genetic mosaicism in human autosomes. *Am. J. Hum. Genet.* **96**, 487–497 (2015).
13. 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).
14. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2014).
15. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near *TCL1A*. *Nat. Genet.* (2016).
16. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
17. Schick, U. M. *et al.* Confirmation of the reported association of clonal chromosomal mosaicism with an increased risk of incident hematologic cancer. *PLoS ONE* **8**, e59823 (2013).
18. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
19. Lee, J. T. Gracefully ageing at 50, X-chromosome inactivation becomes a paradigm for RNA and chromatin control. *Nat. Rev. Mol. Cell Biol.* **12**, 815–826 (2011).
20. Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Mol. Cell* **5**, 695–705 (2000).
21. Wolf, S. F., Jolly, D. J., Lunnen, K. D., Friedmann, T. & Migeon, B. R. Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. *Proc. Natl Acad. Sci. USA* **81**, 2806–2810 (1984).
22. Gendrel, A. V. & Heard, E. Fifty years of X-inactivation research. *Development* **138**, 5049–5055 (2011).
23. Augui, S., Nora, E. P. & Heard, E. Regulation of X-chromosome inactivation by the X-inactivation centre. *Nat. Rev. Genet.* **12**, 429–442 (2011).
24. Jager, N. *et al.* Hypermutation of the inactive X chromosome is a frequent event in cancer. *Cell* **155**, 567–581 (2013).
25. Koren, A. & McCarroll, S. A. Random replication of the inactive X chromosome. *Genome Res.* **24**, 64–69 (2014).
26. Koren, A. *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am. J. Hum. Genet.* **91**, 1033–1040 (2012).
27. Catalan, J., Falck, G. C. & Norppa, H. The X chromosome frequently lags behind in female lymphocyte anaphase. *Am. J. Hum. Genet.* **66**, 687–691 (2000).
28. Razzaghi, H. R. *et al.* Somatic mosaicism for chromosome X and Y aneuploidies in monozygotic twins heterozygous for sickle cell disease mutation. *Am. J. Med. Genet. A* **152A**, 2595–2598 (2010).
29. Russell, L. M., Strike, P., Browne, C. E. & Jacobs, P. A. X chromosome loss and ageing. *Cytogenet. Genome Res.* **116**, 181–185 (2007).
30. Zankl, H., Seidel, H. & Zang, K. D. Cytological and cytogenetical studies on brain tumors. V. Preferential loss of sex chromosomes in human meningiomas. *Humangenetik* **27**, 119–128 (1975).
31. Joo, J. E. *et al.* Human active X-specific DNA methylation events showing stability across time and tissues. *Eur. J. Hum. Genet.* **22**, 1376–1381 (2014).
32. Cotton, A. M. *et al.* Chromosome-wide DNA methylation analysis predicts human tissue-specific X inactivation. *Hum. Genet.* **130**, 187–201 (2011).
33. Sybert, V. P. & McCauley, E. Turner's syndrome. *N. Engl. J. Med.* **351**, 1227–1238 (2004).
34. Papavassiliou, P. *et al.* The phenotype of persons having mosaicism for trisomy 21/Down syndrome reflects the percentage of trisomic cells present in different tissues. *Am. J. Med. Genet. A* **149A**, 573–583 (2009).
35. Messiaen, L. *et al.* Mosaic type-1 NF1 microdeletions as a cause of both generalized and segmental neurofibromatosis type-1 (NF1). *Hum Mutat.* **32**, 213–219 (2011).
36. Macosko, E. Z. & McCarroll, S. A. Exploring the variation within. *Nat. Genet.* **44**, 614–616 (2012).
37. Abkowitz, J. L. Clone wars—the emergence of neoplastic blood-cell clones with aging. *N. Engl. J. Med.* **371**, 2523–2525 (2014).
38. Wang, K. & Bucan, M. Copy number variation detection via high-density SNP genotyping. *CSH Protoc.* **2008**, pdb top46 (2008).
39. Staaf, J. *et al.* Segmentation-based detection of allelic imbalance and loss-of-heterozygosity in cancer cells using whole genome SNP arrays. *Genome Biol.* **9**, R136 (2008).
40. International HapMap, C. The International HapMap Project. *Nature* **426**, 789–796 (2003).

## Acknowledgements

Some individuals, studies, and centers received individual support as follows: Broad Center for Genotyping and Analysis (U01HG04424); Cancer Prevention Study-II (American Cancer Society); Center for Inherited Disease Research (U01HG004438, HHSN268200782096C); Endometrial cancer (R01 CA 134958); Fred Hutchinson Cancer Research Center (funds from the Fred Hutchinson Cancer Research Center and National Institute of Health grants (R35 CA 39779, RO1 CA 75977, RO3 CA 80636, N01 HD 2 3166, K05 CA 92002, CA 105212 and R01 CA87538)); Fudan Lung Cancer Study (Ministry of Health (201002007); Ministry of Science and Technology (2011BAI09B00); National S&T Major Special Project (2011ZX09102-010-01); China National High-Tech Research and Development Program (2012AA02A517, 2012AA02A518); National Science Foundation of China (30890034); National Basic Research Program (2012CB944600); Scientific and Technological Support Plans from Jiangsu Province (BE2010715); Gene-Environment Association Studies (Coordinating Center:U01 HG004446, Manuscript preparation: P01-GM099568); Genes and Environment in Lung Cancer, Singapore Study (National Medical Research Council Singapore grant (NMRC/0897/2004, NMRC/1075/2006); Agency for Science, Technology and Research (A\*STAR) of Singapore); Genetic Epidemiological Study of Lung Adenocarcinoma (National Research Program on Genomic Medicine in Taiwan (DOH98-TD-G-111-015); National Research Program for Biopharmaceuticals in Taiwan (DOH 100-TD-PB-111-TM013); National Science Council, Taiwan (NSC 100-2319-B-400-001)); Guangdong Study (Funding of Guangdong Science and Technology Department (2006B60101010, 2007A032000002, 2011A030400010); Guangzhou Science and Information Technology Bureau (2011Y2-00014); Chinese Lung Cancer Research Foundation; National Natural Science Foundation of China (81101549); Natural Science Foundation of Guangdong Province (S2011010000792)); Hong Kong Study (General Research Fund of Research Grant Council, Hong Kong (781511M)); Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH; Intramural Research Program of the NIH, National Library of Medicine; Intramural Research Program of the National Institute for Occupational Safety and Health; Japanese Female Lung Cancer Collaborative Study (Grants-in-Aid from the Ministry of Health, Labor, and Welfare for Research on Applying Health Technology and for the 3rd-term Comprehensive 10-year Strategy for Cancer Control; National Cancer Center Research and Development Fund; Grant-in-Aid for Scientific Research on Priority Areas and on Innovative Area from the Ministry of Education, Science, Sports, Culture and Technology of Japan; NCI (R01-CA121210); Korea Health Technology R&D Project (Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (H14C0066 )); Lung cancer (Z01CP010200); Lung health (U01HG004738); Ministry of Health (201002007); Ministry of Science and Technology (2011BAI09B00); Melanoma (NCI R29CA70334, R01CA100264, P50CA093459); Multiethnic Cohort Study (Infrastructure grant (U01 CA164973) and U01 CA098758); National Research Foundation of Korea (NRF) (funded by the Korea government (MSIP) (NRF-2014R1A2A2A05003665)); NLCS (China National High-Tech Research and Development Program Grant (2009AA022705); Priority Academic Program Development of Jiangsu Higher Education Institution; National Key Basic Research Program Grant (2011CB503805)); Nurses' Health Study (P01 CA87969, R01 CA49449); Nurses' Health Study II (UMI CA176726, R01 67262); OpPancreatic cancer (Mayo Clinic SPORC in Pancreatic Cancer: P50CA102701); Prostate cancer (U01HG004726, NCI: CA63464, CA54281, CA1326792, RC2 CA148085); Shanghai Women's Health Cohort Study (National Institutes of Health (R37 CA70867); National Cancer Institute intramural research program; NCI Intramural Research Program contract (N02 CP1101066)); Shenyang Lung Cancer Study (National Nature Science Foundation of China (81102194); Liaoning Provincial Department of Education (LS2010168); China Medical Board (00726)); Singapore Chinese Health Study (NIH grants: NCI R01 CA55069, R35 CA53890, R01 CA80205, and R01 CA144034); South Korea Multi-Center Lung Cancer Study (National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (2011-0016106); National R&D Program for Cancer Control, Ministry of Health & Welfare, Republic of Korea (0720550-2); (A010250)); Tianjin Lung Cancer Study (Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT); China (IRT1076); Tianjin Cancer Institute and Hospital, National Foundation for Cancer Research US); University of Texas MD Anderson Cancer Center (institutional support for the Center for Translational and Public Health Genomics (NIH grant P50 CA 91846)); Women's Health Initiative (National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services contracts HHSN268201100046C, HHSN268201100001C, HHSN268201100002C, HHSN268201100003C, HHSN268201100004C, and HHSN271201100004C); Wuhan lung cancer study (National Key Basic Research and Development Program (2011CB503800)) and Yunnan Lung Cancer Study (Intramural program of U.S. National Institutes of Health; National Cancer Institute). Melinda C. Aldrich was supported by NIH/NCI grant K07 CA 172294. We would like to thank the participants and staff of the NHS and NHSII cohorts for their valuable contributions as well as the following state registries for their help: AL, AZ, AR, CA, CO, CT, DE, FL, GA, ID, IL, IN, IA, KY, LA, ME, MD, MA, MI, NE, NH, NJ, NY, NC, ND, OH, OK, OR, PA, RI, SC, TN, TX, VA, WA, WY. This study was approved by the Connecticut Department of Public Health (DPH) Human Investigations Committee. Certain data used in this publication were obtained from the DPH. The authors assume full responsibility for the analyses and interpretation of these data. The Women's Health Initiative would like to acknowledge the following individuals for their participation: Program Office: (National Heart, Lung,

and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Dale Burwen, Joan McGowan, Leslie Ford, and Nancy Geller; Clinical Coordinating Center: (Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg; Investigators and Academic Centers: (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) JoAnn E. Manson; (MedStar Health Research Institute/Howard University, Washington, DC) Barbara V. Howard; (Stanford Prevention Research Center, Stanford, CA) Marcia L. Stefanick; (The Ohio State University, Columbus, OH) Rebecca Jackson; (University of Arizona, Tucson/Phoenix, AZ) Cynthia A. Thomson; (University at Buffalo, Buffalo, NY) Jean Wactawski-Wende; (University of Florida, Gainesville/Jacksonville, FL) Marian Limacher; (University of Iowa, Iowa City/Davenport, IA) Robert Wallace; (University of Pittsburgh, Pittsburgh, PA) Lewis Kuller; (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker. Women's Health Initiative Memory Study: (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker.

## Author contributions

M.D., M.Y. and S.J.C. conceived and designed the analysis. C.C.A., M.C.A., C.A., L.T.A., A.A.A., L.E.B., S.I.B., A.B., W.J.B., C.H.B., P.M.B., L.A.B., H.B.B., L.B., J.E.B., M.A.B., F.C., T.C., K.G.C., I.S.C., N.C., C.C., C.C., K.C., C.C.C., L.S.C., M.C.B., M.C., F.G.D., I.D.V., T.D., J.D., E.J.D., C.G.E., J.H.F., J.D.F., J.F.F., G.M.F., C.S.F., S.G., Y.T.G., S.M.G., M.G.C., M.M.G., J.M.G., G.G.G., E.M.G., E.L.G., L.G., A.M.G., C.A.H., G.H., S.E.H., C.C.H., R.H., E.A.H., Y.C.H., R.N.H., C.A.H., N.H., W.H., D.J.H., A.H., M.J., C.H., K.T.K., H.N.K., Y.H.K., Y.T.K., A.P.K., R.K., W.P.K., L.N.K., C.K., P.K., V.K., R.C.K., A.L.C., Q.L., M.T.L., L.L.M., D.L., X.L., L.M.L., D.L., J.L., J.L., L.L., A.N.N., N.M., K.M., L.H.M., R.R.M., B.S.M., L.M., L.M., S.H.O., I.O., J.Y.P., A.P.G., B.P., U.P., G.M.P., L.P., J.P., L.P.O., M.P.P., Y.L.Q., P.R., F.X.R., E.R., H.A.R., B.R.S., A.M.R., S.A.S., F.S., A.G.S., K.L.S., A.S., V.W.S., G.S., H.S., X.S., M.H.S., X.O.S., D.T.S., M.R.S., V.L.S., R.S.S., D.S., Z.Z.T., P.R.T., L.R.T., G.S.T.,

D.V.D.B., K.V., S.W., J.C.W., Z.W., N.W., W.W., E.W., J.K.W., B.M.W., M.P.W., C.W., T.W., X.W., Y.L.W., J.S.W., L.X., H.P.Y., P.C.Y., K.Y., K.A.Z., A.Z.J., W.Z., B.Z., R.G.Z., L.A.P., N.E.C., N.R., M.T., M.D., M.Y. and S.J.C. contributed samples and provided feedback. B.H., C.D. and C.H. designed and carried out laboratory experiments. M.J.M., W.Z., E.K., J.N.S., N.D.F., Q.Y. and K.B.J. performed and contributed to the analysis. M.J.M., W.Z., M.D., M.Y. and S.J.C. drafted the manuscript. All authors have reviewed and approved the manuscript.

## Additional information

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npng.nature.com/reprintsandpermissions/>

**How to cite this article:** Machiela, M.J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* 7:11843 doi: 10.1038/ncomms11843 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

Mitchell J. Machiela<sup>1</sup>, Weiyin Zhou<sup>1,2</sup>, Eric Karlins<sup>1,2</sup>, Joshua N. Sampson<sup>1</sup>, Neal D. Freedman<sup>1</sup>, Qi Yang<sup>1,2</sup>, Belynda Hicks<sup>1,2</sup>, Casey Dagnall<sup>1,2</sup>, Christopher Hautman<sup>1,2</sup>, Kevin B. Jacobs<sup>2,3</sup>, Christian C. Abnet<sup>1</sup>, Melinda C. Aldrich<sup>4,5</sup>, Christopher Amos<sup>6</sup>, Laufey T. Amundadottir<sup>1</sup>, Alan A. Arslan<sup>7,8,9</sup>, Laura E. Beane-Freeman<sup>1</sup>, Sonja I. Berndt<sup>1</sup>, Amanda Black<sup>1</sup>, William J. Blot<sup>5,10</sup>, Cathryn H. Bock<sup>11</sup>, Paige M. Bracci<sup>12</sup>, Louise A. Brinton<sup>1</sup>, H Bas Bueno-de-Mesquita<sup>13,14,15,16</sup>, Laurie Burdett<sup>1,2</sup>, Julie E. Buring<sup>17</sup>, Mary A. Butler<sup>18</sup>, Federico Canzian<sup>19</sup>, Tania Carreón<sup>18</sup>, Kari G. Chaffee<sup>20</sup>, I-Shou Chang<sup>21</sup>, Nilanjan Chatterjee<sup>1</sup>, Chu Chen<sup>22</sup>, Constance Chen<sup>23</sup>, Kexin Chen<sup>24</sup>, Charles C. Chung<sup>1,2</sup>, Linda S. Cook<sup>25</sup>, Marta Crous Bou<sup>23,26</sup>, Michael Cullen<sup>1,2</sup>, Faith G. Davis<sup>27</sup>, Immaculata De Vivo<sup>23,26</sup>, Ti Ding<sup>28</sup>, Jennifer Doherty<sup>29</sup>, Eric J. Duell<sup>30</sup>, Caroline G. Epstein<sup>1</sup>, Jin-Hu Fan<sup>31</sup>, Jonine D. Figueroa<sup>1</sup>, Joseph F. Fraumeni<sup>1</sup>, Christine M. Friedenreich<sup>32</sup>, Charles S. Fuchs<sup>26,33</sup>, Steven Gallinger<sup>34</sup>, Yu-Tang Gao<sup>35</sup>, Susan M. Gapstur<sup>36</sup>, Montserrat Garcia-Closas<sup>37</sup>, Mia M. Gaudet<sup>36</sup>, J. Michael Gaziano<sup>38,39</sup>, Graham G. Giles<sup>40</sup>, Elizabeth M. Gillanders<sup>41</sup>, Edward L. Giovannucci<sup>26,42</sup>, Lynn Goldin<sup>1</sup>, Alisa M. Goldstein<sup>1</sup>, Christopher A. Haiman<sup>43</sup>, Goran Hallmans<sup>44</sup>, Susan E. Hankinson<sup>26,45</sup>, Curtis C. Harris<sup>46</sup>, Roger Henriksson<sup>47</sup>, Elizabeth A. Holly<sup>12</sup>, Yun-Chul Hong<sup>48</sup>, Robert N. Hoover<sup>1</sup>, Chao A. Hsiung<sup>49</sup>, Nan Hu<sup>1</sup>, Wei Hu<sup>1</sup>, David J. Hunter<sup>23,26,50</sup>, Amy Hutchinson<sup>1,2</sup>, Mazda Jenab<sup>51</sup>, Christoffer Johansen<sup>52,53</sup>, Kay-Tee Khaw<sup>54</sup>, Hee Nam Kim<sup>55</sup>, Yeul Hong Kim<sup>56</sup>, Young Tae Kim<sup>57</sup>, Alison P. Klein<sup>58</sup>, Robert Klein<sup>59</sup>, Woon-Puay Koh<sup>60,61</sup>, Laurence N. Kolonel<sup>62</sup>, Charles Kooperberg<sup>22</sup>, Peter Kraft<sup>23</sup>, Vittorio Krogh<sup>63</sup>, Robert C. Kurtz<sup>64</sup>, Andrea LaCroix<sup>22</sup>, Qing Lan<sup>1</sup>, Maria Teresa Landi<sup>1</sup>, Loic Le Marchand<sup>62</sup>, Donghui Li<sup>65</sup>, Xiaolin Liang<sup>66</sup>, Linda M. Liao<sup>1</sup>, Dongxin Lin<sup>67,68</sup>, Jianjun Liu<sup>69,70</sup>, Jolanta Lissowska<sup>71</sup>, Lingeng Lu<sup>72</sup>, Anthony M. Magliocco<sup>73</sup>, Nuria Malats<sup>74</sup>, Keitaro Matsuo<sup>75</sup>, Lorna H. McNeill<sup>76</sup>, Robert R. McWilliams<sup>77</sup>, Beatrice S. Melin<sup>47</sup>, Lisa Mirabello<sup>1</sup>, Lee Moore<sup>1</sup>, Sara H. Olson<sup>66</sup>, Irene Orlow<sup>66</sup>, Jae Yong Park<sup>78</sup>, Ana Patiño-García<sup>79</sup>, Beata Peplonska<sup>80</sup>, Ulrike Peters<sup>22</sup>, Gloria M. Petersen<sup>20</sup>, Loreall Pooler<sup>81</sup>, Jennifer Prescott<sup>23,26</sup>, Ludmila Prokunina-Olsson<sup>1</sup>, Mark P. Purdue<sup>1</sup>, You-Lin Qiao<sup>82</sup>, Preetha Rajaraman<sup>1</sup>, Francisco X. Real<sup>74,83</sup>, Elio Riboli<sup>84</sup>, Harvey A. Risch<sup>72</sup>, Benjamin Rodriguez-Santiago<sup>83,85,86</sup>, Avima M. Ruder<sup>18</sup>, Sharon A. Savage<sup>1</sup>, Fredrick Schumacher<sup>43</sup>, Ann G. Schwartz<sup>11</sup>, Kendra L. Schwartz<sup>87</sup>,

Adeline Seow<sup>61</sup>, Veronica Wendy Setiawan<sup>43</sup>, Gianluca Severi<sup>40,88</sup>, Hongbing Shen<sup>89,90</sup>, Xin Sheng<sup>81</sup>, Min-Ho Shin<sup>91</sup>, Xiao-Ou Shu<sup>92</sup>, Debra T. Silverman<sup>1</sup>, Margaret R. Spitz<sup>93</sup>, Victoria L. Stevens<sup>36</sup>, Rachael Stolzenberg-Solomon<sup>1</sup>, Daniel Stram<sup>43</sup>, Ze-Zhong Tang<sup>28</sup>, Philip R. Taylor<sup>1</sup>, Lauren R. Teras<sup>36</sup>, Geoffrey S. Tobias<sup>1</sup>, David Van Den Berg<sup>43</sup>, Kala Visvanathan<sup>94</sup>, Sholom Wacholder<sup>1</sup>, Jiu-Cun Wang<sup>95,96</sup>, Zhaoming Wang<sup>1,2</sup>, Nicolas Wentzensen<sup>1</sup>, William Wheeler<sup>97</sup>, Emily White<sup>22</sup>, John K. Wiencke<sup>98</sup>, Brian M. Wolpin<sup>26,33</sup>, Maria Pik Wong<sup>99</sup>, Chen Wu<sup>67,68</sup>, Tangchun Wu<sup>100</sup>, Xifeng Wu<sup>101</sup>, Yi-Long Wu<sup>102</sup>, Jay S. Wunder<sup>103</sup>, Lucy Xia<sup>81</sup>, Hannah P. Yang<sup>1</sup>, Pan-Chyr Yang<sup>104</sup>, Kai Yu<sup>1</sup>, Krista A. Zanetti<sup>41</sup>, Anne Zeleniuch-Jacquotte<sup>9,105</sup>, Wei Zheng<sup>5</sup>, Baosen Zhou<sup>106</sup>, Regina G. Ziegler<sup>1</sup>, Luis A. Perez-Jurado<sup>83,85</sup>, Neil E. Caporaso<sup>1</sup>, Nathaniel Rothman<sup>1</sup>, Margaret Tucker<sup>1</sup>, Michael C. Dean<sup>1</sup>, Meredith Yeager<sup>1,2</sup> & Stephen J. Chanock<sup>1</sup>

<sup>1</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, Maryland 20892, USA. <sup>2</sup>Cancer Genomics Research Laboratory, National Cancer Institute, Division of Cancer Epidemiology and Genetics, Leidos Biomedical Research Inc., Bethesda, Maryland 20892, USA. <sup>3</sup>Bioinformed, LLC, Gaithersburg, Maryland 20877, USA. <sup>4</sup>Department of Thoracic Surgery, Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA. <sup>5</sup>Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. <sup>6</sup>Department of Epidemiology, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. <sup>7</sup>Department of Obstetrics and Gynecology, New York University School of Medicine, New York, New York 10016, USA. <sup>8</sup>Department of Environmental Medicine, New York University School of Medicine, New York, New York 10016, USA. <sup>9</sup>New York University Cancer Institute, New York, New York 10016, USA. <sup>10</sup>International Epidemiology Institute, Rockville, Maryland 20850, USA. <sup>11</sup>Karmanos Cancer Institute and Department of Oncology, Wayne State University School of Medicine, Detroit, Michigan 48201, USA. <sup>12</sup>Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, California 94143, USA. <sup>13</sup>Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), 3721Bilthoven, The Netherlands. <sup>14</sup>Department of Gastroenterology and Hepatology, University Medical Center, 3584 CXUtrecht, The Netherlands. <sup>15</sup>Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London SW7 2AZ, UK. <sup>16</sup>Department of Social and Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur 50603, Malaysia. <sup>17</sup>Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. <sup>18</sup>National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio 45226, USA. <sup>19</sup>Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. <sup>20</sup>Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota 55905, USA. <sup>21</sup>National Institute of Cancer Research, National Health Research Institutes, Zhunan 35053, Taiwan. <sup>22</sup>Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA. <sup>23</sup>Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, Massachusetts 02115, USA. <sup>24</sup>Department of Epidemiology and Biostatistics, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300040, China. <sup>25</sup>University of New Mexico, Albuquerque, New Mexico 87131, USA. <sup>26</sup>Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>27</sup>Department of Public Health Sciences, School of Public Health, University of Alberta, Edmonton, Alberta, Canada T6G 2R3. <sup>28</sup>Shanxi Cancer Hospital, Taiyuan, Shanxi 030013, China. <sup>29</sup>Geisel School of Medicine, Dartmouth College, Lebanon, New Hampshire 03755, USA. <sup>30</sup>Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Bellvitge Biomedical Research Institute, Catalan Institute of Oncology (ICO-IDIBELL), 08908 Barcelona, Spain. <sup>31</sup>Shanghai Cancer Institute, Shanghai 200032, China. <sup>32</sup>Department of Population Health Research, Cancer Control Alberta, Alberta Health Services, Calgary, Alberta, Canada T2N 2T9. <sup>33</sup>Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. <sup>34</sup>Fred A Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada M5G 1X5. <sup>35</sup>Department of Epidemiology, Shanghai Cancer Institute, Renji Hospital, Shanghai Jiaotaong University School of Medicine, Shanghai 200032, China. <sup>36</sup>Epidemiology Research Program, American Cancer Society, Atlanta, Georgia 30303, USA. <sup>37</sup>Division of Genetics and Epidemiology, and Breakthrough Breast Cancer Centre, Institute for Cancer Research, London SM2 5NG, UK. <sup>38</sup>Divisions of Preventive Medicine and Aging, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts 02115, USA. <sup>39</sup>Massachusetts Veterans Epidemiology Research and Information Center/VA Cooperative Studies Programs, Veterans Affairs Boston Healthcare System, Boston, Massachusetts 02130, USA. <sup>40</sup>Cancer Epidemiology Centre, Cancer Council Victoria & Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria 3010, Australia. <sup>41</sup>Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland 20892, USA. <sup>42</sup>Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts 02115, USA. <sup>43</sup>Department of Preventive Medicine, Biostatistics Division, Keck School of Medicine at the University of Southern California, Los Angeles, California 90033, USA. <sup>44</sup>Department of Public Health and Clinical Medicine/Nutritional Research, Umeå University, 901 87Umeå, Sweden. <sup>45</sup>Division of Biostatistics and Epidemiology, University of Massachusetts School of Public Health and Health Sciences, Amherst, Massachusetts 01003, USA. <sup>46</sup>Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, Bethesda, Maryland 20892, USA. <sup>47</sup>Department of Radiation Sciences, Oncology, Umeå University, 901 87 Umeå, Sweden. <sup>48</sup>Department of Preventive Medicine, Seoul National University College of Medicine, Seoul 151-742, Republic of Korea. <sup>49</sup>Institute of Population Health Sciences, National Health Research Institutes, Zhunan 35053, Taiwan. <sup>50</sup>Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA. <sup>51</sup>International Agency for Research on Cancer (IARC-WHO), 69372Lyon, France. <sup>52</sup>Oncology, Finsen Centre, Rigshospitalet, 2100 Copenhagen, Denmark. <sup>53</sup>Unit of Survivorship Research, The Danish Cancer Society Research Centre, 2100 Copenhagen, Denmark. <sup>54</sup>School of Clinical Medicine, University of Cambridge, Cambridge CB2 1TN, UK. <sup>55</sup>Center for Creative Biomedical Scientists, Chonnam National University, Gwangju 500-757, Republic of Korea. <sup>56</sup>Department of Internal Medicine, Division of Oncology/Hematology, College of Medicine, Korea University Anam Hospital, Seoul 151-742, Republic of Korea. <sup>57</sup>Department of Thoracic and Cardiovascular Surgery, Cancer Research Institute, Seoul National University College of Medicine, Seoul 03080, Republic of Korea. <sup>58</sup>Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21287, USA. <sup>59</sup>Program in Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. <sup>60</sup>Duke-NUS Graduate Medical School, Singapore 169857, Singapore. <sup>61</sup>Saw Swee Hock School of Public Health, National University of Singapore, Singapore 119077, Singapore. <sup>62</sup>Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii 96813, USA. <sup>63</sup>Fondazione IRCCS Istituto Nazionale dei Tumori, Milano 20133, Italy. <sup>64</sup>Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. <sup>65</sup>Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. <sup>66</sup>Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. <sup>67</sup>Department of Etiology & Carcinogenesis, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union



Medical College, Beijing 100730, China. <sup>68</sup> State Key Laboratory of Molecular Oncology, Cancer Institute and Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China. <sup>69</sup> Department of Human Genetics, Genome Institute of Singapore 138672, Singapore. <sup>70</sup> School of Life Sciences, Anhui Medical University, Hefei 230032, China. <sup>71</sup> Department of Cancer Epidemiology and Prevention, Maria Skłodowska-Curie Cancer Center and Institute of Oncology, Warsaw 02-781, Poland. <sup>72</sup> Yale School of Public Health, New Haven, Connecticut 06510, USA. <sup>73</sup> H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida 33612, USA. <sup>74</sup> Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain. <sup>75</sup> Division of Molecular Medicine, Aichi Cancer Center Research Institute, Nagoya 464-8681, Japan. <sup>76</sup> Department of Health Disparities Research, Division of OVP, Cancer Prevention and Population Sciences, and Center for Community-Engaged Translational Research, Duncan Family Institute, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. <sup>77</sup> Department of Oncology, Mayo Clinic, Rochester, Minnesota 55905, USA. <sup>78</sup> Lung Cancer Center, Kyungpook National University Medical Center, Daegu 701, Republic of Korea. <sup>79</sup> Department of Pediatrics, University Clinic of Navarra, Universidad de Navarra, IdiSNA, Navarra Institute for Health Research, Pamplona 31080, Spain. <sup>80</sup> Nofer Institute of Occupational Medicine, Lodz 91-348, Poland. <sup>81</sup> University of Southern California, Los Angeles, California 90007, USA. <sup>82</sup> Department of Epidemiology, Cancer Institute (Hospital), Chinese Academy of Medical Sciences, Beijing 100730, China. <sup>83</sup> Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona 08002, Spain. <sup>84</sup> Division of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London SW7 2AZ, UK. <sup>85</sup> Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona 28029, Spain. <sup>86</sup> Quantitative Genomic Medicine Laboratory, qGenomics, Barcelona 08003, Spain. <sup>87</sup> Karmanos Cancer Institute and Department of Family Medicine and Public Health Sciences, Wayne State University School of Medicine, Detroit, Michigan 48201, USA. <sup>88</sup> Human Genetics Foundation (HuGeF), Torino 10126, Italy. <sup>89</sup> Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Nanjing Medical University, Nanjing 210029, China. <sup>90</sup> Department of Epidemiology, Nanjing Medical University School of Public Health, Nanjing 210029, China. <sup>91</sup> Department of Preventive Medicine, Chonnam National University Medical School, Gwanju 501-746, Republic of Korea. <sup>92</sup> Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, Tennessee 37232, USA. <sup>93</sup> Baylor College of Medicine, Houston, Texas 77030, USA. <sup>94</sup> Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21218, USA. <sup>95</sup> Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai 200433, China. <sup>96</sup> State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China. <sup>97</sup> Information Management Services Inc., Calverton, Maryland 20904, USA. <sup>98</sup> University of California San Francisco, San Francisco, California 94143, USA. <sup>99</sup> Department of Pathology, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. <sup>100</sup> Institute of Occupational Medicine and Ministry of Education Key Laboratory for Environment and Health, School of Public Health, Huazhong University of Science and Technology, Wuhan 430400, China. <sup>101</sup> Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. <sup>102</sup> Guangdong Lung Cancer Institute, Guangdong General Hospital & Guangdong Academy of Medical Sciences, Guangzhou 515200, China. <sup>103</sup> Division of Urologic Surgery, Washington University School of Medicine, St Louis, Missouri 63110, USA. <sup>104</sup> Department of Internal Medicine, National Taiwan University College of Medicine, Taipei 10617, Taiwan. <sup>105</sup> Department of Population Health, New York University School of Medicine, New York, New York 10016, USA. <sup>106</sup> Department of Epidemiology, School of Public Health, China Medical University, Shenyang 110001, China.