Mosaic chromosomal alterations is associated with increased lung cancer risk: insight from the INTEGRAL-ILCCO cohort analysis

Chao Cheng[1-3*], Wei Hong[1], Yafang Li[1-3], Xiangjun Xiao[1], James McKay[4], Younghun Han[1,2], Jinyoung Byun[1,2], Bo Peng[1,2], Demetrios Albanes[5], Stephen Lam[6], Adonina Tardon[7], Chu Chen[8], Stig E Bojesen[9-10], Maria T Landi[5], Mattias Johansson[4], Angela Risch[11-13], Heike Bickeböller[14], H-Erich Wichmann[15], David C Christiani[16], Gad Rennert[17], Susanne Arnold[18], Gary Goodman[19], John K Field[20], Michael PA Davies[20], Sanjay S Shete[21-22], Loic Le Marchand[23], Olle Melander[24], Hans Brunnström[24], Geoffrey Liu[25], Rayjean J Hung[26-27], Angeline S Andrew[28], Lambertus A Kiemeney[29], Meng Zhu[30], Hongbing Shen[30], Shan Zienolddiny[31], Kjell Grankvist[32], Mikael Johansson[33], Angela Cox[34], Yun-Chul Hong[35], Jian-Min Yuan[36], Philip Lazarus[37], Matthew B Schabath[38], Melinda C Aldrich[39], Paul Brennan[4], Yong Li[1-3], Olga Gorlova[1-3], Ivan Gorlov[1-3], Christopher I Amos[1-3*]. INTEGRAL-ILCCO lung cancer consortium.

* Corresponding authors
Chao Cheng, Email: chao.cheng@bcm.edu
Christopher I Amos, Email: chris.amos@bcm.edu
Baylor College of Medicine
1 Baylor Plaza, Houston, TX 77030
Tel: 713-798-2102
Fax: 713-798-3658

1. Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX.
2. Section of Epidemiology and Population Sciences, Department of Medicine, Baylor College of Medicine, Houston, TX.
3. Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX.
4. Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France.
5. Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD.
6. Department of Integrative Oncology, University of British Columbia, Vancouver, BC, Canada.
7. Public Health Department, University of Oviedo, ISPA and CIBERESP, Asturias, Spain.
8. Program in Epidemiology, Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA.
9. Department of Clinical Biochemistry, Copenhagen University Hospital, Copenhagen, Denmark.
10. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
11. Thoraxklinik at University Hospital Heidelberg, Heidelberg, Germany.
12. Translational Lung Research Center Heidelberg (TLRC-H), Heidelberg, Germany.
13. University of Salzburg and Cancer Cluster Salzburg, Austria.

14. Department of Genetic Epidemiology, University Medical Center, Georg-August-University Göttingen, Germany.
15. Institute of Medical Statistics and Epidemiology, Technical University Munich, Germany.
16. Departments of Environmental Health and Epidemiology, Harvard TH Chan School of Public Health, Boston, MA.
17. Clalit National Cancer Control Center at Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel.
18. University of Kentucky, Markey Cancer Center, Lexington, Kentucky, USA.
19. Swedish Cancer Institute, Seattle, WA, USA.
20. Department of Molecular and Clinical Cancer Medicine University of Liverpool, Liverpool, United Kingdom.
21. Department of Biostatistics, The University of Texas, M.D. Anderson Cancer Center, Houston, TX.
22. Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, TX USA.
23. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA.
24. Faculty of Medicine, Lund University, Lund, Sweden.
25. University Health Network- The Princess Margaret Cancer Centre, Toronto, CA.
26. Luenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada.
27. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Canada.
28. Departments of Epidemiology and Community and Family Medicine, Dartmouth College, Hanover, NH.
29. Radboud University Medical Center, Nijmegen, The Netherlands.
30. Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, P.R. China.
31. National Institute of Occupational Health, Oslo, Norway.
32. Department of Medical Biosciences, Umeå University, Umeå, Sweden.
33. Department of Radiation Sciences, Umeå University, Umeå, Sweden.
34. Academic Unit of Clinical Oncology University of Sheffield, Weston Park Hospital, Whitham Road, Sheffield, UK
35. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea.
36. Pittsburgh Liver Research Center, University of Pittsburgh & UPMC, UPMC Cancer Pavilion, Suite 4C-470, 5150 Centre Avenue, Pittsburgh, PA 15232, USA
37. Department of Pharmaceutical Sciences, College of Pharmacy, Washington State University, Spokane, Washington, USA.
38. Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, FL.
39. Department of Thoracic Surgery, Division of Epidemiology, Vanderbilt University Medical Center.

Abstract

Mosaic chromosomal alterations (mCAs) detected in blood cells represent a type of clonal hematopoiesis (CH) that is understudied compared to CH-related somatic mutations. A few recent studies indicated its potential link with non-hematological cancers, especially lung cancer. In this study, we investigated the association between mCAs and lung cancer using the high-density genotyping data from the OncoArray study of INTEGRAL-ILCCO, the largest single genetic study of lung cancer with 18,221 lung cancer cases and 14,825 cancer-free controls. We identified a comprehensive list of autosomal mCAs, ChrX mCAs, and mosaic ChrY (mChrY) losses from these samples. Autosomal mCAs were detected in 4.3% of subjects, in addition to ChrX mCAs and mChrY losses detected in 3.6% of females and 9.6% of males, respectively. Multivariable logistic regression analysis indicated that the presence of autosomal mCAs in blood cells was associated with an increased lung cancer risk after adjusting for key confounding factors including age, sex, smoking status, and race. Such an association was mainly driven by a specific type of mCAs -- copy-neutral loss of heterogeneity (CN-LOH) on autosomal chromosomes. The association between autosome CN-LOH and increased risk of lung cancer was further confirmed in two major histological subtypes, lung adenocarcinoma and squamous cell carcinoma. Additionally, we observed a significant increase of ChrX mCAs and mChrY losses in smokers compared to non-smokers, as well racial difference in certain types of mCA events. Our study established a link between mCAs in white blood cells and increased risk of lung cancer.

Introduction

In humans, hematopoietic stem cells reside in bone marrow, maintaining the ability to divide and differentiate into all types of blood cells. With increasing age, irreparable somatic mutations may occur and accumulate in a small fraction of hematopoietic stem cells [1,2]. Some of these mutations confer proliferative or survival advantages and lead to clonal expansion of the hosting cells in blood, a phenomenon called clonal hemopoiesis (CH). While most CH studies have focused on the detection of point mutations and short insertion/deletions (indels), the presence of mosaic chromosomal alterations (mCAs) has become increasingly noticed [2,3].

Recently, two large-scale studies have been performed to identify mCAs from genotyping data of blood-derived DNA using the United Kingdom Biobank (UKBB) [4] and Japan BioBank (BBJ) [5], respectively. These studies revealed that the accumulation of mCAs is a feature of aging with a detection rate of 2-8% in subjects younger than 50 but a rapid increase afterward [4,5]. Particularly, in the BBJ cohort more than 35% of subjects with age ≥ 90 have mCAs [5]. Smokers are more likely to carry mCAs than non-smokers with matched age. In addition, the incidence of mCA in males is significantly higher than in females after adjusting for age and smoking status [5]. In both UKBB

and BBJ studies, a significantly higher all-cause mortality rate has been observed for individuals with mCAs [4,5]. Importantly, it has been reported that mCAs are associated with a variety of human diseases, such as cardiovascular diseases [6], autism spectrum disorder [7] and infectious diseases [8]. As mutations detected in blood cells, mCAs have been found to be associated with hematological cancers [9,10]. Individuals with detected mCAs had a ten times higher risk of developing hematological cancers compared to those without mCAs [11]. Moreover, mCAs involving larger genomic regions tend to be associated with an earlier onset and a higher rate of mortality of patients with hematological malignancy [12].

The association of CH with selected non-hematological cancers has also been reported in previous publications[13,14]. However, most of these studies focused on point mutations and short indels without considering mCA events. The UKBB and BBJ cohorts come from a general population with relatively small number of cancer incidences, which provided limited information for investigating the association between mCAs and specific cancer types. Interestingly, in a multicancer study, genotyping data from 13 cancer genome-wide association datasets were integrated for identifying mCAs in 31,717 cancer cases (including 31,259 non-hematologic cases from over 14 different cancer types) and 26,136 cancer-free controls [10]. This study found that mCAs were more frequently detected in blood samples collected before diagnosis or treatment from subjects with non-hematologic cancers than in controls. When stratified based on cancer types, a significant association was observed in lung cancer. In addition, mosaic loss of chromosome Y (mChrY loss) has been reported to be associated with increased lung cancer risk and prognosis [15,16]. These studies suggested a potential association between mCAs and lung cancer. To further verify this association, a more careful investigation using a large lung cancer cohort is required.

The INTEGRAL (Integrative Analysis of Lung Cancer Etiology and Risk)-ILCCO (International Lung Cancer Consortium) subjects, which is the largest single genetic study of lung cancer [17]. We focused on a major sub-cohort from the OncoArray Consortium Lung Study [18,19], which provides high-density blood genotyping data for 33,046 subjects, including 18,221 lung cancer cases and 14,825 non-cancer controls. Moreover, the data provide high-quality demographic and clinical variables including age, sex, race, smoking status, and histological subtypes, allowing us to investigate the association between mCAs and lung cancer while considering the effect of these confounders.

Results

Systematic identification of mCAs from the OncoArray data

The OncoArray dataset from the INTEGRAL-ILCCO cohort contains blood-derived genotyping array data for a total of 33,046 subjects, including 18,221 lung cancer patients and 14,825 cancer-free controls (Table 1) [19]. We applied the MoChA method [4,12] to identify mCAs presenting on autosomal chromosomes in all subjects and ChrX in female subjects. MoChA harnesses chromosome phase information to combine nearby SNPs and can confidently identify mCAs presenting even in a small fraction of blood cells (cell fraction ≥1%) [12]. For male subjects, MoChA relies on variants in the pseudoautosomal regions (PAR1 and PAR2) of sex chromosomes [16]. However, the OncoArray genotyping platform has only a limited number of SNPs (n=29) in these regions. Therefore, we restricted ChrX-specific mCA detection to female subjects. Nevertheless, frequent mosaic loss of ChrY (mChrY loss) in male blood cells has been reported [16,20–23], and found to be associated with an increased risk of lung cancer [15,24]. As such, we determined the mChrY loss events in our male subjects by using an established method from previous studies [21,25,26].

Table 1. Characteristics of the OncoArray subjects. For *Age*, the mean age and the standard deviation (in the parenthesis) are listed. For other variables, the number and percentage of subject are listed. LUAD: lung adenocarcinoma; LUSC: lung squamous cell carcinoma; SCLC: small cell lung cancer.

Phenotype	Variable	Lung Cancer	Control
	Age	64.9 (±10.0)	62.1 (±10.4)
Sex	Male	11180 (61%)	8915 (60%)
	Female	7041 (39%)	5910 (40%)
Race	White	12896 (90%)	10733 (86%)
	Asian	608 (4.30%)	819 (6.60%)
	Black	346 (2.40%)	576 (4.60%)
	Other	436 (3.10%)	341 (2.70%)
Smoking	Smoker	15967 (89%)	9754 (67%)
	Non-smoker	1984 (11%)	4773 (33%)
Cancer Subtype	LUAD	6852 (38%)	
	LUSC	4408 (24%)	
	SCLC	1648 (9%)	
	Other	6960 (38%)	
	Total	18221	14825

Distribution of mCAs in the human genome

From the OncoArray subjects, we identified a total of 1,808 autosomal mCAs presenting in ≥1% of blood cells. Out of these mCAs, 310 (17.1%), 586 (32.4%), and 763 (42.2%) were confidently categorized as gain, loss, and copy-number neutral loss of heterozygosity (CN-LOH), respectively. The remaining 149 mCAs (8%) were categorized as “undetermined”, because their copy number cannot be explicitly determined. Interestingly, mCAs were not evenly distributed across the genome with Chr11, Chr20, and Chr9 having the largest number of mCAs (Fig. 1A). These 1,808 autosomal mCAs were identified from 1,411 subjects, accounting for about 4.2% of the 33,046

subjects from our cohort. In the 12,951 female subjects, we identified 512 ChrX mCAs involving 397 subjects, which included 181 gain, 143 loss, 123 CN-LOH, and 65 undetermined events (Fig. 1A). Of note, 3.1% of female subjects harbor at least one mCA on ChrX, which is much higher than the detected mCA rate on all individual autosomal chromosomes.

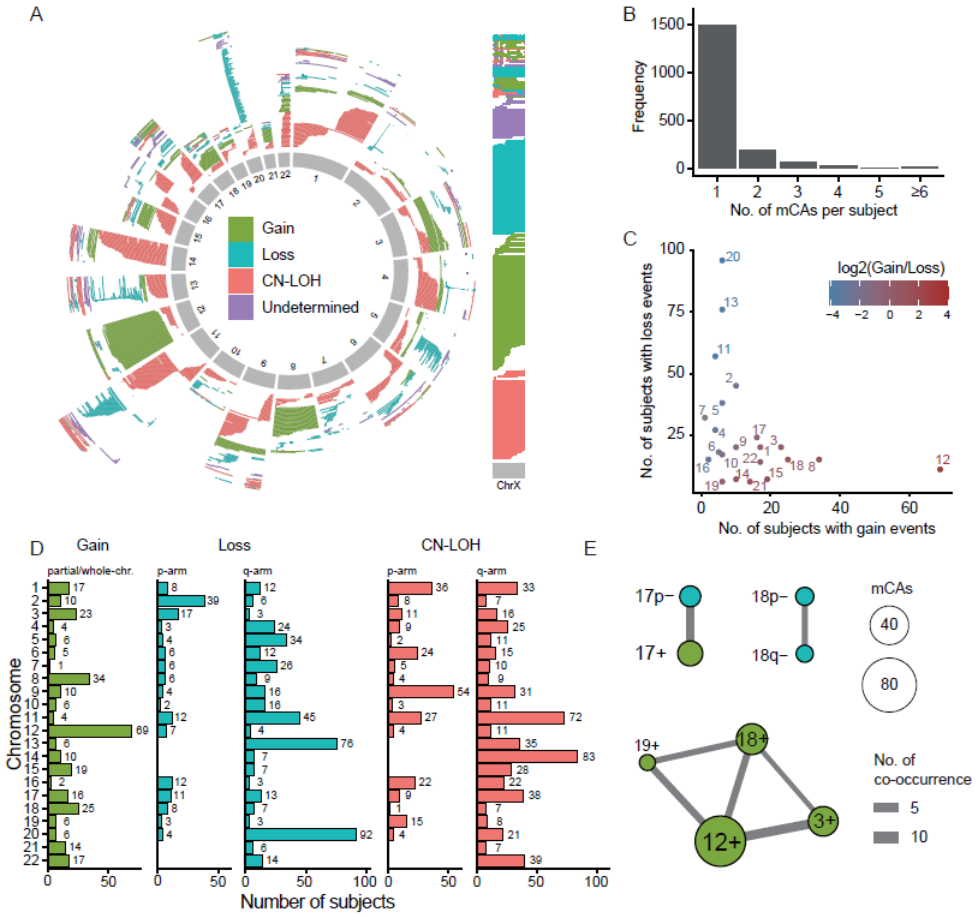

Figure 1. The distribution of mCA events across the human genome. (A) Distribution of mCA events on each autosome and chromosome X. Each mCA event is shown as a line with indicated start and end positions on the corresponding chromosome. (B) Distribution of the number of autosomal and ChrX mCA events detected in each subject. As shown, most subjects have only one mCA event. (C) For each chromosome, the number of subjects with mCA gain (X-axis) and loss (Y-axis) events are counted and shown as a scatterplot. Each dot represents a chromosome. (D) The number of chromosome arm-level mCA events for each chromosome. Mosaic loss and CN-LOH events are further mapped into the long (q-arm) and short arms (p-arm). Most mosaic gain events involve the whole chromosome. (E) The co-occurrence graph for arm-level mCAs. Each edge connects two arm-level mCAs that are significantly co-occurred across subjects (FDR<0.05). Of note, p+/- and q+/- indicate the presence of mCA gain/loss event on the short and long arm, rather than gain/loss of the whole arm.

In the 1,786 mCA-positive subjects, the majority (n=1482, 83%) have only a single autosomal/ChrX mCA event, but a small fraction of subjects presented multiple mCAs (Fig. 1B). Most of the mCAs involved a broad genomic region with a median size of 19.5M bases. We compared the mosaic gain and loss events associated with each autosomal chromosome and found a negative correlation between them (ρ=-0.44). This indicated that most chromosomes or

arms tended to have either gain or loss events (Fig. 1C). Consistent with the UKBB cohort [12] , Chr12 has the largest number of mosaic gain events, while Chr13 and Chr20 were most enriched in mosaic loss events (Fig. 1C).

Most of autosomal mosaic gain events were whole-chromosome events. In contrast, most of the autosomal loss and CN-LOH mCAs involved only certain region of a chromosome. As such, we mapped the loss and CN-LOH mCA events to specific chromosome arms and denoted them as p/q- (loss) or p/q= (CN-LOH). At the arm-level, Chr12q is enriched for mosaic gain events; Chr13q and Chr20q are enriched for mosaic loss events; while Chr11q, Chr14q and Chr9p are enriched for mosaic CN-LOH events (Fig. 1D). At the chromosome/arm level, a small number of subjects (n=155, 9%) harbored multiple mCA events, in which we identified a few mCA pairs with significantly more co-occurrences than what expected by chance (Fig. 1E). Consistent with previous reported results from the UKBB cohort [12], we found a cluster of mosaic gain events on Chr12, Chr3, Chr18 and Chr19 tend to present together. In addition, we found another two pairs of co-occurrences i) mosaic loss of Chr17 short arm (17p-) and mosaic gain of Chr17 (17+), and ii) mosaic loss of Chr18 long (18q-) and short (18p-) arms (Fig. 1E). The occurrences of them have also been observed in the UKBB cohort but not reach the significant threshold [12].

The detection rate of mCAs in blood cells is continuously increased with age

Accumulation of mCAs has been found to be a feature of aging [4,5]. We built a multivariable logistic regression model (Model I, refer to the Methods) to investigate how the presence of mCAs was affected by different subject features including age. Specifically, we investigated autosomal and ChrX mCAs, which were further divided into 3 subtypes (gain, loss, and CN-LOH), as well as mChrY losses. For all mCA types and subtypes, we observed a significant association with age – the probability of a subject being mCA-positive is significantly increase with age (Table S2). As shown in Fig. 2A, the fraction of subjects with autosomal mCAs (in both males and females), ChrX mCAs (in females), and mChrY loss (in males) are continuously increasing with age. It is notable that mChrY loss showed a faster increase than the other mCA types: it was detected in less than 5% of males younger than 60 but in ~18% of males older than 80. We then divided all subjects into a young group (<65) and an old group (≥65), and observed a significantly higher fraction of mCA-positive samples in the old group for all mCA types (Fig. 2B). Our models also identified a sex difference -- males are more likely to have autosomal mCA gains and losses compared to females (Fig. 2C).

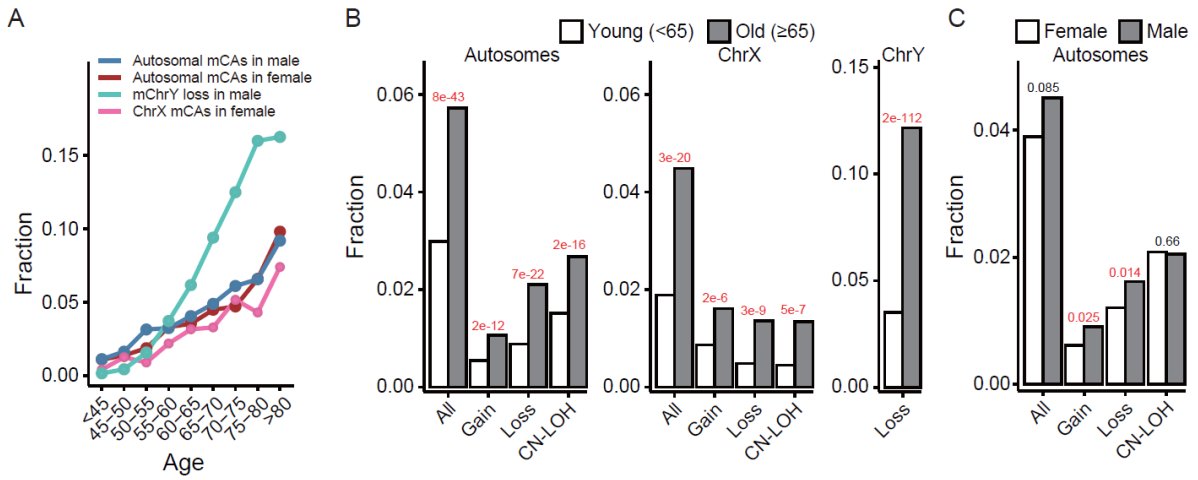

Figure 2. Association of mCAs with age and sex. (A) Fraction of subjects with autosomal mCAs, ChrX mCAs or mChrY loss in each age group. The frequency of all types of mCAs increases with age in both males and females. (B) Comparisons of mCA rate between young (age<65) and old (age≥65) subjects. (C) Comparisons of autosomal mCA rate between males and females. Males tend to have a higher rate of autosomal gains and losses than females.

Significant increase of autosomal CN-LOH in lung cancer patients

Model I indicated that lung cancer cases were more likely to accumulate autosomal mCAs in their blood cells compared to non-cancer controls (Table 1). As shown in Fig. 3A, in both lung cancer cases and controls the fraction of subjects with detected autosome mCAs contiuously increase with age; but, the cases showed an increase starting 5-10 years earlier than in the controls. This suggests that lung cancer patients accumulate mCAs at earlier ages. In other words, the accumulation of autosomal mCAs with age is associated with increased lung cancer risk.

To determine the contribution of mCA events on lung cancer risk while adjusting for major confounding variables (e.g., age, smoking status, etc.), we build another logistic regression model using the lung cancer status as the response variable (Model II, see Methods). Our model indicated that the presence of autosomal mCA events increased the risk of lung cancer by 34% (odds ratio OR=1.34, p=1e-5), after adjusting for age, sex, race, and smoking status (**Table 2**). More specifically, mosaic autosomal loss and CN-LOH is associated with 27% (p=0.03) and 43% (p=1e-4) increased risk of lung cancer, respectively, while mosaic autosomal gain is not significantly associated (Table 2 and Fig. 3B). In contrast, neither ChrX mCAs nor mChrY losses are significantly associated with lung cancer risk (Table 2) after adjusting for increases associated with aging. Furthermore, we examined the three major lung cancer histological types: lung adenocarcinoma (LUAD), squamous cell carcinoma (LUSC), and small cell lung cancer (SCLC). Our results confirmed the association between autosomal mCAs and lung cancer risk in LUAD and LUSC (Table 2) and indicated that the association was mainly driven by mosaic autosomal CN-LOH events. As shown, the presence of autosomal CN-LOH events is associated with 54%

and 40% increased risks of LUAD and LUSC, respectively (Fig. 3C-D, Table 2). While we did not identify significant associations between SCLC and mCAs potentially due to smaller sample sizes, the mosaic autosome CN-LOH events also showed weak correlation with SCLC (p=0.05, Table 2).

Table 2. The associations between different types of mCA and lung cancer while adjusting for age, sex, race, and smoking status. Results are based on the logistic regression model II. Significant associations were highlighted in bold. ALL, LUAD, LUSC, SCLC indicate all lung cancer, lung adenocarcinoma, squamous cell carcinoma, and small cell lung cancer cases, respectively.

mCA		All		LUAD		LUSC		SCLC	
Chr.	*Type*	*Coef.*	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*	*Coef.*	*P-value*
Autosome	**All**	**0.29**	**1e-5**	**0.33**	**7.4e-5**	**0.3**	**0.0018**	0.13	>0.1
	Gain	0.11	>0.1	-0.046	>0.1	0.15	>0.1	-0.61	>0.1
	Loss	**0.24**	**0.03**	0.22	>0.1	0.22	>0.1	0.043	>0.1
	CN-LOH	**0.36**	**1e-4**	**0.43**	**2.1e-4**	**0.34**	**0.012**	0.39	0.05
ChrX	**All**	-0.035	>0.1	0.17	>0.1	-0.029	>0.1	-0.53	>0.1
	Gain	-0.0091	>0.1	0.16	>0.1	-0.15	>0.1	-0.13	>0.1
	Loss	-0.081	>0.1	0.2	>0.1	-0.055	>0.1	-0.76	>0.1
	CN-LOH	0.1	>0.1	0.39	>0.1	0.29	>0.1	-14	>0.1
ChrY	**Loss**	0.024	>0.1	0.11	>0.1	0.014	>0.1	-0.12	>0.1

We compared the occurrences of chromosome/arm level mCAs between the lung cancer and the control group. Some of the mCAs were more likely to present in the cancer group, including the mosaic loss of Chr11q (11q-), CN-LOH of Chr13q (13q=), gain of Chr8 (8+) and gain of Chr3 (3+) (Fig. 3E). Interestingly, no mCAs were enriched in the controls. Deletion of Chr11q was previously reported as one of the most frequent chromosome changes in various cancers [27]. Several tumor suppressor genes such as *ATM* and *CBL* are located at the long arm of Chr11. Somatic inactivating-mutations or loss of these genes are common in various cancers [28,29], and have been found to increase the proliferation rate of cells [30,31]. By enumerating genes in each detected mCA region, we then counted the mosaic copy number alterations of all cancer-related genes across all subjects. At the gene level, we found some cancer-related genes are enriched in the mCA regions in lung cancer versus controls. Among the top ten cancer related genes enriched in the mCA regions in lung cancer versus controls, we found suppressor genes such as *ARHGEF12* [32], *DDX10* [33] and *ATM* were more likely lost in cancer; while, oncogenes such as *BCL6* [34], *LPP* [35] and *MYC* were more likely to be gained in cancer. The oncogene *NRAS* was more likely CN-LOH in lung cancer patients (Fig. 3F).

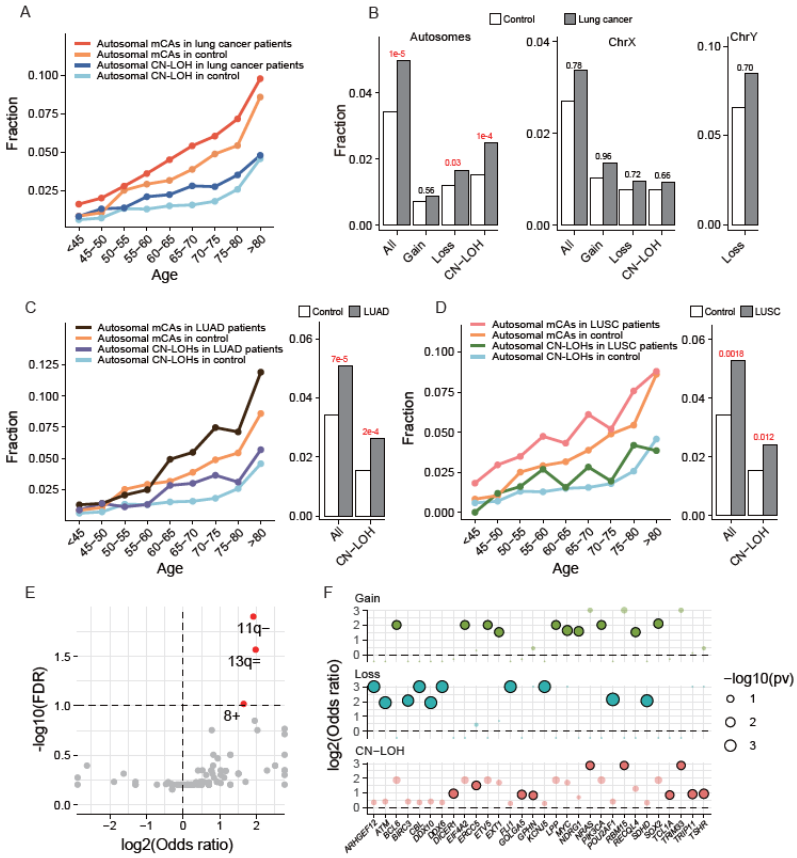

Figure 3. The presence of mCAs is associated with increased risk of lung cancer. (A) Distribution of overall autosomal mCAs and CN-LOHs across age in lung cancer cases and controls. (B) Lung cancer patients show a significantly higher rate of autosomal mCAs, especially CN-LOHs and losses. (C-D) Distribution of overall autosomal mCAs and CN-LOHs across age in two major lung cancer subtypes, lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC). (E) Arm-level autosomal mCAs enriched in lung cancer cases. Significantly enriched mCAs were marked in red. (F) The top 10 most enriched cancer genes in each type of mCAs. Significantly enriched genes were highlighted in a deeper color.

Smokers have a higher rate of ChrX mCAs and mChrY loss

In addition to age and lung cancer status, other clinical factors were also found to be associated with the presence of mCAs in blood cells (Table S1). Specifically, we found that smoking females are 42% more likely to harbor ChrX mCAs in their blood cells than non-smoking females (p=0.01), which was mainly driven by mChrX loss (odds ratio=2.25, p=0.005). In males, smokers had a significantly higher fraction of mChrY loss (odds ratio=2.27, p=1e-12) compared with non-smokers (Table S1). The age-dependent increase of ChrX mCA and mChrY loss for smokers and non-smokers was demonstrated in Fig. 4A. As shown, the fraction of smokers with ChrX mCA and mChrY loss increased with age at a faster rate than non-smokers, especially for mChrY loss. The higher mCA rate of smokers was also shown in Fig. 4B with a significant difference observed for mChrY and mChrX loss. When smokers were further divided into current- and ever-smokers and compared with never-smokers, similar results were observed: the rate of mChrX and mChrY loss

were significantly higher in both current-smokers and ex-smokers than in non-smokers (Supplementary Table S2). Interestingly, while we did not observe any correlation between overall smoking status and autosomal mCAs, current smokers tend to have more autosomal mCAs than ex-smokers (odds ratio=1.16, p=0.043, Supplementary Table S2). A similar trend was also observed in mChrY losses (odds ratio=1.68, p=2.6e-14, Supplementary Table S2), but not found in ChrX mCAs. These results suggested autosomes and ChrY may be more vulnerable to recent smoking harms.

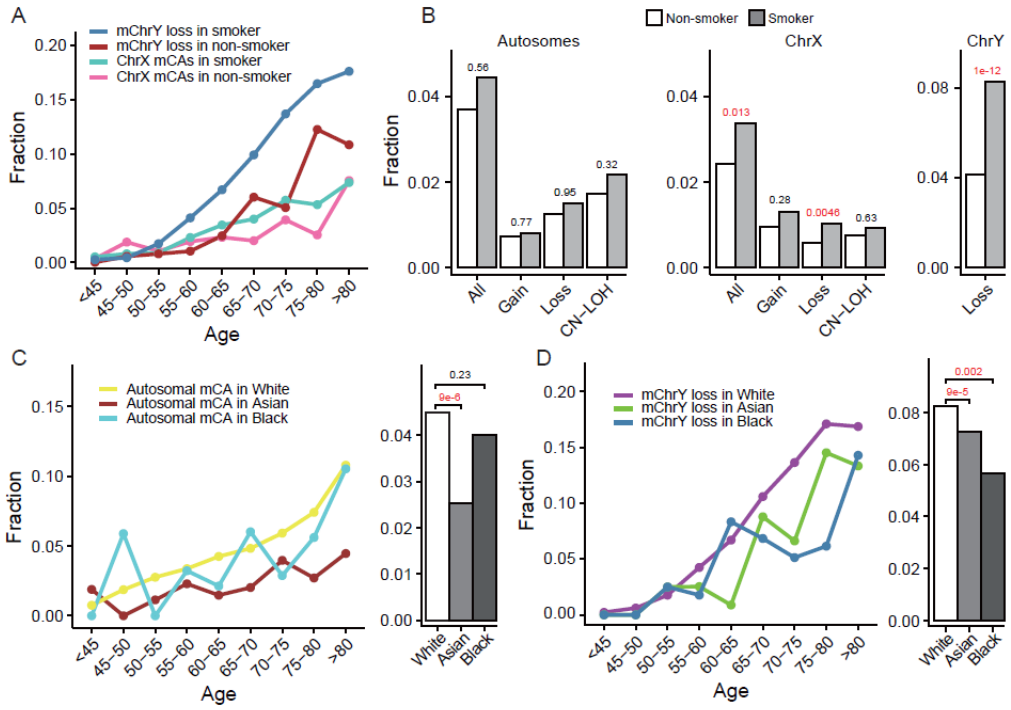

Figure 4. Association of mCAs with smoking status and racial disparity in mCAs. (A) Distribution of mChrX and mChrY losses across age in smokers and non-smokers. (B) Smokers show a significantly higher rate of overall ChrX mCAs (mainly losses) in females and mChrY losses in males. (C-D) Racial difference in the rate of autosomal mCAs and mChrY losses. Asians tend to have less autosomal mCAs but more mChrY losses compared with Whites, while Blacks tend to have less mChrY losses.

Racial disparities in the rate of mCAs

We also observed racial differences in the rate of mCA according to the logistic regression analysis (Model I) (Table S1). Specifically, Asians tended to have a lower rate of autosomal mCAs (odds ratio = 0.46, p=9e-6, Fig. 4C), ChrX mCAs (odds ratio = 0.47, p=0.03) and mChrY loss (odds ratio = 0.57, p=9e-5, Fig. 4D) compared to Whites. In addition, Blacks have a significantly lower rate of mChrY loss than Whites (odds ratio=0.55, p=0.002, Fig. 4D), but no significant difference in the rate of autosomal or ChrX mCAs (Fig. 4C). Of note, the significantly lower rate of mChrY loss in Asians and Blacks compared to Whites is consistent with a previous study based on the UKBB data [36].

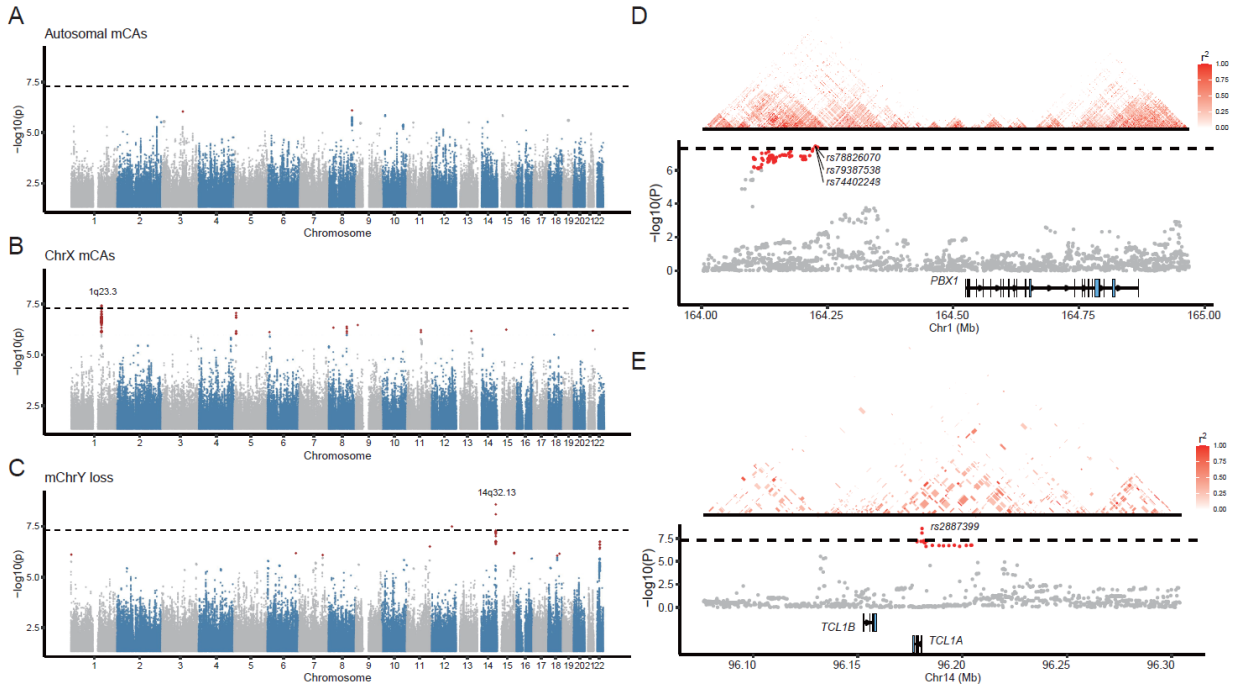

Fig 5. Genetic variants associated with mCA phenotypes. (A-C) Genetic variants associated with autosomal mCAs, ChrX mCAs, and mChrY losses. The dashed line indicates p-value cutoff 5e-8. Genetic variants with p<1e-6 were marked in red. (D-E) The nearest protein coding genes for loci Chr1q23.3 and Chr14q32.13, respectively. Variants with the lowest p-values in each locus were labeled. Heatmaps indicate the pairwise LD r^2 score between variants.

Genetic variants associated with mCA phenotypes

We performed genome-wide association analysis to identify genetic variants associated with the presence of different types of mCA events. At the significance level of p<5e-8, we do not identify any genetic loci that are associated with the presence of autosome mCA events (Fig. 5A). However, we did find that a locus on Chr1q23.3 is significantly associated with the presence of ChrX mCAs events (Fig. 5B), while a locus on Chr14q32.13 is significantly associated with mChrY loss (Fig. 5C). These results suggest that the occurrence of autosome mCAs might be a complex phenotype with different genetic loci contributing to mCAs of different types or different chromosomes. In contrast, the mCAs on sex chromosomes are relatively simple phenotypes, but ChrX mCAs and mChrY loss seem to be controlled by different genetic loci, as also revealed in previous studies [12,25,37]. Particularly, the Chr1q23.3 locus located at ~300kb upstream of *PBX1* gene (Fig. 5D), a cancer hallmark gene which is associated with leukemia [38], non-small cell lung cancer [39] and breast cancer [40]. In addition, the link between Chr14q32.13 locus and mChrY loss has also been identified from independent datasets, with the most significant variant rs2887399 maps to the 5' end of the *TCL1A* gene (Fig. 5E) [25,37]. In addition, we divided autosomal and ChrX mCAs into Gains, Losses, and CN-LOHs, and determined genetic variants associated with these

more specific mCA phenotypes. We identified several loci associated with mosaic autosomal Gains (Chr3p23), ChrX Gains (Chr3q29), and ChrX CN-LOHs (Chr11p15.5) (Fig. S2A and Supplementary Table S3). All the significant variants of locus Chr3p23 are located in the intronic region of *OSBPL10* (Fig. S2B). Circular RNAs derived from *OSBPL10* were found correlated with cell proliferation in cervical and gastric cancers [41,42]. The nearest gene of significant variants at locus Chr3q29 is *XXYLT1* (Fig. S2C), which has been found associated with lung cancer by GWAS [43]. Interestingly, the most significant variant rs76313919 at Chr11p15.5 maps to 5' end of *MOB2* (Fig. S2D), a gene involved in DNA damage response and cell cycle regulation [44].

Discussion

In this study, we investigated the association between mCAs and lung cancer risk using the OncoArray dataset generated by the INTEGRAL-ILCCO cohort. As the largest lung cancer genetics cohort, this dataset contains 18,221 lung cancer cases and 14,825 non-cancer controls. We identified a compressive list of mCAs, including mosaic autosomal/ChrX gain, loss, and CN-LOH as well as mChrY loss. Our analysis indicated that the presence of mCAs was associated with increase lung cancer risk, which was driven by the autosomal CN-LOH events. Stratified analysis confirmed this association was significant in both lung adenosarcoma and squamous lung cancer subjects.

Using the same pipeline, we identified more mCAs in ChrX (with a rate 3.6% in females) than in each individual autosomal chromosomes (with an average rate of 0.25% in all subjects). A similar observation has been reported in previous studies [12,45]. Moreover, ChrX mCAs are more likely to be a whole-chromosome event compared to autosomal mCAs (67.5% vs. 8.2%), suggesting a potential mechanistic difference between the two types of mCAs. While ChrX is a large chromosome and hosts many housekeeping genes, only one copy is active and transcribed in females. Most genes on the inactivated copy of ChrX are packed into heterochromatin, which is not active for transcription. As such, alterations on the ChrX might be less harmful and more likely to accumulate in blood cells than those on autosomal chromosomes. As a matter of fact, it has been experimentally shown that genomic alterations on the inactive ChrX were more likely to be accumulated in the blood [45]. In addition, some genomic alterations on ChrX may contribute to the clonal fitness of the host blood cells, which increases their chance to be detected as mCAs [12,46,47].

This study confirmed previous reports on the association between mChrY loss and smoking staus [16,26,37]. Interestingly, our analysis also revealed a significant association between ChrX mCAs and smoking status. Specifically, smokers had a significantly higher rate of mChrX

loss, but such a correlation was not detected for autosomal mCAs. Association between mChrY loss and lung cancer risk has been investigated in previous studies but reported contradictory results. Qin *et al.* reported that mChrY loss was associated with reduced lung cancer risk in non-smoking Chinese[15]. On the contrary, using the UKBB data Loftfield *et al.* found that individuals with mChrY loss in a high fraction of blood cells were more likely to have lung cancer [24]. As shown in Table 2, no significant association between mChrY loss and lung cancer was observed in the OncoArray data. We also stratified samples based on the blood cell fraction of mCAs using the same threshold setting with Loftfild et al [24], but did not identify the association in either group (Table S4). Stratified analysis based on smoking status indicated a protective effect of mCAs in current smokers but not in non-smokers (Table S4).

GWAS analyses failed to identify genetic loci associated with overall autosome mCA phenotype but identified different genetic loci linked with ChrX mCAs and mChrY loss. Particularly, we verified in our cohort the previously reported association between Chr14q32.13 and mChrY loss [25,37]. In another study, Loh *et al.* performed GWAS to investigate different mCA phenotypes using the UKBB data [12]. Similar to our results, no genetic variants were found to be associated with the overall autosome mCA phenotype, but they identified two genetic loci (SP140L locus on Chr2q37.1 and HLA locus on Chr6p21.33) linked with mChrX losses. While these two loci were not identified in our analysis, we uncovered several genetic loci associated with ChrX mCAs (Chr1q23.3), ChrX Gains (Chr3q29) and ChrX CN-LOHs (Chr11p15.5), respectively. Altogether, our and previous studies may suggest the following insights on genetic regulation of mCAs: i) the autosome and sex chromosome mCAs might be affected by different genetic factors, ii) the overall autosome mCA may be a more complex phenotype compared with ChrX mCA and mChrY loss phenotypes, and iii) the ChrX mCA and mChrY loss phenotypes are linked with different genetic loci.

In summary, we performed a systematic analysis to identify different types of mCAs and investigated their association with lung cancer risk while adjusting for clinical factors. By using the large cohort data from INTEGRAL-ILCCO, our analysis confirmed previously reported associations between mCAs and clinical factors (e.g., age and smoking status). Moreover, we revealed a significant association between mCAs and increased lung cancer risk in both lung adenocarcinoma and squamous lung cancers.

Methods and Materials

The OncoArray data from the INTEGRAL-ILCCO cohort

The OncoArray study is a major part of the INTEGRAL-ILCCO cohort, which provides high-quality genotyping array data and clinical information for a total of 33,046 subjects, 18,221 lung cancer cases and 14,825 controls without lung cancer diagnosis. All of the blood samples were collected before lung cancer diagnosis. The genotyping data were generated by using the Infinium OncoArray-500K BeadChip (Illumina, San Diego, CA) platform, which contains a total of 533,631 customized SNPs for studying cancer genetics [19]. The clinical information includes age, sex, race, smoking status, and lung cancer histological subtype. The OncoArray study has been approved by the institutional review board of all sites accruing participants.

Genotyping using the Oncotype platform

Genotyping and data processing were described by the previous studies [17–19]. Briefly, for the SNP array genotype data, DNA extracted from peripheral white blood cells was genotyped using the OncoArray microarray. We converted all the genotyping intensity files to VCF files with a BCFtools plugin gtc2vcf (https://github.com/freeseek/gtc2vcf). Samples with abnormal heterozygosity rate, sex discordance, <95% completion rates, and unexpected relatedness (identity-by-state > 10%) were discarded.

Identification of autosomal mCAs

We followed the methods of Loh *et al.* [4,12] to detect mosaic chromosomal alterations. Unphased VCF files were firstly split by chromosomes, then we phased each single-chromosome VCF file by SHAPEIT4 [48] with default parameters. The phased output and unphased ChrY data were then concatenated into a single VCF file. We applied a MOsaic CHromosomal Alterations (MoChA) caller to detect mCAs with either B Allele Frequency (BAF) and Log R Ratio (LRR) or allelic depth (AD), with default parameters [4,12]. The highly polymorphic MHC (chr6:27486711-33448264) and KIR (chr19:54574747-55504099) regions were excluded from mCA calling. We then applied a series of filters to exclude potential constitutional duplications and low quality mCA calls. Constitutional duplications have expected deviations in allelic balance (|ΔBAF|) = 1/6, with corresponding LRR ≈ 0.36 [12]. In order to exclude possible constitutional duplications, for mCA events of length > 10 Mb, we excluded events with LRR > 0.35 or with LRR within [0.2, 0.35] and |ΔBAF| > 0.16; for mCA events of length < 10 Mb, we excluded events with LRR > 0.2 or with LRR within [0.1, 0.2] and |ΔBAF| > 0.1. MoChA used a hidden Markov model (HMM) to detect mCAs either based on LRR and BAF or phased BAF (pBAF). LOD scores were used as the measurement of calling quality for model based on LRR and BAF (lod_lrr_baf) or for model based on pBAF (lod_baf_phase). To exclude low-quality mCA calls, we required either lod_lrr_baf or lod_baf_phase to be larger than 10 for mCA events of length > 2 Mb. For mCA events < 2 Mb we required lod_baf_phase > 30 and lod_lrr_baf > 10. In addition, a high-frequency reversion was

found in Chr17q21 [49], which could cause intensively low heterozygosity and induce false calling results. Thus, we removed the mCA events overlapped with Chr17 42-47Mb.

Identification of ChrX mCAs and mChrY losses

The mCAs associated with ChrX were also identified by MoChA. We only identified mCAs in female subjects because MoChA can only call mCAs on diploid homologous chromosome regions. In principle, we can apply MoChA and use the intensities of SNPs located in the pseudo-autosomal regions on sex chromosome (PAR1 and PAR2) to identify ChrX and ChrY mCAs in male subjects. However, the OncoArray genotyping platform contains only a small number of variants (28 SNPs in PAR1 and 1 SNP in PAR2) in the two PARs, which limited the ability of MoChA for phase inference and ChrX/ChrY mCA detection in our male subjects.

Previous studies have reported frequent mosaic loss of ChrY in males, which has been associated with lung cancer [15,16]. We therefore identified mChrY losses in our male subjects by using the method proposed in previous studies [21,25,26]. Briefly, the LRR on non-PAR regions of ChrY was calculated and those with ChrY LRR lower than -0.15 were identified as mChrY loss according to the references [21,37].

Determination of whole-chromosome and arm-level mCAs

We manually inspected the distribution of mCAs on chromosome arms. In autosomes, the vast majority of mosaic gain events were whole-chromosomal, while loss and CN-LOH might only occur at one arm of the chromosome. Thus, we divided autosomal mCAs into five categories: gain (+), loss on short arm (p-) and long arm (q-), CN-LOH on short arm (p=) and long arm (q=). Mosaic ChrX gains, losses and CN-LOHs were not divided into chromosome arm level categories, because most of ChrX mCAs covered nearly the whole chromosome. Altogether, this classification resulted in 103 types of mCA at the whole-chromosome or arm level. We tested the significance of co-occurrence between two mCA events by using the Fisher's exact test. Co-occurred mCA pairs in at least three subjects with an FDR<0.05 were highlighted in the co-occurrence graph.

Multivariable regression model for determine the association of clinical variables with mCAs

To determine the association between clinical variables and mCAs, we constructed a multivariable logistic regression model as the following:

$$Logit(mCA) \sim Age + Sex + Race + Smoking + LungCancer \qquad \text{(Model I)}$$

In the model, the response variable *mCA* is set as binary with 1 indicating the presence of mCAs in a subject, and 0 otherwise. In the independent variables, *Age* is represented as a continuous variable; *Sex* is set 1 for males and 0 for females; *Smoking* is set to 1 for current/ever-smokers and 0 for never-smokers; *Race* is a categorical variable with White as the baseline; and *LungCancer* is set to 1 for lung cancer cases and 0 for controls. The model was separately applied to the 3 mCA types: autosomal mCA, ChrX mCA, and ChrY loss. Of note, only female subjects were used for ChrX mCA analysis and male subjects for mChrY loss analysis, with the "*Sex*" variable removed from the model. The autosomal and ChrX mCAs were further divided into 3 subtypes: gain, loss, and CN-LOH.

Multivariable regression model for determine the contribution of mCAs to lung cancer risk

To quantify the contribution of mCAs to the risk of lung cancer while adjusting for key confounding variable, we constructed the following model:

$$Logit(LungCancer) \sim mCA + Age + Sex + Race + Smoking \qquad \text{(Model II)}$$

The variables were defined in the same way as Model I. In the primary analysis, the model was applied to all lung cancer cases and non-cancer controls. In stratified analysis, the model was applied to three major lung cancer histological subtypes, LUAD, LUSC and SCLC. For each subtype, all non-cancer controls were included in the model for estimating coefficient and significance.

Genetic variants associated with mCA phenotypes

Prior to GWAS analysis, genotype imputation was performed for all subjects in our cohort by using 32,470 reference samples from the Haplotype Reference Consortium (HRC) [50]. Low quality variants and subjects were then filtered out following the method described in Byun *et al.* [51]. To minimize the bias from genetic structure, we only include White/Caucasian subjects in the association analyses. Rare variants with minor allele frequency (MAF) ≤ 1% were excluded from the analysis. For each variant, we separately performed the Hardy-Weinberg equilibrium (HWE) test in lung cancer patients and controls. The variants that significantly deviated from HWE (p-value < 5e-8, Chi-square test) in either lung cancer patients or controls were then excluded. We applied a logistic regression model to identify genetic variants associated with each category of mCA events. Present of mCAs (with/without mCA) in each subject was regarded as the dependent variable and genotype of each SNP as independent variables. Sex, age, lung cancer status, smoking and the first three principal components were included in the model as covariates. We calculated the correlation between mCA status and each SNP by the "glm" option of plink 2.0

[52]. To improve the statistical power, we required the sample size for each genotype ≥ 3 and the total sample size ≥ 30. The cutoff of p-value was set to 5e-8 [18].

REFERENCES

1. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
2. Liu, X., Kamatani, Y. & Terao, C. Genetics of autosomal mosaic chromosomal alteration (mCA). *J. Hum. Genet.* **66**, 879–885 (2021).
3. Guo, X. *et al.* Mosaic loss of human Y chromosome: what, how and why. *Hum. Genet.* **139**, 421–446 (2020).
4. Loh, P.-R., Genovese, G. & McCarroll, S. A. Monogenic and polygenic inheritance become instruments for clonal selection. *Nature* **584**, 136–141 (2020).
5. Terao, C. *et al.* Chromosomal alterations among age-related haematopoietic clones in Japan. *Nature* **584**, 130–135 (2020).
6. Sano, S. *et al.* Hematopoietic loss of Y chromosome leads to cardiac fibrosis and heart failure mortality. *Science* **377**, 292–297 (2022).
7. Sherman, M. A. *et al.* Large mosaic copy number variations confer autism risk. *Nat. Neurosci.* **24**, 197–203 (2021).
8. Zekavat, S. M. *et al.* Hematopoietic mosaic chromosomal alterations increase the risk for diverse types of infection. *Nat. Med.* **27**, 1012–1024 (2021).
9. Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* **44**, 642–650 (2012).
10. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44**, 651–658 (2012).
11. Niroula, A. *et al.* Distinction of lymphoid and myeloid clonal hematopoiesis. *Nat. Med.* **27**, 1921–1927 (2021).
12. Loh, P.-R. *et al.* Insights into clonal haematopoiesis from 8,342 mosaic chromosomal alterations. *Nature* **559**, 350–355 (2018).
13. Kar, S. P. *et al.* Genome-wide analyses of 200,453 individuals yield new insights into the causes and consequences of clonal hematopoiesis. *Nat. Genet.* **54**, 1155–1166 (2022).
14. Coombs, C. C. *et al.* Therapy-related clonal hematopoiesis in patients with non-hematologic cancers Is common and associated with adverse clinical outcomes. *Cell Stem Cell* **21**, 374-382.e4 (2017).
15. Qin, N. *et al.* Association of mosaic loss of chromosome Y with lung cancer risk and prognosis in a Chinese population. *J. Thorac. Oncol.* **14**, 37–44 (2019).
16. Thompson, D. J. *et al.* Genetic predisposition to mosaic Y chromosome loss in blood. *Nature* **575**, 652–657 (2019).
17. Byun, J. *et al.* Trans-ethnic genome-wide meta-analysis of 35,732 cases and 34,424 controls identifies novel genomic cross-ancestry loci contributing to lung cancer

susceptibility. *medRxiv* 2020.10.06.20207753 (2020) doi:10.1101/2020.10.06.20207753.

18. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**, 1126–1132 (2017).
19. Amos, C. I. *et al.* The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prev.* **26**, 126–135 (2017).
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. Loftfield, E. *et al.* Predictors of mosaic chromosome Y loss and associations with mortality in the UK Biobank. *Sci. Rep.* **8**, 12316 (2018).
22. Hirata, T. *et al.* Investigation of chromosome Y loss in men with schizophrenia. *Neuropsychiatr. Dis. Treat.* **14**, 2115–2122 (2018).
23. Graham, E. J. *et al.* Somatic mosaicism of sex chromosomes in the blood and brain. *Brain Res.* **1721**, 146345 (2019).
24. Loftfield, E. *et al.* Mosaic Y loss is moderately associated with solid tumor risk. *Cancer Res.* **79**, 461–466 (2019).
25. Wright, D. J. *et al.* Genetic variants associated with mosaic Y chromosome loss highlight cell cycle genes and overlap with cancer susceptibility. *Nat. Genet.* **49**, 674–679 (2017).
26. Dumanski, J. P. *et al.* Smoking is associated with mosaic loss of chromosome Y. *Science* **347**, 81–83 (2015).
27. Kou, F., Wu, L., Ren, X. & Yang, L. Chromosome abnormalities: new insights into their clinical significance in cancer. *Mol. Ther. - Oncolytics* **17**, 562–570 (2020).
28. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
29. Loh, M. L. *et al.* Mutations in CBL occur frequently in juvenile myelomonocytic leukemia. *Blood* **114**, 1859–1863 (2009).
30. Niemeyer, C. M. *et al.* Germline CBL mutations cause developmental abnormalities and predispose to juvenile myelomonocytic leukemia. *Nat. Genet.* **42**, 794–800 (2010).
31. Westphal, C. H. *et al.* Genetic interactions between atm and p53 influence cellular proliferation and irradiation-induced cell cycle checkpoints. *Cancer Res.* **57**, 1664–1667 (1997).
32. Ong, D. C. T. *et al.* LARG at chromosome 11q23 has functional characteristics of a tumor suppressor in human breast and colorectal cancer. *Oncogene* **28**, 4189–4200 (2009).
33. Gai, M., Bo, Q. & Qi, L. Epigenetic down-regulated DDX10 promotes cell proliferation through Akt/NF-κB pathway in ovarian cancer. *Biochem. Biophys. Res. Commun.* **469**,

1000–1005 (2016).
34. Phan, R. T. & Dalla-Favera, R. The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells. *Nature* **432**, 635–639 (2004).
35. Ngan, E. *et al.* LPP is a Src substrate required for invadopodia formation and efficient breast cancer lung metastasis. *Nat. Commun.* **8**, 15059 (2017).
36. Lin, S. H. *et al.* Mosaic chromosome Y loss is associated with alterations in blood cell counts in UK Biobank men. *Sci. Rep.* **10**, 2–11 (2020).
37. Zhou, W. *et al.* Mosaic loss of chromosome Y is associated with common variation near TCL1A. *Nat. Genet.* **48**, 563–568 (2016).
38. Shimabe, M. *et al.* Pbx1 is a downstream target of Evi-1 in hematopoietic stem/progenitors and leukemic cells. *Oncogene* **28**, 4364–4374 (2009).
39. Mo, M.-L. *et al.* Detection of E2A-PBX1 fusion transcripts in human non-small-cell lung cancer. *J. Exp. Clin. Cancer Res.* **32**, 29 (2013).
40. Ao, X. *et al.* PBX1 is a valuable prognostic biomarker for patients with breast cancer. *Exp. Ther. Med.* **20**, 385–394 (2020).
41. Yang, S. *et al.* FOXA1-induced circOSBPL10 potentiates cervical cancer cell proliferation and migration through miR-1179/UBE2Q1 axis. *Cancer Cell Int.* **20**, 389 (2020).
42. Wang, S. *et al.* Circular RNA profile identifies circOSBPL10 as an oncogenic factor and prognostic marker in gastric cancer. *Oncogene* **38**, 6985–7001 (2019).
43. Yoon, K.-A. *et al.* A genome-wide association study reveals susceptibility variants for non-small cell lung cancer in the Korean population. *Hum. Mol. Genet.* **19**, 4948–4954 (2010).
44. Gomez, V. *et al.* Regulation of DNA damage responses and cell cycle progression by hMOB2. *Cell. Signal.* **27**, 326–339 (2015).
45. Machiela, M. J. *et al.* Female chromosome X mosaicism is age-related and preferentially affects the inactivated X chromosome. *Nat. Commun.* **7**, 11843 (2016).
46. Skowyra, A., Allan, L. A., Saurin, A. T. & Clarke, P. R. USP9X limits mitotic checkpoint complex turnover to strengthen the spindle assembly checkpoint and guard against chromosomal instability. *Cell Rep.* **23**, 852–865 (2018).
47. Dunford, A. *et al.* Tumor-suppressor genes that escape from X-inactivation contribute to cancer sex bias. *Nat. Genet.* **49**, 10–16 (2017).
48. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
49. Alves, J. M. *et al.* Reassessing the evolutionary history of the 17q21 inversion polymorphism. *Genome Biol. Evol.* **7**, 3239–3248 (2015).
50. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat.*

Genet. **48**, 1279–1283 (2016).

51. Byun, J. *et al.* Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. *Nat. Genet.* **54**, 1167–1177 (2022).
52. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

Supplementary Figures

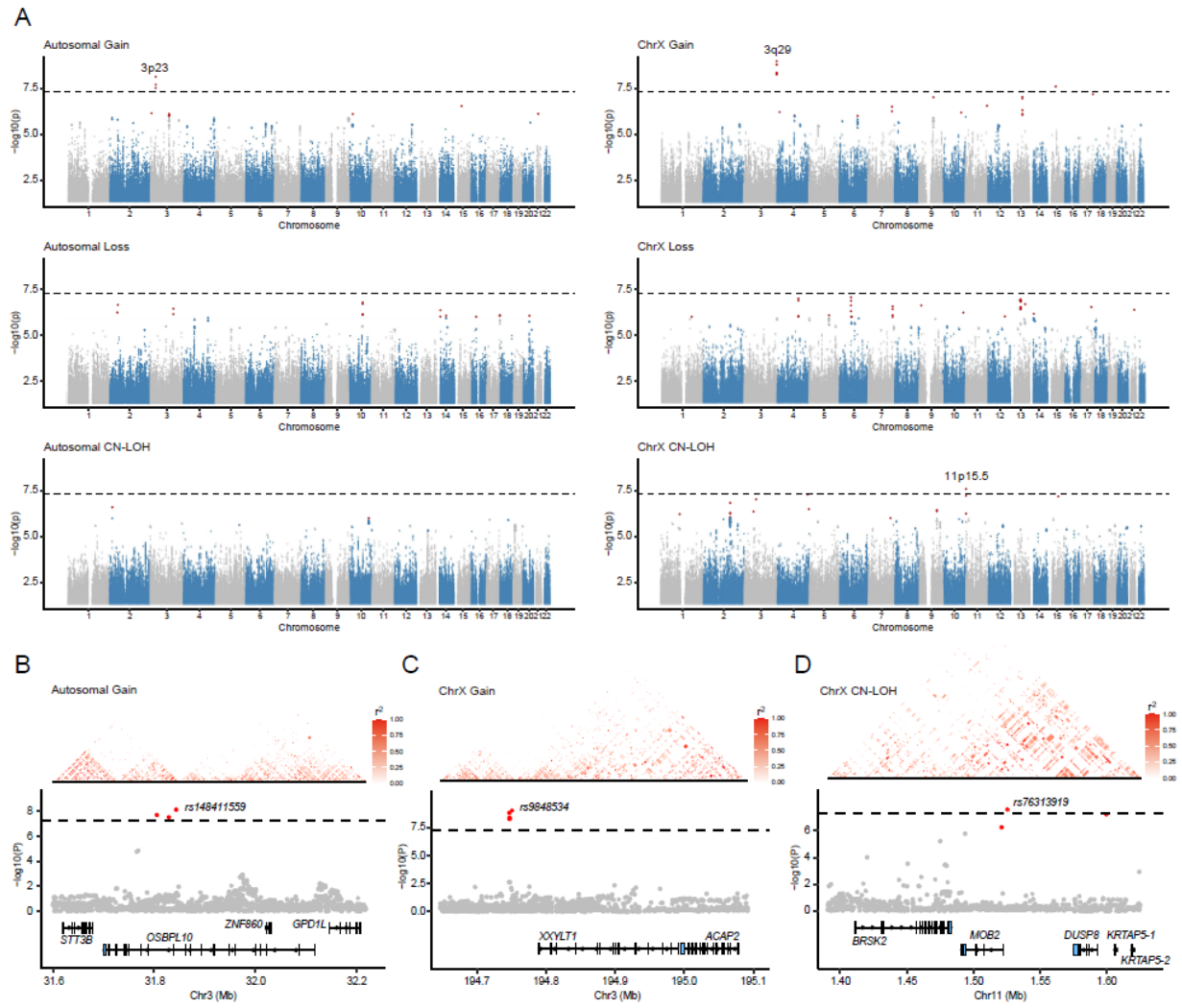

Figure S1. Manhattan plot of genetic variants associated with mosaic gain, loss and CN-LOH in autosomes and ChrX. A) Dashed line indicated p-value cutoff 5e-8. Genetic variants with p-value<1e-6 were marked in red. B-D) show the nearest protein coding genes of loci Chr3p23, Chr3q29 and Chr11p15.5, respectively. Variants with the lowest p-values in each locus were labeled. Heatmaps indicate the pairwise LD r^2 score between variants.

Supplementary Tables

Table S1. The associations between different types of mCA and clinical phenotypes. Results are based on the logistic regression model I. Significant associations were highlighted in bold.

mCA		Age		Sex		Lung Cancer		Smoking		Race-Asian		Race-Black		Race-Other	
Chr.	*Type*	*Coef.*	*P*-value	*Coef.*	*P*-value	*Coef.*	*P*-value	*Coef.*	*P*-value	*Coef.*	*P*-value	*Coef.*	*P*-value	*Coef.*	*P*-value
Autosome	All	**0.045**	**7.8e-43**	0.11	0.085	**0.28**	**2e-5**	0.05	>0.1	**-0.77**	**9.3e-6**	-0.21	>0.1	-0.15	>0.1
	Gain	**0.054**	**1.9e-12**	**0.35**	**0.025**	0.093	>0.1	-0.055	>0.1	**-1.3**	**9.6e-3**	0.19	>0.1	0.079	>0.1
	Loss	**0.054**	**7.2e-22**	**0.28**	**0.014**	**0.23**	**0.039**	-0.0091	>0.1	**-0.56**	**0.033**	-0.68	0.06	-0.44	>0.1
	CN-LOH	**0.038**	**2.2e-16**	-0.04	>0.1	**0.36**	**1.4e-4**	0.12	>0.1	**-0.84**	**1.5e-3**	-0.083	>0.1	0.093	>0.1
ChrX	All	**0.058**	**3.2e-20**	-	-	-0.057	>0.1	**0.35**	**0.013**	**-0.74**	**0.027**	0.15	>0.1	0.31	>0.1
	Gain	**0.045**	**1.7e-6**	-	-	-0.024	>0.1	0.23	>0.1	**-1.4**	**0.048**	0.42	>0.1	0.42	>0.1
	Loss	**0.07**	**3.2e-9**	-	-	-0.12	>0.1	**0.81**	**4.6e-3**	-0.59	>0.1	-0.44	>0.1	-0.89	>0.1
	CN-LOH	**0.058**	**5.3e-7**	-	-	0.075	>0.1	0.12	>0.1	-0.4	>0.1	0.51	>0.1	0.59	>0.1
ChrY	Loss	**0.079**	**1.8e-112**	-	-	-0.0066	>0.1	**0.82**	**1.2e-12**	**-0.56**	**9e-5**	**-0.59**	**0.002**	**-0.52**	**0.01**

Table S2. The associations between different types of mCA and smoking status while adjusting for age, sex, race, and lung cancer. Results are based on the logistic regression model I. Significant associations were highlighted in bold.

mCA		Current-smoker vs Non-smoker		Ever-smoker vs Non-smoker		Current-smoker vs Ever-smoker	
Chr.	*Type*	*Coef.*	*P*	*Coef.*	*P*	*Coef.*	*P*
Autosome	All	**0.29**	**1e-5**	**0.33**	**7.4e-5**	**0.3**	**1.8e-3**
	Gain	0.11	>0.1	-0.046	>0.1	0.15	>0.1
	Loss	**0.24**	**0.03**	0.22	>0.1	0.22	>0.1
	CN-LOH	**0.36**	**1e-4**	**0.43**	**2.1e-4**	**0.34**	**0.012**
ChrX	All	-0.035	>0.1	0.17	>0.1	-0.029	>0.1
	Gain	-0.0091	>0.1	0.16	>0.1	-0.15	>0.1
	Loss	-0.081	>0.1	0.2	>0.1	-0.055	>0.1
	CN-LOH	0.1	>0.1	0.39	>0.1	0.29	>0.1
ChrY	Loss	0.024	>0.1	0.11	>0.1	0.014	>0.1

Table S3. Genetic variants associated with autosomal mCA, ChrX mCA, and mChrY loss phenotypes. Genetic variants with p-value<1e-6 were included.

(not shown due to large file size)

Table S4. The association of mChrY loss with lung cancer when stratified by cell fraction and smoking status. Results are based on the logistic regression model II, adjusting for age, sex, race and smoking status (removed when stratified by smoking status). Significant associations were highlighted in bold.

Group		*Coef.*	*P*-value
Cell fraction of mChrY loss	High cell fraction vs No mChrY loss	0.0075	>0.1
	Low cell fraction vs No mChrY loss	0.011	>0.1
Smoking status	All	0.024	>0.1
	Non-smoker	0.35	>0.1
	Smoker	-0.0045	>0.1
	Current-smoker	**-0.21**	**0.017**
	Ever-smoker	0.1	>0.1