Genetic variation associated with chromosomal aberration frequency: A genome-wide association study

Running title: Genetic variation associated with CA frequency

Key words: Chromosome type aberrations, Chromatid type aberrations, GWAS, Single-nucleotide polymorphism

Yasmeen Niazi (1,14)#*; Hauke Thomsen (1)#; Bozena Smolkova (2); Ludmila Vodickova (3,4,5); Sona Vodenkova (3,4,6); Michal Kroupa (3,5); Veronika Vymetalkova (3,4); Alena Kazimirova (7); Magdalena Barancokova (7); Katarina Volkovova (7); Marta Staruchova (7); Per Hoffmann (8,9); Markus M. Nöthen (8,10); Maria Dusinska (11); Ludovit Musak (12); Pavel Vodicka (3,4,5); Kari Hemminki (1,13); Asta Försti (1,13)

1 Department of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany

2 Department of Molecular Oncology, Cancer Research Institute, Biomedical Research Center, Slovak Academy of Sciences, Dubravska cesta 9, 84505 Bratislava, Slovakia

3 Department of Molecular Biology of Cancer, Institute of Experimental Medicine, The Czech Academy of Sciences, Videnska 1083, 142 00 Prague, Czech Republic

4 Institute of Biology and Medical Genetics, First Faculty of Medicine, Charles University, Albertov 4, 128 00 Prague, Czech Republic

5 Biomedical Centre, Faculty of Medicine in Pilsen, Charles University in Prague, Pilsen, Czech Republic

6 Department of Medical Genetics, Third Faculty of Medicine, Charles University, Prague, Czech Republic

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

7 Department of Biology, Faculty of Medicine, Slovak Medical University, Limbova 12, 833 03 Bratislava, Slovakia

8 Institute of Human Genetics, University of Bonn, D-53127 Bonn, Germany

9 Division of Medical Genetics, Department of Biomedicine, University of Basel, 4003 Basel, Switzerland

10 Department of Genomics, Life & Brain Center, University of Bonn, D-53127 Bonn, Germany

11 Health Effects Laboratory, Department of Environmental Chemistry, NILU-Norwegian Institute for Air Research, Instituttveien 18, 2007 Kjeller, Norway

12 Clinic of Occupational Medicine and Toxicology, Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava and University Hospital Martin, Kollarova 2, 03601 Martin, Slovakia

13 Center of Primary Health Care Research, Clinical Research Center, Lund University, Malmö, Sweden

14 Medizinische Fakultät, Universität Heidelberg, Im Neuenheimer Feld 672, 69120 Heidelberg

# Contributed equally

* Corresponding author:

Yasmeen Niazi,

Department of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 580, 69120 Heidelberg, Germany.

Phone: +49 6221 1805

Email: y.niazi@dkfz.de

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Abstract

Chromosomal aberrations (CAs) in human peripheral blood lymphocytes (PBL) measured with the conventional cytogenetic assay have been used for human biomonitoring of genotoxic exposure for decades. CA frequency in peripheral blood is a marker of cancer susceptibility. Previous studies have shown associations between genetic variants in metabolic pathway, DNA repair and major mitotic checkpoint genes and CAs. We conducted a genome-wide association study on 576 individuals from the Czech Republic and Slovakia followed by a replication in 2 different sample sets of 482 (replication 1) and 1288 (replication 2) samples. To have a broad look at the genetic susceptibility associated with CA frequency, the sample sets composed of individuals either differentially exposed to smoking, occupational/environmental hazards, or they were untreated cancer patients. Phenotypes were divided into chromosome and chromatid type aberrations (CSAs and CTAs, respectively) and total chromosomal aberrations (CAtot). The arbitrary cutoff point between individuals with high and low CA frequency was 2% for CAtot and 1% for CSA and CTA. The data were analyzed using age, sex, occupation/cancer and smoking history as covariates. Altogether 11 loci reached the p-value of 10^{-5} in the GWAS. Replication 1 supported the association of rs1383997 (8q13.3) and rs2824215 (21q21.1) in CAtot and rs983889 (5p15.1) in CTA analysis. These loci were found to be associated with genes involved in mitosis, response to environmental and chemical factors and genes involved in syndromes linked to chromosomal abnormalities. Identification of new genetic variants for the frequency of CAs offers prediction tools for cancer risk in future.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Introduction

Chromosomal aberrations (CAs) encompass structural and numerical chromosomal anomalies. Structural CAs include specific, recurrent deletions, translocations, and inversions that can only be detected by molecular cytogenetics such as fluorescent in situ hybridization and sequencing techniques [Albertini *et al.* 2000] CAs that can be cytologically distinguished at metaphase are non-specific and they can be divided into two main groups: chromosome type aberrations (CSAs) and chromatid type aberrations (CTAs) [Hagmar *et al.* 2001; Bignold 2009; Hemminki *et al.* 2015b; Heng *et al.* 2016]. CSAs arise mainly as a result of direct DNA damage during G_o/G_1 phase, by clastogens such as ionizing radiation and bleomycin. The lesions acquired during G_o/G_1 phase later result in the damage of both chromatids of a chromosome which may create di-centric and ring chromosomes [Albertini *et al.* 2000]. Apart from direct DNA damage, another important contributor to CSA frequency is telomere dysfunction. In somatic cells with critically short telomeres and low telomerase activity, telomeres shorten, become eroded and poorly end-capped. These eroded ends can be recognized by non-homologous end joining repair and become attached to non-homologous chromosomes resulting in fused, ring or fragmented chromosomes [Gostissa *et al.* 2011; Jones *et al.* 2012]. Association between relative telomere length and CA frequency, particularly that of CSAs, has been documented in our previous study [Hemminki *et al.* 2015a]. CTAs are a result of damage by environmental or chemical clastogens during S/G_2 phase or due to replication on a damaged DNA template and involve only one chromatid of a chromosome. Examples of CTAs are chromatid breaks and exchanges [Durante *et al.* 2013].

Conventional cytogenetic examination for CAs in individuals exposed to mutagens and potential carcinogens has been used for decades as a surveillance mechanism for genotoxic effect [Carrano and Natarajan 1988]. Many malignant and benign human tumors exhibit chromosomal abnormalities [Mitelman 2000] and an increase in the frequency of CAs has been found in the incident cancer patients thus closely linking CAs with cancer development [Vodenkova *et al.* 2015]. Some of the CAs observed are also generated during the course of cancer development,

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

nevertheless CA frequency in peripheral blood lymphocytes (PBLs) is considered to be an early marker of cancer susceptibility based on the hypothesis that genetic damage in PBLs reflects similar damage in other body cells undergoing carcinogenesis [Rossner *et al.* 2005].

Inter-individual variation in the frequency of CAs, both in unexposed and in exposed individuals, has raised the question of genetic predisposition to CAs. Studies exploring the genetic causes of increased CA frequency have mainly focused on mitotic checkpoint, DNA repair and metabolic genes and found different variants that are associated with the frequency of CAs [Hemminki *et al.* 2015b; Vodicka *et al.* 2015, 2018; Försti *et al.* 2016]. Despite these findings, there still is a great need to explore the genetic basis of CAs. To achieve this goal we designed a genome-wide association study (GWAS), which is the first GWAS of this nature. The study included not only individuals exposed to potential occupational and environmental carcinogens but also newly diagnosed cancer patients, who may represent a population with increased susceptibility to CAs, and individuals with no recorded exposure to carcinogens. Our aim was to find novel genetic variants predisposing to CAs and potentially to cancer and to elucidate the possible functional effects of these variants by *in silico* predictions.

Materials and Methods

All the samples and the information in the study were obtained with written consent of the participants. The project was carried out according to the rules of the Declaration of Helsinki and ethical approval was obtained from the Ethics Committee of the Institute of Preventive and Clinical Medicine (later Slovak Medical University), the Ethics Committee of the Jessenius Faculty of Medicine in Martin, Comenius University in Bratislava, the Ethics Committee of the Institute for Clinical and Experimental Medicine and Thomayer Hospital, Czech Republic and the Ethics Committee of the VFN (General University Hospital in Prague).

Study Subjects

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

The GWAS sample set consisted of 639 healthy individuals; approximately 89% of them were recruited in Slovakia, while 11% came from the Czech Republic. Blood samples were taken from approximately equal number of males and females. About 57% of these individuals were exposed to some form of genotoxins due to the nature of their professions and their exposure was assessed by personal dosimeters, the rest were office workers and local residents; 30% were self-reported smokers (table I). The genotoxic substances included mainly small organic compounds, such as vinyl chloride, epichlorohydrine and ethylene oxide, anesthetics, heavy metals and styrene [Vodicka *et al.* 2004a; b, 2015, Musak *et al.* 2008, 2013]. The replication was conducted on 2 different sample sets. The first replication set (replication 1) consisted of 482 individuals (Czech, n=449, Slovak, n=33) and the second set (replication 2) was composed of 1288 individuals (all Slovak), (table I). About 46% of the individuals in replication 1 were newly diagnosed primary cancer patients recruited by the Department of Radiotherapy and Oncology, Faculty Hospital Kralovske Vinohrady, Prague, Czech Republic, including breast, colorectal and lung cancer patients; 29% of the individuals in this sample set were self-reported smokers [Vodicka *et al.* 2010; Vodenkova *et al.* 2015]. Blood samples from the cancer patients were drawn before any treatment to avoid any treatment-related increase in the number of CAs. The healthy individuals were recruited by the Blood Center of Faculty Hospital Kralovske Vinohrady, Prague, Czech Republic.

Replication 2 included participants from six molecular-epidemiological studies. In these studies, the effects of environmental/lifestyle factors including smoking, alcohol consumption, nutrition, professional exposure (asbestos, stone wool, glass fibers and radiation), and risk factors such as obesity and aging on health outcomes were assessed by several biomarkers, including chromosomal instability. Altogether 23% of individuals were occupationally exposed with exposure assessed by personal dosimeters and 24% were self-reported smokers. The numbers of participants from individual studies are shown in table I; details of each study were published elsewhere [Dušinská *et al.* 2003, 2004a; b, 2012, Kažimírová *et al.* 2004, 2006, 2009; Tulinska *et al.* 2004; Szabová *et al.* 2012].

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Cytogenetic assay

Cytogenetic analysis was performed on cultured PBLs. Two short term PBL cultures were set up for each sample. For this purpose, 0.5 ml of whole blood was added to 4.5 ml of RPMI (Roswell Park Memorial Institute) medium along with L-glutamine and $NaHCO_3$ (Gibco) supplemented with 20% fetal calf serum (Gibco), and antibiotics (penicillin and streptomycin, Gibco). Phytohaemagglutinin (0.18 mg/ml, PHA, Murex) was added as a proliferation stimulant. Incubation was done at 37°C with 5% CO_2 for 48 hours. Colchicine (0.75 µg/ml, Sigma) was added 2 hours before harvesting the PBLs. Harvesting was followed by centrifugation of cells and hypotonic shock treatment in 0.075 M KCl for 20 min at 37 °C. PBLs were fixed twice in methanol:glacial acetic acid (3:1) and air-dried preparations were made. Staining of slides was carried out using 5% Giemsa-Romanowski solution for 5 min. Cultured PBLs were analyzed in metaphase stage under a light microscope [Dušinská *et al.* 2004b; Kažimírová *et al.* 2004; Musak *et al.* 2013]. For each person 100 mitoses were analyzed in a double blind fashion and the frequency of different types of CAs (CSAs, CTAs) was recorded [Vodicka *et al.* 2010; Musak *et al.* 2013].

Phenotypes analyzed in the association analysis were divided into three categories: CAtot (total chromosomal aberrations), CSAs, and CTAs. For logistic regression analysis the samples were divided into high CA frequency group (CA_{high}) and low CA frequency group (CA_{low}) on the basis of frequency of aberrations. The threshold for inclusion into CA_{high} in case of CAtot was CA frequency ≥2% while for CSA and CTA it was ≥1%. This arbitrary assignment to CA_{high} and CA_{low} groups is based on previous experience with human genotoxic monitoring in the Czech and Slovak populations [Dušinská *et al.* 2003, 2004a; b; Šrám *et al.* 2004; Vodicka *et al.* 2010; Musak *et al.* 2013].

GWAS and quality control

Genotyping of the 639 individuals was done using Illumina HumanOmniExpressExome8v1.3 array comprising nearly 1 million SNPs throughout the genome. General genotyping quality control

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

assessment was done as previously described by Andersson *et al*. 2010 [Anderson *et al.* 2010]. Individuals with discordant gender information, outlying heterozygosity and genotype call rates <95% were excluded. Relatedness between samples was detected by identity-by-state measures. Population stratification was assessed using principal component analysis. SNPs with one or more of the following criteria were excluded: <95% genotype call rate, minor allele frequency <5% or Hardy–Weinberg equilibrium exact P-value $<10^{-5}$. After quality control, 576 samples and 626.004 SNPs remained. Genotypes for common variants across the genome were then imputed using data from the combined UK10K - 1000 Genomes Project (phase 3, Oct. 2014) with IMPUTE2 v2.3.2 [Howie *et al.* 2011] after pre-phasing with SHAPEIT software v2.12 [Delaneau *et al.* 2011]. We set thresholds for imputation quality to retain both potential common and rare variants for validation. Specifically, poorly imputed SNPs defined by an information metric I < 0.70 were excluded. All genomic locations are given in NCBI Build 37/UCSC hg19 coordinates. All SNPs having a MAF< 5% were excluded. After imputation, the SNP set consisted of 10.258.281 genotyped and imputed SNPs. This SNP set consisting of both genotyped and imputed SNPs was used for association analysis.

Association analysis

The consecutive association analysis of the GWAS was conducted by SNPTEST using univariate and multivariate logistic regression and linear regression models for each of the three phenotypes by including the relevant covariates, age, sex, occupational exposure and smoking status. The association data was visualized through Miami plots with generally accepted suggestive significance threshold of $P = 5.0 \times 10^{-5}$ and the genome-wide significance threshold of $P = 5.0 \times 10^{-8}$, using the Genetic analysis package (gap) for CRAN R 2.15 and odds ratios (ORs), effect sizes and 95% confidence intervals (CIs) were obtained for the effective allele in the additive model.

***In silico* analysis**

In silico analysis was done using different bioinformatics tools to examine functional consequences of the highly associated SNPs. These tools included Locus zoom to plot the locus of interest, to see the orientation of genes in the region, linkage disequilibrium between the SNPs and recombination

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

rate, [Pruim *et al.* 2010], UCSC genome browser [Rosenbloom *et al.* 2015] and Haploreg [Ward and Kellis 2012] to investigate the presence of any regulatory elements like promotors, enhancers and transcription factor binding sites, and to see the potential functions and expression effects from eQTL studies of all highly linked SNPs on candidate target genes. Regulome DB was used to predict the likely cell types of action, variant scores, regions of DNase hypersensitivity, and histone modifications [Boyle *et al.* 2012]. A total of eleven SNPs were selected for replication as a result of *in silico* analysis.

Validation and replication

The selected SNPs were validated in a small sample set of 149 individuals from the GWAS and replicated in 2 different replication sets. Validation and replication were carried out through TaqMan (Thermo Fisher Scientific, Darmstadt, Germany) allelic discrimination genotyping assays. Genotype detection was performed using ViiA 7 Real-Time PCR System (Thermo Fisher Scientific).

Statistical analysis

Post replication analysis was performed using PLINK v1.90b3.30 [Purcell *et al.* 2007] (http://pngu.mgh.harvard.edu/purcell/plink/). Effect sizes, 95% CIs and corresponding P-values were calculated by using logistic and linear regression models. All models were corrected for the same covariates as with the model above; in the analysis of replication 1, cancer status was included to the covariates. A meta-analysis for the GWAS and the two replication sets was performed using the GWAMA software [Mägi and Morris 2010]. Heterogeneity was assessed by the I^2 statistics (interpreted as low <0.25, moderate 0.50 and high > 0.75).

Results

The number of subjects in CA_{high} and CA_{low} for CAtot, CTA and CSA, and their distribution among the covariates (age, sex, smoking, occupational exposure and cancer status) for the three study sample sets (GWAS, replication 1 and replication 2) is summarized in table I. These covariates

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

were chosen because they were proven to exert a significant effect on CA frequency in previous studies [Vodicka *et al.* 2010; Hemminki *et al.* 2015b; Vodenkova *et al.* 2015]. Association of background variables of sex, age, smoking, occupational exposure and cancer status with CA frequency was tested with logistic regression model on CAtot. According to this analysis occupational exposure significantly influenced CA frequency in the GWAS ($p = 1.21 \times 10^{-9}$). In replication 1 the most significant variable affecting CA frequency was cancer status ($p = 7.56 \times 10^{-6}$) while in replication 2 the effect of occupational exposure was moderate ($p = 0.009$). The effect of age was moderate in the GWAS ($p = 0.01$) and replication 1 ($p = 0.01$) sample sets but significant in the replication 2 ($p = 2.59 \times 10^{-7}$). Smoking history had a significant association with CA frequency in the replication 1 ($p = 0.001$) and replication 2 ($p = 0.006$) but not in the GWAS ($p = 0.96$). Gender was not associated with CA frequency in the GWAS and replication 1 ($p = 0.57$ and 0.13 respectively) but it was moderately associated in replication 2 ($p = 0.02$). As shown in table II, there were some differences in the median, mean, minimum and maximum values of CAtot, CSA and CTA among the three data sets. In the GWAS and replication 1, the mean frequency of CAtot was about 2% and it was about 1% for CSA and CTA, while in replication 2, the frequencies were about 1% and 0.5%, respectively. Linear mixed model was also used to test the association of covariates with CA frequency and the results were very similar to those from logistic model.

Both logistic and linear regression models were applied for analysis of the phenotypes CAtot, CTA and CSA. Altogether 11 loci, 6 from the CAtot and 5 from the CTA analysis, were chosen for replication and the most significant SNPs with $P < 1x10^{-5}$ from these loci were selected on the basis of *in silico* analyses (table III). Selected SNPs were genotyped in the replication sets. All selected loci contained at least one directly genotyped SNP and the genotyping accuracy was confirmed in a small subset of GWAS samples. No SNP associations at the level of $P < 1x10^{-5}$ were found in CSA.

In case of CAtot, the logistic regression model showed more significant associations than the linear model, however, almost all the loci showed similar trends in the linear model as well. This is

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

evident from the Miami plot (fig.1). Two SNPs in chromosome 9 (rs12002628 and rs16931167) almost reached the genome-wide significance level of $5x10^{-8}$ ($p = 4.78x10^{-7}$ and $2.66x10^{-7}$, respectively, table IV). For all SNPs, except for rs16931167, replication 1 showed ORs on the same direction as in the GWAS and the strongest associations in the meta-analysis were for rs1383997 at 8q13.3 (OR 0.6, 95%CI 0.49-0.73, $P = 3.44x10^{-7}$) and rs2824215 at 21q21.1 (OR 1.57, 95%CI 1.29-1.91, $P = 8.7x10^{-6}$). Replication 2 did not give much support for the GWAS associations, and the strongest association in the meta-analysis of all populations with $P = 4.01x10^{-5}$ was for rs12002628 at 9q21.13. The heterogeneity between the three populations was moderate to high as indicated by the I^2 values (table IV).

For the CTA analysis, on the other hand, higher associations were found in the linear model as compared to the logistic model (fig 2). In CTA, 5 SNPs showed an association at the suggestive level of significance (table V). Here also, the GWAS and replication 1 showed more similar associations than the GWAS and replication 2 and the heterogeneity measured by the I^2 values was high. In the meta-analysis of the GWAS and replication 1, one association, rs983889 at 5p15.1 remained statistically significant at the suggestive level ($P = 1.06x10^{-5}$) and no significant associations were observed in the meta-analysis of all three populations. Although the SNPs from the GWAS were selected based on the linear model, we also calculated the ORs and 95%CIs in the logistic model (table V). For the most significant SNP, rs9838889, the OR was 0.65 (95%CI 0.52-0.80) in the meta-analysis of the GWAS and replication 1.

Discussion

Inherited genetic variation may cause inter-individual differences in the susceptibility to CAs. We tested this hypothesis at the whole-genome level in three populations composed of differentially exposed individuals through smoking, occupational or environmental factors, and untreated cancer patients. In general, occupational exposure and being a newly diagnosed cancer patient had a strong influence on the frequency of CAs, while the genetic factors seemed to play a minor role. The fact

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

that the GWAS and replication 1 showed more similar associations with CAs than the GWAS and the replication 2 may be explained by the composition of these three sample sets.

In the GWAS set more than 50% of the subjects were occupationally exposed to different compounds such as small organic compounds, anesthetics, styrene and heavy metals [Somorovská *et al.* 1999; Vodicka *et al.* 2004a; b, 2015, Musak *et al.* 2008, 2013; Hemminki *et al.* 2015b; Försti *et al.* 2016]. Logistic regression analysis on this sample set showed a highly significant association of increasing CA frequency with occupation and a moderate association with age. In replication 1, in which 46% of individuals were incident cancer patients, a strong association between the frequency of CAs and cancer status was observed. These results are also in consistence with previous studies [Hemminki *et al.* 2015b; Vodicka *et al.* 2015]. For replication 2, age was the most significantly associated factor with smoking and occupational exposure showing a moderate effect. In replication 2 the effect of occupational environment was less significant than in the GWAS sample, probably because the proportion of individuals who were occupationally exposed was only 23%. The proportion of individuals exposed to asbestos, which is equally genotoxic as the chemical compounds in the GWAS [IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. *et al.* 2008; IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. and International Agency for Research on Cancer. 2012] was less than 10% of the study group. The rest of the occupationally exposed individuals were exposed to stone wool and glass fibers which are comparatively less genotoxic [Dušinská *et al.* 2004a; b; Baan and Grosse 2004]. Also, a significant part of the sample consisted of individuals from the aging, obesity and specific food preference studies, who were nominally unexposed to genotoxic agents [Dušinská *et al.* 2003; Kažimírová *et al.* 2004, 2006, 2009; Szabová *et al.* 2012]. These differences in the composition of the study populations were also reflected in the CA frequencies, which were about twice as high in the GWAS and replication 1 as compared to replication 2.

Due to these population and CA frequency differences and because the CAs are measured as a number of aberrations per 100 cells, we used both the logistic and the linear regression models to

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

evaluate the associations between the genetic variants and the frequencies of CAs. We analyzed the CAs as three phenotypic categories, CAtot, CSA and CTA. However, no loci were found to be associated with CSA frequency. CSAs are also known to be affected to a lesser extent by chemical mutagens as compared to CTAs [Natarajan 1993]. The SNPs in CAtot category were selected from the logistic model as it provided stronger associations. In case of CTA, linear model offered better associations than the logistic model; this difference could be attributed to the difference in CA_{high}/CA_{low} cutoff point for CTAs (1%) and CAtot (2%). A linear regression is built on continuous variables as outcome. It has a higher precision, and it provides more statistical power with a smaller sample set [MacCallum *et al.* 2002].

In spite of differences in the study populations, 5 out of 6 CAtot variants showed ORs in the same direction both in the GWAS and replication 1. Two SNPs (rs1383997 and rs2824215) reached the suggestive significance p-value of 1×10^{-5} in the meta-analysis between the GWAS and replication 1. Addition of replication 2 to the meta-analysis resulted in only one marginal association (rs12002628, P-value 4.01×10^{-5}). Variants involved in CTA did not have significant support from the replications either. Only one variant (rs983889) reached the significance level of $P = 1 \times 10^{-5}$ in the meta-analysis of the GWAS and replication 1. Similar to any GWAS, all associated loci were located in the non-coding region of the genome, and we evaluated their potential functional consequences using several *in silico* tools and the existing literature data.

One of the most strongly associated SNPs in the CAtot analyses was located in the gene related to transient receptor potential (TRP) cation channels. Rs1383997 is mapped to the 8q13.3 locus that codes for an antisense transcript. This natural antisense transcript (NAT) is antisense to the musculin (*MSC*) gene and the TRP cation channel subfamily A member 1 (*TRPA1*) gene. NATs are known to regulate the expression of their corresponding sense transcript [Wight and Werner 2013]. Another SNP rs12002628 on chromosome 9q21.13, which reached the suggestive level of significance in meta-analysis between GWAS and replication 1, is present at 5.8kb 5' to another TRP family protein *TRPM3* gene with many linked SNPs in the introns of *TRPM3*. TRP channels

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

regulate the Ca2+ ions homeostasis in response to environmental and chemical factors. Any deregulation in Ca2+ distribution patterns can promote the signs of cancer development such as proliferation, enhanced survival and invasion [Shapovalov *et al.* 2016]. MSC, also known as activated B-cell factor-1 (ABF-1), is a member of basic helix loop helix (bHLH) family of transcription factors which are involved in cell fate determination in several developmental processes like haemopoiesis and myogenesis [Murre *et al.* 1994]. It is mainly expressed in activated B cells in humans and EBV-transformed lymphoblastoid cell lines [Massari *et al.* 1998]. ABF-1 is capable of inhibiting the transactivation capability of E47 in mammalian cells. E47 is involved with several chromosomal translocations and diminished activity of E47 can lead to lymphoid malignancies [Herblot *et al.* 2002].

The second SNP from the CAtot analysis, rs2824215 (21q21.1) is located in a long intergenic non-coding RNA (LiNC), and deletion in this locus has been linked to autistic features with complex chromosomal rearrangements [Haldeman-Englert *et al.* 2010]. Interestingly, two other SNPs, which we selected for replication, rs17215792 (2q33.3) and rs2837619 (21q22.2) are located in the genes associated with autism and Down syndrome, *KLF7* (Kruppel like factor 7) [Pescucci *et al.* 2003; Jang *et al.* 2015] and *DSCAM* (Down Syndrome Cell Adhesion Molecule), respectively [Yamakawa *et al.* 1998; Cvetkovska *et al.* 2013]. Chromosomal abnormalities are an important feature of both diseases [Liao *et al.* 2013]. Both of these SNPs, however, showed only weak, if any, association with CAs in the replication sets.

The only SNP from the CTA analysis with a suggestive level of association in the GWAS and replication 1 meta-analysis, rs983889, is an intronic SNP in the F-box and leucine-rich repeat protein 7 (*FBXL7*) gene. FBXL7 belongs to F-box proteins, which are involved in phosphorylation–dependent ubiquitination of proteins and which display proapoptotic activity [Zheng *et al.* 2016]. Incidentally, one of the targets of FBXL7 is aurora kinase A (*AURKA*), a known oncogene, involved in regulation of mitosis [Tang *et al.* 2015]. During late G2 phase AURKA is recruited to centrosomes [Hanisch *et al.* 2006] and later on promotes centrosome maturation and bipolar spindle

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

formation [Gruss *et al.* 2001]. Since CTAs also arise during S/G2 phase [Durante *et al.* 2013], an indirect involvement of AURKA can be anticipated to affect the frequency of CTAs.

In conclusion, our GWAS identified eleven SNPs associated with CA frequency, from which three were replicated at the suggestive level of significance. *In silico* predictions of functional consequences of the identified SNPs and their loci revealed that they were directly or indirectly related to different cancers. They included genes encoding TRP cation channel proteins, which regulate the Ca2+ ions homeostasis in response to environmental and chemical factors, genes involved in autism and Down syndrome, two syndromes linked to chromosomal abnormalities, and *FBXL7*, which interacts with AURKA, an important regulator of mitosis. Although due to sample size the results of this GWAS are not definitive in terms of pointing out the exact rationale behind CAs development but they certainly point towards the probable loci that could be involved in the elevated frequency of CAs in the presence of environmental stress.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Statement of Author Contributions

PV, KH and AF designed the study; HT and YN analyzed the data; YN and BS performed the genotyping; PV, LV, LM, SV, MK and VV provided samples and data for GWAS and replication 1; MD, BS, AK, MB, KV and MS provided samples and data for replication 2; PH, MMN were responsible for the GWAS; YN wrote the first draft of the manuscript; AF, KH, PV, LV, MD and BS critically revised the manuscript; all authors read and approved the final manuscript.

Acknowledgements

In the Czech Republic, the study was supported by the National Science Foundation, Grant Numbers: 15-14789S and 18-09709S; Charles University in Prague, UNCE 204022, and PROGRES Q 28; Medical Faculty in Pilsen, Charles University in Prague, National Sustainainability Programme I, Nr.LO 1503 and Charles University Research Centre program UNCE/MED/006.

In the Slovak Republic the study was supported by EC contracts QLK4-CT-1999-01629, ERBICI 15-CT96-1012, CIPA-CT94-0129, Slovak Grant Agency APVT-21 013202, APVT-21-017704 and grants from the Ministry of Health, Slovak Republic 2005/43-SZU-21 and 2006/07-SZU-02 MZ SR, 2005/42-SZU-20, SZU.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

References

Albertini RJ, Anderson D, Douglas GR, Hagmar L, Hemminki K, *et al.* 2000. IPCS guidelines for the monitoring of genotoxic effects of carcinogens in humans. International Programme on Chemical Safety. Mutat. Res. 463: 111–72.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, *et al.* 2010. Data quality control in genetic case-control association studies. Nat. Protoc. 5: 1564–1573.

Baan RA, Grosse Y. 2004. Man-made mineral (vitreous) fibres: evaluations of cancer hazards by the IARC Monographs Programme. Mutat. Res. Mol. Mech. Mutagen. 553: 43–58.

Bignold LP. 2009. Mechanisms of clastogen-induced chromosomal aberrations: A critical review and description of a model based on failures of tethering of DNA strand ends to strand-breaking enzymes. Mutat. Res. Mutat. Res. 681: 271–298.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, *et al.* 2012. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 22: 1790–7.

Carrano A V, Natarajan AT. 1988. International Commission for Protection Against Environmental Mutagens and Carcinogens. ICPEMC publication no. 14. Considerations for population monitoring using cytogenetic techniques. Mutat. Res. 204: 379–406.

Cvetkovska V, Hibbert AD, Emran F, Chen BE. 2013. Overexpression of Down syndrome cell adhesion molecule impairs precise synaptic targeting. Nat. Neurosci. 16: 677–82.

Delaneau O, Marchini J, Zagury J-F. 2011. A linear complexity phasing method for thousands of genomes. Nat. Methods 9: 179–181.

Durante M, Bedford JS, Chen DJ, Conrad S, Cornforth MN, *et al.* 2013. From DNA damage to chromosome aberrations: Joining the break. Mutat. Res. Toxicol. Environ. Mutagen. 756: 5–13.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Dušinská M, Kažimírová A, Barančoková M, Beňo M, Smolková B, *et al.* 2003. Nutritional supplementation with antioxidants decreases chromosomal damage in humans. Mutagenesis 18: 371–376.

Dušinská M, Barančoková M, Kažimírová A, Harrington V, Volkovová K, *et al.* 2004.a. Does occupational exposure to mineral fibres cause DNA or chromosome damage? In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*,, pp. 103–110.

Dušinská M, Collins A, Kažimírová A, Barančoková M, Harrington V, *et al.* 2004.b. Genotoxic effects of asbestos in humans. In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*,, pp. 91–102.

Dušinská M, Staruchova M, Horska A, Smolkova B, Collins A, *et al.* 2012. Are glutathione S transferases involved in DNA damage signalling? Interactions with DNA damage and repair revealed from molecular epidemiology studies. Mutat. Res. - Fundam. Mol. Mech. Mutagen. 736: 130–137.

Försti A, Frank C, Smolkova B, Kazimirova A, Barancokova M, *et al.* 2016. Genetic variation in the major mitotic checkpoint genes associated with chromosomal aberrations in healthy humans. Cancer Lett. 380: 442–446.

Gostissa M, Alt FW, Chiarle R. 2011. Mechanisms that Promote and Suppress Chromosomal Translocations in Lymphocytes. Annu. Rev. Immunol. 29: 319–350.

Gruss OJ, Carazo-Salas RE, Schatz CA, Guarguaglini G, Kast J, *et al.* 2001. Ran induces spindle assembly by reversing the inhibitory effect of importin α on TPX2 activity. Cell 104: 83–93.

Hagmar L, Strömberg U, Tinnerberg H, Mikoczy Z. 2001. The usefulness of cytogenetic biomarkers as intermediate endpoints in carcinogenesis. Int. J. Hyg. Environ. Health 204: 43–47.

Haldeman-Englert CR, Chapman KA, Kruger H, Geiger EA, McDonald-McGinn DM, *et al.* 2010.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

A de novo 8.8-Mb deletion of 21q21.1-q21.3 in an autistic male with a complex rearrangement involving chromosomes 6, 10, and 21. Am. J. Med. Genet. Part A 152A: 196–202.

Hanisch A, Wehner A, Nigg EA, Silljé HHW. 2006. Different Plk1 functions show distinct dependencies on Polo-Box domain-mediated targeting. Mol Biol Cell 17: 448–459.

Hemminki K, Rachakonda S, Musak L, Vymetalkova L, Halasova E, *et al.* 2015.a. Telomere length in circulating lymphocytes: Association with chromosomal aberrations. Genes Chromosom. Cancer 54: 194–196.

Hemminki K, Frank C, Försti A, Musak L, Kazimirova A, *et al.* 2015.b. Metabolic gene variants associated with chromosomal aberrations in healthy humans. Genes Chromosom. Cancer 54: 260–266.

Heng HHQ, Regan SM, Liu G, Ye CJ. 2016. Why it is crucial to analyze non clonal chromosome aberrations or NCCAs? Mol. Cytogenet. 9: 15.

Herblot S, Aplan PD, Hoang T. 2002. Gradient of E2A activity in B-cell development. Mol. Cell. Biol. 22: 886–900.

Howie B, Marchini J, Stephens M. 2011. Genotype imputation with thousands of genomes. G3 (Bethesda). 1: 457–70.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans., World Health Organization., International Agency for Research on Cancer. 2008. *1,3-Butadiene, ethylene oxide, and vinyl halides (vinyl flouride, vinyl chloride, and vinyl bromide).* International Agency for Research on Cancer.

IARC Working Group on the Evaluation of Carcinogenic Risks to Humans., International Agency for Research on Cancer. 2012. *A review of human carcinogens. Arsenic, metals, fibres, and dusts.* International Agency for Research on Cancer.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Jang D-H, Chae H, Kim M. 2015. Autistic and Rett-like features associated with 2q33.3-q34 interstitial deletion. Am. J. Med. Genet. Part A 167: 2213–2218.

Jones CH, Pepper C, Baird DM. 2012. Telomere dysfunction and its role in haematological cancer. Br. J. Haematol. 156: 573–587.

Kažimírová A, Barančoková M, Volkovová K, Staruchová M, Krajcovicova-Kudlackova M, *et al.* 2004. Does a vegetarian diet influence genomic stability? Eur. J. Nutr. 43: 32–38.

Kažimírová A, Barančoková M, Krajčovičová-Kudláčková M, Volkovová K, Staruchová M, *et al.* 2006. The relationship between micronuclei in human lymphocytes and selected micronutrients in vegetarians and non-vegetarians. Mutat. Res. - Genet. Toxicol. Environ. Mutagen. 611: 64–70.

Kažimírová A, Barančoková M, Džupinková Z, Wsólová L, Dušinská M. 2009. Micronuclei and chromosomal aberrations, important markers of ageing: Possible association with XPC and XPD polymorphisms. Mutat. Res. Mol. Mech. Mutagen. 661: 35–40.

Liao H-M, Gau SS-F, Tsai W-C, Fang J-S, Su Y-C, *et al.* 2013. Chromosomal Abnormalities in Patients With Autism Spectrum Disorders From Taiwan. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 162: 734–741.

MacCallum RC, Zhang S, Preacher KJ, Rucker DD. 2002. On the practice of dichotomization of quantitative variables. Psychol. Methods 7: 19–40.

Mägi R, Morris AP. 2010. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11: 288.

Massari ME, Rivera RR, Voland JR, Quong MW, Breit TM, *et al.* 1998. Characterization of ABF-1, a novel basic helix-loop-helix transcription factor expressed in activated B lymphocytes. Mol. Cell. Biol. 18: 3130–9.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Mitelman F. 2000. Recurrent chromosome aberrations in cancer. Mutat. Res. 462: 247–53.

Murre C, Bain G, Dijk MA van, Engel I, Furnari BA, *et al.* 1994. Structure and function of helix-loop-helix proteins. Biochim. Biophys. Acta 1218: 129–35.

Musak L, Soucek P, Vodickova L, Naccarati A, Halasova E, *et al.* 2008. Chromosomal aberrations in tire plant workers and interaction with polymorphisms of biotransformation and DNA repair genes. Mutat. Res. Mol. Mech. Mutagen. 641: 36–42.

Musak L, Smerhovsky Z, Halasova E, Osina O, Letkova L, *et al.* 2013. Chromosomal damage among medical staff occupationally exposed to volatile anesthetics, antineoplastic drugs, and formaldehyde. Scand. J. Work. Environ. Heal. 39: 618–630.

Natarajan AT. 1993. Mechanisms for induction of mutations and chromosome alterations. Environ. Health Perspect. 101 Suppl: 225–9.

Pescucci C, Meloni I, Bruttini M, Ariani F, Longo I, *et al.* 2003. Chromosome 2 deletion encompassing the MAP2 gene in a patient with autism and Rett-like features. Clin. Genet. 64: 497–501.

Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, *et al.* 2010. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 26: 2336–2337.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81: 559–75.

Rosenbloom KR, Armstrong J, Barber GP, Casper J, Clawson H, *et al.* 2015. The UCSC Genome Browser database: 2015 update. Nucleic Acids Res. 43: D670–D681.

Rossner P, Boffetta P, Ceppi M, Bonassi S, Smerhovsky Z, *et al.* 2005. Chromosomal aberrations in lymphocytes of healthy subjects and risk of cancer. Environ. Health Perspect. 113: 517–20.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Shapovalov G, Ritaine A, Skryma R, Prevarskaya N. 2016. Role of TRP ion channels in cancer and tumorigenesis. Semin. Immunopathol. 38: 357–369.

Somorovská M, Jahnová E, Tulinská J, Zámecníková M, Sarmanová J, *et al.* 1999. Biomonitoring of occupational exposure to styrene in a plastics lamination plant. Mutat. Res. 428: 255—269.

Šrám RJ, Rössner P, Šmerhovský Z. 2004. Cytogenetic analysis and occupational health in the Czech Republic. Mutat. Res. 566: 21–48.

Szabová M, Jahnová E, Horváthová M, Ilavská S, Pružincová V, *et al.* 2012. Changes in immunologic parameters of humoral immunity and adipocytokines in obese persons are gender dependent. Hum. Immunol. 73: 486–492.

Tang A, Gao K, Chu L, Zhang R, Yang J, *et al.* 2015. Aurora kinases: novel therapy targets in cancers. Oncotarget 8: 23937–23954.

Tulinska J, Jahnova E, Dušinská M, Kuricova M, Liskova A, *et al.* 2004. Immunomodulatory effects of mineral fibres in occupationally exposed workers. In: *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*,, pp. 111–124.

Vodenkova S, Polivkova Z, Musak L, Smerhovsky Z, Zoubkova H, *et al.* 2015. Structural chromosomal aberrations as potential risk markers in incident cancer patients. Mutagenesis 30: 557–563.

Vodicka P, Tuimala J, Stetina R, Kumar R, Manini P, *et al.* 2004.a. Cytogenetic markers, DNA single-strand breaks, urinary metabolites, and DNA repair rates in styrene-exposed lamination workers. Environ. Health Perspect. 112: 867–871.

Vodicka P, Kumar R, Stetina R, Musak L, Soucek P, *et al.* 2004.b. Markers of individual susceptibility and DNA repair rate in workers exposed to xenobiotics in a tire plant. Environ. Mol. Mutagen. 44: 283–292.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Vodicka P, Polivkova Z, Sytarova S, Demova H, Kucerova M, *et al.* 2010. Chromosomal damage in peripheral blood lymphocytes of newly diagnosed cancer patients and healthy controls. Carcinogenesis 31: 1238–1241.

Vodicka P, Musak L, Frank C, Kazimirova A, Vymetalkova V, *et al.* 2015. Interactions of DNA repair gene variants modulate chromosomal aberrations in healthy subjects. Carcinogenesis 36: 1299–1306.

Vodicka P, Musak L, Vodickova L, Vodenkova S, Catalano C, *et al.* 2018. Genetic variations of acquired structural chromosomal aberrations. Mutat. Res. - Genet. Toxicol. Environ. Mutagen.

Ward LD, Kellis M. 2012. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 40: D930–D934.

Wight M, Werner A. 2013. The functions of natural antisense transcripts. Essays Biochem. 54.

Yamakawa K, Huot YK, Haendelt MA, Hubert R, Chen XN, *et al.* 1998. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. Hum. Mol. Genet. 7: 227–37.

Zheng N, Wang Z, Wei W. 2016. Ubiquitination-mediated degradation of cell cycle-related proteins by F-box proteins. Int. J. Biochem. Cell Biol. 73: 99–110.

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

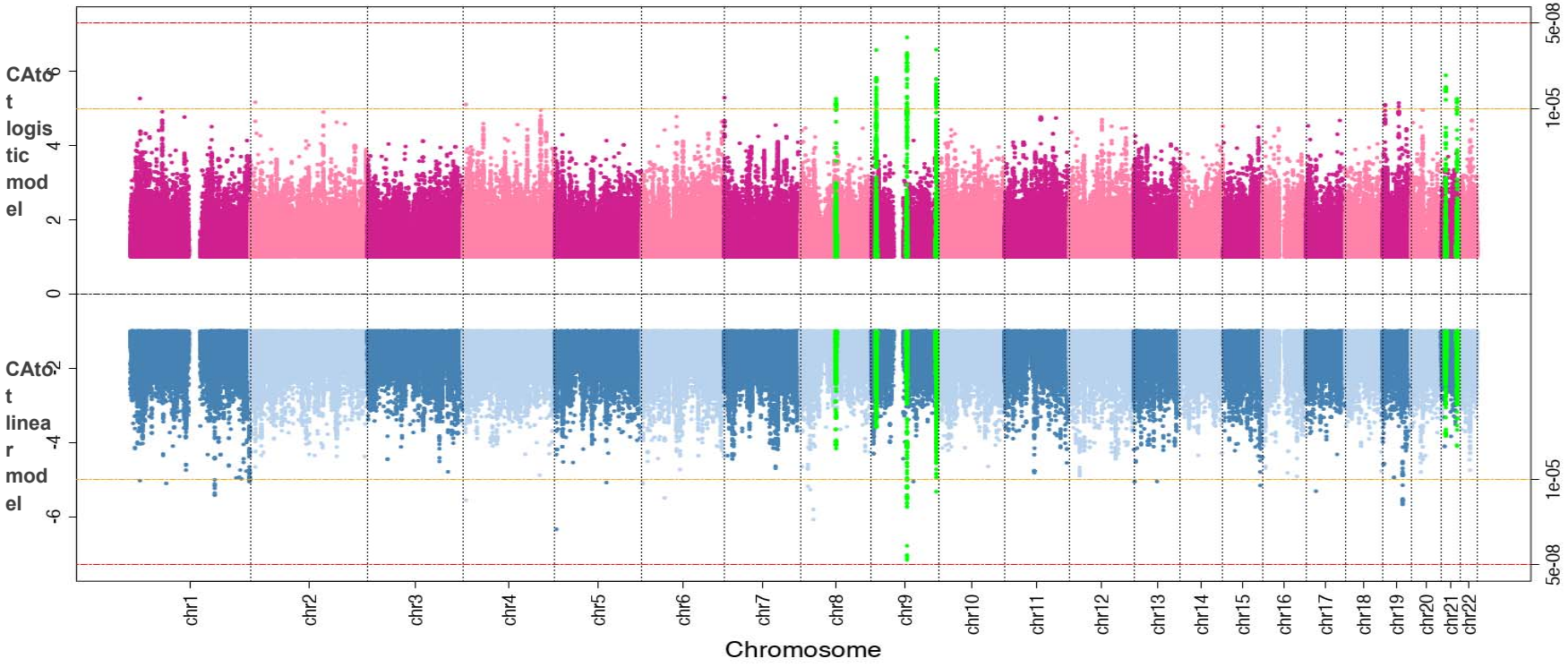
Figure Legends

Fig. 1. Miami plot for CAtot logistic and linear models. The y-axis shows the $-\log_{10}$ P – value of each SNP and the X-axis shows their chromosomal position. Loci selected for replication are highlighted in green. The red horizontal line represents the genome-wide significance threshold of P =5.0×10^{-8} and the yellow horizontal line represents the significance threshold of P = 5.0×10^{-5} as suggestive level of significance.

Fig. 2. Miami plot for CTA linear and logistic models. The y-axis shows the $-\log_{10}$ P – value of each SNP and the X-axis shows their chromosomal position. Loci selected for replication are highlighted in green. The red horizontal line represents the genome-wide significance threshold of P =5.0×10^{-8} and the yellow horizontal line represents the significance threshold of P = 5.0×10^{-5} as suggestive level of significance

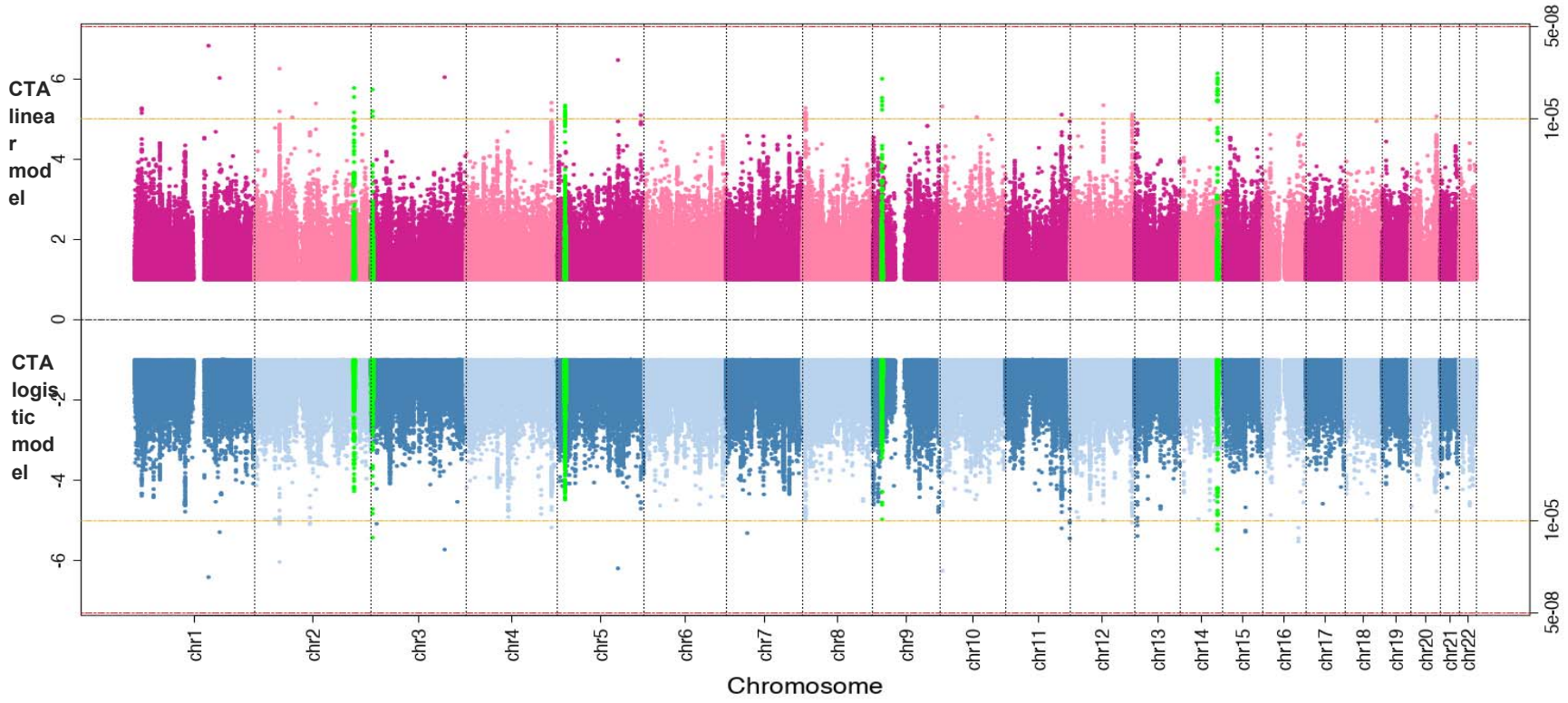
This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Fig. 1. Miami plot for CAtot logistic and linear models. The y-axis shows the $-\log_{10}$ p – value of each SNP and the X-axis shows their chromosomal position. Loci selected for replication are highlighted in green. The red horizontal line represents the genome-wide significance threshold of $p = 5.0 \times 10^{-8}$ and the yellow horizontal line represents the significance threshold of $p = 5.0 \times 10^{-5}$ as suggestive level of significance.



This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Fig. 2. Miami plot for CTA linear and logistic models. The y-axis shows the $-\log_{10}$ p – value of each SNP and the X-axis shows their chromosomal position. Loci selected for replication are highlighted in green. The red horizontal line represents the genome-wide significance threshold of $p = 5.0 \times 10^{-8}$ and the yellow horizontal line represents the significance threshold of $P = 5.0 \times 10^{-5}$ as suggestive level of significance.



This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Table I. Characterization of the study population including distribution of high (CA_{high}) and low chromosomal aberration level (CA_{low}) among CAtot, CSA and CTA categories and according to major confounders, age, sex, smoking, occupational exposure and cancer status.

	GWAS	P[l]	Replication 1	P[l]	Replication 2	P[l]
Mean age ±SD	44.64±12.70	0.01	59.63±12.88	0.014	43.8 ±15.57	2.59E-07
Female/Male %	51.7/48.3	0.57	53.6/46.4	0.13	57.7/42.3	0.02
Smoking Yes/No %	30.5/69.5	0.96	29.3/70.7	0.001	24.1/75.9	0.006
CAtot[a] no. (CA_{high}[b]/CA_{low}[c])	351/225[d]		304/178		295/993	
CTA[e] no. (CA_{high}/CA_{low})	367/209[d]		350/132		387/901	
CSA[f] no. (CA_{high}/CA_{low})	349/227[d]		268/214		367/921	
Occupational exposure %	57.6	1.21E-09			23.1	0.009
Small organic compounds%	21.7					
Anesthetics%	15.5					
Heavy metals%	12					
Styrene%	8.5					
Radiation (pilots)%					5.8	
Asbestos%					4.6	
Stone wool%					7.1	
Glass fibers%					5.8	
Cancer %			46.1	7.56E-06		
Breast cancer%			23.4			
Colorectal cancer%			14.3			
Lung cancer%			8.3			
Others	42.4[g]		53.9[h]		10.7[g]/18.8[i]/20.5[j]/26.9[k]	

[a](Total chromosomal aberrations)

[b](No. of individuals in high chromosomal aberrations group)

[c](No. of individuals in low chromosomal aberrations group)

[d](No. of individuals after the quality control)

[e](Chromatid type aberrations)

[f](Chromosome type aberrations)

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

[g](Office workers and local residents)

[h](Blood donors)

[i](Obesity study)

[j](Aging study)

[k](Vegetarians and nutrition studies)

[l](P values based on binary regression model exhibiting the modulation in frequency of total chromosomal aberrations by major confounders age, sex, smoking and occupational exposure).

Table II. Distribution of CAtot, CSA and CTA in the three data sets.

	No. of Individuals[b]		Mean±SD[a]	Median	Range
GWAS	576	CAtot	1.94 ±1.29	2	0-7
		CSA	0.98 ±1.04	1	0-6
		CTA	0.97 ±0.99	1	0-6
Replication 1	482	CAtot	2.21 ±1.57	2	0-8
		CSA	0.88 ±1.036	1	0-5
		CTA	1.32 ±1.22	1	0-7
Replication 2	1288	CAtot	0.95 ±1.35	0	0-11
		CSA	0.51 ±1.02	0	0-11
		CTA	0.44 ±0.78	0	0-6

[a] Mean number of aberrations per hundred cells

[b]Number of individuals in the study

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Table III. *In silico* predictions for the selected variants. Functional annotations from the ENCODE based tool Haploreg v4.1 (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) and eQTL analysis according to GTEx Portal (http://www.gtexportal.org/home/)

Chromosome	SNP	No. of SNPs with r2>0.8	Promotor histon marks	Enhancer histone marks	Proteins bound	DNAse	Motifs changed	eQTL hits	Gencode genes	Type
2	rs17215792	7	BRST[a], BRN[b]	14 tissues (BLD[c])		ESDR[d]	HMG-IY, Irf, Nkx2		KLF7	3'-UTR
3	rs340828	2		4 tissues (BLD)			16 altered motifs	1 hit	IL5RA	3'-UTR
5	rs983889	27	5 tissues	10 tissues		5 tissues	TCF12, TCF4		FBXL7	Intronic
8	rs1383997	37	21 tissues (BLD)	21 tissues (BLD)	CTCF, SMC3, EGR1, TBP	12 tissues (BLD)	BDP1, CTCF, NF-I	2 hits	RP11-383H13.1	Intronic
9	rs12002628	17					Hoxa4, Hoxb8		TRPM3	5' -UTR
9	rs7025089	53	ESDR, BLD, LNG[e]	21 tissues (BLD)	CTCF	9 tissues	10 altered motifs	2 hits	MED27	Intronic
9	rs16931167	1		LIV[f], HRT[g], PANC			5 altered motifs		PTPRD	Intronic
9	rs7033729	9	ESDR, GI[h]	6 tissues		PLCNT[i]	Arid5b, NRSF	4 hits	FAM154A	Intronic
14	rs8003642	26	SPLN[j]	11 tissues			Irf, Sox		RP11-725G5.2	Intronic
21	rs2824215	18							AF212831.2	5' -UTR

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

21	rs2837619	13			ERalpha-a, Irf, p300	DSCAM	Intronic

[a](Breast)

[b](Brain)

[c](Blood)

[d](ESC_Derived)

[e](Lung)

[f](Liver)

[g](Heart)

[h](Gastrointestinal tract)

[I](Placenta)

[j](Spleen)

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Table IV. SNPs selected from CAtot model and their corresponding OR, 95% CI, and p-values in the logistic model in three sample sets and meta-analyses.

	Logistic Model								
	SNP	**Locus**	**Minor allele**	**Major allele**	Model	OR	95% CI	P	I^2
GWAS	rs1383997	8q13.3	T	C	CAtot	0.56	0.44-0.71	5.44E-06	
Replication 1						0.67	0.49-0.91	0.01	
Replication 2						1.09	0.89-1.34	0.39	
Meta GWAS+ Replication 1[a]						0.6	0.49-0.73	3.44E-07	0
meta all[b]						0.8	0.70-0.93	0.002	0.89
GWAS	rs12002628	9q21.13	T	C	CAtot	0.47	0.35-0.63	4.78E-07	
Replication 1						0.93	0.64-1.33	0.67	
Replication 2						0.79	0.60-1.03	0.08	
Meta GWAS+ Replication 1						0.61	0.48-0.78	7.68E-05	0.89
meta all						0.68	0.57-0.82	4.01E-05	0.81
GWAS	rs7025089	9q34.13	C	A	CAtot	0.55	0.42-0.70	7.97E-06	
Replication 1						0.86	0.63-1.18	0.35	
Replication 2						1.19	0.95-1.47	0.13	
Meta GWAS+ Replication 1						0.66	0.53-0.81	1.02E-04	0.8
meta all						0.87	0.75-1.02	0.08	0.9
GWAS	rs16931167	9p23	T	C	CAtot	2.55	1.76-3.71	2.66E-07	
Replication 1						0.77	0.52-1.12	0.17	

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Replication 2						0.96	0.73-1.28	0.8	
Meta GWAS+ Replication 1						1.41	1.07-1.86	0.02	0.95
meta all						1.17	0.96-1.42	0.13	0.92
GWAS	rs2824215	21q21.1	C	A	CAtot	1.8	1.40-2.30	3.27E-06	
Replication 1						1.27	0.94-1.7	0.12	
Replication 2						1.02	0.83-1.24	0.88	
Meta GWAS+ Replication 1						1.57	1.29-1.91	8.70E-06	0.73
meta all						1.26	1.10-1.45	0.001	0.84
GWAS	rs2837619	21q22.2	G	A	CAtot	0.61	0.47-0.77	5.51E-06	
Replication 1						0.94	0.71-1.24	0.64	
Replication 2						0.97	0.8-1.18	0.78	
Meta GWAS+ Replication 1						0.71	0.58-0.85	3.65E-04	0.86
meta all						0.83	0.72-0.95	0.01	0.84

[a](Meta-analysis between the GWAS and replication 1)

[b](Meta-analysis between all three sample sets i.e. GWAS sample set, replication 1 and replication 2)

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Table V. SNPs selected from CTA model, their beta and p-values in the linear model and the corresponding OR, 95% CI, and p-values in the logistic model in three sample sets and meta-analyses.

	Linear Model									Logistic Model			
	SNP	Locus	Minor	Major	Model	Beta	SE	P	I^2	OR	95% CI	P	I^2
GWAS	rs340828	3p26.2	A	G	CTA	0.26	0.06	6.45E-06		1.83	1.40-2.38	1.48E-05	
Replication 1						-0.03	0.03	0.39		0.87	0.64-1.2	0.4	
Replication 2						-0.004	0.005	0.42		1.01	0.83-1.22	0.96	
Meta GWAS+ Replication 1[a]						0.06	0.02	0.02	0.94	1.19	0.96-1.46	0.11	0.84
meta all[b]						-0.002	0.005	0.75	0.91	1.08	0.94-1.25	0.26	0.74
GWAS	rs8003642	14q32.13	C	A	CTA	-0.33	0.07	1.36E-06		0.52	0.39-0.69	1.77E-05	
Replication 1						-0.02	0.04	0.66		0.91	0.63-1.32	0.62	
Replication 2						-0.003	0.006	0.58		0.86	0.69-1.08	0.2	
Meta GWAS+ Replication 1						-0.1	0.03	3.70E-04	0.91	0.7	0.53-0.91	0.008	0.77
meta all						-0.01	0.01	0.21	0.91	0.79	0.66-0.94	0.01	0.65
GWAS	rs17215792	2q33.3	C	A	CTA	0.47	0.1	1.70E-06		2.52	1.54-4.12	1.11E-04	
Replication 1						0.01	0.06	0.82		1.08	0.59-1.99	0.81	
Replication 2						0.003	0.008	0.73		0.96	0.70-1.33	0.83	
Meta GWAS+ Replication 1						0.15	0.04	7.43E-04	0.91	1.65	1.15-2.36	0.006	0.64
meta all						0.01	0.01	0.34	0.91	1.23	0.97-1.56	0.09	0.74

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

GWAS	rs983889	5p15.1	G	T	CTA	-0.26	0.06	4.82E-06		0.61	0.48-0.77	3.85E-05	
Replication 1						-0.06	0.03	0.07		0.77	0.57-1.04	0.09	
Replication 2						0.008	0.005	0.09		1.05	0.88-1.26	0.61	
Meta GWAS+ Replication 1						-0.11	0.02	1.06E-05	0.82	0.65	0.52-0.80	6.61E-05	0.56
meta all						0.004	0.005	0.4	0.93	0.86	0.75-0.98	0.03	0.85
GWAS	rs7033729	9p22.1	A	G	CTA	0.53	0.11	9.99E-07		3.53	1.92-6.48	1.05E-05	
Replication 1						0.005	0.05	0.93		1.03	0.62-1.72	0.9	
Replication 2						-0.01	0.009	0.23		0.85	0.60-1.21	0.38	
Meta GWAS+ Replication 1						0.12	0.04	0.003	0.93	1.49	1.05-2.12	0.03	0.74
meta all						-0.005	0.01	0.56	0.92	1.12	0.88-1.44	0.35	0.77

[a](Meta-analysis between the GWAS and replication 1)

[b](Meta-analysis between all three sample sets i.e. GWAS sample set, replication 1 and replication 2)

This is the peer reviewed version of the following article: Niazi, Y. et al. 2018: Genetic variation associated with chromosomal aberration frequency: A genome-wide association study, Environmental and Molecular Mutagenesis, which has been published in final form at https://doi.org/10.1002/em.22236. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.