nature genetics

Detectable clonal mosaicism and its relationship to aging and cancer

In an analysis of 31,717 cancer cases and 26,136 cancer-free controls from 13 genome-wide association studies, we observed large chromosomal abnormalities in a subset of clones in DNA obtained from blood or buccal samples. We observed mosaic abnormalities, either aneuploidy or copy-neutral loss of heterozygosity, of >2 Mb in size in autosomes of 517 individuals (0.89%), with abnormal cell proportions of between 7% and 95%. In cancer-free individuals, frequency increased with age, from 0.23% under 50 years to 1.91% between 75 and 79 years ($P = 4.8 \times 10^{-8}$). Mosaic abnormalities were more frequent in individuals with solid tumors (0.97% versus 0.74% in cancer-free individuals; odds ratio (OR) = 1.25; $P = 0.016$), with stronger association with cases who had DNA collected before diagnosis or treatment (OR = 1.45; $P = 0.0005$). Detectable mosaicism was also more common in individuals for whom DNA was collected at least 1 year before diagnosis with leukemia compared to cancer-free individuals (OR = 35.4; $P = 3.8 \times 10^{-11}$). These findings underscore the time-dependent nature of somatic events in the etiology of cancer and potentially other late-onset diseases.

Classically, genetic mosaicism is defined as the coexistence of cells with two or more distinct karyotypes within an individual, resulting from a post-zygotic event during development that can occur in both somatic and germline cells[1,2]. Errors in chromosomal duplication and subsequent transmission to daughter cells may lead to aneuploidy, the gain or loss of chromosomes or segments of chromosomes, and reciprocal gain and loss events that appear as copy-neutral loss of heterozygosity or acquired uniparental disomy. Somatic mosaicism has been established as a cause of miscarriage, birth defects, developmental delay and cancer[3–9]. Because mosaicism can be benign or may occur with diverse clinical phenotypes, there are no accurate estimates of its frequency in the general population[3,6]. On rare occasions, the propensity to develop chromosomal abnormalities is inherited and leads to multiple phenotypic abnormalities, including cancer predisposition, as reported in families with mutations in the *BUB1B* and *CEP57* genes[10,11]. Recently, two groups have identified somatic mosaic mutations in *IDH1* and *IDH2* in tumors of individuals with Ollier disease and Maffucci syndrome[12,13], and another group has characterized somatic mosaicism of an *HRAS* mutation in an individual with urothelial cancer and epidermal nevus[14]. Recent work in a population of twins has suggested that the detection of somatic structural variants in blood increases with age and may be related to a reduction in blood cell clonality[15]. In this report, we broadly define mosaic chromosomal abnormalities as the presence of both normal karyotypes as well as those with large structural genomic events resulting in alteration of copy number or loss of heterozygosity in distinct and detectable subpopulations of cells, regardless of the clonal or developmental origin of the subpopulations.

Recently, we reported on 1,991 individuals from the Spanish Bladder Cancer Study (SBCS) population-based case-control study in which we had performed a genome-wide association study (GWAS) of adult-onset bladder cancer using DNA obtained from blood or buccal samples[16]. The SNP array data generated for the GWAS were subsequently used to detect clonal mosaic abnormalities in the autosomes of 1.7% of study subjects, suggesting a higher frequency of these abnormalities in adults than was previously suspected. Even though somatic mosaicism has been implicated in several cancers, this study did not find a significant difference in the frequency of large chromosomal abnormalities between cases and controls. We used a computational algorithm to detect 42 large mosaic events involving two or more distinct clones in DNA extracted from blood or buccal samples, and we experimentally validated the findings using multiplex ligation-dependent probe amplification (MLPA) and microsatellite analysis (as well as FISH in one subset), establishing the robustness of the software detection method. We found a similar proportion of cells carrying each event for five of six events (four individuals with bladder cancer in whom three had one event and one had three separate events) in which it was possible to examine more than one tissue (whole blood and bladder mucosa), suggesting an early embryonic origin of the somatic mutation leading to the observed mosaic chromosomal abnormalities[16].

RESULTS

Study overview

In this report, we extended our analysis of clonal mosaic abnormalities in the autosomes to 57,853 individuals (including those previously published[16]). We tested 31,717 cancer cases and 26,136 cancer-free controls for evidence of mosaic abnormalities using genome-wide SNP array data generated as part of 13 distinct cancer GWAS drawn from 48 epidemiological case-control and case-cohort studies (**Supplementary Table 1**). DNA samples were extracted from blood or buccal samples using a variety of collection and extraction techniques and were genotyped using one or more Infinium Human SNP array from Illumina (including versions of Hap300, Hap240, Hap550, Hap610, Hap660, Hap1, Omni Express and Omni1;

A full list of authors and affiliations appears at the end of the paper.

Received 29 September 2011; accepted 9 April 2012; published online 6 May 2012; doi:10.1038/ng.2270

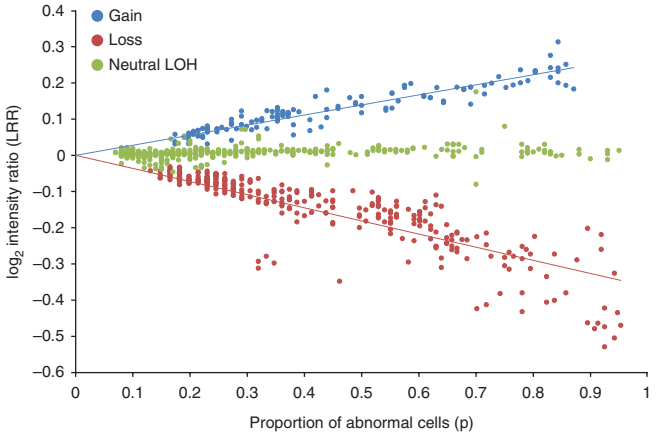

Figure 1 Characteristics of detectable clonal mosaic events. Detectable clonal mosaic events plotted by proportion of abnormal cells (p) and LRR for 681 events in 517 individuals.

see **Supplementary Data**). Genotype clusters were empirically estimated in 45 batches to optimize accuracy while minimizing potential batch effects (Online Methods).

Detection of clonal mosaic events was based on assessment of allelic imbalance and copy-number changes. We used the B-allele frequency (BAF) measurement, derived from the ratio of probe values relative to the locations of the estimated genotype-specific clusters, for initial segmentation using the mosaic alteration detection (MAD) algorithm implemented in R Genomic Alteration Detection Analysis (R-GADA) software with modifications[17,18]. The BAF and $\log_2$ relative probe intensity ratio (LRR), which provides data on copy number, were used to classify each event as copy altering (gain or loss) or neutral (reciprocal gain and loss resulting in loss of heterozygosity, LOH) and to assign the proportions of abnormal (p) and normal (1–p) cells. Mosaic proportions were required to deviate from levels expected from constitutional (non-mosaic) changes in order to exclude homozygous chromosomal segments inherited identical by descent and non-mosaic instances of trisomy, monosomy and uniparental disomy. A minimum event size threshold was set to detect only clonal mosaic events greater than 2 Mb to minimize the false discovery of constitutional copy-number variants. Copy-neutral LOH and copy-loss events could be detected for mosaic proportions between 7% and 95% (**Fig. 1**), with sensitivity that was affected by the signal-to-noise ratio characteristic of each microarray assay and sample quality. There was reduced sensitivity to distinguish between copy-neutral LOH and copy-loss events for mosaic proportions less than 15% across the autosomes. The magnitude of BAF differences for single-copy gain events was one-third of the magnitude of that for copy-neutral LOH or copy-loss events, reducing the sensitivity for calling copy-gain events. As a result, single copy-gain events could only be reliably detected for mosaic proportions between 22% and 88%, with ambiguity in distinguishing copy-gain from copy-neutral LOH for mosaic proportions of less than 20%. Because DNA was obtained for the purpose of performing a GWAS, it was not possible to further explore the developmental and clonal characteristics of mosaic events detected in these individuals (for example, by studying DNA from fractionated blood and other tissue types or determining cell composition of buccal samples or effect of DNA collection and extraction methods on detection and accuracy of the estimation of mosaic proportions). We report only autosomal chromosomal abnormalities, as analysis of the sex chromosomes presents distinct technical and interpretative challenges.

We observed 681 mosaic segments of size greater than 2 Mb on 641 autosomal chromosomes in 517 individuals, for an overall frequency of individuals with detectable mosaicism of 0.87% (**Tables 1** and **2**). The most frequent type of event observed was copy-neutral LOH (48.2%), whereas copy gains and copy losses constituted 15.1% and 34.8% of mosaic events, respectively (**Table 1**). A small proportion (1.9%) of mosaic chromosomes were complex, harboring more than one type of event. Of mosaic chromosomal events, 18.7% spanned the entire chromosome, including 62 complete trisomies, predominantly of chromosomes 8, 12 and 15. We found that 47.9% of mosaic chromosomal events began at a telomere and extended across some portion of the chromosomal arm (**Fig. 2** and **Table 1**). The majority of telomeric events were mosaic copy-neutral LOH (85.7%), most frequently on chromosome 9p (**Table 3**). The remaining mosaic chromosomal events were interstitial (31.5%), spanning neither telomere nor centromere, whereas an additional small proportion (1.8%) spanned the centromere or had more complex structure (for example, distinct events involving both telomeres but not the whole chromosome). The majority of interstitial events were mosaic copy loss (91.6%), which was most frequently observed within specific regions of chromosomes 13q and 20q (**Fig. 2**). We observed 69 individuals (46 cancer cases and 23 cancer-free controls) with clonal mosaic events on multiple chromosomes. Among cancer-free individuals, the greatest number of mosaic chromosomal events observed was five, whereas six individuals with cancer had more than five events, including two individuals with gastric cancer who each had 20 events. A list of mosaic events with phenotype data is provided (**Supplementary Data**).

Mosaic abnormalities increase with age

The strongest predictor of mosaic autosomal abnormalities was age at DNA collection. We examined the effect of increased age on the frequency of mosaicism across all studies, which predominantly included individuals over the age of 50. The frequency of cancer-free individuals with detectable clonal mosaic events increased with age from 0.23% for those under 50 to 1.91% ($P = 4.8 \times 10^{-8}$) for those between the ages of 75 and 79, with slightly higher frequencies occurring in individuals with cancer (**Fig. 3**). In individuals with early-onset cancer (under age 40), which constituted less than 5% of analyzed cases (for example, testicular cancer and osteogenic sarcoma), we did not observe an increase in mosaic abnormalities. Further studies are needed to investigate the relationship between mosaic abnormalities and cancer in children and young adults, particularly because of the strong association between mosaicism and many developmental disorders. There did

Table 1 Count and frequency of mosaic chromosomal events by event type and location

	Mosaic chromosome count					Mosaic chromosome frequency (%)				
Event location	Gain	Loss	CN LOH	Mixed	Total	Gain	Loss	CN LOH	Mixed	Total
Chromosome	62	11	42	5	120	9.7	1.7	6.6	0.8	18.7
Telomeric p	11	13	114	1	139	1.7	2.0	17.8	0.2	21.7
Telomeric q	9	10	149	0	168	1.4	1.6	23.2	0.0	26.2
Interstitial	14	185	2	1	202	2.2	28.9	0.3	0.2	31.5
Span centromere	1	1	2	0	4	0.2	0.2	0.3	0.0	0.6
Complex	0	3	0	5	8	0.0	0.5	0.0	0.8	1.2
Total	97	223	309	12	641	15.1	34.8	48.2	1.9	

CN LOH, copy-neutral loss of heterozygosity.

Table 2 Count and frequency of individuals with detectable clonal mosaic events for cancer-free individuals and by first diagnosed cancer site

	Mosaic counts			Non-mosaic counts			Mosaic frequency (%)		
Site of first cancer	Likely untreated	Possibly treated	Total	Likely untreated	Possibly treated	Total	Likely untreated	Possibly treated	Overall
Overall[a]			498			57,201			0.86
Cancer free			194			25,942			0.74
First non-hematologic cancer	185	119	304	13,865	17,394	31,259	1.32	0.68	0.96
Bladder	37	6	43	2,240	973	3,213	1.62	0.61	1.32
Breast	4	8	12	1,060	1,753	2,813	0.38	0.45	0.42
Endometrium	3	6	9	247	624	871	1.20	0.95	1.02
Esophagus	1	6	7	53	1,855	1,908	1.85	0.32	0.37
Glioma	7	2	9	1,279	441	1,720	0.54	0.45	0.52
Kidney	21	3	24	1,241	325	1,566	1.66	0.91	1.51
Lung	73	26	99	4,647	2,605	7,252	1.55	0.99	1.35
Osteosarcoma	0	3	3	0	760	760		0.39	0.39
Ovary	1	3	4	260	283	543	0.38	1.05	0.73
Pancreas	2	29	31	379	3,513	3,892	0.52	0.82	0.79
Prostate	32	11	43	2,116	1,410	3,526	1.49	0.77	1.20
Stomach	2	13	15	99	2,194	2,293	1.98	0.59	0.65
Testis	2	0	2	144	503	647	1.37	0.00	0.31
Other sites	0	3	3	100	155	255	0.00	1.90	1.16
Any hematologic cancer	8	11	19	34	62	96	19.05	15.07	16.52
Leukemia	8	9	17	34	11	45	19.05	45.00	27.42
Lymphoid	4	5	9	14	5	19	22.22	50.00	32.14
Myeloid	4	3	7	16	5	21	20.00	37.50	25.00
Other[b]	0	1	1	4	1	5	0.00	50.00	16.67
Lymphoma	0	2	2	0	42	42		4.55	4.55
Multiple myeloma	0	0	0	0	9	9		0.00	0.00

Non-hematological cancers are listed by first cancer site and exclude individuals diagnosed with a hematological cancer, who are shown separately.
[a]Overall total of cancer-free individuals and those with non-hematologic cancers. [b]Subtype not specified in leukemia diagnosis code.

not seem to be a relationship between age at DNA collection and the number or size of mosaic events or the proportion of abnormal cells (**Supplementary Figs. 1** and **2**).

We regressed the presence of detectable clonal mosaicism in 26,136 cancer-free individuals on age at DNA collection (in 5-year intervals), sex (male versus female), DNA source (buccal cells versus blood), smoking (ever versus never) and admixture coefficients for African and east Asian ancestry in a logistic model to identify additional factors that influence the frequency of detectable clonal mosaicism. The source of DNA was known for 87% of individuals, with 19% of samples derived from buccal cells and the remainder derived from blood. DNA source was not significantly associated with mosaicism (OR = 0.83, 95% confidence interval (CI) = 0.55–1.26; $P = 0.39$). In admixture analysis, 75% of subjects were determined to be of European ancestry, 9% of African ancestry and 16% of east Asian ancestry. Although power was limited, we determined that cancer-free individuals with African admixture were at a lower risk of being mosaic (OR = 0.43, 95% CI = 0.20–0.92; $P = 0.03$), but those with east Asian admixture were not (OR = 0.60, 95% CI = 0.32–1.15; $P = 0.12$). We did not observe an association between smoking and the frequency of mosaic abnormalities (OR = 1.04, 95% CI = 0.75–1.44; $P = 0.81$).

In 26,136 cancer-free controls and 23,093 cancer cases drawn from cancer sites that were non-sex specific and non-hematological (excluding 8,470 individuals with leukemia, lymphoma, multiple myeloma and cancers of the breast, endometrium, ovary, testis and prostate), we observed a higher frequency of males with mosaic abnormalities than females. In cancer-free individuals, we observed mosaic events in 0.56% of females and 0.87% of males (OR = 1.35, 95% CI = 0.98–1.88; $P = 0.07$); for individuals with cancer, we observed mosaic events in 0.79% of females and 1.21% of males (OR = 1.48, 95% CI = 1.08–2.03; $P = 0.015$); and, overall, 0.65% of females and 1.04% of males had mosaic events (OR = 1.42, 95% CI = 1.14–1.80; $P = 0.002$) in logistic models adjusted for cancer diagnosis (if applicable), age at DNA collection, ancestry, DNA source and smoking. These differences could be due to a true sex-specific effect akin to different sex-specific mutation and recombination rates[19]; however, the complex and heterogeneous nature of the inclusion of individual studies and the differences in their entry and selection criteria could result in spurious associations. Although this observation was consistent across cancer types, it should be confirmed in additional studies that are better designed to address this question.

Mosaic abnormalities and cancer risk

To determine the relationship between detectable mosaic autosomal abnormalities and non-hematological cancers, we regressed the presence of detectable clonal mosaicism on cancer diagnosis, age, sex, DNA source, smoking and ancestry in a logistic model. We observed a modest increase in cancer risk for mosaic individuals (OR = 1.27, 95% CI = 1.05–1.52; $P = 0.012$) (**Table 2** and **Supplementary Table 2**). Notable associations were observed in stratified analyses of lung (OR = 1.56, 95% CI = 1.18–2.08; $P = 0.002$) and kidney (OR = 1.98, 95% CI = 1.27–3.06; $P = 0.002$) cancers, both of which are tobacco-associated malignancies. However, no cancer site–specific associations were observed for bladder, esophageal, stomach and pancreatic cancers, which are also typically associated with tobacco use. There was no significant association in non-hematological cancer cases overall between smoking (ever versus never) and the

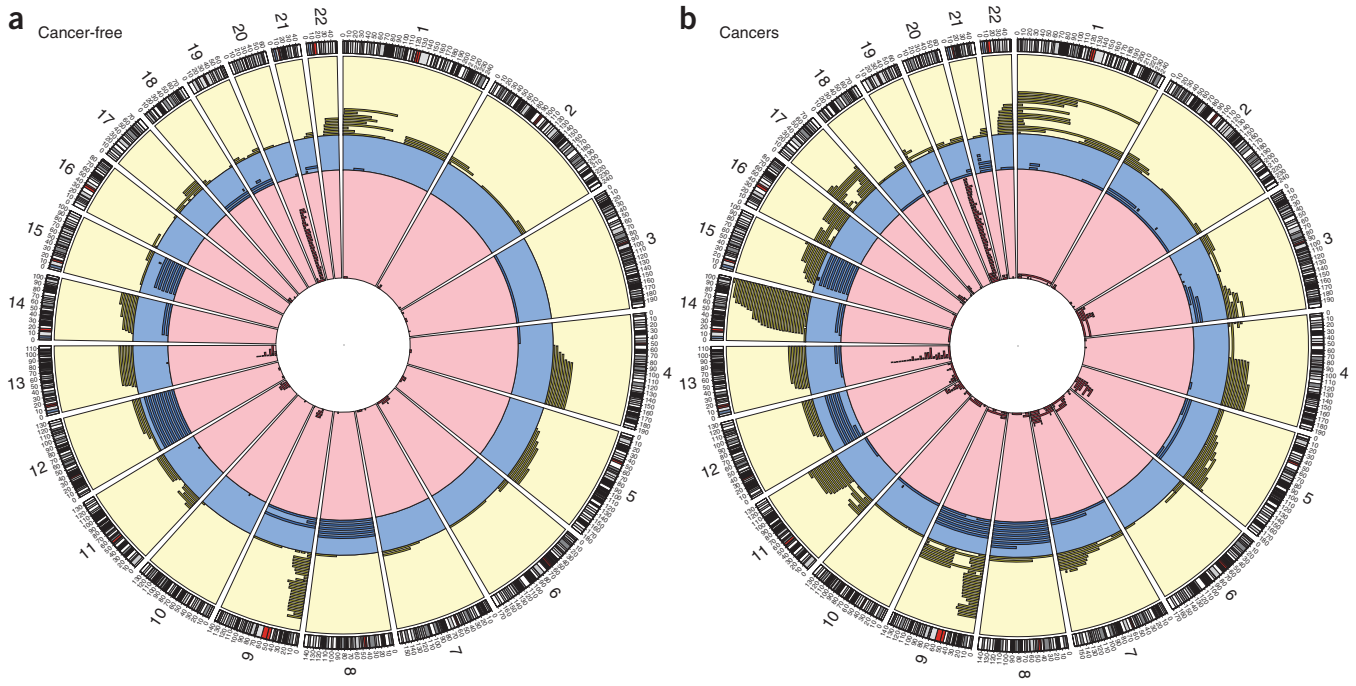

Figure 2 Circular genomic plot of detectable clonal mosaic events. Genomic location of detectable clonal mosaic events. Outer rings are the autosomes 1 to 22. Yellow region, events of copy-neutral LOH; blue region, copy-gain events; red region, copy-loss events. (**a**) Events in cancer-free controls. (**b**) Events in cancer cases. The distribution of the number of clonal mosaic chromosomal events per individual is shown in **Supplementary Table 3.**

frequency of mosaicism (OR = 1.19, 95% CI = 0.92–1.54; $P = 0.19$) or when stratified by cancer site (data not shown).

In an analysis of the subset of 14,050 individuals with cancer for whom it was possible to determine that DNA was likely obtained before or at the time of diagnosis and before treatment with radiation or chemotherapy for a primary tumor (designated as likely untreated), we observed a stronger association between mosaic abnormalities and diagnosis with non-hematological cancer (OR = 1.45, 95% CI = 1.18–1.80; $P = 0.0005$). The associations for lung and kidney cancers were also increased in significance (**Table 2**). It is notable that the evidence for association with non-hematological cancer was diminished in individuals who were potentially treated (OR = 1.03,

Table 3 Distribution and frequency of recurrent detectable clonal mosaic events

	Mosaic counts						Mosaic frequency (%)				
	13q (del)	20q (del)	9p (CN LOH)	14q (CN LOH)	Other	Total	13q (del)	20q (del)	9p (CN LOH)	14q (CN LOH)	Other
Overall	33	77	56	35	480	681	5	11	8	5	70
Cancer free	10	30	28	7	150	225	4	13	12	3	67
Cancer diagnosis	23	47	28	28	330	456	5	10	6	6	72
First non-hematologic cancer											
Bladder	2	4	2	7	35	50	4	8	4	14	70
Breast	1	1	0	1	13	16	6	6		6	81
Endometrium	0	1	3	1	6	11		9	27	9	55
Esophagus	0	1	1	1	6	9		11	11	11	67
Glioma	0	2	2	0	6	10		20	20		60
Kidney	1	3	0	4	20	28	4	11		14	71
Lung	9	14	10	7	90	130	7	11	8	5	69
Osteosarcoma	0	0	1	0	6	7			14		86
Ovary	0	2	0	0	2	4		50			50
Pancreas	1	4	1	1	29	36	3	11	3	3	81
Prostate	4	10	4	1	33	52	8	19	8	2	63
Stomach	0	1	4	3	55	63		2	6	5	87
Testis	0	0	0	0	4	4					100
Other sites	0	0	0	1	3	4				25	75
Any hematologic cancer											
Leukemia	5	3	0	1	20	29	17	10		3	69
Lymphoma	0	1	0	0	2	3		33			67

Del, deletion; CN LOH, copy-neutral loss of heterozygosity.

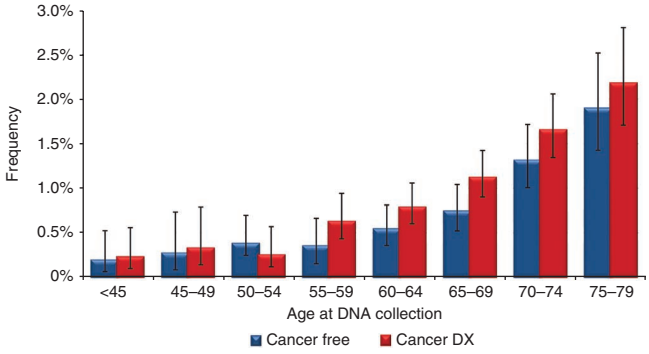

Figure 3 Frequency of detectable clonal mosaic events by age and cancer status. Analysis excluded 1,000 individuals with unknown age at DNA collection. Shown are 95% confidence intervals.

95% CI = 0.81–1.30; $P = 0.80$). We had approached this analysis with the hypothesis that there could be an increased frequency in detectable clonal mosaicism in non-hematological cancers induced by chemotherapy or radiotherapy. Thus, the observation that the frequency of mosaic events was reduced by treatment to virtually the same level seen in the cancer-free population was unexpected. Although this observed attenuation could have many explanations (for example, related to the diagnosis and treatment of a solid tumor leading to a decrease in the populations of cells with mosaic alteration), we had limited capacity to model and control for treatment effects, as many of the studies did not provide any information on treatment or only provided incomplete, retrospective ascertainment of the specifics. Although many of the participating studies were prospectively ascertained cohorts, DNA collection often occurred after cancer diagnosis. Additional studies are needed in prospectively ascertained cohorts, in addition to longitudinal studies in which multiple DNA samples are collected before and after diagnosis, in order to explore the effects of treatment and disease on mosaicism.

For the 43 individuals with hematological cancers for whom DNA was obtained at least 1 year before diagnosis, the frequency of detectable clonal mosaicism was 20% for those with myeloid leukemia and 22% for those with lymphocytic leukemia (predominantly chronic lymphocytic leukemia; **Table 2**) compared to 0.74% in 26,136 cancer-free controls (overall OR = 35.4, 95% CI = 14.7–76.6; Fisher's exact $P = 3.8 \times 10^{-11}$). Of the eight mosaic individuals with leukemia for whom DNA samples were collected at least 1 year before diagnosis, four were diagnosed with chronic lymphocytic leukemia (CLL), of whom two had a mosaic deletion in a region of chromosome 13q14 previously described to be deleted in CLL[20]. DNA was obtained more than 5 years before diagnosis for six mosaic individuals, with the longest interval being 14 years, suggesting that detectable clonal mosaicism could be a marker of hematological cancer or its precursors, such as monoclonal B cell lymphocytosis (MBL) for CLL and myelodysplastic syndrome for acute myelogenous leukemia. Recent work shows that the majority of individuals with MBL have mono- or biallelic 13q14 abnormalities[21]. However, further studies will be needed, preferably with serial sampling before and after diagnosis, to investigate the predictive nature of detectable clonal mosaicism, especially for events involving regions of chromosomes 13 and 20 in leukemia risk[20].

We further explored the four most recurrently altered regions , those detected in more than 20 individuals, which also harbored well-known cancer genes (as noted in the Catalogue of Somatic Mutations in Cancer (COSMIC)[22] and Mitelman databases; these were on chromosomes 9p (copy-neutral LOH), 13q (deletion), 14 (copy-neutral LOH) and 20q (deletion) (**Table 3**). Notably, the most recurrent mosaic events were observed in cancer-free individuals as well as across multiple individuals with solid tumors. We observed a comparable frequency of mosaic events in non-hematologic cancer cases and cancer-free controls for three of the regions, whereas chromosome 14 copy-neutral LOH abnormalities were more frequent in non-hematological cancer cases (OR = 3.32, 95% CI = 1.42–9.00, Fisher's exact $P = 0.003$), particularly in individuals with bladder or kidney cancer. Copy-neutral LOH in this region of chromosome 14 has been associated with increased susceptibility to sporadic cancers, and this region includes imprinted genes, such as the tumor suppressing non-coding RNA, *MEG3* (encoding maternally expressed gene 3)[8,23]. The recurrent segmental deletion of 13q14 was observed in 5 leukemia cases but also in 18 individuals with solid tumors (9 with lung cancer and 4 with prostate cancer) and in 10 cancer-free individuals. This region includes the tumor suppressor gene *DLEU7* (encoding deleted in leukemia 7) and the related genes, *DLEU1* and *DLEU2*, the latter of which encodes two microRNAs within one of its introns (*miR-15a* and *miR-16-1*)[24–26]. The retinoblastoma gene *RB1* was also included within the affected region in a subset of cases with a mosaic deletion of 13q14. It cannot be ruled out that these individuals have either undiagnosed CLL or MBL. We observed the 20q deletion in two individuals with myeloid leukemia, as has been described previously[27], but also in cancer-free individuals and in individuals with solid tumors.

The accuracy of our software methods in detecting clonal mosaic abnormalities has previously been addressed, and we were able to validate 100% of 42 events in 34 individuals from the Spanish Bladder Cancer Study using confirmatory cytogenetic assays[16] (**Supplementary Fig. 3**). We also performed a comparison of mosaic events in samples from the EAGLE and PLCO lung cancer studies, which were independently analyzed as part of the Gene-Environment Association Studies (GENEVA) consortium report on mosaic events[28]. A total of 83 mosaic events in individuals from the Environment and Genetics in Lung Cancer Etiology Study (EAGLE) and Prostate, Lung, Colorectal, Ovarian Cancer Screening Trial (PLCO) lung cancer studies were detected in common, 20 additional events of less than 2 Mb in size and 8 events of greater than 2 Mb in size were detected by GENEVA and not by our study, and we detected 20 additional events (of >2 Mb in size) that were not detected by GENEVA. Although additional cytogenetic or molecular validation was not performed, neither method detected notable false positive events, according to manual review of the data. The concordance rate was 75% if considering events of >2 Mb in size (the cutoff for this analysis) or 63% if considering all events, both of which are considerably better than the 25–50% concordance rates observed across CNV detection methods[29–31]. Our method is more conservative in the size of events detected, whereas the GENEVA method is more conservative with respect to sample quality but provides calls for smaller events when assay quality is sufficient. Better approaches are needed to accurately characterize events of smaller size as either mosaic or constitutional and to estimate their frequency. Further improvements to data normalization, segmentation and event classification methods will also likely reduce false negative detection rates.

DISCUSSION

Our study has important implications for the design and analysis of molecular epidemiology studies in cancer, as well as the somatic characterization of cancer genomes, such as The Cancer Genome Atlas[32] and the International Cancer Genome[33]. Investigators will need to carefully analyze samples used as exemplars of germline

DNA for somatic alterations, such as detectable clonal mosaicism. Otherwise, comparisons between "somatic cells" (used as a surrogate for germline) and tumors may result in implausible somatic changes (for example, large gains of heterozygosity), and it may be impossible to determine whether somatic events predate changes secondary to driver mutations. Because the method to detect mosaic events with next-generation sequencing technologies is neither routine nor well understood, for the near future, it may be prudent to continue to use SNP microarrays for such analyses. As a result of the increased frequency of detectable clonal mosaicism with age, this will be particularly important for the analysis of epithelial cancers, which characteristically occur in the older population. For future large-scale GWAS in prospective studies, it may be wise to consider analyzing the earliest DNA samples before diagnosis and to consider the time from collection to diagnosis in the analysis of longitudinally collected biospecimens.

We have extended our initial observation that detectable clonal mosaicism of the autosomes is present in the population with unexpectedly high frequency, particularly in the aging genome. A recent study of detectable clonal mosaicism in twins reported an increase in frequency with age and suggested that this could lead to a less diverse blood cell population and immune system[15]. These emerging data identify a number of critical issues in the mechanisms underlying a possible shift in the repertoire of clones with large structural abnormalities. For example, cells with abnormal karyotypes could have an early developmental origin in which a somatic event in a single stem cell progenitor during embryogenesis could become apparent when cellular diversity decreases with age and cell populations become increasingly oligoclonal. Higher rates of detectable clonal mosaicism in older cancer-free individuals could also be due to increased rates of somatic mutation or diminished capacity for genomic maintenance, such as with telomere attrition[34], leading to proliferation of somatically altered cell populations. A survival bottleneck of cellular progenitors could also lead to observable mosaic alterations that were previously below the threshold of detection but subsequently expanded due to positive selection. Further work is required to unravel the underlying mechanisms that result in mosaic abnormalities, particularly in respect to how and when altered clones are created, tissue specificity and the timing of expansion of distinct populations of cells with age. Finally, these findings underscore the importance of considering the role and time-dependent nature of somatic events in the etiology of cancer and other late-onset diseases.

URLs. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, http://cgap.nci.nih.gov/Chromosomes/Mitelman; Database of Genomic Variants, http://projects.tcag.ca/variation/; R, http://www.R-project.org/; GLU, http://code.google.com/p/glu-genetics/; supplementary data for this study, http://cgf.nci.nih.gov/supplementarydata/NGA31644R2SupplementaryData.zip.

METHODS

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

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

This research was supported by the Intramural Research Program and by contract number HHSN261200800001E of the US National Institutes of Health (NIH), NCI. Support for each contributing study is listed in the **Supplementary Note**. We thank C. Laurie and B. Weir for constructive discussion and a comparison of methodology and results for the GENEVA study. The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the National Cancer Institute, the National Institute for Occupational Safety and Health or the Maryland Cancer Registry.

AUTHOR CONTRIBUTIONS

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

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

Published online at http://www.nature.com/doifinder/10.1038/ng.2270.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Youssoufian, H. & Pyeritz, R.E. Mechanisms and consequences of somatic mosaicism in humans. *Nat. Rev. Genet.* **3**, 748–758 (2002).
2. Notini, A.J., Craig, J.M. & White, S.J. Copy number variation and mosaicism. *Cytogenet. Genome Res.* **123**, 270–277 (2008).
3. Hsu, L.Y. *et al.* Proposed guidelines for diagnosis of chromosome mosaicism in amniocytes based on data derived from chromosome mosaicism and pseudomosaicism studies. *Prenat. Diagn.* **12**, 555–573 (1992).
4. Menten, B. *et al.* Emerging patterns of cryptic chromosomal imbalance in patients with idiopathic mental retardation and multiple congenital anomalies: a new series of 140 patients and review of published reports. *J. Med. Genet.* **43**, 625–633 (2006).
5. Lu, X.Y. *et al.* Genomic imbalances in neonates with birth defects: high detection rates by using chromosomal microarray analysis. *Pediatrics* **122**, 1310–1318 (2008).
6. Conlin, L.K. *et al.* Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Hum. Mol. Genet.* **19**, 1263–1275 (2010).
7. Heim, S. & Mitelman, F. Nonrandom chromosome abnormalities in cancer—an overview. in *Cancer Cytogenetics* (eds. Mitelman, F. & Heim, S.) (John Wiley & Sons, Hoboken, New Jersey, 2009).
8. Tuna, M., Knuutila, S. & Mills, G.B. Uniparental disomy in cancer. *Trends Mol. Med.* **15**, 120–128 (2009).
9. Solomon, D.A. *et al.* Mutational inactivation of *STAG2* causes aneuploidy in human cancer. *Science* **333**, 1039–1043 (2011).
10. Rio Frio, T. *et al.* Homozygous *BUB1B* mutation and susceptibility to gastrointestinal neoplasia. *N. Engl. J. Med.* **363**, 2628–2637 (2010).
11. Snape, K. *et al.* Mutations in *CEP57* cause mosaic variegated aneuploidy syndrome. *Nat. Genet.* **43**, 527–529 (2011).
12. Amary, M.F. *et al.* Ollier disease and Maffucci syndrome are caused by somatic mosaic mutations of *IDH1* and *IDH2*. *Nat. Genet.* **43**, 1262–1265 (2011).
13. Pansuriya, T.C. *et al.* Somatic mosaic *IDH1* and *IDH2* mutations are associated with enchondroma and spindle cell hemangioma in Ollier disease and Maffucci syndrome. *Nat. Genet.* **43**, 1256–1261 (2011).
14. Hafner, C., Toll, A. & Real, F.X. *HRAS* mutation mosaicism causing urothelial cancer and epidermal nevus. *N. Engl. J. Med.* **365**, 1940–1942 (2011).
15. Forsberg, L.A. *et al.* Age-related somatic structural changes in the nuclear genome of human blood cells. *Am. J. Hum. Genet.* **90**, 217–228 (2012).
16. Rodíguez-Santiago, B. *et al.* Mosaic uniparental disomies and aneuploidies as large structural variants of the human genome. *Am. J. Hum. Genet.* **87**, 129–138 (2010).

17. González, J.R. *et al.* A fast and accurate method to detect allelic genomic imbalances underlying mosaic rearrangements using SNP array data. *BMC Bioinformatics* **12**, 166 (2011).
18. Pique-Regi, R., Caceres, A. & Gonzalez, J.R. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics* **11**, 380 (2010).
19. Hedrick, P.W. Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution* **61**, 2750–2771 (2007).
20. Döhner, H. *et al.* Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* **343**, 1910–1916 (2000).
21. Lanasa, M.C. *et al.* Immunophenotypic and gene expression analysis of monoclonal B-cell lymphocytosis shows biologic characteristics associated with good prognosis CLL. *Leukemia* **25**, 1459–1466 (2011).
22. Forbes, S.A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945–D950 (2011).
23. Benetatos, L., Vartholomatos, G. & Hatzimichael, E. *MEG3* imprinted gene contribution in tumorigenesis. *Int. J. Cancer* **129**, 773–779 (2011).
24. Lee, S. *et al.* Forerunner genes contiguous to *RB1* contribute to the development of *in situ* neoplasia. *Proc. Natl. Acad. Sci. USA* **104**, 13732–13737 (2007).
25. Migliazza, A. *et al.* Nucleotide sequence, transcription map, and mutation analysis of the 13q14 chromosomal region deleted in B-cell chronic lymphocytic leukemia. *Blood* **97**, 2098–2104 (2001).
26. Pekarsky, Y., Zanesi, N. & Croce, C.M. Molecular basis of CLL. *Semin. Cancer Biol.* **20**, 370–376 (2010).
27. Gurvich, N. *et al.* L3MBTL1 polycomb protein, a candidate tumor suppressor in del(20q12) myeloid disorders, is essential for genome stability. *Proc. Natl. Acad. Sci. USA* **107**, 22552–22557 (2010).
28. Laurie, C.C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nat. Genet.* published online, doi:10.1038/ng.2271 (6 May 2012).
29. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* **29**, 512–520 (2011).
30. Dellinger, A.E. *et al.* Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res.* **38**, e105 (2010).
31. Marenne, G. *et al.* Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. *Hum. Mutat.* **32**, 240–248 (2011).
32. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
33. Hudson, T.J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
34. Sahin, E. & Depinho, R.A. Linking functional decline of telomeres, mitochondria and stem cells during ageing. *Nature* **464**, 520–528 (2010).

Kevin B Jacobs[1,2], Meredith Yeager[1,2], Weiyin Zhou[1,2], Sholom Wacholder[1], Zhaoming Wang[1,2], Benjamin Rodriguez-Santiago[3–5], Amy Hutchinson[1,2], Xiang Deng[1,2], Chenwei Liu[1,2], Marie-Josephe Horner[1], Michael Cullen[1,2], Caroline G Epstein[1], Laurie Burdett[1,2], Michael C Dean[6], Nilanjan Chatterjee[1], Joshua Sampson[1], Charles C Chung[1], Joseph Kovaks[1], Susan M Gapstur[7], Victoria L Stevens[7], Lauren T Teras[7], Mia M Gaudet[7], Demetrius Albanes[1], Stephanie J Weinstein[1], Jarmo Virtamo[8], Philip R Taylor[1], Neal D Freedman[1], Christian C Abnet[1], Alisa M Goldstein[1], Nan Hu[1], Kai Yu[1], Jian-Min Yuan[9,10], Linda Liao[1], Ti Ding[11], You-Lin Qiao[12], Yu-Tang Gao[13], Woon-Puay Koh[14], Yong-Bing Xiang[13], Ze-Zhong Tang[11], Jin-Hu Fan[12], Melinda C Aldrich[15,16], Christopher Amos[17], William J Blot[16,18], Cathryn H Bock[19,20], Elizabeth M Gillanders[21], Curtis C Harris[22], Christopher A Haiman[23], Brian E Henderson[23], Laurence N Kolonel[24], Loic Le Marchand[24], Lorna H McNeill[25,26], Benjamin A Rybicki[27], Ann G Schwartz[19,20], Lisa B Signorello[16,18,28], Margaret R Spitz[29], John K Wiencke[30], Margaret Wrensch[30], Xifeng Wu[17], Krista A Zanetti[21,22], Regina G Ziegler[1], Jonine D Figueroa[1], Montserrat Garcia-Closas[1,31], Nuria Malats[32], Gaelle Marenne[32], Ludmila Prokunina-Olsson[1], Dalsu Baris[1], Molly Schwenn[33], Alison Johnson[34], Maria Teresa Landi[1], Lynn Goldin[1], Dario Consonni[35,36], Pier Alberto Bertazzi[35,36], Melissa Rotunno[1], Preetha Rajaraman[1], Ulrika Andersson[37], Laura E Beane Freeman[1], Christine D Berg[38], Julie E Buring[39,40], Mary A Butler[41], Tania Carreon[41], Maria Feychting[42], Anders Ahlbom[42], J Michael Gaziano[40,43,44], Graham G Giles[45,46], Goran Hallmans[47], Susan E Hankinson[48], Patricia Hartge[1], Roger Henriksson[37,49], Peter D Inskip[1], Christoffer Johansen[50], Annelie Landgren[1], Roberta McKean-Cowdin[23], Dominique S Michaud[51,52], Beatrice S Melin[37], Ulrike Peters[53,54], Avima M Ruder[41], Howard D Sesso[40], Gianluca Severi[45,46], Xiao-Ou Shu[16,28], Kala Visvanathan[55], Emily White[53,54], Alicja Wolk[56], Anne Zeleniuch-Jacquotte[57], Wei Zheng[16,28], Debra T Silverman[1], Manolis Kogevinas[58–61], Juan R Gonzalez[59–61], Olaya Villa[3–5], Donghui Li[62], Eric J Duell[63], Harvey A Risch[64], Sara H Olson[65], Charles Kooperberg[53], Brian M Wolpin[48,66], Li Jiao[67], Manal Hassan[62], William Wheeler[68], Alan A Arslan[69–71], H Bas Bueno-de-Mesquita[72,73], Charles S Fuchs[48,66], Steven Gallinger[74], Myron D Gross[75], Elizabeth A Holly[76], Alison P Klein[77], Andrea LaCroix[53], Margaret T Mandelson[53,78], Gloria Petersen[79], Marie-Christine Boutron-Ruault[80], Paige M Bracci[76], Federico Canzian[81], Kenneth Chang[82], Michelle Cotterchio[83], Edward L Giovannucci[48,84,85], Michael Goggins[77,86,87], Judith A Hoffman Bolton[55], Mazda Jenab[88], Kay-Tee Khaw[89], Vittorio Krogh[90], Robert C Kurtz[91], Robert R McWilliams[92], Julie B Mendelsohn[1], Kari G Rabe[79], Elio Riboli[51], Anne Tjønneland[50], Geoffrey S Tobias[1], Dimitrios Trichopoulos[84,93], Joanne W Elena[38], Herbert Yu[64], Laufey Amundadottir[1], Rachael Z Stolzenberg-Solomon[1], Peter Kraft[48,84], Fredrick Schumacher[23], Daniel Stram[23], Sharon A Savage[1], Lisa Mirabello[1], Irene L Andrulis[74,94], Jay S Wunder[74,94], Ana Patiño García[95], Luis Sierrasesúmaga[95], Donald A Barkauskas[23], Richard G Gorlick[96,97], Mark Purdue[1], Wong-Ho Chow[1], Lee E Moore[1], Kendra L Schwartz[98], Faith G Davis[99], Ann W Hsing[1], Sonja I Berndt[1], Amanda Black[1], Nicolas Wentzensen[1], Louise A Brinton[1], Jolanta Lissowska[100], Beata Peplonska[101], Katherine A McGlynn[1], Michael B Cook[1], Barry I Graubard[1], Christian P Kratz[1,102], Mark H Greene[1], Ralph L Erickson[103],

David J Hunter[48,84], Gilles Thomas[104], Robert N Hoover[1], Francisco X Real[3,105], Joseph F Fraumeni Jr[1], Neil E Caporaso[1], Margaret Tucker[1], Nathaniel Rothman[1], Luis A Pérez-Jurado[3,4,106] & Stephen J Chanock[1,106]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), Rockville, Maryland, USA. [2]Core Genotyping Facility, SAIC-Frederick, Inc, NCI-Frederick, Frederick, Maryland, USA. [3]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain. [4]Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain. [5]Quantitative Genomic Medicine Laboratory, qGenomics, Barcelona, Spain. [6]Laboratory of Experimental Immunology, Center for Cancer Research, NCI-Frederick, Frederick, Maryland, USA. [7]Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA. [8]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland. [9]Department of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, Minnesota, USA. [10]University of Pittsburgh Cancer Institute, Pittsburgh, Pennsylvania, USA. [11]Shanxi Cancer Hospital, Taiyuan, People's Republic of China. [12]Department of Epidemiology, Cancer Institute (Hospital), Chinese Academy of Medical Sciences, Beijing, People's Republic of China. [13]Shanghai Cancer Institute, Shanghai, People's Republic of China. [14]Saw Swee Hock School of Public Health, National University of Singapore, Singapore. [15]Department of Thoracic Surgery, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [16]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University Medical Center, Nashville, Tennessee, USA. [17]Department of Epidemiology, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [18]International Epidemiology Institute, Rockville, Maryland, USA. [19]Karmanos Cancer Institute, Wayne State University, Detroit, Michigan, USA. [20]Department of Oncology, Wayne State University School of Medicine, Detroit, Michigan, USA. [21]Division of Cancer Control and Population Sciences, NCI, Bethesda, Maryland, USA. [22]Laboratory of Human Carcinogenesis, Center for Cancer Research, NCI, Bethesda, Maryland, USA. [23]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, USA. [24]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA. [25]Department of Health Disparities Research, Division of Office of the Vice-President, Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [26]Center for Community-Engaged and Translational Research, Duncan Family Institute, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [27]Department of Public Health Sciences, Henry Ford Hospital, Detroit, Michigan, USA. [28]Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [29]Dan L Duncan Cancer Center, Baylor College of Medicine, Houston, Texas, USA. [30]Department of Neurological Surgery, University of California, San Francisco, California, USA. [31]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [32]Genetics and Molecular Epidemiology Group, Centro Nacional de Investigaciones Oncológicas (CNIO), Madrid, Spain. [33]Maine Cancer Registry, Augusta, Maine, USA. [34]Vermont Cancer Registry, Burlington, Vermont, USA. [35]Department of Occupational and Environmental Health, University of Milan, Milan, Italy. [36]Unit of Epidemiology, Fondazione Istituto di Ricevero e Cura a Carattere Scientifico (IRCCS), Ospedale Maggiore Policlinico, Mangiagalli e Regina Elena, Milan, Italy. [37]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden. [38]Clinical and Translational Epidemiology Branch, Division of Cancer Control and Population Sciences, NCI, Bethesda, Maryland, USA. [39]Department of Ambulatory Care and Prevention, Harvard Medical School, Boston, Massachusetts, USA. [40]Division of Preventive Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [41]National Institute for Occupational Safety and Health, Centers for Disease Control and Prevention, Cincinnati, Ohio, USA. [42]Division of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [43]Division of Aging, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [44]Massachusetts Veteran's Epidemiology, Research and Information Center, Geriatric Research Education and Clinical Center, Veterans Affairs Boston Healthcare System, Boston, Massachusetts, USA. [45]Cancer Epidemiology Centre, The Cancer Council of Victoria, Melbourne, Victoria, Australia. [46]Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, University of Melbourne, Melbourne, Victoria, Australia. [47]Department of Public Health and Clinical Medicine, Umeå University, Umeå, Sweden. [48]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [49]Department of Oncology, Karolinska University Hospital, Stockholm, Sweden. [50]Institute of Cancer Epidemiology, Danish Cancer Society, Copenhagen, Denmark. [51]Division of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [52]Department of Epidemiology, Division of Biology and Medicine, Brown University, Providence, Rhode Island, USA. [53]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [54]Department of Epidemiology, University of Washington, Seattle, Washington, USA. [55]Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [56]Division of Nutritional Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [57]Department of Environmental Medicine, Division of Epidemiology, New York University School of Medicine, New York, New York, USA. [58]National School of Public Health, Athens, Greece. [59]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain. [60]Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain. [61]Centro de Investigacion Biomedica en Red Epidemiologia y Salud Publica (CIBERESP), Barcelona, Spain. [62]Department of Gastrointestinal Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [63]Catalan Institute of Oncology (ICO), Institut d'Investigació Biomèdica de Bellvitge (IDIBELL), Barcelona, Spain. [64]Yale University School of Public Health, New Haven, Connecticut, USA. [65]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [66]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [67]Department of Medicine, Baylor College of Medicine, Houston, Texas, USA. [68]Information Management Services, Silver Spring, Maryland, USA. [69]Department of Obstetrics and Gynecology, New York University School of Medicine, New York, New York, USA. [70]Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA. [71]New York University Cancer Institute, New York, New York, USA. [72]Department of Gastroenterology and Hepatology, University Medical Center Utrecht, Utrecht, The Netherlands. [73]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. [74]Fred A Litwin Centre for Cancer Genetics, Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada. [75]Department of Laboratory Medicine and Pathology, School of Medicine, University of Minnesota, Minneapolis, Minnesota, USA. [76]Department of Epidemiology and Biostatistics, University of California, San Francisco, California, USA. [77]Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [78]Group Health Center for Health Studies, Seattle, Washington, USA. [79]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA. [80]Institut National de la Santé et de la Recherche Médicale (INSERM), Paris-Sud University, Institut Gustave-Roussy, Villejuif, France. [81]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [82]Comprehensive Digestive Disease Center, University of California Irvine Medical Center, Orange, California, USA. [83]Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [84]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [85]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. [86]Department of Medicine, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [87]Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [88]International Agency for Research on Cancer (IARC)/World Health Organization (WHO), Lyon, France. [89]Department of Public Health and Primary Care, Clinical Gerontology, Addenbrooke's Hospital, University of Cambridge, Cambridge, UK. [90]Nutritional Epidemiology Unit, IRCCS, Istituto Nazionale dei Tumori, Milan, Italy. [91]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [92]Division of Medical Oncology, Mayo Clinic, Rochester, Minnesota, USA. [93]Bureau of Epidemiologic Research, Academy of Athens, Athens, Greece. [94]Mount Sinai Hospital Christopher Sharp Centre for Surgery and Oncology, Toronto, Ontario, Canada. [95]Department of Pediatrics, Clínica Universidad de Navarra, Pamplona, Spain. [96]Department of Molecular Pharmacology, Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, USA. [97]Department of Pediatrics, Albert Einstein College of Medicine of Yeshiva University, Bronx, New York, USA. [98]Department of Family Medicine and Public Health Sciences, Wayne State University, Detroit, Michigan, USA. [99]Division of Epidemiology and Biostatistics, School of Public Health, University of Illinois at Chicago, Chicago, Illinois, USA. [100]Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland. [101]Nofer Institute of Occupational Medicine, Lodz, Poland. [102]Zentrum für Kinderheilkunde und Jugendmedizin, Klinik für Pädiatrische Hämatologie und Onkologie, Medizinische Hochschule Hannover, Hannover, Germany. [103]Walter Reed Army Institute of Research, Silver Spring, Maryland, USA. [104]Synergie-Lyon-Cancer, Universite Lyon 1, Centre Leon Berard, Lyon, France. [105]Epithelial Carcinogenesis Group, CNIO, Madrid, Spain. [106]These authors jointly directed this work. Correspondence should be addressed to S.J.C. (chanocks@mail.nih.gov).

ONLINE METHODS

Study design. The US National Cancer Institute's Division of Cancer Epidemiology and Genetics (DCEG) and Core Genotyping Facility (CGF) have performed multiple GWAS in collaboration with many groups from around the world to detect heritable genetic risk factors for a range of solid tumor cancer sites (**Supplementary Table 1**). We extended analysis of large autosomal abnormalities to a bladder cancer study from Spain, which included laboratory confirmation of all large mosaic events[16], to examine the autosomes of 64,789 human DNA samples from 57,853 individuals who were genotyped using Illumina Infinium BeadArray assays (including the 1,991 individuals previously studied). Although relatively rare, large-scale chromosomal abnormalities were routinely detected previously; those samples were excluded from analysis to find common SNPs that contribute to cancer susceptibility. An analysis of sex chromosomes is not presented in this report. All subjects included were recruited under the supervision of an Institutional Review Board and provided informed consent though their respective study centers.

Data analysis. Analysis began using fluorescence signals imported into Illumina GenomeStudio software to compute affinity-normalized probe intensities and to call genotypes using standard methods. Genotype clusters were empirically derived from study data in 45 batches to maximize the power to estimate accurate clusters for each genotype state and to minimize batch effects due to arrays, DNA collection and laboratory processing (see ref. 35 for more details). Data on called genotype, genotype call quality and genotype probe intensities for each assay were exported from GenomeStudio and imported for analysis into a custom software pipeline.

Detection of mosaic chromosomal events used information on copy-number changes and allelic imbalance. The LRR value for each SNP or CNV assay provided data on probe intensity relative to that of the estimated genotype-specific cluster location. Information on allelic ratio was provided by BAF, which is derived from the ratio of probe values relative to the locations of the estimated genotype-specific cluster locations. LRR and BAF were estimated by the GenomeStudio software, but these estimates can suffer from bias due to the properties of the assay chemistry and fluorescent dyes used in the probes, which can reduce precision in estimating copy-number and allelic imbalances. We implemented a method similar to one that was previously described[36] to re-estimate LRR and BAF after applying quantile normalization with an enhanced multiple regression model, incorporating within-chip signal rescaling terms and a polynomial correction for GC and CpG waves. The correction model was an extension to a previously published method[37] with terms for multiple window sizes for the proportion of GC and CpG content around the genomic location of each set of probes. The CG and CpG correction model was estimated per sample, as the phenomenon is modulated primarily by the concentration of DNA input. Finally, LRR and BAF were recomputed using the resulting quantile-normalized and GC- and CpG-corrected values, as described[38]. Reduction in variance of the LRR values is shown (**Supplementary Fig. 3**). Additional mathematical details of these methods are available in the **Supplementary Note**.

Each subject, sample and assay included in this analysis had (i) sufficient informed consent to participate in this analysis and were not withdrawn from the study at the request of the principal investigators; (ii) identity consistent with duplicate samples; (iii) diagnosis as cancer free or available information on the first site of diagnosed cancer; (iv) a minimum array completion rate of 88% for only SNP assays that passed GenomeStudio thresholds for calling; and (v) assays with LRR s.d. (σ_{LRR}) < 0.33 and heterozygote BAF s.d. ($\sigma_{BAF\ AB}$) < 0.05 after quantile normalization, GC and CpG wave correction and BAF and LRR re-estimation. These liberal thresholds were chosen to exclude low-quality assays that would result in large numbers of false positive mosaic segment calls and yet retain samples with one or more highly abnormal chromosome that may have resulted in inflated missing genotype call rates and LRR and/or BAF variance.

Renormalized LRR and BAF values from qualifying assays were then analyzed with the MAD method implemented in the R-GADA software package[16,17] to detect whole-chromosome and large segmental events greater than 2 Mb in size. The MAD method applies the sparse Bayesian learning (SBL) algorithm to the B-deviation, which is derived from the BAF and genotype states. R-GADA was modified to compute the B-deviation, as originally defined[39], and used the minimum of the heterozygote and homozygote B-deviation values for probes with uncalled genotype, differing from the default version that excludes values with uncalled genotype. After initial segmentation, a backward elimination step then ranked the statistical significance of each breakpoint and excluded segments with weak evidence to control false positive detection rates. We applied the MAD method with the following parameters: *T* statistic of >9, SBL hyperparameter of 0.8 and segment length of ≥ 75 contiguous probes.

The default method implemented in R-GADA assigns event type based on LRR value and, separately, the proportion of mosaic and normal cells from Bdeviation values. This approach can be quite variable, so a Gaussian mixture model was fit to the BAF values of each segment with 2–4 Gaussian components, and the best-fitting model was chosen using the Akaike information criterion[40]. Each event was assigned a copy-number state and mosaic proportion based on fitting a Gaussian mixture model to estimate the location of heterozygote BAF bands and applying a classification method to choose the state that minimized the absolute distance between each state model and expected values, given the observed LRR mean and location of BAF bands for the segment. Additional details are available in the **Supplementary Note**. This mixture model approach was applied, because it is conservative when differentiating mosaic events from constitutional monosomy, trisomy, uniparental disomy, segmental copy-number variants and loss of heterozygosity due to chromosome segments inherited identical by descent.

Each event detected was classified as copy altering (gain or loss) or neutral (reciprocal gain and loss resulting in loss of heterozygosity), and the mosaic proportion of abnormal cells was estimated. False positive calls due to noisy assay data and non-mosaic copy-number variants and events resulting in constitutional loss of heterozygosity due to homozygous segments inherited by descent and uniparental disomy were also excluded from analysis based on manual review by two blinded, independent investigators; in rare circumstances, putative events were compared to established CNVs in the Database of Genomic Variants and subsequently excluded. We edited 33 putative events as a result of this manual review, typically to adjust segment boundaries or the estimated mosaic proportions.

Circular genomic plots of all mosaic events were generated using Circos software[41] for cancer-free individuals and those with cancer. Confidence intervals on frequencies were reported using asymmetric 95% confidence bounds from the Wilson Score Interval[42] method. Unadjusted analysis of count data and frequencies was performed using Fisher's Exact test for contingency tables, as implemented in the R software package. Logistic regression models were fit using the GLU software package, which was also used to estimate admixture coefficients for each subject. Admixture analysis was performed after removing chromosomes that contained putative mosaic events and was used three reference populations drawn from the International HapMap Project[43]: (i) 45 Han Chinese individuals from Beijing, China, and 45 Japanese from Tokyo, Japan (CHB and JPT); (ii) 60 unrelated Yoruba individuals from Ibadan, Nigeria (YRI) and (iii) 60 unrelated individuals of Northern and Western European ancestry from Utah (CEU). This analysis resulted in estimates of admixture coefficients for east Asian, African and European ancestry and gave almost identical results to those from principal-components analysis, also used to estimate ancestry and population structure.

To determine the relationship between individuals having one or more mosaic event and their sex, age at DNA collection, smoking behavior, DNA source, ancestry and cancer diagnosis, we fit several models that regressed the presence of mosaicism for each individual on relevant covariates in a logistic model. The following covariate terms were defined for each individual: (i) sex (0 for females, 1 for males); (ii) age at DNA collection (in 5-year intervals, as a series of categorical variables with terms for ages between 45–49, 50–54, 55–59, 60–64, 65–69, 70–74, and 75 or older; the referent category included all individuals under age 45); (iii) smoking (1 for individuals who ever smoked (current, former or occasional) and 0 otherwise); (iv) unknown smoking status (1 for individuals with unknown smoking behavior and 0 otherwise); (v) DNA source (1 for individuals who contributed DNA derived from a buccal sample and 0 otherwise); (vi) unknown DNA source (1 for individuals for whom DNA was obtained from an unknown tissue type and 0 otherwise); (vii) east Asian ancestry (a continuous measure of admixture estimate); (viii) African ancestry (a continuous measure of admixture estimate); (ix) cancer diagnosis (1 for individuals diagnosed with one or more cancers and 0 if no cancer diagnosis was provided by study) and (x) possibly treated

(0 if subject had DNA collected at least 1 year before diagnosis of their first cancer or before treatment with chemotherapy or radiation could have occurred and 1 otherwise). This latter variable is highly heuristic due to the lack of available data on treatment. It is intentionally conservative, in the sense that many subjects will be listed as possibly treated, even though it is unlikely that they would have had sufficient time or stage/grade of cancer to warrant treatment with radiation or chemotherapy.

When modeling mosaicism in controls, adjustments terms were included for age at DNA collection, sex, study, DNA source, smoking and ancestry. When modeling the relationships between mosaicism and sex, individuals diagnosed with sex-specific cancer sites were excluded. In models to determine the association between mosaicism and articular sex-specific cancer sites, only controls relevant to the cancer site were included in the model (for example, only male controls were used in models of prostate cancer cases). Models of mosaicism for early-onset cancer sites (testis and osteogenic sarcoma) excluded controls over the age of 45.

35. Petersen, G.M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* **42**, 224–228 (2010).
36. Staaf, J. *et al.* Normalization of Illumina Infinium whole-genome SNP data improves copy number estimates and allelic intensity ratios. *BMC Bioinformatics* **9**, 409 (2008).
37. Diskin, S.J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
38. Peiffer, D.A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–1148 (2006).
39. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009).
40. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
41. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
42. Agresti, A. & Coull, B.A. Approximate is better than "exact" for interval estimation of binomial proportions. *Am. Stat.* **52**, 119–126 (1998).
43. Frazer, K.A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).