nature genetics

Detectable clonal mosaicism from birth to old age and its relationship to cancer

Cathy C Laurie[1,38,39], Cecelia A Laurie[1,38], Kenneth Rice[1], Kimberly F Doheny[2], Leila R Zelnick[1], Caitlin P McHugh[1], Hua Ling[2], Kurt N Hetrick[2], Elizabeth W Pugh[2], Chris Amos[3], Qingyi Wei[3], Li-e Wang[3], Jeffrey E Lee[4], Kathleen C Barnes[5], Nadia N Hansel[5], Rasika Mathias[5], Denise Daley[6], Terri H Beaty[7], Alan F Scott[8], Ingo Ruczinski[9], Rob B Scharpf[10], Laura J Bierut[11], Sarah M Hartz[11], Maria Teresa Landi[12], Neal D Freedman[12], Lynn R Goldin[12], David Ginsburg[13,14], Jun Li[15], Karl C Desch[16], Sara S Strom[17], William J Blot[18], Lisa B Signorello[18], Sue A Ingles[19], Stephen J Chanock[12], Sonja I Berndt[12], Loic Le Marchand[20], Brian E Henderson[19], Kristine R Monroe[19], John A Heit[21], Mariza de Andrade[22], Sebastian M Armasu[22], Cynthia Regnier[23,24], William L Lowe[25], M Geoffrey Hayes[25], Mary L Marazita[26], Eleanor Feingold[27], Jeffrey C Murray[28], Mads Melbye[29], Bjarke Feenstra[29], Jae H Kang[30], Janey L Wiggs[31], Gail P Jarvik[32], Andrew N McDavid[33], Venkatraman E Seshan[34], Daniel B Mirel[35], Andrew Crenshaw[35], Nataliya Sharopova[36], Anastasia Wise[37], Jess Shen[1], David R Crosslin[1], David M Levine[1], Xiuwen Zheng[1], Jenna I Udren[1], Siiri Bennett[1], Sarah C Nelson[1], Stephanie M Gogarten[1], Matthew P Conomos[1], Patrick Heagerty[1], Teri Manolio[37,39], Louis R Pasquale[31,39], Christopher A Haiman[19,39], Neil Caporaso[12,39] & Bruce S Weir[1,39]

We detected clonal mosaicism for large chromosomal anomalies (duplications, deletions and uniparental disomy) using SNP microarray data from over 50,000 subjects recruited for genome-wide association studies. This detection method requires a relatively high frequency of cells with the same abnormal karyotype (>5–10%; presumably of clonal origin) in the presence of normal cells. The frequency of detectable clonal mosaicism in peripheral blood is low (<0.5%) from birth until 50 years of age, after which it rapidly rises to 2–3% in the elderly. Many of the mosaic anomalies are characteristic of those found in hematological cancers and identify common deleted regions with genes previously associated with these cancers. Although only 3% of subjects with detectable clonal mosaicism had any record of hematological cancer before DNA sampling, those without a previous diagnosis have an estimated tenfold higher risk of a subsequent hematological cancer (95% confidence interval = 6–18).

Chromosomal mosaicism is the presence of different karyotypes in two or more cell lineages within an individual derived from a single zygote[1,2]. This karyotypic variation may arise early in development and involve both the soma and the germline or may occur later and be restricted to one or more specific cell types. In cancer, chromosomal anomalies can initiate a neoplastic clone or arise during clonal evolution and serve as clonal markers[3]. Here, we consider such clonal variation as a form of mosaicism, as the cancer cells may have acquired one or more chromosomal abnormalities, whereas other cells in the same tissue or elsewhere in the body retain a normal karyotype. Chromosomal mosaicism in humans has been well studied in embryos[4,5], fetuses from spontaneous abortions[6], children with birth defects or developmental delay[7,8] and individuals with cancer[9]. However, little is known about the type, frequency and age distribution of acquired chromosomal anomalies in large samples from the general population[9,10].

Data from genome-wide association studies provide an opportunity to detect chromosomal variation in tens of thousands of people of all ages and to investigate the association of mosaicism with disease. SNP microarray data are used routinely to detect chromosomal anomalies (copy-number variants (CNVs) and uniparental disomy (UPD)) in clinical cytogenetic laboratories[11,12] and to detect small CNVs in population studies[13–15]. However, the analytical methods used in population studies are not optimal for detecting large anomalies or mosaicism. Therefore, we developed an efficient method to identify and map large (50-kb to whole-chromosome) anomalies and mosaicism within a single DNA sample. This method requires a relatively high frequency of cells with the same abnormal karyotype (>5–10%; presumably of clonal origin) in the presence of normal cells. Therefore, we use the term 'detectable clonal mosaicism' rather than 'chromosomal mosaicism', to emphasize the observation of clones of cells with abnormal karyotype that occurred at a frequency sufficient for detection using SNP microarray data.

To detect clonal mosaicism, we analyzed DNA samples (primarily from peripheral blood) from over 50,000 people genotyped for the

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

Received 12 September 2011; accepted 9 April 2012; published online 6 May 2012; doi:10.1038/ng.2271

Table 1 Summary of GENEVA study characteristics

GENEVA study	Illumina array[a]	Relatedness	Design	Ancestry[b]	Mean age[c]	Male (%)	*n*[d]
Melanoma	Omni1M	Mostly unrelated[e]	Case-control	European	52	58	2,947
Lung Health	660W	Mostly unrelated	Cohort	European	54	63	4,087
Cleft Lip/Palate	610	Trios	Case-parent trio	European and Asian	33	52	6,860
Addiction	1M	Mostly unrelated	Case-control	European and African-American	39	46	2,790
Lung Cancer	550	Mostly unrelated	Case-control	European	66	72	5,518
Blood Clotting	Omni1M	Sibling pairs	Population sample	European	21	38	1,158
Prostate Cancer Japanese/Latino	660W	Mostly unrelated	Case-control	Asian and Hispanic	71	100	4,281
Prostate Cancer African-American	1M	Mostly unrelated	Case-control	African-American	69	100	4,338
Venous Thrombo-embolism	660W	Mostly unrelated	Case-control	European	55	49	2,591
Birth Weight Afro-Caribbean	1M	Mother-offspring duos	Population sample	Afro-Caribbean	25	25	2,254
Birth Weight European	610	Mother-offspring duos	Population sample	European	31	25	2,712
Birth Weight Hispanic	1M	Mother-offspring duos	Population sample	Hispanic	29	21	1,419
Dental Caries	610	Families and singletons	Population sample	European	36	45	3,841
Prematurity	660W	Mother-offspring duos	Case-control	European	30	26	3,725
Glaucoma	660W	Mostly unrelated	Case-control	European	67	42	1,977

[a]A full description of the array type is provided in **Supplementary Table 1**. [b]Predominant ancestry; most studies have small numbers of other ancestry groups. [c]Mean age of participants older than 15 years. The Cleft Lip/Palate, Birth Weight, Dental Caries and Prematurity studies also have substantial numbers of infants and children less than 15 years old. [d]Total number of subjects with genotyped samples analyzed in this study. [e]Unrelated is equivalent to less relatedness than second-degree relatives, based on estimation of identity-by-descent coefficients.

Gene-Environment Association Studies (GENEVA) consortium[16]. The GENEVA studies include individuals of all ages, from birth to old age, who are from several common ancestry groups and have a variety of different health conditions or are healthy controls (**Table 1**, **Supplementary Fig. 1** and **Supplementary Table 1**). Here, we characterize the types of chromosomal anomalies detected, show how the prevalence of detectable clonal mosaicism within blood cells increases with age and examine the association between mosaic anomalies and hematological cancer.

RESULTS

Types of anomalies detected

This report is focused on autosomal anomalies, defined here as deviations from the normal biparental disomic state. Anomalies were detected using log R ratio (LRR) and B-allele frequency (BAF)[17]. LRR is a measure of relative signal intensity ($\log_2$ of the ratio of observed to expected intensity, where the expectation is based on data from other samples). BAF is an estimate of the frequency of the B allele of a given SNP in the population of cells from which the DNA was extracted. In a normal cell, the BAF at any locus is either 0 (AA), 1/2 (AB) or 1 (BB), and the expected LRR is 0. Both copy-number changes and copy-neutral changes from biparental to uniparental disomy (UPD) result in changes in BAF, and copy-number changes also affect LRR (**Figs. 1** and **2**). Our detection method identifies both non-mosaic (constitutional) and clonal mosaic anomalies, which were distinguished using standards based on parent-offspring transmission in family studies and polymorphic CNVs in non-family studies. We detected three types of clonal mosaic anomalies: mixtures of disomic and monosomic cells (deletions), mixtures of disomic and trisomic cells (duplications) and copy-neutral mixtures of biparental and acquired uniparental disomy (aUPD) (examples are shown in **Fig. 3** and **Supplementary Fig. 2**). The aUPD primarily occurred at terminal segments, as expected for an origin in mitotic crossing over (**Supplementary Fig. 3**), whereas some cases of whole-chromosome aUPD might result from aneuploidy rescue (**Supplementary Fig. 4**).

Using a method optimized to detect large anomalies (50 kb to whole chromosome), we identified at least one non-mosaic anomaly (large CNV) in 75% of all subjects, at least one clonal mosaic anomaly in 0.80% of subjects and both types in 0.69% of subjects. The median size of all anomalies detected was 150 kb (**Supplementary Fig. 5**), and the mean number per subject was 1.5 (range 0 to 13). There were 514 mosaic anomalies in 404 of 50,222 subjects analyzed.

The reproducibility (in 568 duplicate sample pairs) of all anomalies analyzed for mosaic status was 82% (with >80% overlap; Online Methods and **Supplementary Table 2**). For clonal mosaic anomalies in duplicate samples, the reproducibility was 68% (15/22), and all discordant calls seemed to be false negatives, based on examination of BAF and LRR plots. We also assessed the reproducibility of clonal mosaic anomaly calls in comparison to those reported by Jacobs *et al.* in an accompanying paper[18], where the same raw data were analyzed for 5,510 subjects from the GENEVA Lung Cancer study. Whereas the methods in both studies detected 83 mosaic anomalies, the GENEVA method described here detected an additional 28 mosaic anomalies (8 of >2 Mb in size), and the Jacobs method detected an additional 20 mosaic anomalies (all of >2 Mb). The overall reproducibility was 63% and was 75% when considering only anomalies greater than 2 Mb in size (the size detection limit of the Jacobs method). Both estimates are considerably greater than the 25–50% reproducibility across methods estimated for several common CNV-calling algorithms[19]. All findings of mosaic events that were discordant seemed to be due to false negatives. The Jacobs method is more conservative with respect to size threshold (2 Mb), whereas our method is more conservative with respect to sample quality (but calls mosaic anomalies involving segments less than 2 Mb in size when sample quality is sufficient). Therefore, the false negative rates of both methods seem high, and the prevalence of clonal mosaic anomalies detected here is likely to be underestimated. Mosaic detection is difficult when the fraction of abnormal cells is extreme, when the anomaly length is small or when sample quality is low (with high BAF and/or LRR variability).

The clonal mosaic anomalies detected in GENEVA subjects were classified as 15.6% duplications, 50.4% deletions and 34.0% aUPD. All three classes of mosaic anomalies were large (**Fig. 4** and **Supplementary Fig. 6**). Median lengths were 34.1 Mb for duplications, 3.8 Mb for deletions and 39.8 Mb for aUPD. Mosaic aneuploidies included +8, +9, +12, +14, +15, +18, +19, −21 and +22, and whole-chromosome mosaic UPD events were found on chromosomes 2, 3, 13, 14 and 15. Plots of the breakpoints of all mosaic anomalies are provided (**Supplementary Fig. 7**), and genomic coordinates (along with other information) are given (**Supplementary Table 3**).

There was a highly significant excess of subjects with multiple clonal mosaic events compared to the Poisson distribution expected

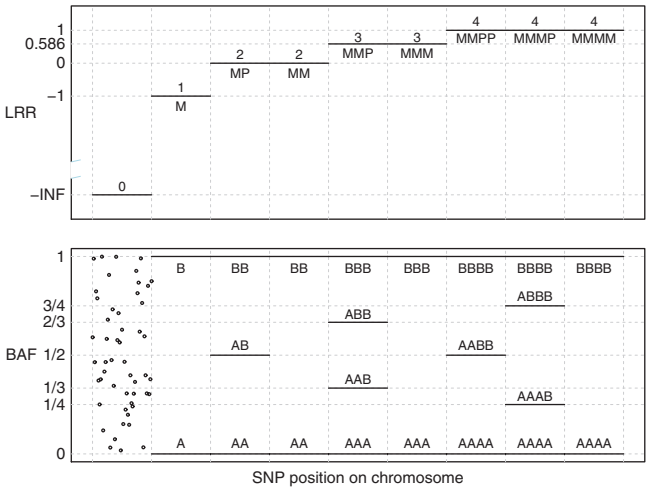


Figure 1 Expected values of BAF and LRR for discrete copy-number states. Mosaic anomalies have intermediate positions between these discrete states. Copy number is given above the horizontal lines in the LRR plot, and SNP genotypes are given in the BAF plot. M, maternal chromosome; P, paternal chromosome. States with M and P reversed are also possible. The scatter of points for copy number = 0 (homozygous deletion) represents background signal noise. INF, infinity.

if the anomalies occurred independently ($P < 2 \times 10^{-16}$). Multiple clonal anomalies in individual subjects occurred in two ways: (i) as compound sets of anomalies adjacent to one another on a single chromosome, suggesting a single event or related mechanism of origin (**Supplementary Fig. 2g**), and (ii) as non-adjacent sets. Among the 404 mosaic subjects, 64 had multiple mosaic anomalies of one or both types (with 2.6 expected), and 55 had only non-adjacent sets (with 2.4 expected). The excess of multiple mosaic anomalies was observed for both CNVs and aUPD. The age of the subjects with multiple anomalies was not significantly different than that of subjects with a single anomaly ($P = 0.99$).

The frequency of detectable clonal mosaicism increases with age

The observed frequency of subjects with one or more clonal mosaic anomalies detected (the mosaic status) is shown (**Fig. 5** and **Supplementary Table 4**). It was low (<0.5%) in subjects less than 50 years of age but increased with age to 2.7% in subjects over 80. The mosaic frequency was 0.2% in the groups including subjects from 0–14 years old (15/8,535) and from 15–29 years old (16/6,739), despite the fact that approximately half of the subjects that were 0–14 years old had a phenotypic abnormality (non-syndromic cleft lip/palate or prematurity or low birth weight). Excluding subjects less than 15 years of age in multiple logistic regression of mosaic status on age at DNA sampling and adjusting for several covariates (study, sex, DNA source and ancestry), we found that age was a highly significant predictor of mosaic status ($P = 2 \times 10^{-16}$; odds ratio (OR) = 1.05, 95% confidence interval (CI) = 1.04–1.07). Among the covariates, only 'study' (defined in **Table 1**) was significant ($P = 0.01$), and a subsequent test of age-by-study interaction was not significant. It is notable that the DNA source (92% from blood and 8% from saliva/buccal swabs) was not a significant predictor ($P = 0.45$). When only blood samples were analyzed, the age-effect estimate was the same (to three decimal places), and the P value was only slightly higher (4×10^{-15}) than in the combined analysis of both DNA sources. Copy-number mosaic and aUPD anomalies, when tested separately, each had a significant age effect and similar odds ratios (P for gain = 0.01; P for loss = 5×10^{-11}; P for aUPD = 6×10^{-8}; OR (95% CI) for gain = 1.032 (1.005–1.061); OR (95% CI) for loss = 1.057 (1.039–1.075); OR (95% CI) for aUPD = 1.056 (1.035–1.077)).

This age effect was specific for mosaic anomalies. The same logistic regression performed with non-mosaic anomalies did not show a significant age effect ($P = 0.11$), and the regression coefficient estimate was negative (−0.002), whereas that for mosaic anomalies was positive (0.050) (**Supplementary Fig. 8**). This result indicates that our classification method distinguishes effectively between acquired and constitutional anomalies.

To further explore the robustness of the age effect on clonal mosaicism, we performed additional analyses with each of the seven studies having more than 1,000 subjects over 50 years of age (using both blood and saliva/buccal samples). Only the age effect was significant ($P = 8 \times 10^{-16}$) in a combined logistic regression of mosaic status on study, sex, DNA source, ancestry and smoking status (separately testing either ever or never smokers). When only controls from these studies were analyzed together, the age effect remained highly significant ($P = 7 \times 10^{-11}$). We also analyzed each study separately with age and the case status specific to that study. A meta-analysis showed a highly significant effect of age (**Fig. 6**), which was very robust to differences in both study and subject characteristics.

These cross-sectional analyses strongly suggest that most of the mosaic anomalies detectable by SNP microarrays appear late in life, because they arise more frequently and/or because they are more readily detected due to clonal expansion. This suggestion is supported by longitudinal observation for one GENEVA subject (the only subject sampled twice who had mosaicism in at least one sample). This subject was sampled at age 66 and again at age 72 (both times with DNA from saliva). No mosaic anomalies were detected in the earlier sample, but the later sample contained five mosaic deletions, each on a different chromosome. Additional studies with subjects sampled at multiple ages are needed to evaluate the temporal origin and stability of mosaic anomalies.

In some GENEVA subjects, anomalies seemed to have occurred early enough in development to be mosaic in both the soma and germline. In 35 parent-offspring pairs in which a mosaic anomaly was detected in the parent, there were three cases in which the offspring was non-mosaic for the same anomaly (one deletion and two duplications), and there was no corresponding anomaly (mosaic or otherwise) in the remaining 32 offspring. Although this result suggests that a fairly large proportion of individuals have mosaicism shared by the germline and soma, it may not be representative of the more frequent mosaic anomalies

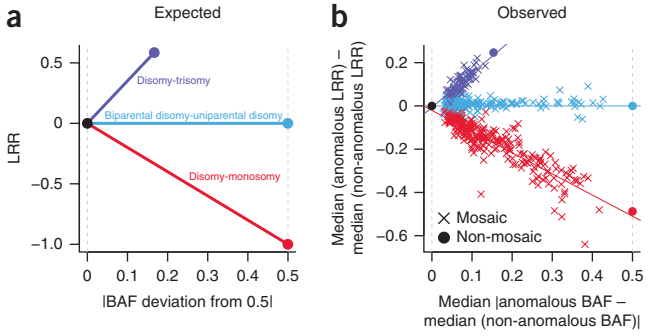


Figure 2 BAF and LRR metrics for clonal mosaic anomalies detected in GENEVA subjects. (**a**,**b**) Values are shown for expected (**a**) and observed (**b**) ($n = 514$). Anomalous measures are for regions within the anomaly; non-anomalous measures are for autosomal regions from the same sample. The purple and red circles represent the mean values of non-mosaic anomalies, and the black and cyan circles represent theoretical values.

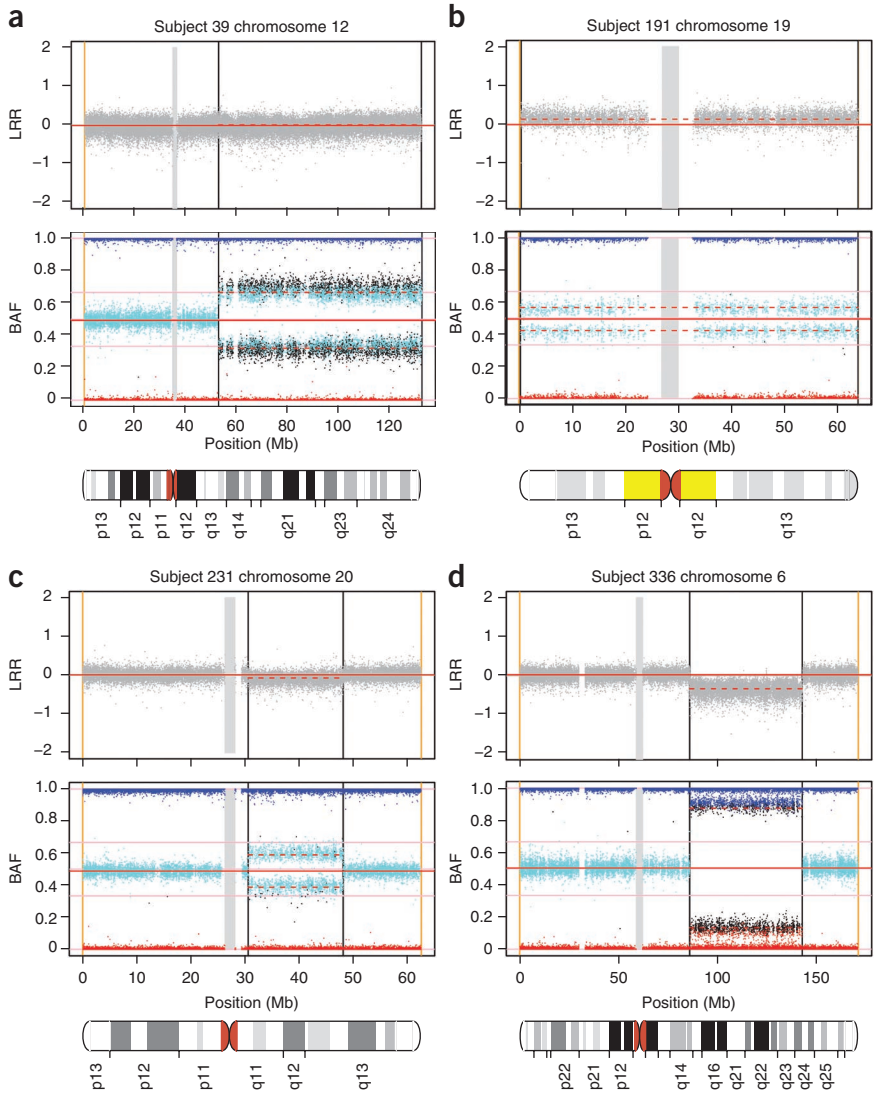


Figure 3 BAF and LRR plots of four representative mosaic anomalies. Each pair of plots is for a different sample-chromosome combination, and each data point in the plots is a single SNP. Data points in BAF plots are colored by genotype (red, AA; cyan, AB; purple, BB; black, missing call). The vertical black lines indicate the breakpoint(s) of the anomaly. The vertical gray rectangle is the centromeric gap. The solid horizontal red line in each plot is the median value for non-anomalous regions of the autosomes. The horizontal dashed red line is the median value within the anomaly. In the BAF plot, the horizontal pink lines at 0, 1/2 and 1 represent the expected positions of disomic AA, AB and BB genotypes, respectively, and those at 1/3 and 2/3 represent the expected positions of AAB and ABB trisomic genotypes, respectively. (**a**) Mosaic aUPD for distal 12q is indicated by the split in the intermediate BAF band along with the lack of change in LRR. A non-mosaic uniparental isodisomy would have only two BAF bands (at 0 and 1). (**b**) Mosaic trisomy for chromosome 19 is indicated by a narrow split in the intermediate BAF band along with a small elevation of LRR. A non-mosaic trisomy would have a wider BAF split (at 1/3 and 2/3) and a larger increase in LRR. (**c**) A mosaic deletion at 20q is indicated by a narrow split in the intermediate BAF band along with a small decrease in LRR. A non-mosaic heterozygous deletion would have no intermediate BAF bands and a larger decrease in LRR. (**d**) A mosaic deletion at 6q is indicated by a wide split in the intermediate BAF band along with a large decrease in LRR. The mosaic in **d** has a greater proportion of cells containing the deletion than the one in **c**. For additional examples, see **Supplementary Figure 2**.

that occur in older subjects, as parents in the family studies were sampled in their 20s and 30s (**Table 1**). The mosaic events that occur in subjects less than 50 years of age may have different origins than those that occur later, when the frequency of events increases rapidly.

Mosaic anomalies characteristic of hematological cancers

The clonal mosaic anomalies detected in this study tended to cluster in location within and among chromosome arms (**Fig. 4** and **Supplementary Figs. 7** and **9**). Regions with multiple overlapping anomalies frequently coincided with regions of copy-number change or aUPD that are characteristic of hematological cancers. Using the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, we found that 222 of 669 recurrent duplications and deletions found in hematological cancers have >80% overlap with at least one mosaic CNV in GENEVA subjects. Also, 77% of GENEVA mosaic CNVs have >80% overlap with the aberrations in the Mitelman database, and 48% overlap both cytological bands defining the limits of the corresponding aberration. The most common overlaps were of 20q–, 13q–, 11q–, 17p–, 12+ and 8+.

Common deleted regions (CDRs) of mosaic anomalies in different GENEVA subjects often mapped to genes previously associated with hematological cancers (**Supplementary Fig. 7**, chromosomes 4, 13, 20 and 22). First, at 13q, 31 deletions have a CDR of 299 kb containing only 1 gene, *DLEU7*, which is thought to be a tumor suppressor[20]. In addition, 18 deletions at 13q include *RB1*, and 24 include *MIR15A* and *MIR16-1*.

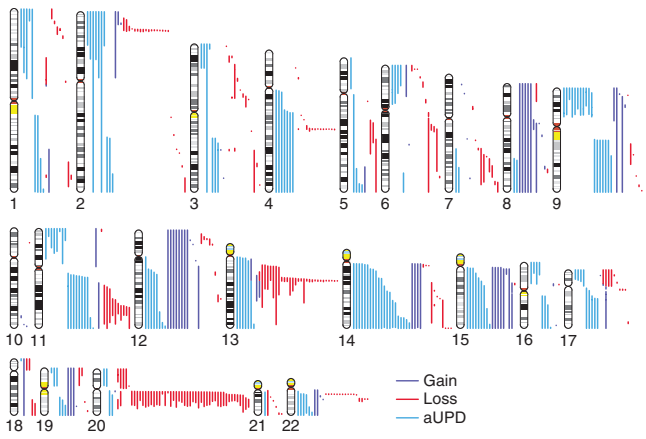


Figure 4 The lengths and chromosomal positions of the 514 clonal mosaic anomalies detected in GENEVA subjects. An ideogram of each autosome is shown, with scaled and color-coded representations of each mosaic anomaly to the right.

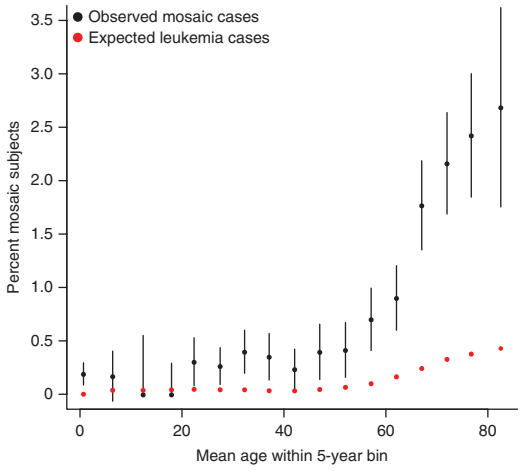
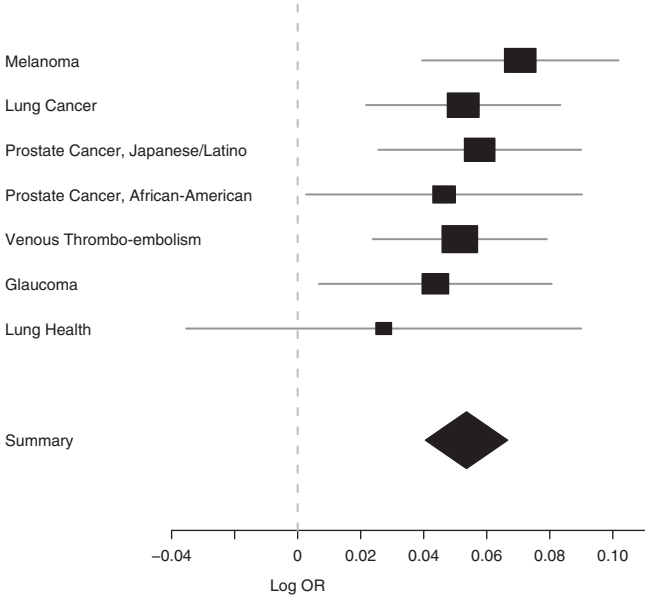


Figure 5 The percentage of subjects having one or more mosaic anomaly within 5-year age bins. Vertical bars are 95% confidence intervals. For two cells with zero counts, the upper bar connects zero to the frequency with a lower 95% confidence interval of zero, given the sample size. Expected leukemia values are given for reference and were calculated using age- and sex-specific prevalence estimates from the Surveillance Epidemiology and End Results (SEER) database.

Deletions in this region (13q14) represent the most common cytogenetic abnormality in chronic lymphocytic leukemia (CLL)[21], which is the most common leukemia in older adults. Second, at 4q, 14 deletions have a CDR of 214 kb containing only 1 gene, *TET2*, which is commonly deleted in myelodysplastic syndrome (MDS), myeloproliferative disorder (MPD) and acute myeloid leukemia (AML)[22]. Third, at 2p, 17 deletions have a CDR of 194 kb, which contains 2 genes, one of which is *DNMT3A*, recently found to be commonly mutated in AML-M5 (ref. 23). Fourth, at 22q, 11 deletions have a CDR of 153 kb, which includes 3 genes, 1 of which is *PRAME*, which is frequently deleted in CLL[24]. Fifth, at 20q, 46 deletions have a CDR of 965 kb containing 7 genes, including *L3MBTL1*, which is a candidate tumor suppressor in del(20q12) myeloid disorders[25].

Long (multi-megabase) segments of aUPD are frequently observed in cancers of many types[26]. In most cases, the anomaly causing UPD occurs on a terminal segment of one arm, consistent with this alteration originating from a single mitotic crossover that is followed by outgrowth of one of the daughter cells. aUPD is frequently observed in hematological cancers, such as MDS, MPD and AML, and is associated with homozygosity of mutations in several tumor suppressors and oncogenes[27,28]. All autosomes (except chromosome 10) had at least one clonal mosaic aUPD anomaly in GENEVA subjects. Chromosomes 9 (with 24 anomalies), 14 (with 21) and 11 (with 19) have the most aUPD anomalies, greatly exceeding the expected number based on arm length (**Supplementary Fig. 9**).

Despite the observation that many of the clonal mosaic anomalies identified here are characteristic of hematological cancer, the proportion of subjects with one or more mosaic anomalies who had a record of hematological cancer before DNA sampling was low. This proportion was estimated to be 2.8% (95% CI = 1.0–4.7%) in 291 mosaic subjects (with DNA from blood from 13 GENEVA studies, using medical records, self-reported conditions and study exclusion criteria as described in the **Supplementary Note**).



Hematological cancer incidence

We investigated whether detectable clonal mosaicism predisposes to incident hematological cancer after DNA sampling by using three GENEVA studies that included cohorts with cancer diagnosis records both before and after DNA sampling. We analyzed 8,562 subjects who had DNA derived from blood and no record of hematological cancer before DNA sampling from (i) the Glaucoma study, with subjects from the Nurses Health Study (NHS, $n = 363$) and the Health Professionals Follow-up Study (HPFS, $n = 285$); (ii) the Lung Cancer study, with subjects from the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO, $n = 1,600$); and (iii) the Prostate Cancer study, with subjects from the Multiethnic Cohort (MEC, $n = 6,314$). Among the 8,562 subjects analyzed for incident hematological cancer, 8,323 were non-mosaic with no events, 90 were non-mosaic with events, 134 were mosaic with no events, and 15 were mosaic with events (where an event was a hematological cancer diagnosis).

To test for an association between mosaic status and incident hematological cancer, we used a cause-specific Cox proportional hazards model to analyze time to a hematological cancer diagnosis from the date of DNA sampling, with right censoring at death or the endpoint of follow-up data. We performed a stratified analysis of the four cohorts, which included mosaic status and adjusted for age at DNA sampling, non-hematological cancer status (as a time-dependent covariate), ancestry (two principal components) and sex (within the PLCO stratum). The hazard ratio estimate for mosaic status was 10.1 (95% CI = 5.8–17.7) with a P value of 3×10^{-10}. A meta-analysis showed consistent results among cohorts and gave a very similar effect estimate (**Supplementary Fig. 10**). These results estimate that the risk of hematological cancer is tenfold higher for mosaic than for non-mosaic individuals.

Because the incidences of cancer and the clonal mosaic anomalies detected in this study both increase with age, adjustment for age at time of DNA sampling in the Cox regression model is critical. We modeled the age covariate as either a linear or nonlinear effect (spline

Figure 6 Fixed-effects meta-analysis for effect of age at DNA sampling on mosaic status. Effect estimates are from logistic regression of mosaic status on age at DNA sampling, with adjustment for case status specific to each study. The summary estimate of the log OR is 0.05 (95% CI = 0.04–0.07), and the corresponding OR is 1.06 (95% CI = 1.04–1.07). Cochran's Q test of heterogeneity gave $P = 0.89$. The sizes of the black boxes are proportional to the inverse of the squared standard error, and the gray lines are 95% confidence intervals. The horizontal corners of the diamond span the 95% confidence interval of the summary estimate. For study descriptions, see **Table 1**.

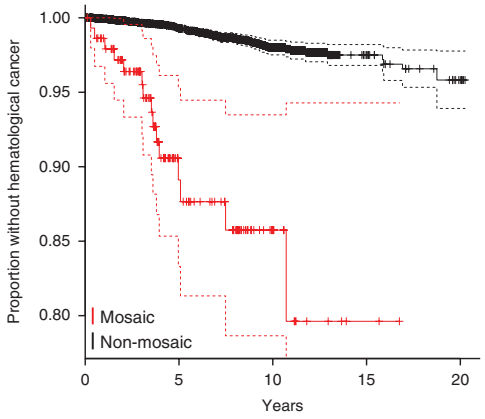


Figure 7 A Kaplan-Meier plot of the proportion of living subjects who remain free of hematological cancer as a function of time since the time of DNA sampling and determination of mosaic status. Estimates for mosaic and non-mosaic subjects are given separately (solid lines), with each shown with 95% confidence intervals (dashed lines). The vertical ticks represent censoring times. For the 15 mosaic subjects with incident cancer, the times between DNA sampling and diagnosis were 3.5, 6.1, 12.7, 18.8, 25.0, 36.9, 37.5, 42.9, 44.0, 46.2, 48.0, 60.4, 61.8, 91.1 and 130.5 months.

smoothing with 5 degrees of freedom) and found that the mosaic effect estimates and *P* values were essentially identical.

Among the 15 mosaic subjects who had a hematological diagnosis after DNA sampling, 4 had myeloid leukemia, 6 had chronic lymphocytic leukemia, 1 had multiple myeloma, 1 had MDS, 1 had MPD, and 2 had non-Hodgkin lymphoma. Thus, the 15 cases are approximately evenly divided between mature B-cell neoplasms and myeloid malignancies. Not unexpectedly, the leukemias were over-represented in mosaic relative to non-mosaic subjects ($P = 0.005$; **Supplementary Tables 5** and **6**). A variety of chromosomal anomalies were found in the mosaic subjects (**Supplementary Table 7**). Deletions covering the described CDRs were found in several of these subjects: at 13q– in five CLL cases, at 4q– in one chronic myelogenous leukemia (CML) case, at 20q– in one multiple myeloma case and one AML case and at 22q– in one CLL case. Five of the 15 mosaic subjects with incident hematological cancer had more than one mosaic anomaly, which is a higher proportion than that found in the remaining subjects within these cohorts (25/134), although this increase is not significant ($P = 0.18$).

Although the risk of incident hematological cancer was estimated as tenfold higher for mosaic than for non-mosaic subjects (95% CI = 5.8–17.7), it is important to note that the incidence rate in mosaic individuals is low (10-year event rate of 0.143, 95% CI = 0.065–0.214; **Fig. 7**) and that only a small fraction of mosaic GENEVA subjects have a record of hematological cancer before DNA sampling (2.8%, 95% CI = 1.0–4.7%). The period between the first appearance of detectable clonal mosaicism and the incidence of hematological cancer is of interest but cannot be estimated from our data, as mosaicism was present for an unknown period of time before DNA sampling. However, the median time of 3.5 years between DNA sampling and hematological cancer diagnosis provides a very rough minimum estimate (range 3.5 months to 10.7 years with $n = 15$; **Fig. 7**).

Non-hematological cancer

To investigate the relationship between mosaic status and non-hematological cancer, we performed two types of analyses. First, in each of the three GENEVA case-control cancer studies (Lung Cancer, Prostate Cancer and Melanoma), we performed logistic regression of mosaic status on case status and age at DNA sampling. Case status was not a significant factor in any of the three studies or in a meta-analysis (one-tailed $P = 0.06$). The estimated odds of having a clonal mosaic anomaly was higher among cancer cases than controls in the lung and prostate cancer studies, but lower in the melanoma study (**Supplementary Fig. 11**). Second, in the cohort studies (PLCO, HPFS, NHS and MEC), we performed logistic regression of mosaic status on whether or not the subject had a non-hematological cancer before DNA sampling (excluding any hematological cancer cases). In these analyses, the relationship was consistently positive but small and not significant overall (one-tailed $P = 0.11$; **Supplementary Fig. 12**). In summary, the evidence hints at a positive relationship between mosaic status and non-hematological cancer but lacks statistical significance. Therefore, further work is needed in larger sets of non-hematological cancer studies, including data on potential exposure, disease and treatment effects.

DISCUSSION

Here, we have shown that the frequency of subjects with detectable clonal mosaicism for large chromosomal anomalies in peripheral blood is low (<0.5%) from birth until 50 years of age, after which it rises rapidly. This relationship between mosaicism and age is very robust to both study and subject characteristics. Among the covariates of sex, ancestry, smoking status and disease status (exclusive of hematological cancer), none had a significant effect on mosaic status. The age effect in GENEVA subjects is consistent with a recent study showing that acquired differences in structural chromosome variants between members of monozygotic twin pairs (including clonal mosaic anomalies) are observed in pairs of >55 years of age but not in younger pairs[29]. Nevertheless, longitudinal studies are required to rule out the possibility that a trend in environmental exposures across birth cohorts might contribute to the increase in mosaicism with age.

The observed increase in detectable clonal mosaicism late in life may be due to a change in the frequency with which chromosomal anomalies occur (increased somatic mutation rate) and/or their ability to form large clones (clonal expansion). Previous work has shown that the occurrence of chromosomal anomalies (rearrangements and aneuploidies) during cell division increases with age in cultured lymphocytes and fibroblasts[30,31], that DNA damage accumulates with age in mouse hematopoietic stem cells[32] and that mitotic recombination (leading to UPD) increases with replicative age in yeast[33]. This apparent increase in somatic mutation may result from age-related decline in genomic maintenance mechanisms (such as telomere attrition[34]). Clonal expansion of cells containing chromosomal anomalies could be due to either positive selection or to random changes in the frequencies of hematopoietic stem cell descendants. In principle, stem cell senescence and age-related decline in replicative function[35] could result in a decrease in the effective population size of stem cells, leading to shifts in clonal composition analogous to random drift in small populations of individuals[36]. However, analyses of the clonal composition of blood cells in healthy women using X-inactivation markers suggest stability over time and between lymphoid and myeloid lineages, even in the elderly[37,38]. Therefore, in most cases, positive selection may be required to establish clones of cells with chromosomal anomalies that are sufficiently large for detection with SNP microarrays. The potential for positive selection may increase with age as somatic mutations accumulate in genes that regulate cellular proliferation. For example, a highly proliferative clone may arise when a recessive tumor suppressor mutation becomes hemizygous in

combination with a deletion or homozygous due to aUPD. This suggestion is supported by the observation that acquired anomalies tend to cluster in certain genomic regions and that common deleted regions map to genes previously associated with hematological cancer.

In the mosaic individuals described in this study, the chromosomally abnormal cells constitute a substantial fraction of white blood cells, as a minimum of 5–10% is required for detection by our method, and many abnormal clones are substantially larger (**Fig. 2**). The blood samples used for DNA extraction were not fractionated by white blood cell type. The abnormal blood cells within an individual might include multiple cell types if the anomaly arose in a multipotent hematopoietic stem cell that became predominant due to senescence or positive selection within the stem cell population. Alternatively, the abnormal cells might include a restricted set of cell types, particularly when the normal composition of blood (which is 60–70% neutrophils and 20–40% lymphocytes[39]) is altered by unregulated proliferation[40].

There is a strong association between the clonal mosaic anomalies detected in our study and hematological cancer. We estimate the risk of acquiring a hematological cancer diagnosis as tenfold higher for subjects with mosaic anomalies. This association is strongly supported by the finding that many of the mosaic anomalies are characteristic of those found in hematological cancers. Nevertheless, the event numbers analyzed here are small, and additional studies are needed across a broader diversity of cohorts to establish the clinical importance of these findings.

Notwithstanding the strong association with hematological cancer, we estimated that ~97% of subjects with clonal mosaic anomalies did not have a record of a hematological cancer before DNA sampling, and the incidence rate of such cancer was low (~14% over 10 years in subjects who survived and were not lost to follow-up analysis during this period). These results suggest that the clonal mosaicism observed in elderly subjects may be an asymptomatic condition with a predisposition to hematological cancer that is often not realized.

It is possible that many of the subjects with detectable clonal mosaicism in our study have monoclonal B-cell lymphocytosis (MBL), an asymptomatic condition with an estimated prevalence of 3–5% in the elderly. MBL is characterized by a clonal population of B lymphocytes with an immunophenotype similar to CLL or other B-cell malignancy[41]. Most, if not all, cases of CLL are preceded by MBL, but most cases of MBL do not progress to malignancy[42,43]. However, 85% of MBL cases detected in population screening studies have a B-cell count below 500 cells/μl (ref. 43), which is less than 10% of the normal white blood cell count. Because 10% is near the lower limit of detection for chromosomal mosaicism using our methods, the two types of clones may not be closely related. Nevertheless, further work on the relationship between B-cell immunophenotypes and mosaic anomalies is warranted.

Although it seems that most of the clonal mosaicism observed in GENEVA subjects represents a non-malignant condition, further work is needed to evaluate the fraction of subjects who might have unrecorded malignant conditions, such as MDS and MPD, or undiagnosed CLL. MDS and MPD were added to the Surveillance, Epidemiology, and End Results (SEER) cancer registries in 2001 and may still be under-recorded, because they are often managed outside of the hospital setting[44]. Therefore, accurate prevalence data from widespread populations are not available, but local population estimates (0.1% MDS[45] and 0.5% MPD[46] in the elderly) are substantially less than the ~2.5% of GENEVA subjects with mosaic anomalies who were over 75 years old.

This survey is the first large-scale study of acquired chromosomal anomalies in people of all ages and various states of health. Previously, the extent of chromosomal variation within developmentally normal individuals in the absence of overt cancer was largely unknown. The results presented here indicate that a substantial fraction of blood cells in people without a history of hematological cancer may contain large chromosomal anomalies, including multi-megabase deletions, duplications and aUPD anomalies. The frequency of people with such clonal anomalies in a mosaic state is low up to approximately 50 years of age and then increases rapidly up to 2–3% with advanced age. We find that these anomalies are associated with an approximately tenfold higher risk of hematological cancer, but subjects with detectable clonal mosaicism may survive for years without having a hematological cancer diagnosis. Further work is needed to determine the stability of the mosaic state over time, to replicate and improve estimates of predisposition to hematological cancer and to identify anomalies associated with asymptomatic cancer precursor conditions. It also will be important to explore the health consequences of these anomalies for conditions other than cancer, such as altered immune system function.

URLs. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, http://cgap.nci.nih.gov/Chromosomes/Mitelman; Surveillance Epidemiology and End Results (SEER), http://seer.cancer.gov/; R, http://www.R-project.org/.

METHODS

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

Accession numbers. Data from these studies are available at dbGaP (phs000187.v1.p1, phs000335.v2.p2, phs000094.v1.p1, phs000092.v1.p1, phs000093.v2.p2, phs000304.v1.p1, phs000306.v3.p1, phs000289.v2.p1, phs000096.v4.p1, phs000095.v1.p1, phs000103.v1.p1 and phs000308.v1.p1) (see also **Supplementary Table 1**).

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

ACKNOWLEDGMENTS

The GENEVA Consortium thanks the subjects and the staff of all GENEVA studies for their important contributions. We thank the following state cancer registries for their help: Alaska, Arizona, Arkansas, California, Colorado, Connecticut, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kentucky, Louisiana, Maine, Maryland, Massachusetts, Michigan, Nebraska, New Hampshire, New Jersey, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, Rhode Island, South Carolina, Tennessee, Texas, Virginia, Washington and Wyoming. We thank C. Laird and G. Marti for helpful comments on the manuscript and B. Wakimoto and D. Gottschling for enlightening discussions. We also thank K. Jacobs for exchanging ideas and for working with us to estimate cross-method concordance of mosaic detection using the PLCO/GENEVA Lung Cancer study. Support for the GENEVA genome-wide association studies was provided through the US National Institutes of Health (NIH) Genes, Environment and Health Initiative (GEI). Some studies also received support from individual NIH Institutes. The grant numbers are: Melanoma (NCI R29CA70334, R01CA100264 and P50CA093459); Lung Health (U01HG004738); Cleft Lip/Palate (National Institute Dental and Craniofacial Research (NIDCR): U01DE018993 and NIH contract: HHSN268200782096C); Addiction (U01HG004422, National Institute on Alcohol Abuse and Alcoholism (NIAAA): U10AA008401, National Cancer Institute (NCI): P01CA089392, National Institute on Drug Abuse (NIDA): R01DA013423 and R01DA019963); Lung Cancer (Z01CP010200); Blood Clotting (R37 HL 039693); Prostate Cancer (U01HG004726, NCI: CA63464, CA54281, CA1326792 and RC2 CA148085); Venous Thromboembolism (U01HG004735); Birth Weight (U01HG004415); Dental Caries (NIDCR: U01DE018903 and R01DE014899, NIH Center for Inherited Disease Research (CIDR) contract: HHSN268200-782096C); Prematurity (U01HG004423); Glaucoma (U01HG004728, National Eye Institute (NEI): R01EY015473 and R01EY015872); GENEVA Coordinating Center

(U01 HG004446); CIDR (U01HG004438 and HHSN268200782096C); Broad Center for Genotyping and Analysis (U01HG04424); the Intramural Research Program of the NIH, the National Library of Medicine; and the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, NCI, NIH. L.R.P. was also supported by a Physician Scientist award from Research to Prevent Blindness in NYC and an Ophthalmology Scholar Award from Harvard Medical School and from the Harvard Glaucoma Center of Excellence. L.R.Z. was supported by the NCI (T32 CA09168).

AUTHOR CONTRIBUTIONS

K.F.D., H.L., K.N.H. and E.W.P. initiated the detection of chromosomal anomalies in GENEVA GWAS data. C.A.L. developed the automated methods of anomaly detection, with assistance from C.C.L., L.R.Z., C.P.M., V.E.S. and A.N.M. C.C.L., C.A.L., K.R., L.R.Z., C.P.M., J.S., D.R.C., D.M.L., X.Z., S.C.N., S.M.G., M.P.C., J.I.U. and S.B. performed data analyses. C.A., Q.W., L.W., J.E.L., K.C.B., N.N.H., R.M., T.H.B., A.F.S., L.J.B., M.T.L., L.R.G., D.G., K.C.D., S.S.S., W.J.B., L.B.S., S.A.I., S.J.C., S.I.B., L.L.M., B.E.H., J.A.H., S.M.A., C.R., W.L.L., M.L.M., J.C.M., M.M., B.F., J.H.K., J.L.W., L.R.P., C.A.H. and N.C. contributed sample collections and phenotypic data. K.F.D., H.L., K.N.H., E.W.P., D.B.M. and A.C. performed genotyping. L.R.P., J.H.K., N.C., C.A.H., B.E.H. and K.R.M. provided data and interpretation for analysis of incident hematological cancer. C.C.L., C.A.L., L.R.Z., K.F.D., K.R., C.A., D.D., T.H.B., A.F.S., I.R., R.B.S., L.J.B., S.M.H., N.D.F., J.L., B.E.H., K.R.M., M.d.A., W.L.L., M.G.H., M.L.M., E.F., J.C.M., M.M., B.F., J.L.W., A.W., C.P.M., J.S., D.R.C., D.M.L., X.Z., J.I.U., S.B., S.C.N., S.M.G., P.H., G.P.J., A.N.M., V.E.S., H.L., K.N.H., E.W.P., D.B.M., A.C., N.S., T.M., L.R.P., C.A.H., N.C. and B.S.W. contributed ideas and advice during regular discussions of the project. C.C.L. coordinated the study and wrote the first draft of the manuscript, with guidance from a writing committee consisting of C.A.L., K.R., K.F.D., T.M., L.R.P., N.C. and B.S.W. All authors contributed to review and revision 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.2271.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Miller, O.J. & Therman, E. *Human Chromosomes* (Springer-Verlag, New York, 2001).
2. Strachan, T. & Read, A.P. *Human Molecular Genetics* (Wiley-Liss, New York, 1996).
3. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23–28 (1976).
4. Delhanty, J.D. Mechanisms of aneuploidy induction in human oogenesis and early embryogenesis. *Cytogenet. Genome Res.* **111**, 237–244 (2005).
5. Vanneste, E. *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nat. Med.* **15**, 577–583 (2009).
6. Hassold, T. Mosaic trisomies in human spontaneous abortions. *Hum. Genet.* **61**, 31–35 (1982).
7. 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).
8. Ballif, B.C. *et al.* Detection of low-level mosaicism by array CGH in routine diagnostic specimens. *Am. J. Med. Genet. A.* **140**, 2757–2767 (2006).
9. Heim, S. & Mitelman, F. Nonrandom chromosome abnormalities in cancer—an overview. in *Cancer Cytogenetics* (eds. Mitelman, F. & Heim, S.) 25–44 (John Wiley & Sons, Hoboken, New Jersey, 2009).
10. Gardner, R.J.M. & Sutherland, G.R. *Chromosome Abnormalities and Genetic Counseling* (Oxford University Press, Oxford, 2004).
11. Maciejewski, J.P., Tiu, R.V. & O'Keefe, C. Application of array-based whole genome scanning technologies as a cytogenetic tool in haematological malignancies. *Br. J. Haematol.* **146**, 479–488 (2009).
12. Dougherty, M.J. *et al.* Implementation of high resolution single nucleotide polymorphism array analysis as a clinical test for patients with hematologic malignancies. *Cancer Genet.* **204**, 26–38 (2011).
13. McCarroll, S.A. & Altshuler, D.M. Copy-number variation and association studies of human disease. *Nat. Genet.* **39**, S37–S42 (2007).
14. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
15. 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).
16. Cornelis, M.C. *et al.* The Gene, Environment Association Studies consortium (GENEVA): maximizing the knowledge obtained from GWAS by collaboration across studies of multiple conditions. *Genet. Epidemiol.* **34**, 364–372 (2010).
17. Peiffer, D.A. *et al.* High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Res.* **16**, 1136–1148 (2006).
18. Jacobs, K.B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* published online, doi:10.1038/ng.2270 (6 May 2012).
19. 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).
20. Pekarsky, Y., Zanesi, N. & Croce, C.M. Molecular basis of CLL. *Semin. Cancer Biol.* **20**, 370–376 (2010).
21. Döhner, H. *et al.* Genomic aberrations and survival in chronic lymphocytic leukemia. *N. Engl. J. Med.* **343**, 1910–1916 (2000).
22. Bejar, R., Levine, R. & Ebert, B.L. Unraveling the molecular pathophysiology of myelodysplastic syndromes. *J. Clin. Oncol.* **29**, 504–515 (2011).
23. Yan, X.J. *et al.* Exome sequencing identifies somatic mutations of DNA methyltransferase gene *DNMT3A* in acute monocytic leukemia. *Nat. Genet.* **43**, 309–315 (2011).
24. Gunn, S.R. *et al.* Array CGH analysis of chronic lymphocytic leukemia reveals frequent cryptic monoallelic and biallelic deletions of chromosome 22q11 that include the *PRAME* gene. *Leuk. Res.* **33**, 1276–1281 (2009).
25. 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).
26. Tuna, M., Knuutila, S. & Mills, G.B. Uniparental disomy in cancer. *Trends Mol. Med.* **15**, 120–128 (2009).
27. O'Keefe, C., McDevitt, M.A. & Maciejewski, J.P. Copy neutral loss of heterozygosity: a novel chromosomal lesion in myeloid malignancies. *Blood* **115**, 2731–2739 (2010).
28. Raghavan, M., Gupta, M., Molloy, G., Chaplin, T. & Young, B.D. Mitotic recombination in haematological malignancy. *Adv. Enzyme Regul.* **50**, 96–103 (2010).
29. 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).
30. Vorobtsova, I., Semenov, A., Timofeyeva, N., Kanayeva, A. & Zvereva, I. An investigation of the age-dependency of chromosome abnormalities in human populations exposed to low-dose ionising radiation. *Mech. Ageing Dev.* **122**, 1373–1382 (2001).
31. Mukherjee, A.B. & Thomas, S. A longitudinal study of human age-related chromosomal analysis in skin fibroblasts. *Exp. Cell Res.* **235**, 161–169 (1997).
32. Rossi, D.J. *et al.* Hematopoietic stem cell quiescence attenuates DNA damage response and permits DNA damage accumulation during aging. *Cell Cycle* **6**, 2371–2376 (2007).
33. Lindstrom, D.L., Leverich, C.K., Henderson, K.A. & Gottschling, D.E. Replicative age induces mitotic recombination in the ribosomal RNA gene cluster of *Saccharomyces cerevisiae*. *PLoS Genet.* **7**, e1002015 (2011).
34. Sahin, E. & Depinho, R.A. Linking functional decline of telomeres, mitochondria and stem cells during ageing. *Nature* **464**, 520–528 (2010).
35. Sharpless, N.E. & DePinho, R.A. How stem cells age and why this makes us grow old. *Nat. Rev. Mol. Cell Biol.* **8**, 703–713 (2007).
36. Crow, J.F. & Kimura, M. *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
37. Prchal, J.T. *et al.* Clonal stability of blood cell lineages indicated by X-chromosomal transcriptional polymorphism. *J. Exp. Med.* **183**, 561–567 (1996).
38. Swierczek, S.I. *et al.* Hematopoiesis is not clonal in healthy elderly women. *Blood* **112**, 3186–3193 (2008).
39. Fischbach, F. & Dunning, M.B. *A Manual of Laboratory and Diagnostic Tests* (Lippincott, Williams and Wilkins, Philadelphia, 1992).
40. Vandewoestyne, M.L. *et al.* Laser microdissection for the assessment of the clonal relationship between chronic lymphocytic leukemia/small lymphocytic lymphoma and proliferating B cells within lymph node pseudofollicles. *Leukemia* **25**, 883–888 (2011).
41. Marti, G.E. *et al.* Diagnostic criteria for monoclonal B-cell lymphocytosis. *Br. J. Haematol.* **130**, 325–332 (2005).
42. Landgren, O. *et al.* B-cell clones as early markers for chronic lymphocytic leukemia. *N. Engl. J. Med.* **360**, 659–667 (2009).
43. Shanafelt, T.D., Ghia, P., Lanasa, M.C., Landgren, O. & Rawstron, A.C. Monoclonal B-cell lymphocytosis (MBL): biology, natural history and clinical management. *Leukemia* **24**, 512–520 (2010).
44. Cogle, C.R., Craig, B.M., Rollison, D.E. & List, A.F. Incidence of the myelodysplastic syndromes using a novel claims-based algorithm: high number of uncaptured cases by cancer registries. *Blood* **117**, 7121–7125 (2011).
45. Neukirchen, J. *et al.* Incidence and prevalence of myelodysplastic syndromes: data from the Dusseldorf MDS-registry. *Leuk. Res.* **35**, 1591–1596 (2011).
46. Ma, X., Vanasse, G., Cartmel, B., Wang, Y. & Selinger, H.A. Prevalence of polycythemia vera and essential thrombocythemia. *Am. J. Hematol.* **83**, 359–362 (2008).

[1]Department of Biostatistics, University of Washington, Seattle, Washington, USA. [2]The Center for Inherited Disease Research, Johns Hopkins University, Baltimore, Maryland, USA. [3]Department of Epidemiology, Division of Cancer Prevention and Population Sciences, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [4]Department of Surgical Oncology, Division of Surgery, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [5]Department of Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [6]Department of Medicine, University of British Columbia, Vancouver, British Columbia, Canada. [7]Department of Epidemiology, School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA. [8]Institute of Genetic Medicine, School of Medicine, Johns Hopkins University, Baltimore, Maryland, USA. [9]Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. [10]Department of Oncology, Johns Hopkins University, Baltimore, Maryland, USA. [11]Department of Psychiatry, School of Medicine, Washington University School of Medicine, Saint Louis, Missouri, USA. [12]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [13]Howard Hughes Medical Institute, University of Michigan, Ann Arbor, Michigan, USA. [14]Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA. [15]Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, USA. [16]Department of Pediatrics and Communicable Diseases, University of Michigan, Ann Arbor, Michigan, USA. [17]Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [18]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt University, Nashville, Tennessee, USA. [19]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, California, USA. [20]Epidemiology Program, Cancer Research Center, University of Hawaii, Honolulu, Hawaii, USA. [21]Department of Internal Medicine, Mayo Clinic, Rochester, Minnesota, USA. [22]Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA. [23]Division of Nephrology and Hypertension, Mayo Clinic, Rochester, Minnesota, USA. [24]Mayo Hyperoxaluria Center, Mayo Clinic, Rochester, Minnesota, USA. [25]Division of Endocrinology, Metabolism and Molecular Medicine, Northwestern University, Chicago, Illinois, USA. [26]Center for Craniofacial and Dental Genetics, Department of Oral Biology School of Dental Medicine, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [27]Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA. [28]Department of Pediatrics, University of Iowa, Iowa City, Iowa, USA. [29]Department of Epidemiology Research, Statens Serum Institut, Copenhagen, Denmark. [30]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [31]Department of Ophthalmology, Massachusetts Eye and Ear Infirmary, Harvard Medical School, Boston, Massachusetts, USA. [32]Division of Medical Genetics, University of Washington, Seattle, Washington, USA. [33]Cancer Prevention Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [34]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [35]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [36]National Center for Biotechnology Information, Bethesda, Maryland, USA. [37]Office of Population Genomics, National Human Genome Research Institute at the National Institutes of Health, Bethesda, Maryland, USA. [38]These authors contributed equally to this work. [39]These authors jointly directed this work. Correspondence should be addressed to C.C.L. (cclaurie@u.washington.edu).

ONLINE METHODS

Study subjects, phenotypic data and genotyping. Subjects were recruited for 15 different studies belonging to the GENEVA Consortium[16] (**Table 1**). Each study was approved by the institutional review board of the study investigator's institution, and all subjects provided written informed consent for participation in the study. Phenotypes are described in the **Supplementary Note**. Genotyping for each study was performed on one of five different Illumina array types at the Center for Inherited Disease Research (CIDR), the Broad Institute Center for Genotyping and Analysis or the University of Southern California (**Supplementary Table 1**). DNA samples were derived from blood (92%) or saliva/buccal swabs (8%). No lymphoblastoid cell lines or samples for which the whole genome was amplified were included in the analyses described here. Because cell lines may have artifactual mosaic anomalies[47], misidentification of DNA source is a concern. However, only the addiction study had both cell line and non-cell line samples, and the non-cell line samples analyzed here did not have an unusual frequency of mosaic anomalies. Genotype data cleaning and the calculation of BAF and LRR are described in the **Supplementary Note**. Sample sizes for analyses varied, because data on age at DNA sampling or other variables were missing for a small proportion of the subjects.

Anomaly detection and quality control. The method of anomaly detection is described in detail in the **Supplementary Note** and summarized here. Detection of anomalies (both mosaic and non-mosaic) was based on BAF and LRR metrics. The primary focus for detecting anomalies was BAF, because we wanted to identify copy-neutral events (mosaic UPD anomalies) and because BAF has a higher signal-to-noise ratio and is less prone to artifacts (such as GC waves[48]) than LRR. The main approach was to detect a split in the BAF intermediate band, which in normal (biparental disomic) samples is centered at 1/2 and corresponds to AB heterozygotes (**Fig. 1**). In trisomic samples, this band splits into two components (AAB and ABB) at BAF = 1/3 and 2/3. In disomic-trisomic mosaic anomalies, the width of the split varies from 0 to 1/3, and LRR varies from 0 to a theoretical value of $\log_2$ (3/2). In disomic-monosomic mosaic anomalies, the width of the split varies from 0 to 1, and LRR varies from 0 to a theoretical value of $\log_2$ (1/2). In biparental-uniparental disomic mosaic anomalies, the width of the split varies from 0 to 1, and LRR remains constant at 0. These transitions are shown as deviations from the expected in **Figure 2**. In chromosomal regions containing heterozygous SNPs, BAF alone can detect duplications (both mosaic and non-mosaic), mosaic deletions, mosaic uniparental disomy and homozygous deletions. LRR is required to detect monosomic regions and duplications in regions lacking heterozygosity. Therefore, we implemented two separate but complementary methods, called BAF and LOH (the latter for LRR change detection in regions lacking heterozygosity). Anomalies detected by the BAF method were classified as mosaic or non-mosaic. Anomalies detected by the LOH method were used here only to define the BAF and LRR values of heterozygous deletions and not for mosaic detection. We did not attempt to identify non-mosaic segments of uniparental isodisomy, which have no heterozygosity and normal LRR.

In the BAF method, circular binary segmentation (CBS)[49] was used to detect change points in a metric modified from one previously published[15]: sqrt (min (BAF, 1–BAF, |BAF – median (BAF)|)) for SNPs called as missing or heterozygotes (excluding homozygotes). The use of missing calls allows detection of wide splits (**Fig. 3d**). In the LOH method, CBS was applied to LRR values and combined with overlapping runs of homozygosity. By focusing on regions of homozygosity, we avoided a high false positive rate associated with a genome-wide search for changes in LRR. In both methods, the identification of anomalous segments involved establishing a non-anomalous baseline, choosing anomalous segments based on deviation from baseline and applying quality control filters. Computations were performed using the Bioconductor packages DNACopy and GWASTools. The latter was developed by our group; relevant functions are described in the **Supplementary Note**.

Quality control filtering was performed at the sample and anomaly level. Low-quality samples (with high variance of BAF and/or LRR metrics or a high level of segmentation) were filtered differently for the two methods. The percentages of samples that passed quality control for the BAF (mean = 99.1%) and LOH (86.8%) methods are shown in **Supplementary Table 1**. In some studies, a high fraction of samples failed quality control for LOH detection (maximum 47%), but the failure rates for BAF detection (from which all mosaic anomalies were identified) were all low (maximum 8%). Anomaly-level quality control involved several steps, including manual curation of all anomalies designated as mosaic and all other anomalies greater than 2 Mb in length (**Supplementary Note**). Manual curation involved evaluation of BAF and LRR plots (**Fig. 3** and **Supplementary Fig. 2**). Note that a sample of eight of the smallest mosaic deletions is shown (**Supplementary Fig. 2m–t**). Features that distinguish mosaic from non-mosaic anomalies are described in the legend to **Figure 3**.

The reproducibility of anomaly detection was assessed using samples genotyped in duplicate (n = 568 pairs). For each sample pair, we defined a unit of observation as a contiguous chromosomal region containing an anomaly in one or both samples. Each unit was given a score equal to the length of the intersection divided by the length of the union of anomalies in that unit. A reproducibility measure was defined as the fraction of units with a score greater than either 0.30 or 0.80 (chosen for comparison with published CNV studies). We also calculated the average of the scores that were greater than zero. These quantities are summarized for each study (**Supplementary Table 2**). For BAF, the mean reproducibility measure was 90% (with a 30% overlap threshold) and 82% (80% overlap). For LOH, the means were 71% (30% overlap) and 67% (80% overlap). The mean of scores greater than zero was 95% for BAF- and 96% for LOH-detected anomalies (30% threshold), indicating that when an anomaly is detected in both scans, the breakpoints are highly reproducible. These reproducibility estimates are higher than the 40–60% that is typical for detecting CNVs by hidden Markov model (HMM)[19,50], perhaps in large part because we did not attempt to detect small anomalies (the 5th percentile of anomalies we detected is 35 kb in length). In our experience, standard methods of CNV detection, such as PennCNV[51], tend to break up large anomalies into many segments and to miss large mosaic anomalies.

Identifying and classifying clonal mosaic anomalies. Clonal mosaic anomalies were identified in GENEVA family studies by using transmitted anomalies to characterize the bivariate distribution of BAF and LRR expected for non-mosaic (constitutional) anomalies. For transmitted anomalies, this distribution was approximately bivariate normal within a study, and we used this distribution to estimate a 95% prediction ellipse[52], which defines an area likely to contain most of the constitutional anomalies (**Supplementary Fig. 13**). Among the anomalies used to identify mosaic anomalies, the majority are 3N duplications. There is also a small cluster of 4N anomalies, but we did not attempt to detect mosaics consisting of 3N and 4N cells. Anomalies outside of these two clusters contain mosaics and artifacts. The latter consist of false positives and anomalies with inaccurate breakpoints (which distort the median BAF and LRR values). To distinguish between the mosaics and artifacts, we performed a manual review of BAF and LRR plots for all anomalies that fell outside of the 95% prediction ellipse and below the mean LRR for anomalies used to define the ellipse. The non-family studies were analyzed in a similar way, except that we replaced the class of transmitted anomalies with polymorphic CNVs. The latter were defined by hierarchical clustering to identify sets of anomalies with similar breakpoints. We then defined polymorphic sets as those with at least four members (but excluding sets with mean anomaly length greater than 10 Mb). We also included in the mosaic class three whole-autosome anomalies (chromosomes 12, 8 and 22) that fell within the 3N ellipse, because constitutional trisomies for these chromosomes are not compatible with normal development[1]. Although we did not have access to biospecimens necessary for experimental validation of mosaics (live cells or those preserved for cytology), all anomalies classified as mosaics were manually reviewed, and the BAF and LRR patterns that we observed are very similar to those reported in previous studies[7,17,53], which performed cytological validation for a variety of mosaic types.

Classification of clonal mosaic anomalies as duplication, deletion or aUPD was performed using the median LRR and BAF deviations from non-anomalous segments (**Fig. 2b**). Deviations from non-anomalous segments within the same sample were used to control for overall LRR variation among samples and for BAF asymmetry that occurred in some samples. Anomalies that were either terminal segments or whole chromosome and that had an LRR deviation within a neutral zone (|LRR| < 0.05) were classified as aUPD. This neutral zone was chosen because it included nearly all of the wide splits

(BAF deviation > 0.25) that had much smaller LRR deviations than expected for disomic-trisomic or disomic-monosomic transitions and included very few interstitial anomalies (**Supplementary Fig. 3**). All other anomalies (except for a few outliers) were classified as either duplications or deletions, depending on the sign of their LRR deviation. There was some ambiguity in classifying anomalies near the tip of the arrow, where the three transition zones intersect. This ambiguity is noted as intensity.flag (**Supplementary Table 3**). Mixture proportions in mosaics can be estimated as position along the transitional line that connects the two constitutional states (**Fig. 2** and **Supplementary Note**).

All anomalies discussed here are autosomal in the reference genome. Detection of X-chromosome mosaic anomalies is complicated by the fact that LRR is a measure of the intensity of a sample relative to other samples. LRR values for the X chromosome (calculated in the standard way) are affected by the sex ratio in the sample set and are not comparable to those for the autosomes.

Statistical analysis. All statistical analyses were carried out in the R statistical package using the functions described in the **Supplementary Note**.

47. Simon-Sanchez, J. *et al.* Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. *Hum. Mol. Genet.* **16**, 1–14 (2007).
48. Diskin, S.J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008).
49. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
50. Lin, P. *et al.* Copy number variation accuracy in genome-wide association studies. *Hum. Hered.* **71**, 141–147 (2011).
51. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
52. Tracy, N.D., Young, J.C. & Mason, R.L. Multivariate control charts for individual observations. *Journal of Quality Technology* **24**, 88–95 (1992).
53. Rodríguez-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).

doi:10.1038/ng.2271