# Genome-wide identification of copy neutral loss of heterozygosity reveals its possible association with spatial positioning of chromosomes

Kim, Hyeonjeong
Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University

Suyama, Mikita
Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University



出版情報：Human Molecular Genetics. 32 (7), pp.1175-1183, 2022-11-09. Oxford University Press (OUP)
バージョン：

*Human Molecular Genetics*, 2023, Vol. 32, 7, 1175–1183
https://doi.org/10.1093/hmg/ddac278
Advance access publication date 9 November 2022
**Original Article**

# Genome-wide identification of copy neutral loss of heterozygosity reveals its possible association with spatial positioning of chromosomes

Hyeonjeong Kim and Mikita Suyama*

Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Fukuoka 812-8582, Japan

*To whom correspondence should be addressed. Email: mikita@bioreg.kyushu-u.ac.jp

## Abstract

Loss of heterozygosity (LOH) is a genetic alteration that results from the loss of one allele at a heterozygous locus. In particular, copy neutral LOH (CN-LOH) events are generated, for example, by mitotic homologous recombination after monoallelic defection or gene conversion, resulting in novel homozygous locus having two copies of the normal counterpart allele. This phenomenon can serve as a source of genome diversity and is associated with various diseases. To clarify the nature of the CN-LOH such as the frequency, genomic distribution and inheritance pattern, we made use of whole-genome sequencing data of the three-generation CEPH/Utah family cohort, with the pedigree consisting of grandparents, parents and offspring. We identified an average of 40.7 CN-LOH events per individual taking advantage of 285 healthy individuals from 33 families in the cohort. On average 65% of them were classified as gonosomal-mosaicism-associated CN-LOH, which exists in both germline and somatic cells. We also confirmed that the incidence of the CN-LOH has little to do with the parents' age and sex. Furthermore, through the analysis of the genomic region including the CN-LOH, we found that the chance of the occurrence of the CN-LOH tends to increase at the GC-rich locus and/or on the chromosome having a relatively close inter-homolog distance. We expect that these results provide significant insights into the association between genetic alteration and spatial position of chromosomes as well as the intrinsic genetic property of the CN-LOH.

## Introduction

Genetic alteration is the source of genome diversity and has the potential to give rise to genomic evolution. Loss of heterozygosity (LOH), one such genetic alteration, is a homozygotization that results from the loss of the heterozygous state in diploid cells. In particular, if an LOH is generated by mitotic homologous recombination or gene conversion, the locus becomes copy neutral LOH (CN-LOH) (1).

CN-LOH has been reported to be implicated in malignancies as illustrated by Knudson's Two-Hit hypothesis and also known to be associated with other human diseases. The event often has adverse effects on cells by eliminating a copy of a normal allele and duplicating a copy of a mutated allele in a specific locus such as an oncogene and tumor suppressor gene. For example, myeloproliferative disorder was reported to be caused by gain-of-function mutations in *JAK2* gene followed by CN-LOH (2). Subsequently, studies showed that various diseases like myelodysplastic syndrome occur and deteriorate via CN-LOH as assessed using technologies for genetic analysis, such as SNP array and next-generation sequencing (3–6).

Although the CN-LOH is a functionally important genetic event, little is known about its nature; for example, its frequency and genomic distribution remain unclear. If we attempted to identify the CN-LOH using genomic data obtained only from a single individual, it would be difficult to discriminate between the CN-LOH and a normal homozygous allelic state. Therefore, data from single individuals and their parents are required to perform this type of analysis with precision. The same holds true for the identification of *de novo* mutations (DNMs), which are also difficult to distinguish from the normal heterozygous allelic state. Sasani recently conducted an investigation of the inheritance patterns of DNMs as well as identification using the whole-genome sequencing data of CEPH/Utah families (7). Their analysis motivated us to use the same data to identify the CN-LOH. This dataset comprises large three-generation family units consisting of grandparents, parents and several offspring in Utah in the United States (8). In particular, the Utah population has a traditionally high birth rate, which is a very powerful advantage in this type of analysis (7).

Here, using 285 individuals from 33 families in Utah in the United States, we identified the CN-LOH events over the whole-genome. We then assessed the characteristics of the CN-LOH, which is an exclusively postzygotic event, unlike DNM, by examining the correlation between the incidence of the event and the parents' features, such as age and sex. Moreover, the CN-LOH events were classified into two groups, those existing in the germline and somatic cells, by investigating the mode of inheritance of the CN-LOH events. We thus confirmed that the CN-LOH events are transmitted to progeny, showing that the CN-LOH events are also present in the germline cells and contributes to the

**Received:** September 8, 2022. **Revised:** October 17, 2022. **Accepted:** November 2, 2022

genetic diversity. Finally, we found that the occurrence of the CN-LOH might be associated with the inter-homolog distances, which reflect the chromosome territory (CT).

## Results

### Identification of CN-LOH events that is generated by mitotic homologous recombination using single nucleotide variant data

We investigated CN-LOH that is generated via mitotic homologous recombination or gene conversion using 33 large families from Utah in the United States (Fig. 1A). Although the original CEPH/Utah family dataset comprised 603 individuals from 33 large families who exhibited a blood relation (7), we selected 285 individuals from 33 immediate family units based on the married couples in the second generation for this study (Supplementary Material, Fig. S1), because the second-generation individuals have data pertaining to both biological parents and offspring, thus allowing the direct identification of such CN-LOH and examination of the inheritance of these events. To identify the CN-LOH, first, we detected the CN-LOH-defining SNV (L-SNV), which was considered to be included in the CN-LOH region because it indicated the genotype of the CN-LOH by comparing the genotype of second-generation individuals with those of their parents (first-generation individuals). Simply, we regarded the homozygous variants that indicated an impossible combination from the parents' genotype as L-SNVs (Fig. 1A); for example, although the possible combinations resulting from parents with A/T and A/A genotypes are A/A or A/T, if the offspring's genotype is T/T, then this SNV is deemed to be an L-SNV (see section Materials and Methods for details). Then, we estimated the minimum and maximum size of the CN-LOH events according to L-SNVs and adjacent heterozygous variants (Supplementary Material, Fig. S2). Because the CN-LOH is homozygous tracts, the heterozygous variants can be regarded as the mark of the boundary of the CN-LOH region. The minimum size corresponds to the distance from the first L-SNV to the last L-SNV in a CN-LOH event, and the maximum size corresponds to the distance between two nucleotides just before the first heterozygous variant upstream and downstream from a CN-LOH event. Finally, to examine the mode of the inheritance of the CN-LOH, we tracked the transmission of the CN-LOH of second-generation individuals to their offspring (third-generation individuals) by phasing the genotype throughout all three generations in the pedigree.

We detected 4296 L-SNVs in the entire cohort of 66 second-generation individuals from the 33 Utah families. The mean and median numbers of L-SNVs identified in an individual were 65.1 and 58.5, respectively. Then, we identified 2684 CN-LOH events based on the L-SNVs and heterozygous SNVs (Fig. 1B, Supplementary Material, Fig. S2). The number of L-SNVs included in a CN-LOH was 1 to 43. Moreover, 7 (sample ID 557) to 68 (sample ID 461) CN-LOH events were observed in each, and the mean and median numbers of events were 40.7 and 39, respectively. The length range of the minimum size was 1–155 999 bp, with a median size of 1 bp. In turn, the length range of the maximum size was 13–652 474 bp, with a median size of 8909 bp (Supplementary Material, Table S1). The incidence and scale of the CN-LOH events vary according to the individual.

We validated the identification of L-SNV because there is a concern about false positive calls due to genotyping error, which can easily lead to violation of Mendelian inheritance at a random locus (Supplementary Material, Fig. S3). To validate the identification of the L-SNV, we adopted a simulation-based approach to calculate a false positive rate. First, we arbitrarily choose a trio family that consists of a second-generation individual and one's parents (first-generation). In this analysis, we used sample ID 4, which was identified with 64 L-SNVs that are comparable to the mean number of the identified L-SNV, and the individual's parents (sample IDs 7 and 8). Then, we randomly marked 1% out of all variants of the three individuals as genotyping errors, respectively. Although it is known that the genotyping error rate of Genome Analysis Toolkit, which is a variant caller used in the CEPH/Utah cohort, is 0.1–1% (9), we adopted the 1% error rate to measure the false positive rate under stringent conditions. Finally, the false positive rate was calculated as the proportion of marked variants identified to be L-SNV. We carried out the above procedures 100 times repeatedly. The mean number and the standard deviation of the false positive calls of the sample ID 4's L-SNV are 1.93 and 1.43, respectively. Hence, there can be about a 3.0% (1.93/64) false positive rate for the identification of the L-SNV in each individual. The false positive rate calculated here must be the maximum because the reported genotyping calling error rate of gatk is 0.1–1% (9).

### Relationship between CN-LOH and parental age and sex

The emergence of postzygotic variants, which are genetic alterations that occur after fertilization, seems to be less affected by the condition of the parents compared with that of the variants that occur before fertilization (10–14). For example, about 10% of postzygotic DNMs did not correlate with the parents' sex and age, whereas about 90% of gonadal DNMs correlated with the parental parameters (7). Therefore, we hypothesized that there is little correlation between the occurrence of CN-LOH and parental sex and age, because the events occur between inter-homolog after fertilization.

To address this issue, first, we attempted to estimate the correlation between the incidence of CN-LOH and parental age. Among the 132 first-generation individuals, the age range of the first-generation male at childbirth (second-generation individuals) was 18.4 (sample ID 147) to 47.2 (sample ID 389) years, whereas the age range of the first-generation female was 16.4 (sample ID 577) to 37.1 (sample ID 442) years. We found that there was no significant correlation between the incidence of CN-LOH and paternal age ($r = 0.061$, $P = 0.63$, Fig. 2A). A similar trend was observed for maternal age ($r = -0.09$, $P = 0.47$, Fig. 2B). This result corroborates the findings of previous studies of postzygotic variation.

To examine the effect of parental sex on CN-LOH in terms of the mechanism underlying its occurrence, we classified the CN-LOH region according to chromosome and discriminated their gamete of origin. In general, regarding CN-LOH, the chromosome region at which a deletion occurred is termed 'recipient'; concomitantly, the counterpart chromosome region that repairs the defect of the homolog is termed 'donor' (15) (Fig. 2C). We investigated the origin of the CN-LOH region in the second-generation individuals. We observed that there was no significant difference between the number of sperm-originated CN-LOH events and the number of egg-originated CN-LOH events (Wilcoxon test, $P = 0.94$) (Fig. 2D). This implies that the incidence of the CN-LOH is not affected by a gamete bias. Taken together, these findings lead us to propose that the emergence of CN-LOH is not affected by the parents' age and sex, similar to that observed for postzygotic DNM.

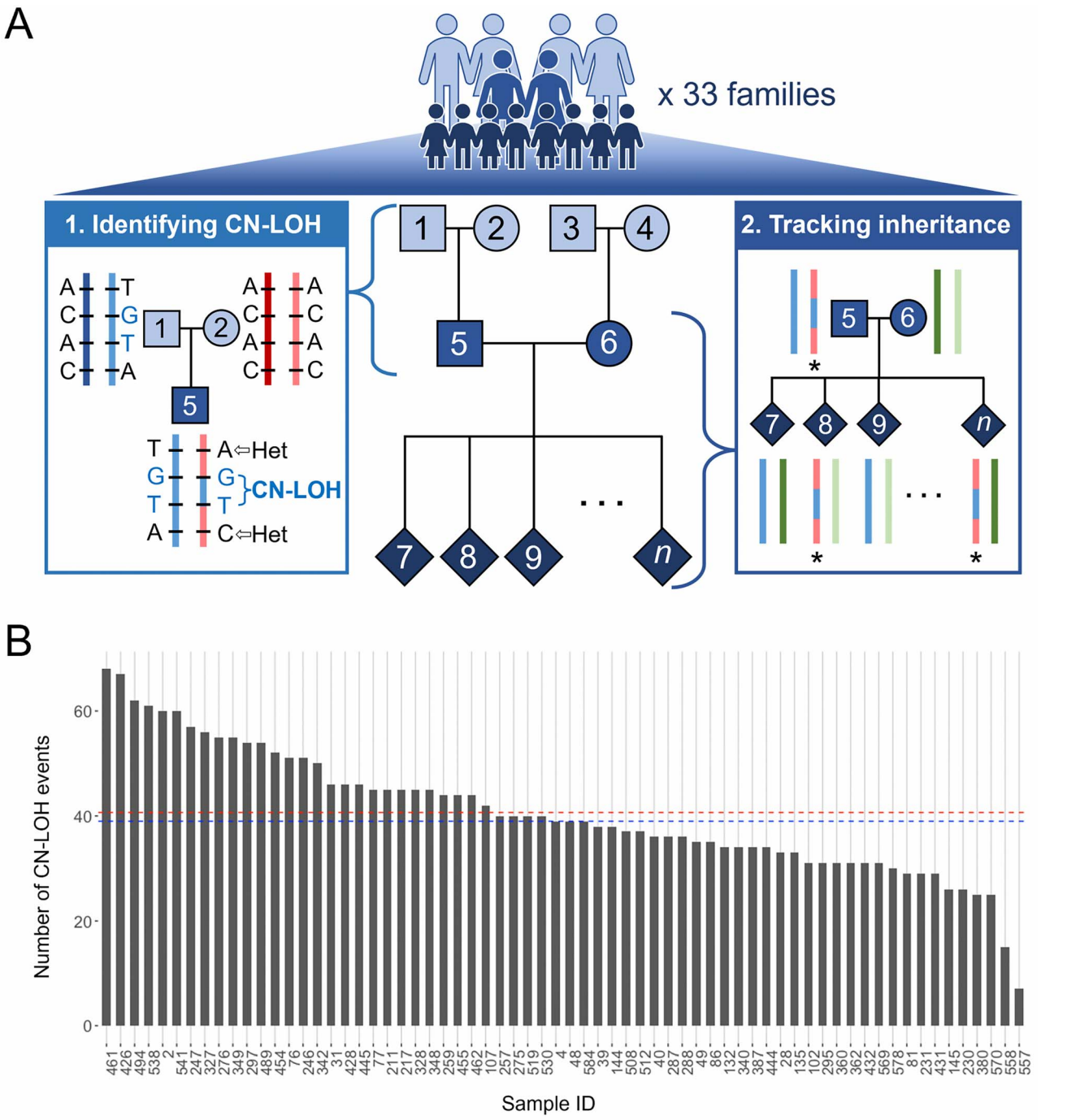


**Figure 1.** Identification of CN-LOH events using the SNV data from the CEPH/Utah families. (**A**) Schematic workflow of the study. Families in the CEPH/Utah dataset consist of three generations. The method of division of a large family is described in Supplementary Material, Figure S1. The CN-LOH events were identified by comparing the genotype of first- and second-generation individuals. The events presented a homozygous state that originated from the genotype that existed only in one parent (blue letter in the left box). The size of CN-LOH was restricted by the closest heterozygous variant that existed both upstream and downstream. The detailed strategy to determine the size of CN-LOH is presented in Supplementary Material, Figure S2. Subsequently, the mode of inheritance of the events was tracked by comparing the sequence block in the second- and third-generation individuals. (**B**) Total number of CN-LOH events identified in 66 second-generation individuals. The bar graph depicts the number of events. The blue and red dashed lines indicate the median and mean numbers of the events, respectively.

## Distinction between gonosomal-mosaicism-associated CN-LOH and Somatic-mosaicism-associated CN-LOH

The postzygotic variants are distributed in various ways in the human body and sometimes affect not only the carrier individuals but also their progeny via transmission (10,12,13,16–18). To examine the distribution of CN-LOH events in germline and somatic cells, we investigated the transmission of the events to offspring (Fig. 3A). Briefly, we set a 10-kbp window including heterozygote SNVs around the CN-LOH events and confirmed the presence of the window in offspring. If the window including the recipient was observed in more than one offspring individual (third-generation individuals), it was considered a gonosomal-mosaicism-associated CN-LOH, which is presented both in

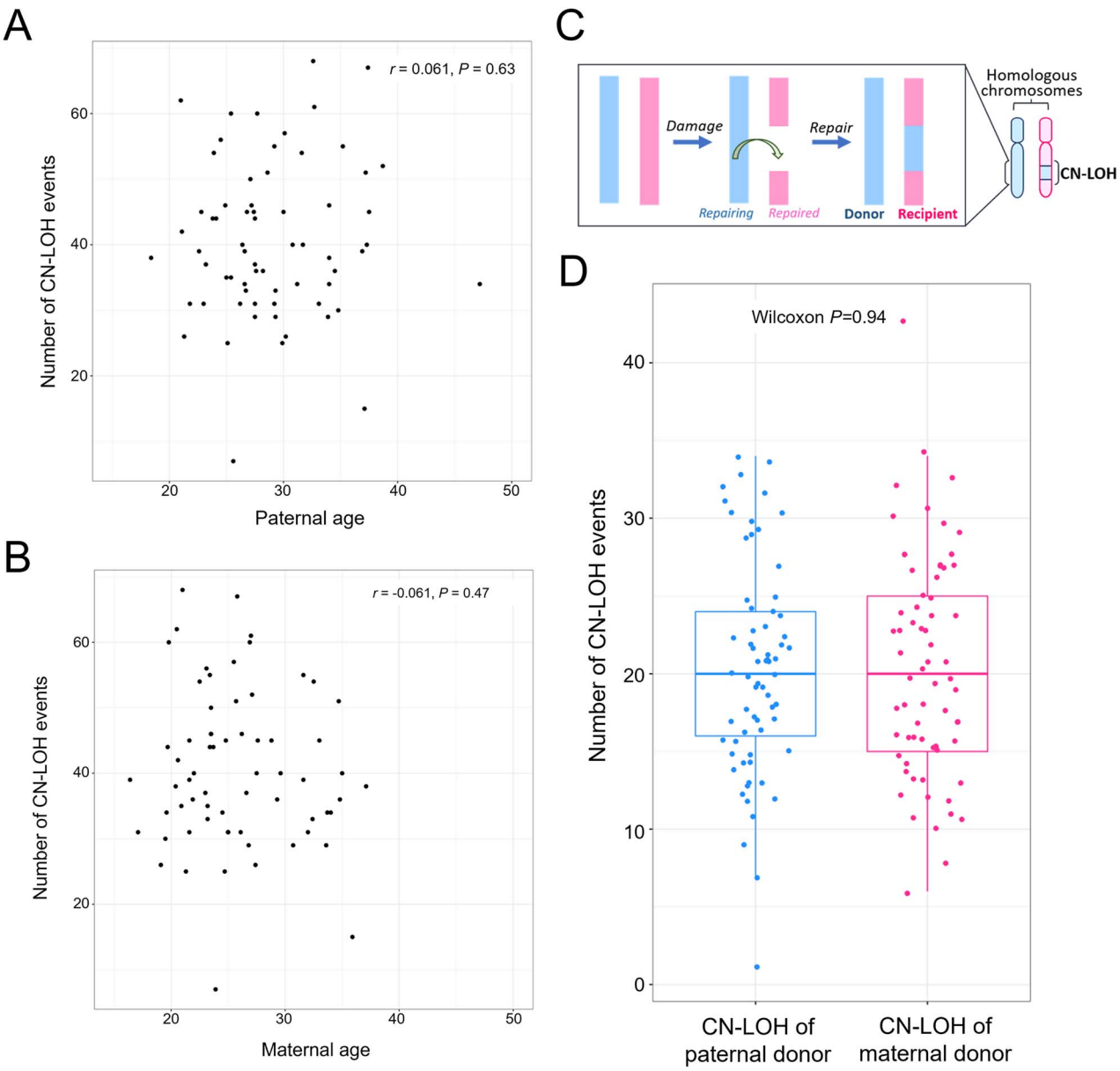


**Figure 2.** Relationships between CN-LOH and parental age/sex. (**A**) Scatterplot between the number of CN-LOH events and the paternal age (first-generation males). (**B**) The relationship between the number of CN-LOH events and the maternal age (first-generation females). (**C**) Conceptual diagram of the 'donor' and 'recipient' of homologous chromosomes. The segment in cyan indicates the part of the 'donor' chromosome that repairs the damage of the 'recipient.' The segment in magenta indicates the damaged part of the 'recipient' chromosome that is repaired by the 'donor.' (**D**) The boxplot indicates the number of CN-LOH events that originated from paternal or maternal chromosomes.

germline and somatic cells. In contrast, if the recipient before the occurrence of CN-LOH, was similar to that of the progenitor (first-generation individual), and/or the donor window was observed, this was considered a somatic-mosaicism-associated CN-LOH, which is present only in somatic cells (10,18).

We were able to track the transmission of 1870 CN-LOH events to offspring. There were 1214 gonosomal-mosaicism-associated CN-LOH events and 656 somatic-mosaicism-associated CN-LOH events (Fig. 3B). The average proportion of gonosomal versus somatic mosaicism-associated CN-LOH events was 0.65 versus 0.35 among the individuals studied here. Notably, the incidence of gonosomal-mosaicism-associated CN-LOH was approximately twice that of somatic-mosaicism-associated CN-LOH in normal healthy individuals. It is probable that gonosomal-mosaicism-associated CN-LOH, which can be observed in the blood cells of parents and offspring, occurs at a very early embryonic stage. Overall, given this variation in the inheritance mode of CN-LOH among the individuals, we speculated that this yields different genome sequences among siblings and, hence, potentially has a deleterious effect on diseases such as tumorigenesis in individuals.

## Relationship between the occurrence of CN-LOH and chromosomal features

The occurrence of HR repair during mitosis is typically associated with the degree of chromatin compactness and spatial distance of inter-homologs (19–22). HR repair tends to be suppressed at heterochromatin more than it is at euchromatin, and the chance of repair increases at closer inter-homolog distances (20–22). Similarly, regarding CN-LOH events, we wondered whether

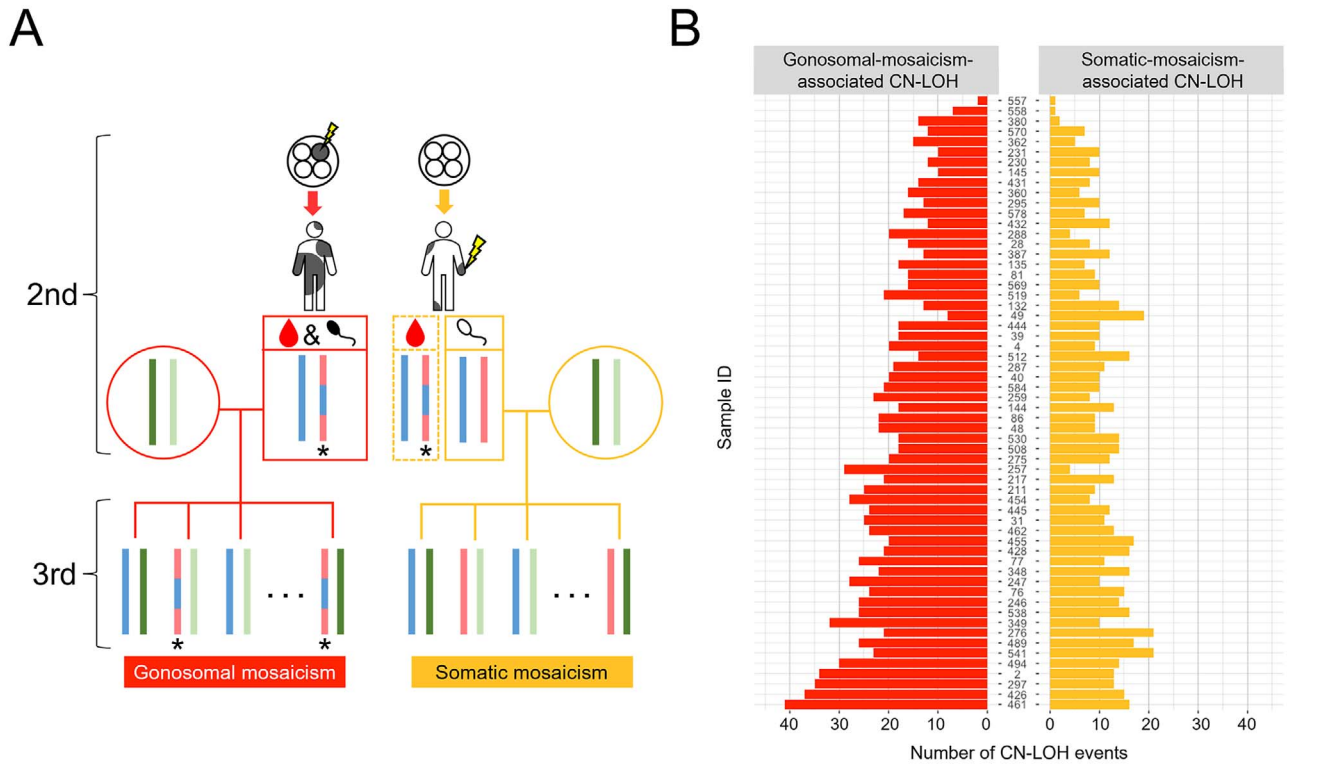


**Figure 3.** Distinction of gonosomal-mosaicism-associated CN-LOH from somatic-mosaicism-associated CN-LOH. (**A**) Schematic concept of gonosomal and somatic mosaicism. In the case of gonosomal mosaicism, the recipient is present both in germline (black sperm icon) and somatic cells (blood icon) because the CN-LOH occurs at a very early embryonic stage. Therefore, recipient can be observed in one or more third-generation siblings (asterisk). In the case of somatic mosaicism, the recipient is not present in germline cells (white sperm icon). LOH was not observed in any of the third-generation siblings. (**B**) The identified CN-LOH was classified into gonosomal-mosaicism-associated CN-LOH (red bar) and somatic-mosaicism-associated CN-LOH (gold bar). The sample ID (at the center) is arranged in the order of increasing count, which is the sum of the two types of CN-LOH.

they exhibit similar tendencies in light of the fact that these events occur between inter-homologs.

To address this question, we measured the GC content of the genomic region including the CN-LOH. The GC content of this region has been shown to strongly correlate with chromatin compactness (23). We first investigated the enrichment of all 2684 CN-LOH events by comparing the frequency of CN-LOH and GC content on the reference genome fraction, which was cleaved at 1 kbp as a window. At all CN-LOH events, the enrichment of the events was likely to increase as the GC content increased, from about 45% (Fig. 4A). In particular, we observed that CN-LOH tended to be enriched in GC-rich windows, from around >60% to ≤65%. This result corroborates the assumption that HR repair seems to be suppressed at heterochromatin in the genome.

In the human nucleus, gene-rich chromosomes, such as chromosome 19, are localized in the internal part of the nucleus, whereas gene-poor chromosomes, such as chromosome 18, are positioned at the periphery of the nucleus, near the nuclear lamina (22,24–28). If chromosomes are located in the internal part of the nucleus, they become spatially closer than if they are located at the periphery of the nucleus; accordingly, the inter-homolog distances are reduced and the chance of HR repair increases (22). Therefore, we hypothesized that CN-LOH occurs at gene-rich chromosomes more frequently than it does at gene-poor chromosomes because of the shorter inter-homolog distances.

Here, we observed that chromosome 19 was strikingly high in both gene density and mean number of CN-LOH events, whereas chromosome 18 was low in both features (Fig. 4B). It is well known that although inter-homolog distances are usually larger than inter-heterolog distances, the inter-homolog distance of chromosome 19 is small (24–26,28). Chromosomes 1, 4, 10, 14, 16 and 18 seem to reflect the fact that the inter-homolog distances are larger than the inter-heterolog distances (22). Chromosomes 21 and 22 exhibited a relatively high mean number of CN-LOH events, even though their gene density is lower than 30 genes/Mb. This result seems to confirm the results of a previous study that showed that shorter chromosomes tend to exist in the interior of the nucleus (25,27,29–31). In particular, chromosome 21 is preferentially situated in the internal region of the nucleus, because this chromosome is acrocentric and includes nucleolar organizer regions (22,30,32,33). We further confirmed that there are no mutations on the *LMNA* gene, which is known to disrupt chromosome positioning in nucleus (34–36), in all samples. Taken together, these results suggest that the chance of the emergence of CN-LOH depends on the gene density associated with chromatin compactness and inter-homolog distances, as illustrated by the radial position of a chromosome associated with CT.

## Discussion

We quantitatively identified CN-LOH events in germline as well as somatic cells by taking advantage of the SNV data of three-generation families. We identified approximately 40.7 CN-LOH events per individual. This number is comparable to the number of DNMs per individual, i.e. 70, reported previously (7). Intriguingly, more than half of the CN-LOH events exhibited gonosomal mosaicism. Genome variants tend to be patrolled and/or repaired more strictly in germline cells, as they are transferable to progeny,

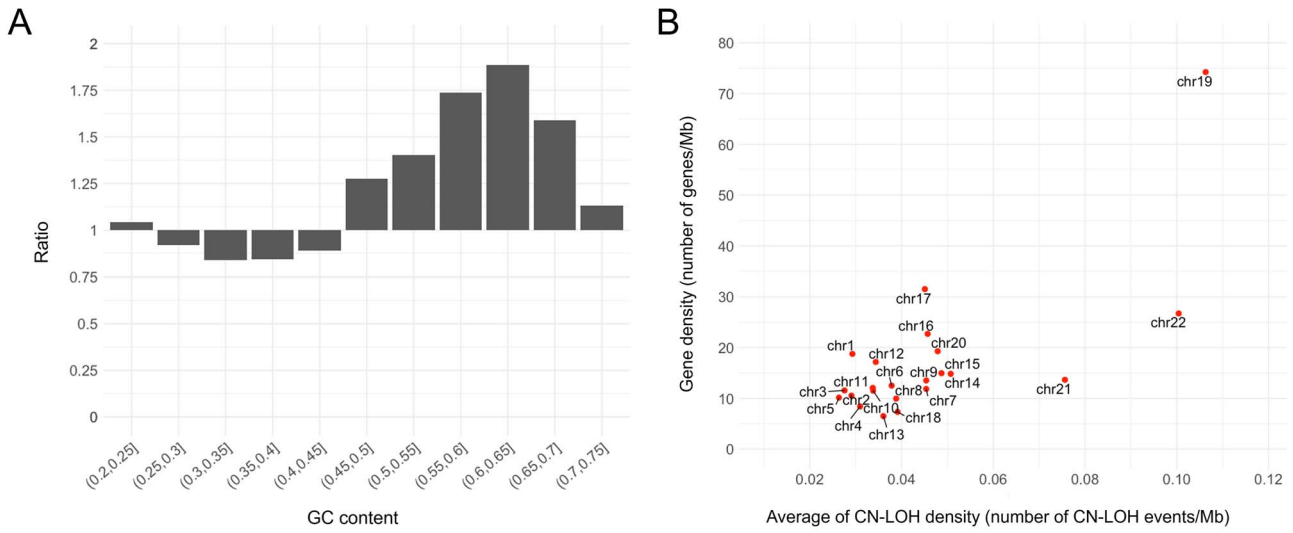


**Figure 4.** Relationship between the occurrence of CN-LOH and chromosomal features. (**A**) Histogram of CN-LOH enrichment according to GC content. The bar indicates the ratio of the observed CN-LOH to the expected CN-LOH according to the GC content of the genome regions. A ratio <1 means that LOH events are less enriched at that GC content, whereas a ratio >1 means that CN-LOH events are enriched at that GC content. (**B**) Scatterplot of the gene density per chromosome versus the average number of CN-LOH events per chromosome.

compared with in disposable somatic cells (37). The fact that the DNM rate in germline cells is lower than that detected in disposable somatic cells is a representative example of such a phenomenon (38,39). In contrast, considering that CN-LOH results from DNA repair after genome defection, it is possible to infer that the incidence of CN-LOH exhibits an opposite trend to that observed for DNM. Furthermore, we revealed the possible association between CT and the occurrence of CN-LOH by investigating the genomic features of the CN-LOH regions. Specifically, the incidence of CN-LOH is inversely proportional to the interhomolog distances, which depend on several parameters, such as gene density and chromosome length.

Two issues should be addressed in the interpretation of our findings. The first point is that the number of CN-LOH events identified in this study may have been underestimated for the following two reasons. (1) We applied stringent filtering to remove repeat regions, unassembled regions, and LCR. In fact, studies of genetic alterations have mentioned that HR spontaneously occurs in repeat regions (40). Nevertheless, to ensure the accuracy of our results, we excluded such regions, because they may have a low sequencing quality. (2) The raw data used here were sequenced from peripheral blood cells exclusively; i.e. tissue-specific CN-LOH events occurring in tissues other than blood were not included in our results. For example, for the onset of myeloproliferative disorders, the CN-LOH has to occur in T cells (2). Therefore, we suggest that the CN-LOH identified in this study may have been sufficiently frequent for detection by occurring in hematopoietic stem cells or in very early embryonic stem cells.

The second point is that the proportion of CN-LOH associated with gonosomal mosaicism observed in this study may actually be higher in reality. In this study, CN-LOH was classified into two groups, i.e. gonosomal-mosaicism-associated CN-LOH and somatic-mosaicism-associated CN-LOH, by investigating whether the CN-LOH was transmitted to offspring. The number of progenies is particularly important in this regard because it is directly related to the accuracy of the results (12,41,42). If, for instance, the third-generation includes 4 individuals, the probability that the CN-LOH that occurred in the parent is not transmitted to any offspring is (1/2) ^ 4, which is not negligible. Although human beings inherently do not have many offspring compared with other organisms, the Utah population used here, fortunately, has a relatively high birth rate because of their religious beliefs. Therefore, we included only families that had 7 or more third-generation individuals in our study, to take full advantage of the characteristics of the cohort [average of about 8.97 children (third-generation) per family].

Inherited variants in tumor suppressor genes are thought to be the main cause of hereditary cancer (43,44). In some tumors, the inherited pattern causes tumors more frequently than does the sporadic pattern. For example, in some cases, hematologic malignancies are incurred from the duplication of the mutated allele in specific gene such as *JAK2* gene via CN-LOH (5). Although cancers with a hereditary susceptibility represent 5–10% of all types of cancer (44), the nature of CN-LOH that is the trigger of the disease as the second hit has received little attention. We expect that the findings pertaining to the incidence of CN-LOH obtained in this study will provide clues to infer the onset rate of cancer among families with a hereditary cancer susceptibility.

## Materials and Methods

### Data set

VCF files were downloaded from the National Center for Biotechnology Information (NCBI)'s dbGaP, and the dataset title was "Genome sequencing of large, multigenerational CEPH/Utah families" (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001872.v1.p1). The dataset consisted of 603 individuals from 33 large families in Utah in the United States. The families comprise three biological generations, including offspring, parents, and grandparents, i.e. these data do not contain non-paternity. Pedigree information of these families was also obtained from the same study. There is no information about the consanguinity rate of the families in the study. Here, we used 285 individuals from 33 three-generation immediate families, which are a subset of the 33 large families and met the following conditions: (1) an intact family with usable SNV data for all family members (Supplementary Material, Fig. S1), and (2) presence of 7 or more siblings in the third-generation of a family.

## Identification of L-SNVs and CN-LOH

First, we performed the following quality control of the VCF files to identify L-SNVs that were considered to be included in the CN-LOH. The variant had to satisfy the condition of GATK HaplotypeCaller (45) as 'PASS', a read depth ≥ 12, and a Phred-scaled genotype quality ≥20 (42). The DNA sequences that corresponded to repeat regions and LCR were excluded [the data were downloaded from RepeatMasker (46) (Genome Reference Consortium Human Build 37; GRCh37) (https://www.repeatmasker.org) of the UCSC genome browser (47) and https://raw.githubusercontent.com/lh3/varcmp/master/scripts/LCR-hs37d5.bed.gz (48,49), respectively]. We only used variants located in autosomes to avoid a bias originating from sex. To discriminate from copy loss LOH, which is caused by direct deletion, we have eliminated the variants represented, for example, as 'REF: A; ALT: *', which means 'reference allele A is deleted in the individual', from the VCF file.

An L-SNV was defined as a homozygous locus with a genotype originating from a single parent that did not exist in combinations of parents, as assessed by referring to the genotype fields of the VCF format (Supplementary Material, Fig. S4). For example, if parents' genotypes are 1/2 and 2/2 and the progeny's genotype is 1/1, the progeny's SNV is regarded as an L-SNV. If, however, the progeny's genotype is 2/2, the SNV is excluded as a normal homozygous variant because the SNV is included in the possible combinations from the parent's genotype. If the progeny's genotype is 3/3, this SNV is excluded as a de novo variant or a variant call error. This is because any parents do not have allele 3, although the genotype is not included in the possible combinations from the parents' genotype. To discriminate between the homologous 'recipient' region, in which the CN-LOH literally occurred, and the 'donor' region, which acts as a template for DNA repair from L-SNV, we phased the filtered SNVs to comply with Mendelian inheritance using Beagle version 4.0 (50). We then phased the L-SNVs manually because of their relatively low phasing accuracy, which is attributable to the characteristics of Mendelian inheritance errors. Finally, the CN-LOH was inferred from the phased L-SNVs and SNVs. We defined the CN-LOH regions as homozygous tracts including L-SNVs and optionally homozygous SNVs that were restricted by the nearest heterozygous SNVs on both sides (Supplementary Material, Fig. S2).

## Assessment of the effect of parental age and sex on the occurrence of CN-LOH

To assess the effect of parental age on the occurrence of CN-LOH in offspring, we investigated the correlation between the incidence of CN-LOH in second-generation individuals and the age of the first-generation individuals. We obtained the information of the age and sex of first-generation individuals from https://github.com/quinlan-lab/ceph-dnm-manuscript. The parental age was rounded off to one decimal place in this study. Correlations with a $P$-value < 0.5 (Pearson's coefficient) between the incidence of CN-LOH and parental age were estimated using the default option of the 'ggpubr' package (v. 0.4.0) (https://rpkgs.datanovia.com/ggpubr/index.html) of the R (v.4.0.2) (The R Project for Statistical Computing, Vienna, Austria) software and visualized using the same package. The same approach was employed to assess and visualize the effect of parental sex on CN-LOH. In this case, the Wilcoxon test was used to compare the mean number of CN-LOH events.

## Discrimination between gonosomal-mosaicism-associated CN-LOH and somatic-mosaicism-associated CN-LOH

To discriminate between CN-LOH associated with gonosomal mosaicism and that associated with somatic mosaicism, we tracked the mode of transmission of the CN-LOH to the offspring. For this analysis, we used the 10-kbp window surrounding the CN-LOH that contained the heterozygous as well as the homozygous variants, such as the L-SNV. If the 'recipient' window was not observed in any of the offspring, the CN-LOH in the window was considered to be associated with somatic mosaicism, which is only present in somatic cells. In contrast, if the 'recipient' window was observed in one or more siblings, the CN-LOH was considered to be associated with gonosomal mosaicism, which is present in both germline and somatic cells. During the comparison of the window of second- and third-generation individuals, the variants existing in the window had to completely match each other.

## Estimation of GC content and gene density

We downloaded the information pertaining to chromosome length from NCBI's (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/) GRCh37. The GC content was calculated using the following formula.

$$\text{GC content} = (G + C) / (G + C + A + T).$$

Where A, C, G, and T indicate the number of each nucleotide. During calculation of GC content, we only used a 1-kbp window that contains more than 100 nt after removing the LCR, repeat region, and unassembled region. To measure gene density, we counted the number of 'protein_coding' genes at each chromosome, followed by calculations using the filtered chromosomal length. We used the comprehensive gene annotation of the GENCODE project (https://www.gencodegenes.org/human/release_19.html, GRCh37.p13).

## Supplementary Material

Supplementary Material is available at *HMG* online.

*Conflict of Interest statement*. The authors declared that no competing interests exist.

## Data availability

The whole-genome sequencing datasets utilized in this study are available in NCBI Sequence Read Archive (SRA) and dbGaP with the accession number phs001872.v1.p1. All source data generated in this study are available at https://github.com/rgwluj123/LOH_3generation_Utah.

## Code availability

The code developed for this study are freely available at https://github.com/rgwluj123/LOH_3generation_Utah.

## Author Contributions

M.S. conceived the project. M.S. and H.K. designed the research. H.K performed the research and analyzed the data. H.K. wrote the manuscript. M.S. revised the manuscript and supervised the study.

## Funding

Japan Science and Technology Agency (JST) SPRING (JPMJSP2136).

## References

1. Moynahan, M.E. and Jasin, M. (1997) Loss of heterozygosity induced by a chromosomal double-strand break. *Proc. Natl. Acad. Sci. U. S. A.*, **94**, 8988–8993.
2. Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.-S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M. and Skoda, R.C. (2005) A gain-of-function mutation of JAK2 in myeloproliferative disorders. *N. Engl. J. Med.*, **352**, 1779–1790.
3. Irving, J.A.E., Bloodworth, L., Bown, N.P., Case, M.C., Hogarth, L.A. and Hall, A.G. (2005) Loss of heterozygosity in childhood acute lymphoblastic leukemia detected by genome-wide microarray single nucleotide polymorphism analysis. *Cancer Res.*, **65**, 3053–3058.
4. Knudson, A.G. (2001) Two genetic hits (more or less) to cancer. *Nat. Rev. Cancer*, **1**, 157–162.
5. Maciejewski, J.P. and Mufti, G.J. (2008) Whole genome scanning as a cytogenetic tool in hematologic malignancies. *Blood*, **112**, 965–974.
6. Tlemsani, C., Takahashi, N., Pongor, L., Rajapakse, V.N., Tyagi, M., Wen, X., Fasaye, G.A., Schmidt, K.T., Desai, P., Kim, C. *et al.* (2021) Whole-exome sequencing reveals germline-mutated small cell lung cancer subtype with favorable response to DNA repair-targeted therapies. *Sci. Transl. Med.*, **13**, eabc7488.
7. Sasani, T.A., Pedersen, B.S., Gao, Z., Baird, L., Przeworski, M., Jorde, L.B. and Quinlan, A.R. (2019) Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife*, **8**, e46922.
8. Dausset, J., Cann, H., Cohen, D., Lathrop, M., Lalouel, J.M. and White, R. (1990) Centre d'Etude du polymorphisme humain (CEPH): collaborative genetic mapping of the human genome. *Genomics*, **6**, 575–577.
9. Depristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., Del Angel, G., Rivas, M.A., Hanna, M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
10. Biesecker, L.G. and Spinner, N.B. (2013) A genomic view of mosaicism and human disease. *Nat. Rev. Genet.*, **14**, 307–320.
11. Besenbacher, S., Liu, S., Izarzugaza, J.M.G., Grove, J., Belling, K., Bork-Jensen, J., Huang, S., Als, T.D., Li, S., Yadav, R. *et al.* (2015) Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat. Commun.*, **6**, 1–9.
12. Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Al Turki, S., Dominiczak, A., Morris, A., Porteous, D., Smith, B. *et al.* (2016) Timing, rates and spectra of human germline mutation. *Nat. Genet.*, **48**, 126–133.
13. Jónsson, H., Sulem, P., Arnadottir, G.A., Pálsson, G., Eggertsson, H.P., Kristmundsdottir, S., Zink, F., Kehr, B., Hjorleifsson, K.E., Jensson, B. *et al.* (2018) Multiple transmissions of de novo mutations in families. *Nat. Genet.*, **50**, 1674–1680.
14. Lindsay, S.J., Rahbari, R., Kaplanis, J., Keane, T. and Hurles, M.E. (2019) Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.*, **10**, 1–12.
15. Moynahan, M.E. and Jasin, M. (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nat. Rev. Mol. Cell Biol.*, **11**, 196–207.
16. Campbell, I.M., Stewart, J.R., James, R.A., Lupski, J.R., Stankiewicz, P., Olofsson, P. and Shaw, C.A. (2014) Parent of origin, mosaicism, and recurrence risk: probabilistic modeling explains the broken symmetry of transmission genetics. *Am. J. Hum. Genet.*, **95**, 345–359.
17. Campbell, I.M., Yuan, B., Robberecht, C., Pfundt, R., Szafranski, P., McEntagart, M.E., Nagamani, S.C.S., Erez, A., Bartnik, M., Wiśniowiecka-Kowalnik, B. *et al.* (2014) Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am. J. Hum. Genet.*, **95**, 173–182.
18. Campbell, I.M., Shaw, C.A., Stankiewicz, P. and Lupski, J.R. (2015) Somatic mosaicism: implications for disease and transmission genetics. *Trends Genet.*, **31**, 382–392.
19. Grewal, S.I.S. and Jia, S. (2007) Heterochromatin revisited. *Nat. Rev. Genet.*, **8**, 35–46.
20. Wang, J., Jia, S.T. and Jia, S. (2016) New insights into the regulation of heterochromatin. *Trends Genet.*, **32**, 284–294.
21. Watts, F.Z. (2016) Repair of DNA double-strand breaks in heterochromatin. *Biomol. Ther.*, **6**, 47.
22. Heride, C., Ricoul, M., Kiêu, K., Von Hase, J., Guillemot, V., Cremer, C., Dubrana, K. and Sabatier, L. (2010) Distance between homologous chromosomes results from chromosome positioning constraints. *J. Cell Sci.*, **123**, 4063–4075.
23. Dekker, J. (2007) GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. *Genome Biol.*, **8**, R116.
24. Croft, J.A., Bridger, J.M., Boyle, S., Perry, P., Teague, P. and Bickmore, W.A. (1999) Differences in the localization and morphology of chromosomes in the human nucleus. *J. Cell Biol.*, **145**, 1119–1131.
25. Bridger, J.M., Boyle, S., Kill, I.R. and Bickmore, W.A. (2000) Remodelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr. Biol.*, **10**, 149–152.
26. Boyle, S., Gilchrist, S., Bridger, J.M., Mahy, N.L., Ellis, J.A. and Bickmore, W.A. (2001) The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. *Hum. Mol. Genet.*, **10**, 211–220.
27. Cremer, T. and Cremer, C. (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat. Rev. Genet.*, **2**, 292–301.
28. Cremer, M., Küpper, K., Wagler, B., Wizelman, L., Hase, J.V., Weiland, Y., Kreja, L., Diebold, J., Speicher, M.R. and Cremer, T. (2003) Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. *J. Cell Biol.*, **162**, 809–820.
29. Sun, H.B., Shen, J. and Yokota, H. (2000) Size-dependent positioning of human chromosomes in interphase nuclei. *Biophys. J.*, **79**, 184–190.
30. Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M.R. *et al.* (2005) Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biol.*, **3**, e157.
31. Mora, L., Sánchez, I., Garcia, M. and Ponsà, M. (2006) Chromosome territory positioning of conserved homologous chromosomes in different primate species. *Chromosoma*, **115**, 367–375.
32. Chandley, A.C., Speed, R.M. and Leitch, A.R. (1996) Different distributions of homologous chromosomes in adult human Sertoli cells and in lymphocytes signify nuclear differentiation. *J. Cell Sci.*, **109**, 773–776.
33. Hernandez-Verdun, D. (2006) The nucleolus: a model for the organization of nuclear functions. *Histochem. Cell Biol.*, **126**, 135–148.
34. Filesi, I., Gullotta, F., Lattanzi, G., D'Apice, M.R., Capanni, C., Nardone, A.M., Columbaro, M., Scarano, G., Mattioli, E., Sabatelli, P.

*et al.* (2005) Alterations of nuclear envelope and chromatin organization in mandibuloacral dysplasia, a rare form of laminopathy. *Physiol. Genomics*, **23**, 150–158.
35. Meaburn, K.J., Cabuy, E., Bonne, G., Levy, N., Morris, G.E., Novelli, G., Kill, I.R. and Bridger, J.M. (2007) Primary laminopathy fibroblasts display altered genome organization and apoptosis. *Aging Cell*, **6**, 139–153.
36. Bera, M. and Sengupta, K. (2020) Nuclear filaments: role in chromosomal positioning and gene expression. *Nucleus*, **11**, 99.
37. Vermezovic, J., Stergiou, L., Hengartner, M.O. and D'Adda Di Fagagna, F. (2012) Differential regulation of DNA damage response activation between somatic and germline cells in *Caenorhabditis elegans*. *Cell Death Differ.*, **19**, 1847–1855.
38. Kirkwood, T.B.L. (1977) Evolution of ageing. *Nature*, **270**, 301–304.
39. Moore, L., Cagan, A., Coorens, T.H.H., Neville, M.D.C., Sanghvi, R., Sanders, M.A., Oliver, T.R.W., Leongamornlert, D., Ellis, P., Noorani, A. *et al.* (2021) The mutational landscape of human somatic and germline cells. *Nature*, **597**, 381–386.
40. Read, L.R., Raynard, S.J., Rukść, A. and Baker, M.D. (2004) Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. *Nucleic Acids Res.*, **32**, 1184–1196.
41. Goldmann, J.M., Wong, W.S.W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A.B., Glusman, G., Vissers, L.E.L.M., Hoischen, A., Roach, J.C. *et al.* (2016) Parent-of-origin-specific signatures of de novo mutations. *Nat. Genet.*, **48**, 935–939.
42. Jónsson, H., Sulem, P., Kehr, B., Kristmundsdottir, S., Zink, F., Hjartarson, E., Hardarson, M.T., Hjorleifsson, K.E., Eggertsson, H.P., Gudjonsson, S.A. *et al.* (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, **549**, 519–522.
43. Hodgson, S. (2008) Mechanisms of inherited cancer susceptibility. *J Zhejiang Univ Sci B*, **9**, 1–4.
44. Tsaousis, G.N., Papadopoulou, E., Apessos, A., Agiannitopoulos, K., Pepe, G., Kampouri, S., Diamantopoulos, N., Floros, T., Iosifidou, R., Katopodi, O. *et al.* (2019) Analysis of hereditary cancer syndromes by using a panel of genes: novel and multiple pathogenic mutations. *BMC Cancer*, **19**, 1–19.
45. Van der Auwera, G.A. and B.D.O. (2020) *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*, 1st edn. O'Reilly Media, California, p. 2020.
46. Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O. and Walichiewicz, J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
47. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, and D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
48. Li, H. and Wren, J. (2014) Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, **30**, 2843–2851.
49. Turner, T.N., Coe, B.P., Dickel, D.E., Hoekzema, K., Nelson, B.J., Zody, M.C., Kronenberg, Z.N., Hormozdiari, F., Raja, A., Pennacchio, L.A. *et al.* (2017) Genomic patterns of de novo mutation in simplex autism. *Cell*, **171**, 710–722.
50. Browning, S.R. and Browning, B.L. (2007) Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, **81**, 1084–1097.