Full title: Genome-wide identification of loss of heterozygosity generated by mitotic homologous recombination reveals its possible association with spatial positioning of chromosomes

Short title: Genome-wide identification of loss of heterozygosity

Hyeonjeong Kim[1], Mikita Suyama[1*]

[1] Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, Maidashi 3-1-1, Higashi-ku, Fukuoka, 812-8582, Japan

*Corresponding author

E-mail: mikita@bioreg.kyushu-u.ac.jp

Author Contributions

Conceptualization: Mikita Suyama

Data curation: Hyeonjeong Kim

Formal analysis: Hyeonjeong Kim

Investigation: Hyeonjeong Kim

Methodology: Hyeonjeong Kim

Project administration: Mikita Suyama

Software: Hyeonjeong Kim

Supervision: Mikita Suyama

Validation: Hyeonjeong Kim

Visualization: Hyeonjeong Kim

Writing – original draft: Hyeonjeong Kim

Writing – review & editing: Mikita Suyama

Abstract

Loss of heterozygosity (LOH) is a genetic alteration that results from the loss of one allele at a heterozygous locus. Some LOH events are generated by mitotic homologous recombination after monoallelic defection, then the novel homozygous locus has 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 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 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 LOH, which exists in both germline and somatic cells. We also confirmed that the incidence of the LOH has little to do with the parents' age and sex. Furthermore, through the analysis of the genomic region including the LOH, we found that the chance of the occurrence of the 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 LOH.

Author Summary

Loss of heterozygosity (LOH) is a common genetic alteration that a heterozygous locus becomes a homozygous locus. In some cases, if a monoallelic defection accompanying homologous recombination between inter-homolog occurs, it results in copy-neutral LOH having two copies of the allele. Although the LOH is potentially important in understanding pathogenesis of various diseases, its fundamental features have received scant attention. To characterize the nature of the LOH, data from single individuals and their parents are required. This is because it would be difficult to discriminate between the LOH and a normal homozygous allelic state using genomic data obtained only from a single individual. Whole-genome sequencing data of 33 CEPH/Utah families, which comprise large three-generation family units, motivated us to perform a genome-wide identification of the LOH. Using this dataset, we successfully identified the LOH and analyzed its frequency, genomic distribution, and inheritance pattern. Moreover, we revealed that the occurrence of the LOH was affected by the inter-homolog distances, which reflect the chromosome territory. Our findings pertaining to the LOH provide insight into the association between genetic alteration and spatial positioning of chromosomes.

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 via monoallelic defection in diploid cells. Particularly, if the monoallelic defection is repaired by mitotic homologous recombination using the normal counterpart allele, the locus becomes copy-neutral LOH [1].

The LOH having a neutral copy number 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 LOH [2]. Subsequently, studies showed that various diseases like myelodysplastic syndrome occur and deteriorate via LOH as assessed using technologies for genetic analysis, such as SNP array and next-generation sequencing [3–6].

Although the 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 LOH using genomic data obtained only from a single individual, it would be difficult to discriminate between the 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 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 LOH events over the whole-genome. We then assessed the characteristics of the 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 LOH events were classified into two groups, those existing in the germline and somatic cells, by investigating the mode of inheritance of the LOH events. We thus confirmed that the LOH events are transmitted to progeny, showing that the LOH events are also present in the germline cells and contributes to the genetic diversity. Finally, we found that the occurrence of the LOH, which is generated by mitotic homologous recombination, might be associated with the inter-homolog distances, which reflect the chromosome territory (CT).

Results

Identification of LOH events that is generated by mitotic homologous recombination using single nucleotide variant (SNV) data

We investigated LOH that is generated via gene conversion by mitotic homologous recombination 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 (S1 Fig), because the second-generation individuals have data pertaining to both biological parents and offspring, thus allowing the direct identification of such LOH and examination of the inheritance of these events. To identify the LOH, first, we detected the LOH-defining SNV (L-SNV), which was considered to be included in the LOH region because it indicated the genotype of the 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; 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 (Fig 1A) (see Methods for details). Subsequently, to examine the mode of the inheritance of the LOH, we tracked the transmission of the LOH of second-generation individuals to their offspring (third-generation individuals) by phasing the genotype throughout all three generations in the pedigree.

We detected 4,296 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 2,684 LOH events by merging the L-SNVs depending on the heterozygous SNVs (Fig 1B). The number of L-SNVs included in an LOH was 1 to 43. Moreover, 7 (sample ID 557) to 68 (sample ID 461) LOH events were observed in each, and the mean and median numbers of events were 40.7 and 39, respectively. In addition, we estimated the minimum and maximum size of the events according to L-SNVs and heterozygous variants. The minimum size corresponds to the distance from the first L-SNV to the last L-SNV in an LOH event, and the maximum size corresponds to the distance between two nucleotides just before the first heterozygous variant upstream and downstream

from an LOH event (S2 Fig). Based on this criterion, the length range of the minimum size was 1 to 155,999 bp, with a median size of 1 bp. In turn, the length range of the maximum size was 13 to 652,474 bp, with a median size of 8,909 bp (S1 Table). The incidence and scale of the 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. 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 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 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 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 LOH and paternal age ($r = 0.061$, $P = 0.63$, Fig 2A). A similar trend was observed for maternal age ($r = -0.09$, $P = 0.47$, S3 Fig). This result corroborates the findings of previous studies of postzygotic variation.

To examine the effect of parental sex on LOH in terms of the mechanism underlying its occurrence, we classified the LOH region according to chromosome and discriminated their gamete of origin. In general, regarding 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 2B). We investigated the origin of the LOH region in the second-generation individuals. We observed that there was no significant difference between the number of sperm-originated LOH events and the number of egg-originated LOH events (Wilcoxon test, $P = 0.94$) (Fig 2C). This implies that the DNA lesion that gives rise to the LOH is not affected by a gamete bias. Taken together, these findings lead us to propose that the emergence of LOH is not affected by the parents' age and sex, similar to that observed for postzygotic DNM.

Distinction Between Gonosomal-mosaicism-associated LOH and Somatic-mosaicism-associated 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 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 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 LOH, which is presented both in germline and somatic cells. In contrast,

if the recipient before the occurrence of 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 LOH, which is present only in somatic cells [10,18].

We were able to track the transmission of 1,870 LOH events to offspring. There were 1,214 gonosomal-mosaicism-associated LOH events and 656 somatic-mosaicism-associated LOH events (Fig 3B). The average proportion of gonosomal vs. somatic mosaicism-associated LOH events was 0.65 vs. 0.35 among the individuals studied here. Notably, the incidence of gonosomal-mosaicism-associated LOH was approximately twice that of somatic-mosaicism-associated LOH in normal healthy individuals. It is probable that gonosomal-mosaicism-associated 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 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 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 LOH events, we wondered whether they exhibit similar tendencies in light of the fact that these events are the result of HR repair after the defection of a single chromosome.

To address this question, we measured the GC content of the genomic region including the LOH. The GC content of this region has been shown to strongly correlate with chromatin compactness [23]. We first investigated the enrichment of all 2,684 LOH events by comparing the frequency of LOH and GC content on the reference genome fraction, which was cleaved at 1 kbp as a window. At all 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 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 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 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 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]. Taken together, these results suggest that the chance of the emergence of 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 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 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 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, compared with in disposable somatic cells [34]. 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 [35,36]. In contrast, considering that LOH results from DNA repair after genome defection, it is possible to infer that the incidence of LOH exhibits an opposite trend to that observed for DNM. Furthermore, we revealed the possible association between CT and the occurrence of LOH by investigating the genomic features of the LOH regions. Specifically, the incidence of LOH is inversely proportional to the inter-homolog 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 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 [37]. 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 LOH events occurring in tissues other than blood were not included in our results. For example, for the onset of myeloproliferative disorders, the LOH has to occur in T cells [2]. Therefore, we suggest that the 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 LOH associated with gonosomal mosaicism observed in this study may actually be higher in reality. In this study, LOH was classified into two groups, i.e., gonosomal-mosaicism-associated LOH and somatic-mosaicism-associated LOH, by investigating whether the 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 [12,38,39]. If, for instance, the third-

generation includes 4 individuals, the probability that the 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 [40,41]. 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 LOH [5]. Although cancers with a hereditary susceptibility represent 5%–10% of all types of cancer [41], the nature of 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 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

Dataset

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. Pedigree information of these families was also obtained from the same 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 (S1 Fig), and (2) presence of 7 or more siblings in the third-generation of a family.

Identification of L-SNVs and LOH

First, we performed the following quality control of the VCF files to identify L-SNVs that were considered to be included in the LOH. The variant had to satisfy the condition of GATK HaplotypeCaller [42] as "PASS", a read depth ≥12, and a Phred-scaled genotype quality ≥20[39]. The DNA sequences that corresponded to repeat regions and LCR were excluded [the data were downloaded from RepeatMasker [43] (Genome Reference Consortium Human Build 37; GRCh37) (http://www.repeatmasker.org) of the UCSC genome browser [44] and https://raw.githubusercontent.com/lh3/varcmp/master/scripts/LCR-hs37d5.bed.gz [45,46], respectively]. We only used variants located in autosomes to avoid a bias originating from sex.

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 (S4 Fig). 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 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 [47]. 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 LOH was inferred from the phased L-SNVs and SNVs. We defined the LOH regions as consecutive homologous regions including L-SNVs and optionally homozygous SNVs that were restricted by the nearest heterozygous SNVs on both sides (S2 Fig).

Assessment of the effect of parental age and sex on the occurrence of LOH

To assess the effect of parental age on the occurrence of LOH in offspring, we investigated the correlation between the incidence of 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 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 LOH. In this case, the Wilcoxon test was used to compare the mean number of LOH events.

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

To discriminate between LOH associated with gonosomal mosaicism and that associated with somatic mosaicism, we tracked the mode of transmission of the LOH to the offspring. For this analysis, we used the 10-kbp window surrounding the 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 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 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).

Acknowledgments

This work was supported by JST SPRING, Grant Number JPMJSP2136.

References

1. Moynahan ME, Jasin M. Loss of heterozygosity induced by a chromosomal double-strand break. Proc Natl Acad Sci U S A. 1997;94: 8988–8993. doi:10.1073/pnas.94.17.8988

2. Kralovics R, Passamonti F, Buser AS, Teo S-S, Tiedt R, Passweg JR, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med. 2005;352: 1779–1790. doi:10.1056/NEJMoa051113

3. Irving JAE, Bloodworth L, Bown NP, Case MC, Hogarth LA, Hall AG. Loss of Heterozygosity in Childhood Acute Lymphoblastic Leukemia Detected by Genome-Wide Microarray Single Nucleotide Polymorphism Analysis. Cancer Res. 2005;65: 3053–3058. doi:10.1158/0008-5472.CAN-04-2604

4. Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer. 2001;1: 157–162. doi:10.1038/35101031

5. Maciejewski JP, Mufti GJ. Whole genome scanning as a cytogenetic tool in hematologic malignancies. Blood. 2008;112: 965–974. doi:10.1182/BLOOD-2008-02-130435

6. Tlemsani C, Takahashi N, Pongor L, Rajapakse VN, Tyagi M, Wen X, et al. Whole-exome sequencing reveals germline-mutated small cell lung cancer subtype with favorable response to DNA repair-targeted therapies. Sci Transl Med. 2021;13: eabc7488. doi:10.1126/scitranslmed.abc7488

7. Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, et al. Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. Elife. 2019;8: e46922. doi:10.7554/eLife.46922

8. Dausset J, Cann H, Cohen D, Lathrop M, Lalouel JM, White R. Centre d'Etude du polymorphisme humain (CEPH): Collaborative genetic mapping of the human genome. Genomics. 1990;6: 575–577. doi:10.1016/0888-7543(90)90491-c

9. Depristo MA, Banks E, Poplin R, Garimella K V., Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498. doi:10.1038/ng.806

10. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. Nat Rev Genet. 2013;14: 307–320. doi:10.1038/nrg3424

11. Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. Nat Commun. 2015;6: 1–9. doi:10.1038/ncomms6969

12. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, et al. Timing, rates and spectra of human germline mutation. Nat Genet. 2016;48: 126–133. doi:10.1038/ng.3469

13. Jónsson H, Sulem P, Arnadottir GA, Pálsson G, Eggertsson HP, Kristmundsdottir S, et al. Multiple transmissions of de novo mutations in families. Nat Genet. 2018;50: 1674–1680.

doi:10.1038/s41588-018-0259-9

14. Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME. Similarities and differences in patterns of germline mutation between mice and humans. Nat Commun. 2019;10: 1–12. doi:10.1038/s41467-019-12023-w

15. Moynahan ME, Jasin M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat Rev Mol Cell Biol. 2010;11: 196–207. doi:10.1038/nrm2851

16. Campbell IM, Stewart JR, James RA, Lupski JR, Stankiewicz P, Olofsson P, et al. Parent of Origin, Mosaicism, and Recurrence Risk: Probabilistic Modeling Explains the Broken Symmetry of Transmission Genetics. Am J Hum Genet. 2014;95: 345–359. doi:10.1016/j.ajhg.2014.08.010

17. Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, McEntagart ME, et al. Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. Am J Hum Genet. 2014;95: 173–182. doi:10.1016/j.ajhg.2014.07.003

18. Campbell IM, Shaw CA, Stankiewicz P, Lupski JR. Somatic mosaicism: implications for disease and transmission genetics. Trends Genet. 2015;31: 382–392. doi:10.1016/j.tig.2015.03.013

19. Grewal SIS, Jia S. Heterochromatin revisited. Nat Rev Genet. 2007;8: 35–46. doi:10.1038/nrg2008

20. Wang J, Jia ST, Jia S. New Insights into the Regulation of Heterochromatin. Trends Genet. 2016;32: 284–294. doi:10.1016/j.tig.2016.02.005

21. Watts FZ. Repair of DNA Double-Strand Breaks in Heterochromatin. Biomolecules. 2016;6: 47. doi:10.3390/biom6040047

22. Heride C, Ricoul M, Kiêu K, Von Hase J, Guillemot V, Cremer C, et al. Distance between homologous chromosomes results from chromosome positioning constraints. J Cell Sci. 2010;123: 4063–4075. doi:10.1242/jcs.066498

23. Dekker J. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 2007;8: R116. doi:10.1186/gb-2007-8-6-r116

24. Croft JA, Bridger JM, Boyle S, Perry P, Teague P, Bickmore WA. Differences in the Localization and Morphology of Chromosomes in the Human Nucleus. J Cell Biol. 1999;145: 1119–1131. doi:10.1083/jcb.145.6.1119

25. Bridger JM, Boyle S, Kill IR, Bickmore WA. Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. Curr Biol. 2000;10: 149–152. doi:10.1016/s0960-9822(00)00312-2

26. Boyle S, Gilchrist S, Bridger JM, Mahy NL, Ellis JA, Bickmore WA. The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells. Hum Mol Genet. 2001;10: 211–220. doi:10.1093/hmg/10.3.211

27. Cremer T, Cremer C. Chromosome territories, nuclear architecture and gene regulation in

mammalian cells. Nat Rev Genet. 2001;2: 292–301. doi:10.1038/35066075

28. Cremer M, Küpper K, Wagler B, Wizelman L, Hase J V., Weiland Y, et al. Inheritance of gene density-related higher order chromatin arrangements in normal and tumor cell nuclei. J Cell Biol. 2003;162: 809–820. doi:10.1083/jcb.200304096

29. Sun H Bin, Shen J, Yokota H. Size-dependent positioning of human chromosomes in interphase nuclei. Biophys J. 2000;79: 184–190. doi:10.1016/S0006-3495(00)76282-5

30. Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, et al. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. PLOS Biol. 2005;3: e157. doi:10.1371/journal.pbio.0030157

31. Mora L, Sánchez I, Garcia M, Ponsà M. Chromosome territory positioning of conserved homologous chromosomes in different primate species. Chromosoma. 2006;115: 367–375. doi:10.1007/s00412-006-0064-6

32. Chandley AC, Speed RM, Leitch AR. Different distributions of homologous chromosomes in adult human Sertoli cells and in lymphocytes signify nuclear differentiation. J Cell Sci. 1996;109: 773–776. doi:10.1242/jcs.109.4.773

33. Hernandez-Verdun D. The nucleolus: A model for the organization of nuclear functions. Histochem Cell Biol. 2006;126: 135–148. doi:10.1007/s00418-006-0212-3

34. Vermezovic J, Stergiou L, Hengartner MO, D'Adda Di Fagagna F. Differential regulation of

DNA damage response activation between somatic and germline cells in Caenorhabditis elegans. Cell Death Differ. 2012;19: 1847–1855. doi:10.1038/cdd.2012.69

35. Kirkwood TBL. Evolution of ageing. Nature. 1977;270: 301–304. doi:10.1038/270301a0

36. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, et al. The mutational landscape of human somatic and germline cells. Nature. 2021;597: 381–386. doi:10.1038/s41586-021-03822-7

37. Read LR, Raynard SJ, Rukść A, Baker MD. Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. Nucleic Acids Res. 2004;32: 1184–1196. doi:10.1093/nar/gkh280

38. Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, et al. Parent-of-origin-specific signatures of de novo mutations. Nat Genet. 2016;48: 935–939. doi:10.1038/ng.3597

39. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549: 519–522. doi:10.1038/nature24018

40. Hodgson S. Mechanisms of inherited cancer susceptibility. J Zhejiang Univ Sci B. 2008;9: 1–4. doi:10.1631/jzus.B073001

41. Tsaousis GN, Papadopoulou E, Apessos A, Agiannitopoulos K, Pepe G, Kampouri S, et al. Analysis of hereditary cancer syndromes by using a panel of genes: Novel and multiple

pathogenic mutations. BMC Cancer. 2019;19: 1–19. doi:10.1186/s12885-019-5756-4

42. Geraldine A. Van der Auwera BDO. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. 1st ed. O'Reilly Media; 2020.

43. Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110: 462–467. doi:10.1159/000084979

44. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12: 996–1006. doi:10.1101/gr.229102

45. Li H, Wren J. Toward better understanding of artifacts in variant calling from high-coverage samples. Bioinformatics. 2014;30: 2843–2851. doi:10.1093/bioinformatics/btu356

46. Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, et al. Genomic Patterns of De Novo Mutation in Simplex Autism. Cell. 2017;171: 710–722. doi:10.1016/j.cell.2017.08.047

47. Browning SR, Browning BL. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. Am J Hum Genet. 2007;81: 1084–1097. doi:10.1086/521987

Competing interests

The authors declared that no competing interests exist.

Supporting information

S1 Fig. Example of division of a large family.

S2 Fig. Schematic concept of minimum and maximum size of LOH.

S3 Fig. Relationship between LOH and maternal age.

S4 Fig. Flow chart of L-SNV identification.

S1 Table. List of loss of heterozygosity identified in this study.

Figure legends

Fig 1. Identification of 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 S1 Fig. The 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 LOH was restricted by the closest heterozygous variant that existed both upstream and downstream. The detailed strategy to determine the size of LOH is presented in S2 Fig. 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 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.

Fig 2. Relationships between LOH and parental age/sex.

(A) Scatterplot between the number of LOH events and the paternal age (first-generation males). The relationship between the number of LOH events and the maternal age is presented in S3 Fig. (B) 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.” (C) The boxplot indicates the number of LOH events that originated from paternal or maternal chromosomes.

Fig 3. Distinction of gonosomal-mosaicism-associated LOH from somatic-mosaicism-associated LOH.

(A) Schematic concept of gonosomal and somatic mosaicism. In the case of gonosomal mosaicism, the variant is present both in germline (black sperm icon) and somatic cells (blood icon) because the variant occurs at a very early embryonic stage. Therefore, LOH can be observed in one or more third-generation siblings (yellow rectangle). In the case of somatic mosaicism, the variant is not present in germline cells (white sperm icon). LOH was not observed in any of the third-generation siblings. (B) LOH was

classified into gonosomal-mosaicism-associated LOH (red bar) and somatic-mosaicism-associated 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 LOH.

Fig 4. Relationship between the occurrence of LOH and chromosomal features.

(A) Histogram of LOH enrichment according to GC content. The bar indicates the ratio of the observed LOH to the expected 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 LOH events are enriched at that GC content. (B) Scatterplot of the gene density per chromosome vs. the average number of LOH events per chromosome.

Data reporting

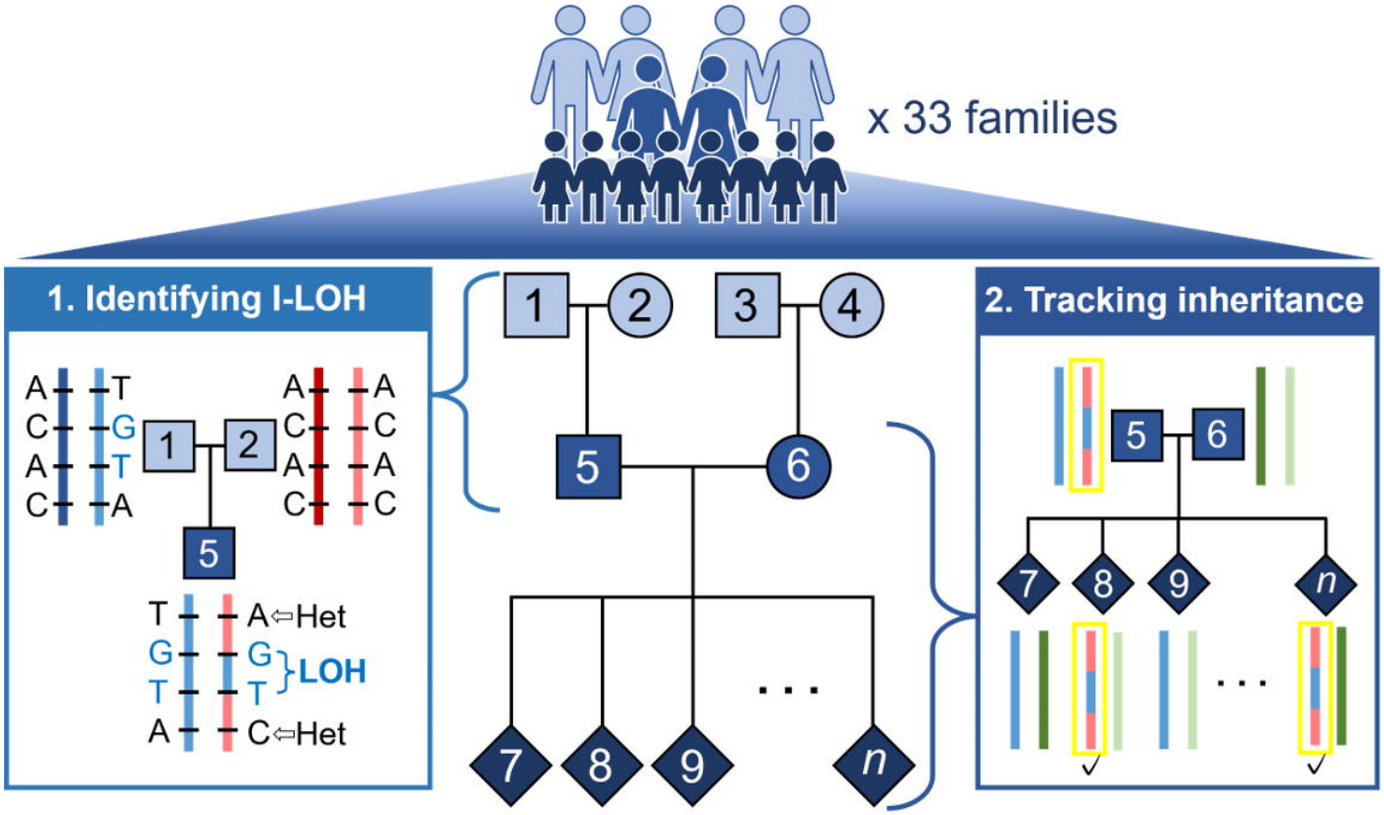
Data availability

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.

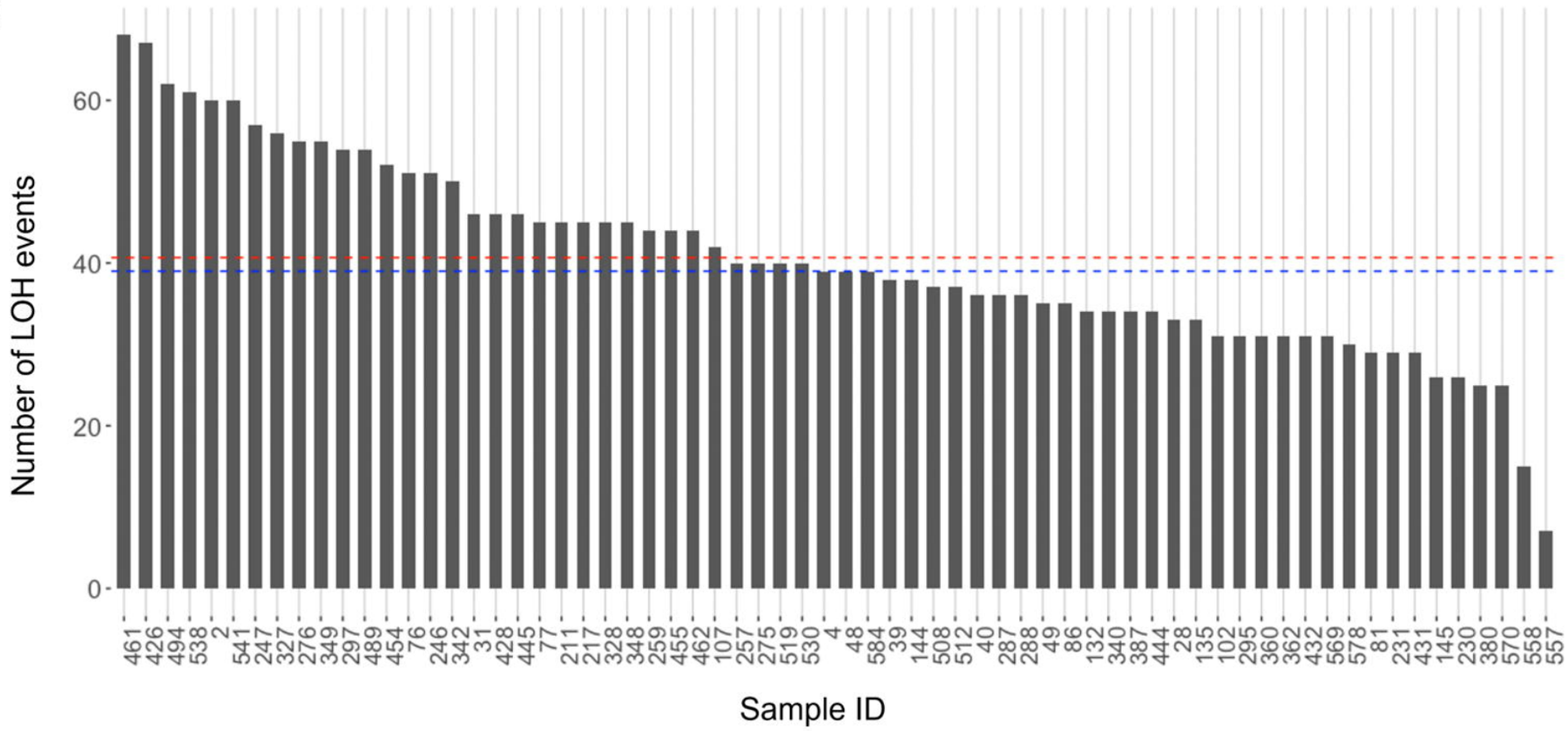
Code availability

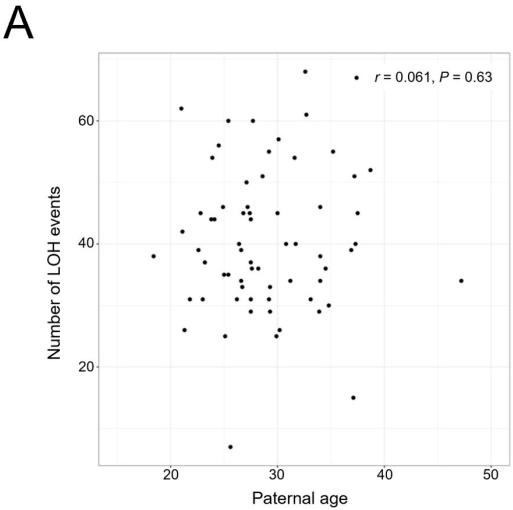
The code developed for this study are freely available at https://github.com/rgwluj123/LOH_3generation_Utah.

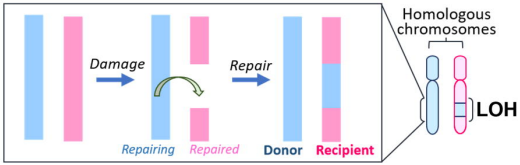
A

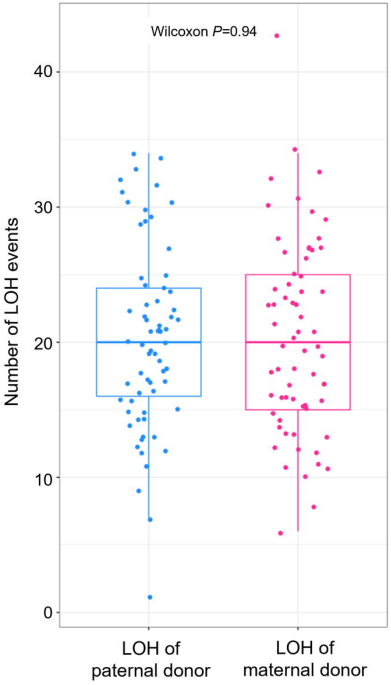

x 33 families
1. Identifying I-LOH
A
C
A
C
T
G
T
A
1
2
5
A
C
A
C
A
C
A
C
T
G
T
A
A⇦Het
G
T
LOH
C⇦Het
1
2
3
4
5
6
7
8
9
n
2. Tracking inheritance
5
6
7
8
9
n

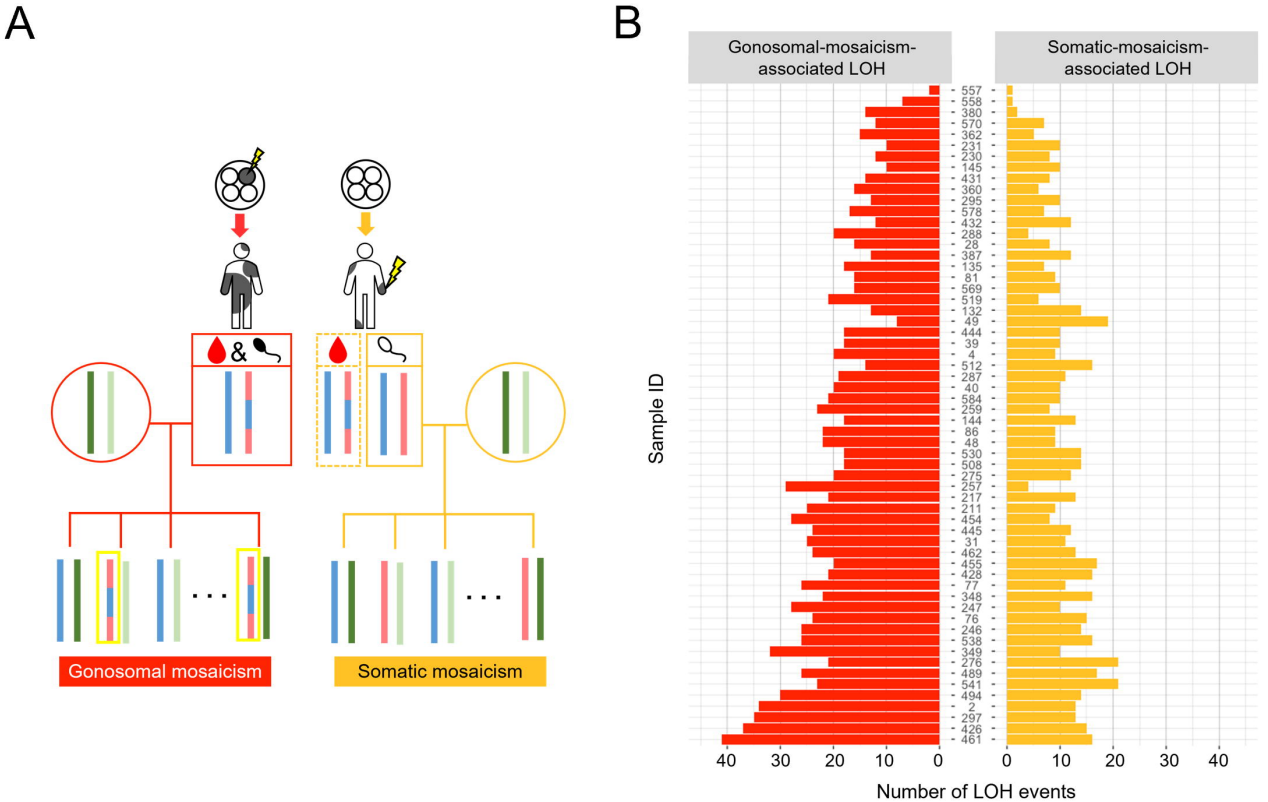
B

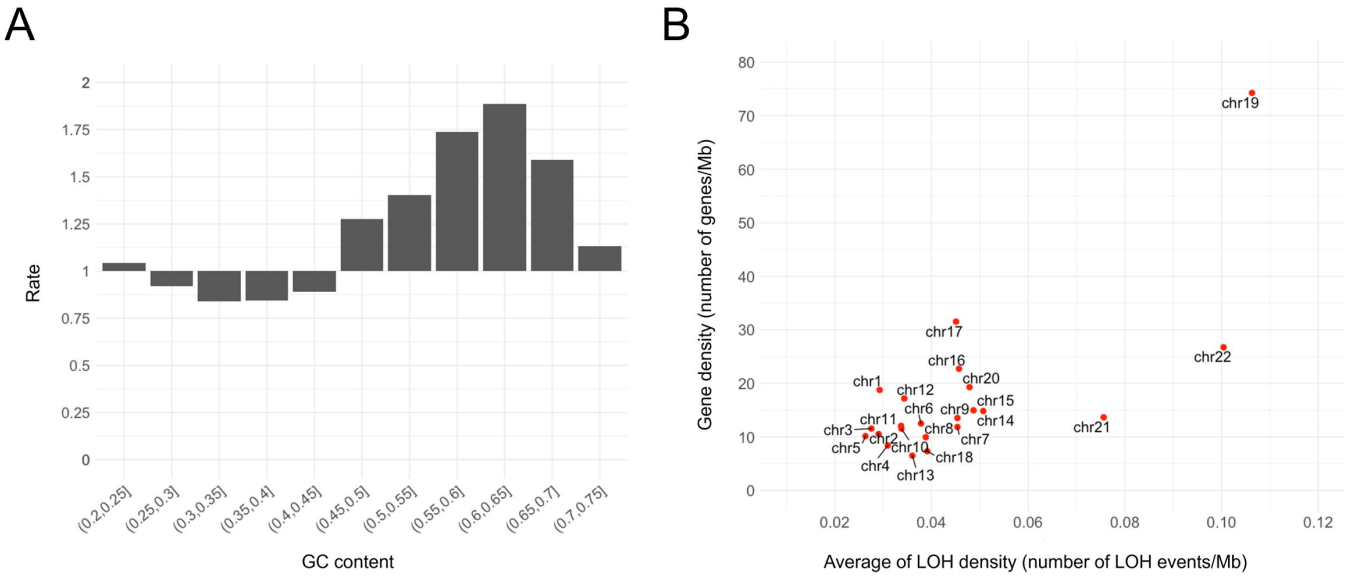

Number of LOH events
0
20
40
60
461 426 494 538 2 541 247 327 276 349 297 489 454 76 246 342 31 428 445 77 211 217 328 348 259 455 462 107 257 275 519 530 4 48 584 39 144 508 512 40 287 288 49 86 132 340 387 444 28 135 102 295 360 362 432 569 578 81 231 431 145 230 380 570 558 557
Sample ID


A
r = 0.061, P = 0.63
Number of LOH events
Paternal age
20
30
40
50
60


B
Damage
Repair
Repairing
Repaired
Donor
Recipient
Homologous chromosomes
LOH


C
Wilcoxon P=0.94
Number of LOH events
0
10
20
30
40
LOH of paternal donor
LOH of maternal donor


A
Gonosomal mosaicism
Somatic mosaicism
B
Gonosomal-mosaicism-associated LOH
Somatic-mosaicism-associated LOH
Sample ID
557
558
380
570
362
231
230
145
431
360
295
578
432
288
28
387
135
81
569
519
132
49
444
39
4
512
287
40
584
259
144
86
48
530
508
275
257
217
211
454
445
31
462
455
428
77
348
247
76
246
538
349
276
489
541
494
2
297
426
461
40
30
20
10
0
0
10
20
30
40
Number of LOH events


A
Rate
2
1.75
1.5
1.25
1
0.75
0.5
0.25
0
(0.2,0.25]
(0.25,0.3]
(0.3,0.35]
(0.35,0.4]
(0.4,0.45]
(0.45,0.5]
(0.5,0.55]
(0.55,0.6]
(0.6,0.65]
(0.65,0.7]
(0.7,0.75]
GC content
B
Gene density (number of genes/Mb)
80
70
60
50
40
30
20
10
0
chr19
chr17
chr22
chr16
chr20
chr1
chr12
chr15
chr6
chr9
chr11
chr14
chr3
chr8
chr21
chr2
chr7
chr5
chr10
chr18
chr4
chr13
0.02
0.04
0.06
0.08
0.10
0.12
Average of LOH density (number of LOH events/Mb)