Title:

Genome-wide identification of loss of heterozygosity reveals its association with spatial positioning of chromosomes

Authors and affiliations:

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

Contact info:

Mikita Suyama

*Correspondence: mikita@bioreg.kyushu-u.ac.jp

Phone: +81-92-642-6384

ORCID: 0000-0001-9526-3193

[1] Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University

Maidashi 3-1-1, Higashi-ku, Fukuoka, 812-8582, Japan

Abstract

Loss of heterozygosity (LOH) is a genetic alteration that results from the loss of one allele at a heterozygous locus. This phenomenon can serve as a crucial source of genome diversity and is associated with diseases such as cancer. We investigated the frequency, genomic distribution, and inheritance pattern of LOH using 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 and observed that 65% of them, on average, were transmitted to offspring because of gonosomal mosaicism. Moreover, we revealed that the occurrence of LOH was affected by the inter-homolog distances, which reflect the chromosome territory. Our findings pertaining to LOH provide insight into the pathogenesis of hereditary cancer, as illustrated by the Two-Hit Hypothesis.

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. After the occurrence of a DNA lesion, such as DNA double-strand breaks, in a chromosome, LOH spontaneously occurs via homologous recombination (HR) from the homologous chromosome (Moynahan and Jasin, 1997).

The LOH has been implicated in carcinogenesis, as illustrated by Knudson's Two-Hit Hypothesis, which was reported in 1971 (Knudson, 1971). This hypothesis was coined to explain the two critical mutations that contribute to the onset of cancer. One explanation for hereditary cancer is that an inherited mutation (first hit) is already present in germline cells and an LOH (second hit) is acquired in somatic cells. The other explanation is for sporadic cancer, both variants (the first and second hits) occur in somatic cells (Knudson, 1971). In humans, such a case was first reported in retinoblastoma (Friend et al., 1986). Knudson also demonstrated his hypothesis by showing that the *Tsc2* gene acquired tumor-predisposing mutations via LOH using the Eker rat, which is a model of dominantly inherited renal tumors (Hino et al., 1993). Subsequently, studies performed using somatic cells showed that various hereditary cancers, such as breast and ovarian cancer, occur via the two-hit phenomenon, as assessed using technologies for genetic analysis, such as DNA markers, microarray assays, and next-generation sequencing (Irving et al., 2005; Knudson, 2001; Maciejewski and Mufti, 2008; Tlemsani et al., 2021).

Although 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 LOH using genomic data obtained only from a single individual, it would be difficult to discriminate between 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 (Sasani et al., 2019). Their analysis motivated us to use the same data to identify LOH. This dataset comprises large three-generation family units consisting of grandparents, parents, and several offspring in Utah in the United States (Dausset et al.,

1990). In particular, the Utah population has a traditionally high birth rate, which is a very powerful advantage in this type of analysis (Sasani et al., 2019).

Here, using 285 individuals from 33 families in Utah in the United States, we identified 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, LOH events were classified into two groups, those existing in the germline and somatic cells, by investigating the mode of inheritance of the LOH. We thus confirmed that LOH is transmitted to progeny, showing that LOH is also present in the germline cells and contributes to the genetic diversity. Finally, we discovered that the occurrence of LOH is significantly associated with the inter-homolog distances, which reflect the chromosome territory (CT).

Results

Identification of LOH events using single nucleotide variant (SNV) data from the CEPH/Utah families

We investigated LOH in 33 large families from Utah in the United States (Figure 1A). LOH is the genetic alteration that occurs between inter-homologs during the repair of a DNA lesion in one homolog via gene conversion of the counterpart homolog. Although the original CEPH/Utah family dataset comprised 603 individuals from 33 large families who exhibited a blood relation (Sasani et al., 2019), we selected 285 individuals from 33 immediate family units based on the married couples in the second-generation for this study (Figure 1–figure supplement 1), because the second-generation individuals have data pertaining to both biological parents and offspring, thus allowing the direct identification of LOH and examination of the inheritance of these events. To identify 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 (Figure 1A). Subsequently, to examine the mode of the inheritance of 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. Then, we identified 2,684 LOH events by merging the L-SNVs depending on the heterozygous SNVs (Figure 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 (Figure 1–figure supplement 2). 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 (Supplementary file 1). The incidence and scale of the LOH events vary according to the individual.

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 (Besenbacher et al., 2015; Biesecker and Spinner, 2013; Jónsson et al., 2018; Lindsay et al., 2019; Rahbari et al., 2016). 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 (Sasani et al., 2019). 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, Figure 2A). A similar trend was observed for maternal age (r = −0.09, P = 0.47, Figure 2–figure supplement 1). 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" (Moynahan and Jasin, 2010) (Figure 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) (Figure 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 (Biesecker and Spinner, 2013; Campbell et al., 2015, 2014a, 2014b; Jónsson et al., 2018; Rahbari et al., 2016) . To examine the distribution of LOH events in germline and somatic cells, we investigated the transmission of the events to offspring (Figure 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 window was not observed in any offspring and the original window, which is the condition of 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 (Biesecker and Spinner, 2013; Campbell et al., 2015).

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 (Figure 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 (Grewal and Jia, 2007; Heride et al., 2010; Wang et al., 2016; Watts, 2016). 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 (Heride et al., 2010; Wang et al., 2016; Watts, 2016). 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 (Dekker, 2007). 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% (Figure 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 (Boyle et al., 2001; Bridger et al., 2000; Cremer et al., 2003; Cremer and Cremer, 2001; Croft et al., 1999; Heride et al., 2010). 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 (Heride et al., 2010). 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 (Figure 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 (Boyle et al., 2001; Bridger et al., 2000; Cremer et al., 2003; Croft et

al., 1999). 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 (Heride et al., 2010). 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 (Bolzer et al., 2005; Bridger et al., 2000; Cremer and Cremer, 2001; Mora et al., 2006; Sun et al., 2000). In particular, chromosome 21 is preferentially situated in the internal region of the nucleus, because this chromosome is acrocentric and includes nucleolar organizer regions (Bolzer et al., 2005; Chandley et al., 1996; Heride et al., 2010; Hernandez-Verdun, 2006). 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 (Sasani et al., 2019). 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 (Vermezovic et al., 2012). 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 (Kirkwood, 1977; Moore et al., 2021). 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 that CT affects 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 (Read et al., 2004). 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 retinoblastoma, which illustrates the Two-Hit Hypothesis, not only mutation (first hit) but also LOH (second hit) has to occur in retinal cells (Knudson, 1971). 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 (Goldmann et al., 2016; Jónsson et al., 2017; Rahbari et al., 2016). 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 (Hodgson, 2008; Tsaousis et al., 2019). In some tumors, the inherited pattern causes tumors more frequently than does the sporadic pattern. For example, in many cases, retinoblastoma is incurred from the duplication of the hereditary defected *RB1* gene via somatic LOH (Friend et al., 1986; Ryland et al., 2015). Although cancers with a hereditary susceptibility represent 5%–10% of all types of cancer (Tsaousis et al., 2019), 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 (Figure 1–figure supplement 1), 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 (Geraldine A. Van der Auwera, 2020) as "PASS", a read depth ≥12, and a Phred-scaled genotype quality ≥20(Jónsson et al., 2017). The DNA sequences that corresponded to repeat regions and LCR were excluded [the data were downloaded from RepeatMasker (Jurka et al., 2005) (Genome Reference Consortium Human Build 37; GRCh37) (http://www.repeatmasker.org) of the UCSC genome browser (Kent et al., 2002) and https://raw.githubusercontent.com/lh3/varcmp/master/scripts/LCR-hs37d5.bed.gz (Li and Wren, 2014; Turner et al., 2017), 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. 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 (Browning and Browning, 2007). 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 (Figure 1–figure supplement 2).

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).

Data and code availability

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

The following previously published data sets were used

Sasani TA, et al, (2019), 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, NCBI dbGaP, phs001872.v1.p1

Acknowledgments

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

References

Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, Huang S, Als TD, Li S, Yadav R, Rubio-García A, Lescai F, Demontis D, Rao J, Ye W, Mailund T, Friborg RM, Pedersen CNS, Xu R, Sun J, Liu H, Wang O, Cheng X, Flores D, Rydza E, Rapacki K, Damm Sørensen J, Chmura P, Westergaard D, Dworzynski P, Sørensen TIA, Lund O, Hansen T, Xu X, Li N, Bolund L, Pedersen O, Eiberg H, Krogh A, Børglum AD, Brunak S, Kristiansen K, Schierup MH, Wang J, Gupta R, Villesen P, Rasmussen S. 2015. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**:1–9. doi:10.1038/ncomms6969

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

Bolzer A, Kreth G, Solovei I, Koehler D, Saracoglu K, Fauth C, Müller S, Eils R, Cremer C, Speicher MR, Cremer T. 2005. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLOS Biol* **3**:e157. doi:10.1371/JOURNAL.PBIO.0030157

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

Bridger JM, Boyle S, Kill IR, Bickmore WA. 2000. Re-modelling of nuclear architecture in quiescent and senescent human fibroblasts. *Curr Biol* **10**:149–152. doi:10.1016/S0960-9822(00)00312-2

Browning SR, Browning BL. 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. doi:10.1086/521987

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

Campbell IM, Stewart JR, James RA, Lupski JR, Stankiewicz P, Olofsson P, Shaw CA. 2014a. Parent of Origin, Mosaicism, and Recurrence Risk: Probabilistic Modeling Explains the Broken Symmetry of Transmission Genetics. *Am J Hum Genet* **95**:345–359. doi:10.1016/J.AJHG.2014.08.010

Campbell IM, Yuan B, Robberecht C, Pfundt R, Szafranski P, McEntagart ME, Nagamani SCS, Erez A, Bartnik M, Wiśniowiecka-Kowalnik B, Plunkett KS, Pursley AN, Kang SHL, Bi W, Lalani SR, Bacino CA, Vast M, Marks K, Patton M, Olofsson P, Patel A, Veltman JA, Cheung SW, Shaw CA, Vissers LELM, Vermeesch JR, Lupski JR, Stankiewicz P. 2014b. Parental Somatic Mosaicism Is Underrecognized and Influences Recurrence Risk of Genomic Disorders. *Am J Hum Genet* **95**:173–182. doi:10.1016/J.AJHG.2014.07.003

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

Cremer M, Küpper K, Wagler B, Wizelman L, Hase J V., Weiland Y, Kreja L, Diebold J, Speicher MR, Cremer T. 2003. Inheritance of gene density-related higher order chromatin arrangements in normal

and tumor cell nuclei. *J Cell Biol* **162**:809–820. doi:10.1083/JCB.200304096

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**:292–301. doi:10.1038/35066075

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

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

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. doi:10.1186/GB-2007-8-6-R116

Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, Dryja TP. 1986. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. *Nature* **323**:643–646. doi:10.1038/323643a0

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

Goldmann JM, Wong WSW, Pinelli M, Farrah T, Bodian D, Stittrich AB, Glusman G, Vissers LELM,

Hoischen A, Roach JC, Vockley JG, Veltman JA, Solomon BD, Gilissen C, Niederhuber JE. 2016. Parent-of-origin-specific signatures of de novo mutations. *Nat Genet* **48**:935–939. doi:10.1038/ng.3597

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

Heride C, Ricoul M, Kiêu K, Von Hase J, Guillemot V, Cremer C, Dubrana K, Sabatier L. 2010. Distance between homologous chromosomes results from chromosome positioning constraints. *J Cell Sci* **123**:4063–4075. doi:10.1242/JCS.066498

Hernandez-Verdun D. 2006. The nucleolus: A model for the organization of nuclear functions. *Histochem Cell Biol* **126**:135–148. doi:10.1007/S00418-006-0212-3/FIGURES/5

Hino O, Mitani H, Knudson AG. 1993. Genetic Predisposition to Transplacentally Induced Renal Cell Carcinomas in the Eker Rat1. *Cancer Res* **53**:5856–5858.

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

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

Jónsson H, Sulem P, Arnadottir GA, Pálsson G, Eggertsson HP, Kristmundsdottir S, Zink F, Kehr B,

Hjorleifsson KE, Jensson B, Jonsdottir I, Marelsson SE, Gudjonsson SA, Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Stacey SN, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A, Halldorsson B V., Helgason A, Gudbjartsson DF, Stefansson K. 2018. Multiple transmissions of de novo mutations in families. *Nat Genet* **50**:1674–1680. doi:10.1038/s41588-018-0259-9

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir Adalbjorg, Jonasdottir Aslaug, Rafnar T, Frigge M, Stacey SN, Th. Magnusson O, Thorsteinsdottir U, Masson G, Kong A, Halldorsson B V., Helgason A, Gudbjartsson DF, Stefansson K. 2017. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**:519–522. doi:10.1038/nature24018

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

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**:996–1006. doi:10.1101/GR.229102

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

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

Knudson AG. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A* **68**:820–823. doi:10.1073/PNAS.68.4.820

Li H, Wren J. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**:2843–2851. doi:10.1093/BIOINFORMATICS/BTU356

Lindsay SJ, Rahbari R, Kaplanis J, Keane T, Hurles ME. 2019. Similarities and differences in patterns of germline mutation between mice and humans. *Nat Commun* **10**:1–12. doi:10.1038/S41467-019-12023-W

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

Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA, Oliver TRW, Leongamornlert D, Ellis P, Noorani A, Mitchell TJ, Butler TM, Hooks Y, Warren AY, Jorgensen M, Dawson KJ, Menzies A, O'Neill L, Latimer C, Teng M, van Boxtel R, Iacobuzio-Donahue CA, Martincorena I, Heer R, Campbell PJ, Fitzgerald RC, Stratton MR, Rahbari R. 2021. The mutational landscape of human somatic and germline cells. *Nature* **597**:381–386. doi:10.1038/s41586-021-03822-7

Mora L, Sánchez I, Garcia M, Ponsà M. 2006. Chromosome territory positioning of conserved homologous chromosomes in different primate species. *Chromosoma* **115**:367–375. doi:10.1007/S00412-006-0064-6/TABLES/2

Moynahan ME, Jasin M. 2010. Mitotic homologous recombination maintains genomic stability and

suppresses tumorigenesis. *Nat Rev Mol Cell Biol* **11**:196–207. doi:10.1038/nrm2851

Moynahan ME, Jasin M. 1997. Loss of heterozygosity induced by a chromosomal double-strand break. *Proc Natl Acad Sci U S A* **94**:8988–8993. doi:10.1073/PNAS.94.17.8988/ASSET/EA5AD652-9DDA-4830-8593-1B2DF46AA905/ASSETS/GRAPHIC/PQ1771018003.JPEG

Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Al Turki S, Dominiczak A, Morris A, Porteous D, Smith B, Stratton MR, Hurles ME. 2016. Timing, rates and spectra of human germline mutation. *Nat Genet* **48**:126–133. doi:10.1038/ng.3469

Read LR, Raynard SJ, Rukść A, Baker MD. 2004. Gene repeat expansion and contraction by spontaneous intrachromosomal homologous recombination in mammalian cells. *Nucleic Acids Res* **32**:1184–1196. doi:10.1093/NAR/GKH280

Ryland GL, Doyle MA, Goode D, Boyle SE, Choong DYH, Rowley SM, Li J, Bowtell DD, Tothill RW, Campbell IG, Gorringe KL. 2015. Loss of heterozygosity: What is it good for? *BMC Med Genomics* **8**:1–12. doi:10.1186/s12920-015-0123-z

Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. 2019. Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**:e46922. doi:10.1101/552117

Sun H Bin, Shen J, Yokota H. 2000. Size-dependent positioning of human chromosomes in interphase

nuclei. *Biophys J* **79**:184–190. doi:10.1016/S0006-3495(00)76282-5

Tlemsani C, Takahashi N, Pongor L, Rajapakse VN, Tyagi M, Wen X, Fasaye GA, Schmidt KT, Desai P, Kim C, Rajan A, Swift S, Sciuto L, Vilimas R, Webb S, Nichols S, Figg WD, Pommier Y, Calzone K, Steinberg SM, Wei JS, Guha U, Turner CE, Khan J, Thomas A. 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. doi:10.1126/SCITRANSLMED.ABC7488/SUPPL_FILE/ABC7488_SM.PDF

Tsaousis GN, Papadopoulou E, Apessos A, Agiannitopoulos K, Pepe G, Kampouri S, Diamantopoulos N, Floros T, Iosifidou R, Katopodi O, Koumarianou A, Markopoulos C, Papazisis K, Venizelos V, Xanthakis I, Xepapadakis G, Banu E, Eniu DT, Negru S, Stanculeanu DL, Ungureanu A, Ozmen V, Tansan S, Tekinel M, Yalcin S, Nasioulas G. 2019. Analysis of hereditary cancer syndromes by using a panel of genes: Novel and multiple pathogenic mutations. *BMC Cancer* **19**:1–19. doi:10.1186/S12885-019-5756-4/FIGURES/6

Turner TN, Coe BP, Dickel DE, Hoekzema K, Nelson BJ, Zody MC, Kronenberg ZN, Hormozdiari F, Raja A, Pennacchio LA, Darnell RB, Eichler EE. 2017. Genomic Patterns of De Novo Mutation in Simplex Autism. *Cell* **171**:710–722. doi:10.1016/J.CELL.2017.08.047

Vermezovic J, Stergiou L, Hengartner MO, 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. doi:10.1038/cdd.2012.69

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

Watts FZ. 2016. Repair of DNA Double-Strand Breaks in Heterochromatin. *Biomolecules* **6**:47. doi:10.3390/BIOM6040047

Author information

Hyeonjeong Kim

Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University

Maidashi 3-1-1, Higashi-ku, Fukuoka, 812-8582, Japan

Contribution: Data curation, Software, Formal analysis, Investigation, Visualization, Methodology, Writing—original draft

Competing interests: No competing interests declared

Mikita Suyama

Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University

Maidashi 3-1-1, Higashi-ku, Fukuoka, 812-8582, Japan

Contribution: Conceptualization, Supervision, Project administration, Writing—review and editing

Competing interests: No competing interests declared

Figure legends

Figure 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 Figure 1–figure supplement 1. 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 Figure 1–figure supplement 2. 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.

Figure 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 Figure 2–figure supplement 1. (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.

Figure 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.

Figure 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.

Figure Supplement legends

Figure 1-figure supplement 1

Example of division of a large family. Immediate families used in this study were selected based on the married couples in the second-generation in a large family. (A) Family x separates into family x_1 (box in red) and family x_2 (box in cyan) by the married couples in the second-generation (box in yellow). (B) Family y separates into family y_1 (box in red) and family y_2 (box in cyan) by the married couples in the second-generation (box in cyan). However, the family y_2 would be excluded in this study because of the lack of data on member (slashed individual).

Figure 1–figure supplement 2

Schematic concept of minimum and maximum size of LOH. The minimum size represents the distance from the first L-SNV to the last L-SNV in an LOH event. The maximum size represents the distance between two nucleotides just before the first heterozygous variant upstream and downstream from an LOH event. (A) This figure shows the case of one L-SNV between heterozygous variants. (B) This figure shows the case of two or more L-SNVs between heterozygous variants. Het, heterozygous variant

Figure 2–figure supplement 1

Relationship between LOH and maternal age. Scatterplot between the number of LOH events and the maternal age (first-generation females).

Figures and data

7 figures and 1 additional file

Figure 1

Figure 1–figure supplement 1

Figure 1–figure supplement 2

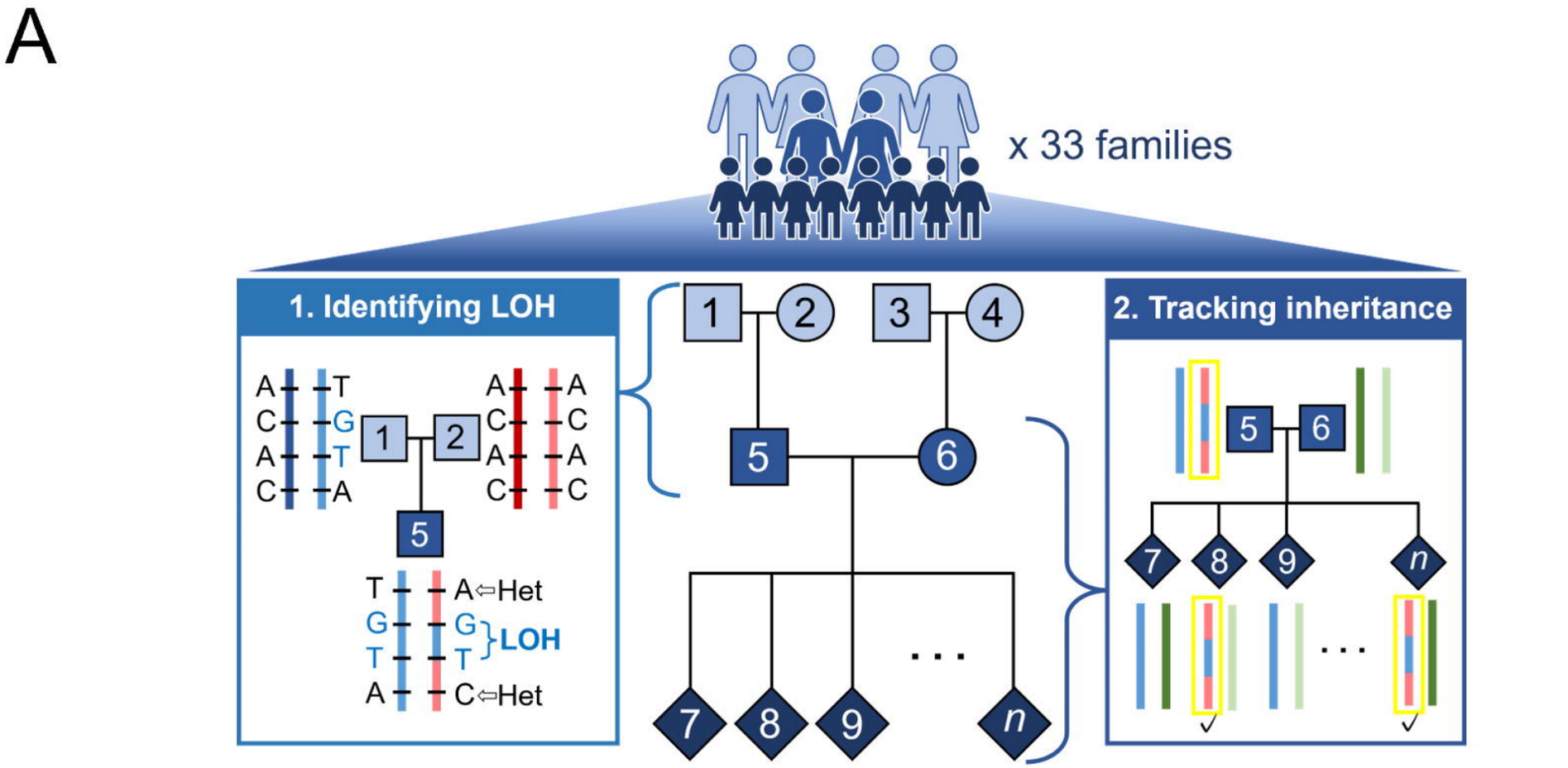
Figure 2

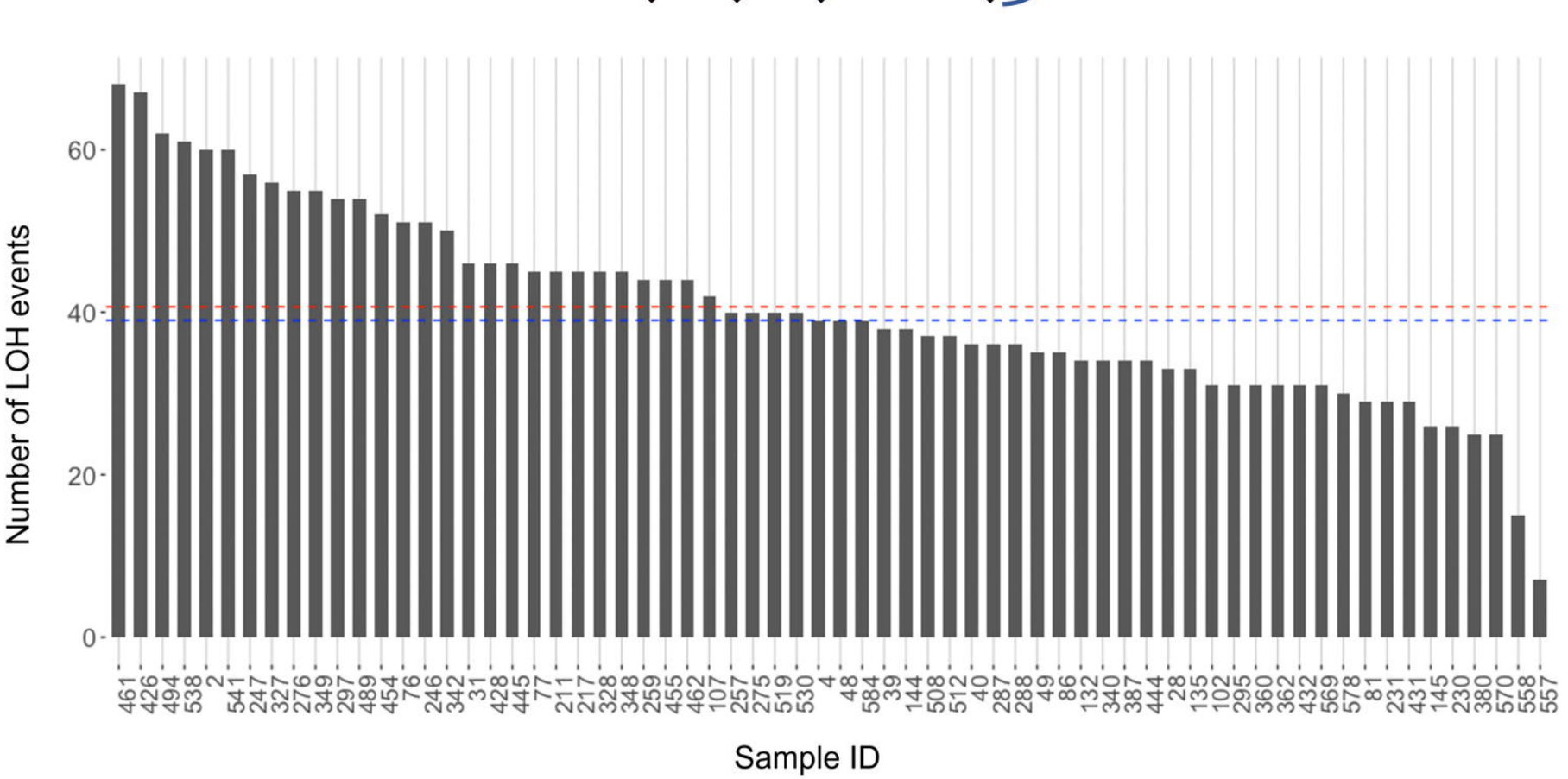
Figure 2–figure supplement 1

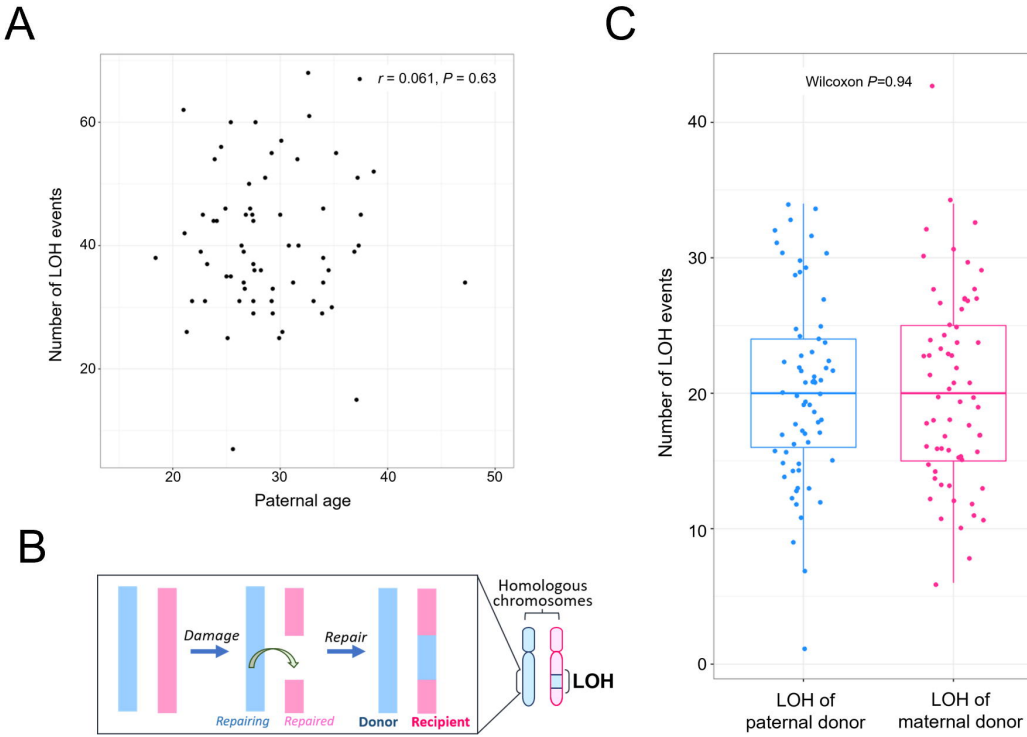
Figure 3

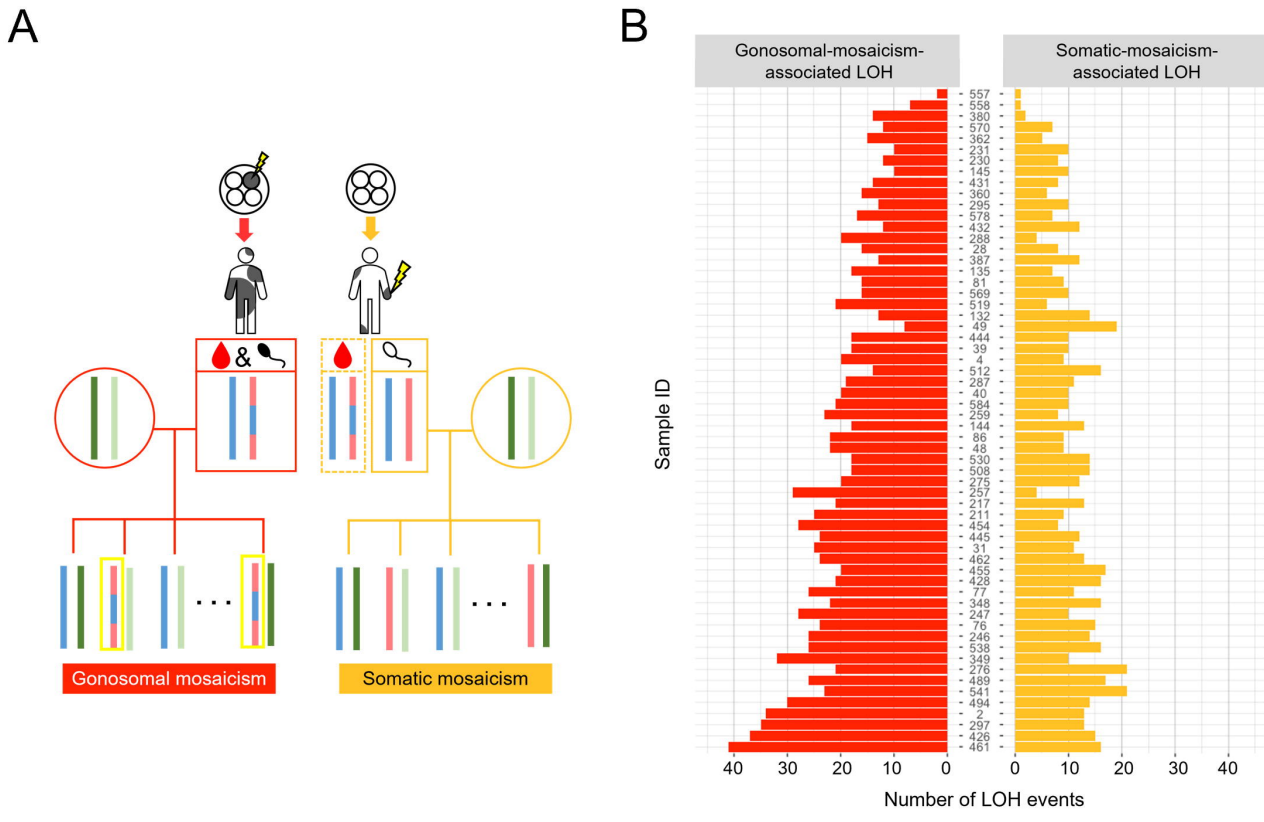
Figure 4

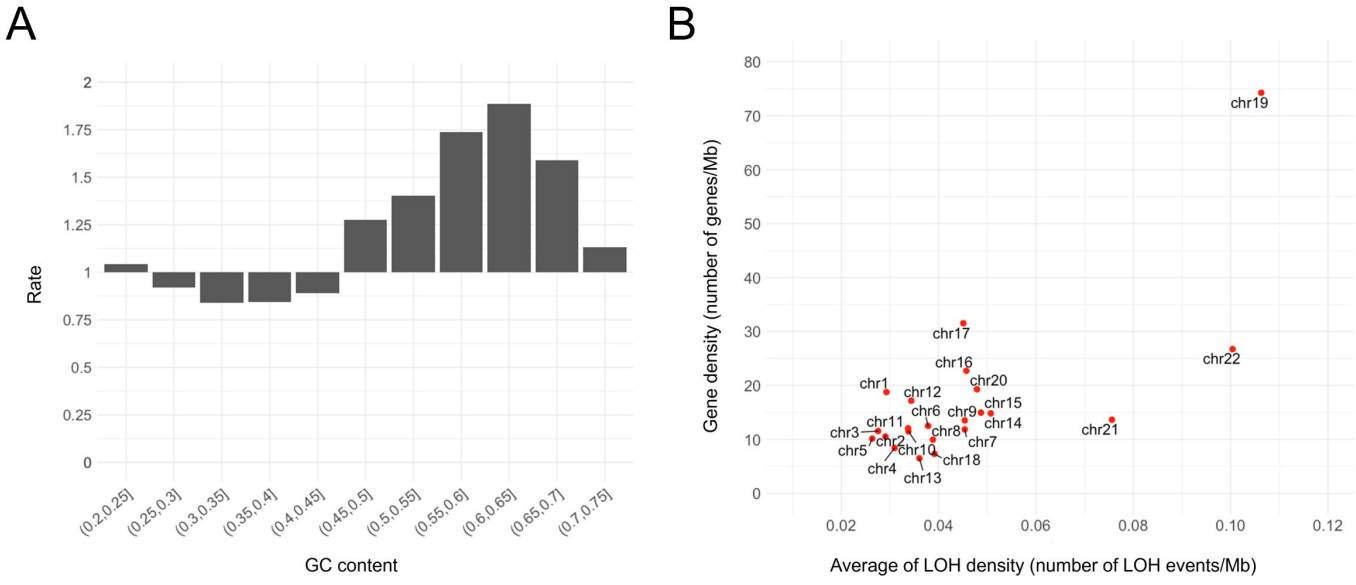
Supplementary File 1


A
x 33 families
1. Identifying 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
60
40
20
0
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
B
Damage
Repair
Repairing
Repaired
Donor
Recipient
Homologous chromosomes
LOH
C
Wilcoxon P=0.94
Number of LOH events
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
chr16
chr22
chr1
chr12
chr20
chr15
chr6
chr9
chr11
chr8
chr14
chr21
chr3
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)