
















Differences between germline genomes of monozygotic twins

Hakon Jonsson¹  , Erna Magnusdottir² , Hannes P. Eggertsson¹ , Olafur A. Stefansson¹, Gudny A. Arnadottir¹ , Ogmundur Eiriksson¹, Florian Zink¹, Einar A. Helgason¹, Ingileif Jonsdottir¹ , Arnaldur Gylfason¹, Adalbjorg Jonasdottir¹, Aslaug Jonasdottir¹, Doruk Beyter¹, Thora Steingrimsdottir², Gudmundur L. Norddahl¹, Olafur Th. Magnusson¹, Gisli Masson¹, Bjarni V. Halldorsson^{1,3} , Unnur Thorsteinsdottir^{1,2}, Agnar Helgason^{1,4} , Patrick Sulem¹ , Daniel F. Gudbjartsson^{1,5}   and Kari Stefansson^{1,2}  

Despite the important role that monozygotic twins have played in genetics research, little is known about their genomic differences. Here we show that monozygotic twins differ on average by 5.2 early developmental mutations and that approximately 15% of monozygotic twins have a substantial number of these early developmental mutations specific to one of them. Using the parents and offspring of twins, we identified pre-twinning mutations. We observed instances where a twin was formed from a single cell lineage in the pre-twinning cell mass and instances where a twin was formed from several cell lineages. CpG>TpG mutations increased in frequency with embryonic development, coinciding with an increase in DNA methylation. Our results indicate that allocations of cells during development shapes genomic differences between monozygotic twins.

A common assumption is that the sequences of the genomes of monozygotic twins are almost identical¹. However, there is a paucity of studies characterizing genomic differences between these twins^{1–5}. The average number of differences between the genomes of monozygotic twins is not known. Furthermore, the types of mutations leading to these differences and their timing are unknown. When DNA isolated from the blood of monozygotic twins is sequenced and compared, some of the differences seen may be due to somatic mutations in blood cells or their precursors. Such mutations are more likely to be in a detectable quantity with the increasing age of the twins due to clonal hematopoiesis⁶ (Fig. 1).

To track mutations that separate monozygotic twins, it is important to take advantage of what we know about the earliest stages of human development. During the first week, the zygote divides several times to form a mass of approximately 16 cells called the morula, which is contained within a glycoprotein shell called the zona pellucida⁷. At the end of the first week, the embryo hatches from the zona pellucida, implants into the uterine lining and forms the blastocyst, a fluid-filled cyst with a lining of cells that covers a portion of its inner wall. These cells are termed the inner cell mass and give rise to the individual or two individuals in the case of identical twins^{7–9}. At 1–2 weeks after blastocyst formation, a set of cells in the embryo are slated to become germ cells (primordial germ cell specification (PGCS))¹⁰.

We set out to time the mutations that separate monozygotic twins, for example, mutations specific to one twin that must have occurred after the initial formation of the zygote. These postzygotic mutations accumulate from early development throughout life^{11,12}. To refine the timing of postzygotic mutations, we determined whether or not these mutations were transmitted to the offspring of

the twins^{13–18}. Postzygotic mutations present in both the germ and somatic cells of twins most likely occurred during early development or, more specifically, before PGCS (Fig. 1). The presence of transmitted mutations in the somatic tissues of the transmitting parent have been used to detect and time postzygotic mutations^{13–19}. However, these approaches have limited power when a postzygotic mutation is present at a high frequency.

To estimate the number and timing of mutations differing between monozygotic twins, we searched for postzygotic mutations present in the somatic tissue of one of the twins but not the other and timed them by comparing whole-genome sequencing (WGS) data from monozygotic twins, their offspring, spouses and parents. In addition, to allow us to probe the differences between monozygotic twins, this approach provides some insights into the earliest events during embryonic development. These early developmental mutations allowed us to characterize the fate of mutated cells and their descendants during early development and demonstrate the stochastic component of cell allocation during the earliest phases of human development. The sharing of early developmental mutations by twins allowed us to divide twin pairs into two groups, one where both twins were formed from the same cell lineages of the pre-twinning cell population and the other where they were not. Primarily, this analysis allowed us to determine the number of mutations that separate monozygotic twins, their type and the timing of their occurrence.

Results

Genomic differences between monozygotic twins. We first estimated the number of discordant postzygotic mutations in pairs of monozygotic twins (381 twin pairs; 2 triplets) by comparing sequence variation in somatic tissues (1 adipose, 204 buccal and

¹deCODE genetics/Amgen, Reykjavik, Iceland. ²Faculty of Medicine, School of Health Sciences, University of Iceland, Reykjavik, Iceland. ³School of Science and Engineering, Reykjavik University, Reykjavik, Iceland. ⁴Department of Anthropology, University of Iceland, Reykjavik, Iceland. ⁵School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. ✉e-mail: hakon.jonsson@decode.is; daniel.gudbjartsson@decode.is; kari.stefansson@decode.is

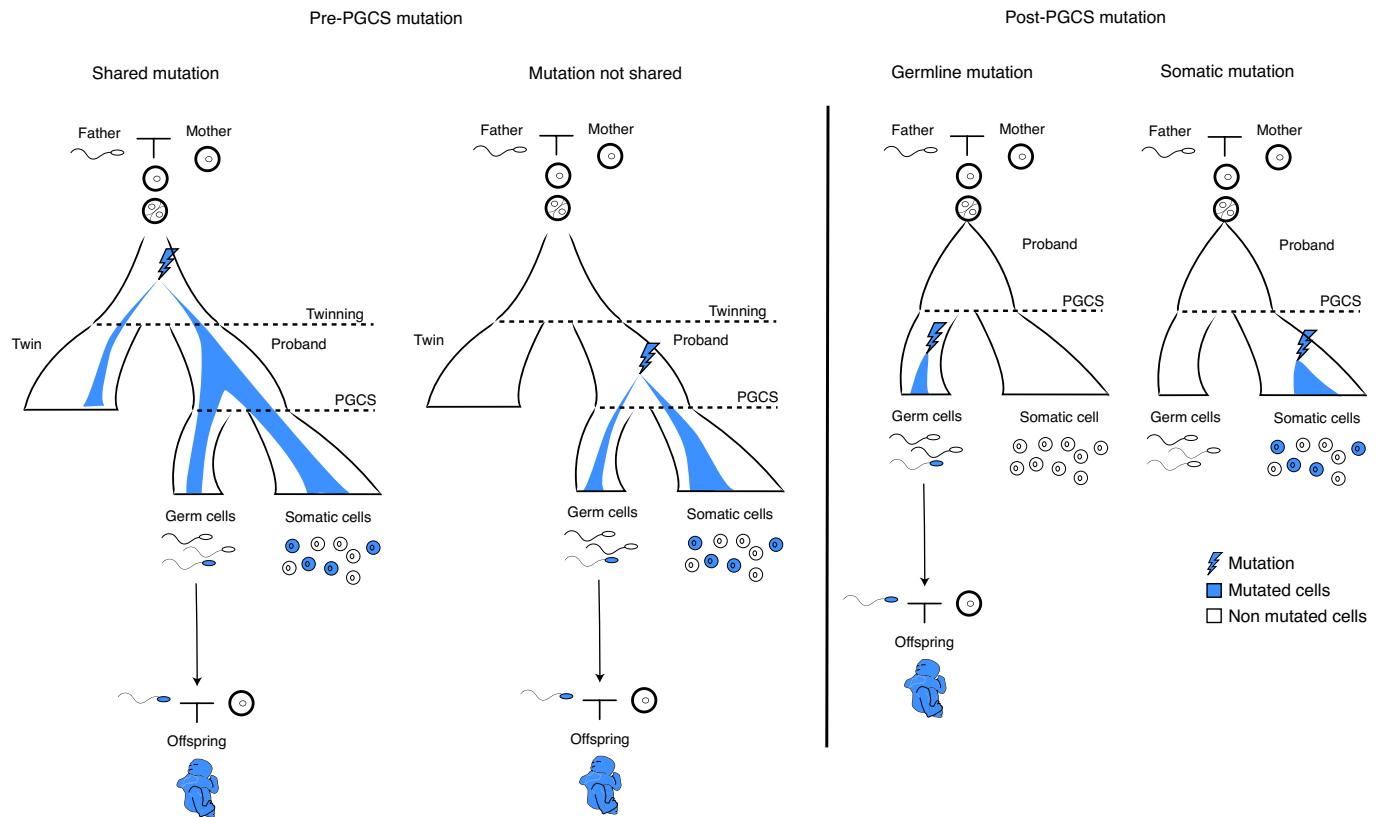


Fig. 1 | The timing of postzygotic mutations. Postzygotic mutations can be classified as pre-PGCS mutations if they are present in both the soma and germline of the proband. If they are present only in the soma or the germline, we classified them as post-PGCS mutations. Note that a pre-PGCS mutation could be misclassified as a post-PGCS mutation if the mutated cells were not detected in the somatic tissue or the germline. Pre-PGCS mutations were classified according to whether they were present in both twins or only one. Post-PGCS mutations are classified according to whether they are present in the germline or the soma of the proband. Discussion and detailed definitions of postzygotic mutations are in the Supplementary Note. Image of offspring adapted from Kevin Dufendach, MD (2008) under a Creative Commons [CC BY 3.0](https://creativecommons.org/licenses/by/3.0/) license.

563 blood samples). To estimate the somatic variant allele frequency (VAF), that is, the fraction of reads supporting the alternative allele, of postzygotic mutations accurately we sequenced a subset of individuals in the twin pairs (239 individuals out of 768) to an average coverage of 152 \times (Extended Data Figs. 1 and 2) and the remaining samples to an average of 38 \times . We found a total of 23,653 postzygotic mutations that were specific to one twin, with a median of 14 postzygotic mutations differing between a pair of twins (median of 48 for high-coverage pairs). We estimated the false positive rate of the postzygotic mutations to be 16% (95% confidence interval (CI) = 14–19%) by read-tracing mutations to nearby heterozygous germline variants. Postzygotic mutations with a VAF > 45% were of higher quality, with a false positive rate of 3% (95% CI = 1–8%). Furthermore, we randomly selected 46 mutations for targeted resequencing; 43 had sufficient coverage and of these 31 were validated, resulting in a false positive rate of 28% (95% CI = 15–44%).

There was considerable variability in the number of postzygotic mutations; for example, 39 twin pairs differed by more than 100 mutations, whereas 38 pairs did not differ at all (5 and 12 twin pairs, respectively, when restricting our analysis to high-coverage samples, over 100 \times average coverage). Furthermore, mutations in individuals with 10 or more mutations had a lower VAF (median VAF = 9.4%) than in individuals with fewer than 10 mutations (median VAF = 19.8%; Wilcoxon signed-rank test, $P = 7.8 \times 10^{-5}$ for samples with high coverage). This extensive accumulation of mutations was more common in blood (median count of 18) than in buccal samples (median count of 6; Wilcoxon signed-rank test, $P = 3.7 \times 10^{-4}$ for samples with high coverage). This indicates that

a considerable part of the difference in the number of postzygotic mutations between monozygotic twins is due to clonal hematopoiesis. Interestingly, at the high end of the VAF spectrum, we found a population of twin pairs where postzygotic mutations were nearly constitutional (VAF > 45%) in one of the monozygotic twins (105 out of 768 individuals; Fig. 2a). This indicates that for these individuals the somatic tissue is made up of a single cell lineage defined by postzygotic mutations. Notably, the low number of mutations (median of 3 mutations) in these individuals indicates that the near-constitutional cell lineages were formed within the first few cell divisions after the zygote was formed.

Postzygotic mutations accumulate throughout the life of the individual from early development to the time when the somatic sample was donated. To time the occurrence of mutations, we exploited the fact that the accumulation of postnatal mutations is generally a function of the age of the individual whereas mutations occurring during early development are not. We found that the number of postzygotic mutations increased with the age of the individual (Fig. 2b). However, when we restrict the analysis to near-constitutional mutations, the effect of age is not seen (Fig. 2c). This indicates that the near-constitutional postzygotic mutations separating pairs of twins must have taken place early during development. Interestingly, if one of monozygotic twins carried a near-constitutional mutation, the other twin carried a different near-constitutional mutation in 42% of cases (Fisher's exact test, $P = 1.6 \times 10^{-11}$). This suggests that postzygotic mutations in one twin can be informative regarding the development of both twins during early development.

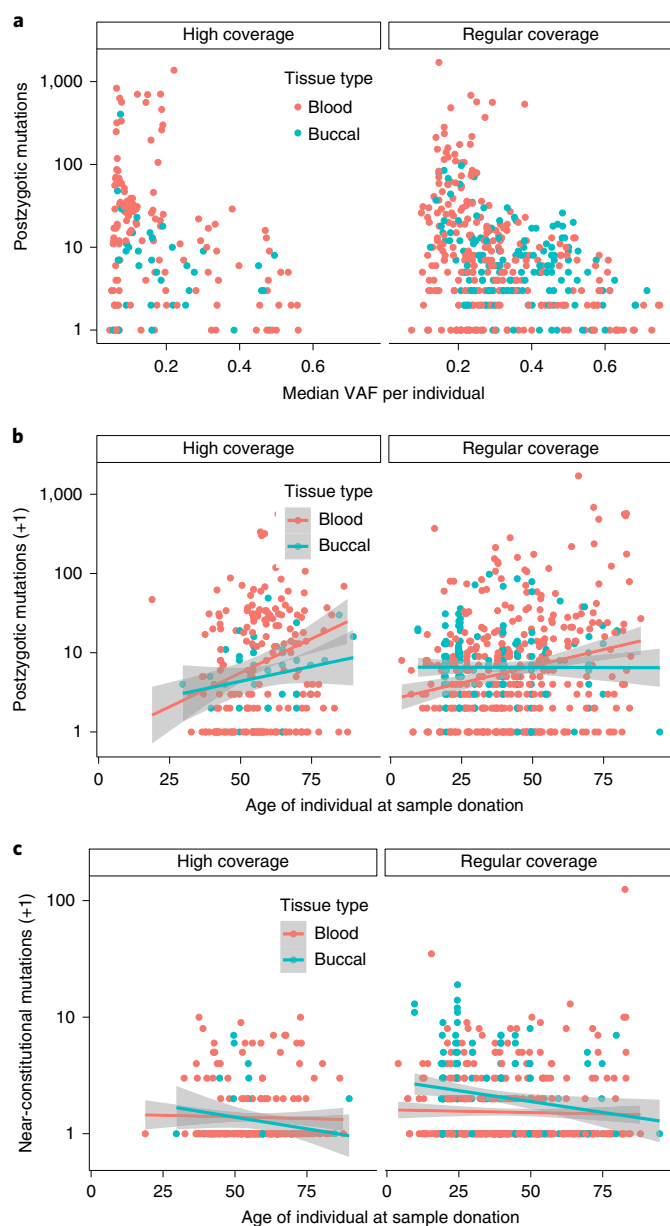


Fig. 2 | Number of postzygotic mutations per individual. **a**, Number of postzygotic mutations as a function of median VAF per individual. **b**, Number of postzygotic mutations as a function of the age of an individual when the somatic sample was donated. **c**, Same as **b** except that mutation counts are restricted to near-constitutional mutations. This analysis was restricted to blood and buccal samples with sampling dates, resulting in 766 individuals from 380 biologically independent monozygotic twin pairs and 2 monozygotic triplets. **b,c**, The lines are regression lines and the error bands are the 95% CIs.

Early developmental mutations and monozygotic twins. In this section, we define a proband as an individual who has both an offspring and a monozygotic twin (Fig. 3) and pre-PGCS mutations as postzygotic mutations detected in both soma and germline of a proband, the latter evidenced by transmission to an offspring. To find the pre-PGCS mutations, we also sequenced the genomes of family members (offspring and spouses/partners) of 181 monozygotic twin pairs, resulting in a total of 451 quadruplets (246 probands) consisting of a proband, a monozygotic twin, a spouse/partner and an offspring (Supplementary Tables 3 and 4, Fig. 3d

and Extended Data Fig. 2). Specifically, we looked for mutations present in the proband's offspring but not in the proband's twin or spouse/partner. These are pre-PGCS mutations in the proband and post-PGCS mutations transmitted to the offspring from the proband and spouse/partner. We found that 27,265 mutations absent from the proband's twin and spouse/partner were transmitted to the proband's offspring (60.5 per offspring). Of these, 582 mutations were found in the somatic tissue of the proband and are therefore most likely pre-PGCS mutations (Fig. 4a,b); on average, 1.3 (95% CI=1.1–1.5) pre-PGCS mutations were transmitted from a proband to each offspring. The 2.1% fraction (1.3 out of 60.5) of pre-PGCS mutations transmitted to the offspring was comparable to an imputed estimate of 2.3% using a recent three-generation dataset, assuming that the somatic VAF in the proband was proportional to the transmission rate ('Comparison to the Sasani et al. data' in the Supplementary Note)¹³. The transmission rates of pre-PGCS mutations from probands were independent of whether DNA was extracted from blood (1.3; 95% CI=1.1–1.5) or buccal swabs (1.2; 95% CI=0.9–1.6). Both male and female probands transmitted 1.3 pre-PGCS mutations (95% CI=1.0–1.5 and 95% CI=1.1–1.6, respectively); hence, each sex contributed the same number of pre-PGCS mutations as reported previously by us¹⁷ and others^{13,15}. Furthermore, these pre-PGCS mutations were distributed equally between the paternal (272) and maternal (310) chromosomes of the probands (binomial test, $P=0.12$) as reported by others¹³.

Since only half of the proband's genome is transmitted to the offspring, we can assume that in the diploid genome there are 2.6 pre-PGCS mutations (twice as many as transmitted ones). The same applies to the proband's twin, resulting in a difference of 5.2 (95% CI=4.4–6.0) pre-PGCS mutations between twins. There is considerable variability in the number of pre-PGCS mutations that separate twins (Fig. 4b), ranging from no transmission of pre-PGCS mutations to 207 offspring, to transmission of 8 pre-PGCS mutations to 3 offspring.

To determine more accurately when during development pre-PGCS mutations that differ between twins arose, we determined their somatic VAFs in the probands under the premise that the higher the pre-PGCS VAF the earlier during development these occurred. Mutations specific to one twin that occur after twinning are generally not constitutional. However, we found that 64 offspring of 36 probands out of 246 (15%) carried a pre-PGCS mutation that was nearly constitutional in the proband (Fig. 4c and Extended Data Figs. 3 and 4). Furthermore, on average these 36 probands transmitted 3.5 pre-PGCS mutations (95% CI=3.1–3.9) to each offspring, compared to 0.9 such transmissions by probands (95% CI=0.8–1.1) lacking a near-constitutional pre-PGCS mutation (block jackknife, $P=1.1 \times 10^{-30}$). In other words, 15% of our probands carried a substantial number of near-constitutional pre-PGCS mutations that were absent from their twins. Since mutations accumulate as cells divide, the high number of pre-PGCS mutations in the offspring of the probands with near-constitutional pre-PGCS mutations suggests that those probands are derived from a single cell from the inner cell mass that accumulated mutations before diverging from the pre-twinning cell population.

The pre-PGCS mutations with high VAF that differed between twins likely occurred during the initial developmental stages, perhaps even before twinning occurred. To determine whether we could find pre-twinning mutations present in both twins (Fig. 3b), we restricted our analysis to a subset of probands with both parents sequenced and mutations present in a proband's offspring but absent from their parents and spouse/partner, resulting in 63 twin pairs in 92 three-generation families (Fig. 5a). We found 187 mutations present in mosaic form in probands with <25% VAF or significant VAF difference between twins ($P<0.001$). Of these pre-PGCS mutations, 112 were present in both twins (shared mutations; Fig. 5b). Out of the 63 twin pairs, 36 (57%) shared at least one

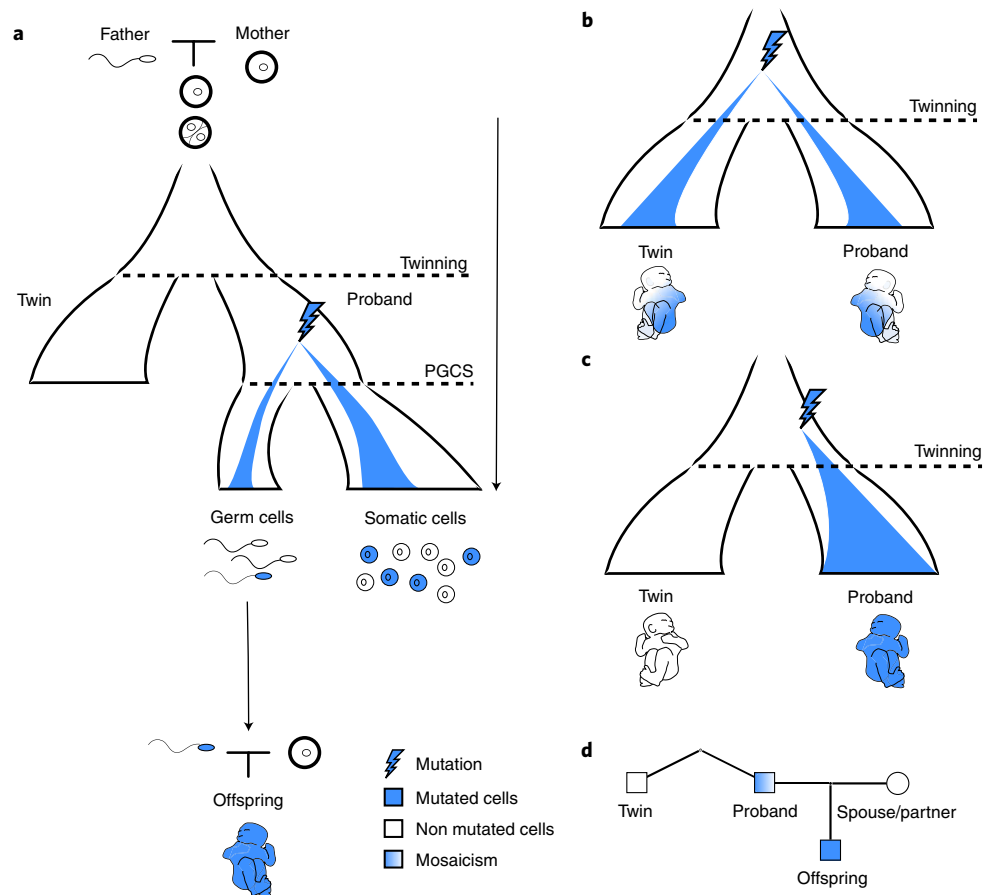


Fig. 3 | Timing of pre-PGCS and pre-twinning mutations in twins. **a**, Schematic overview of a pre-PGCS mutation that occurred after twinning resulting in mosaicism in the soma and germline of the proband but absent from the monozygotic twin. This pre-PGCS mutation was subsequently transmitted to the proband's offspring; thus, the mutation will be in all the cells of the offspring. **b,c**, Pre-PGCS mutations that occurred pre-twinning and contribution of the mutated cell lineage to either both twins (**b**) or one twin (**c**). **d**, Monozygotic twin approach. Extraction of pre-PGCS mutations using the spouse/partner and monozygotic twin of the proband. Note that the mutation can be mosaic in the proband. Image of offspring adapted from Kevin Dufendach, MD (2008) under a Creative Commons CC BY 3.0 license.

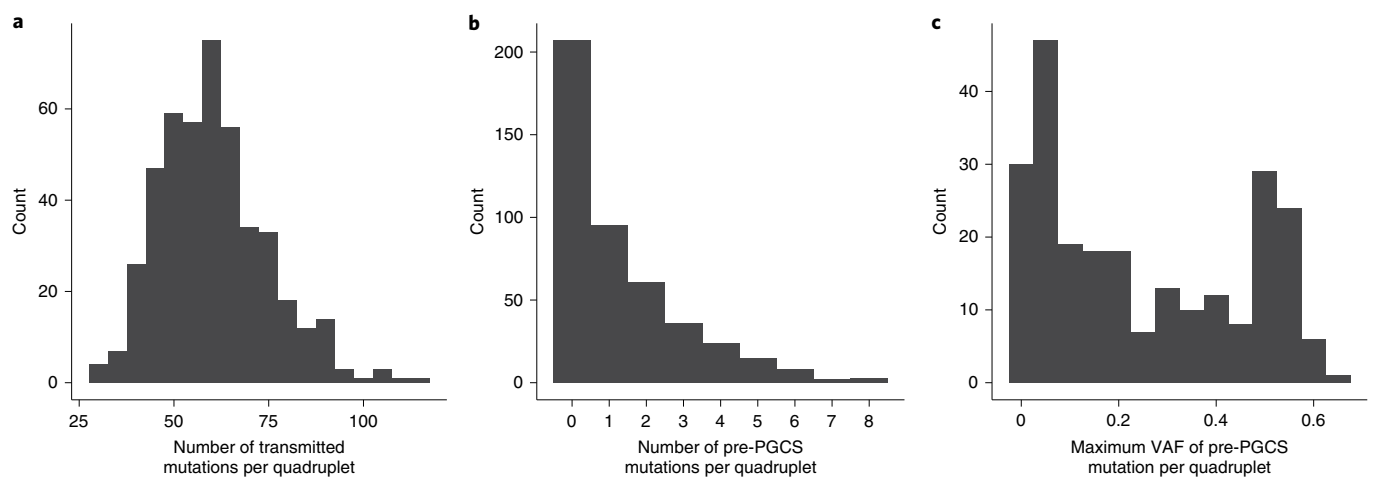


Fig. 4 | Number of mutations transmitted to the offspring and VAF of pre-PGCS mutations. **a**, Number of mutations per quadruplet. **b**, Number of pre-PGCS mutations per quadruplet. **c**, Maximum frequency of pre-PGCS mutations in the proband per quadruplet. To account for the variable number of offspring per proband, we considered the maximum per quadruplet. See Extended Data Fig. 4 for the maximum VAF per proband/spouse pair.

mutation. For 14 twin pairs where both twins had offspring and spouses/partners in the study, as conditioned on a transmission of a pre-PGCS mutation to an offspring of the proband and the

somatic VAF of the mutation in the monozygotic twin, we would expect transmission of 21 pre-PGCS mutations to the offspring of the twins. We identified 21 such transmissions, indicating

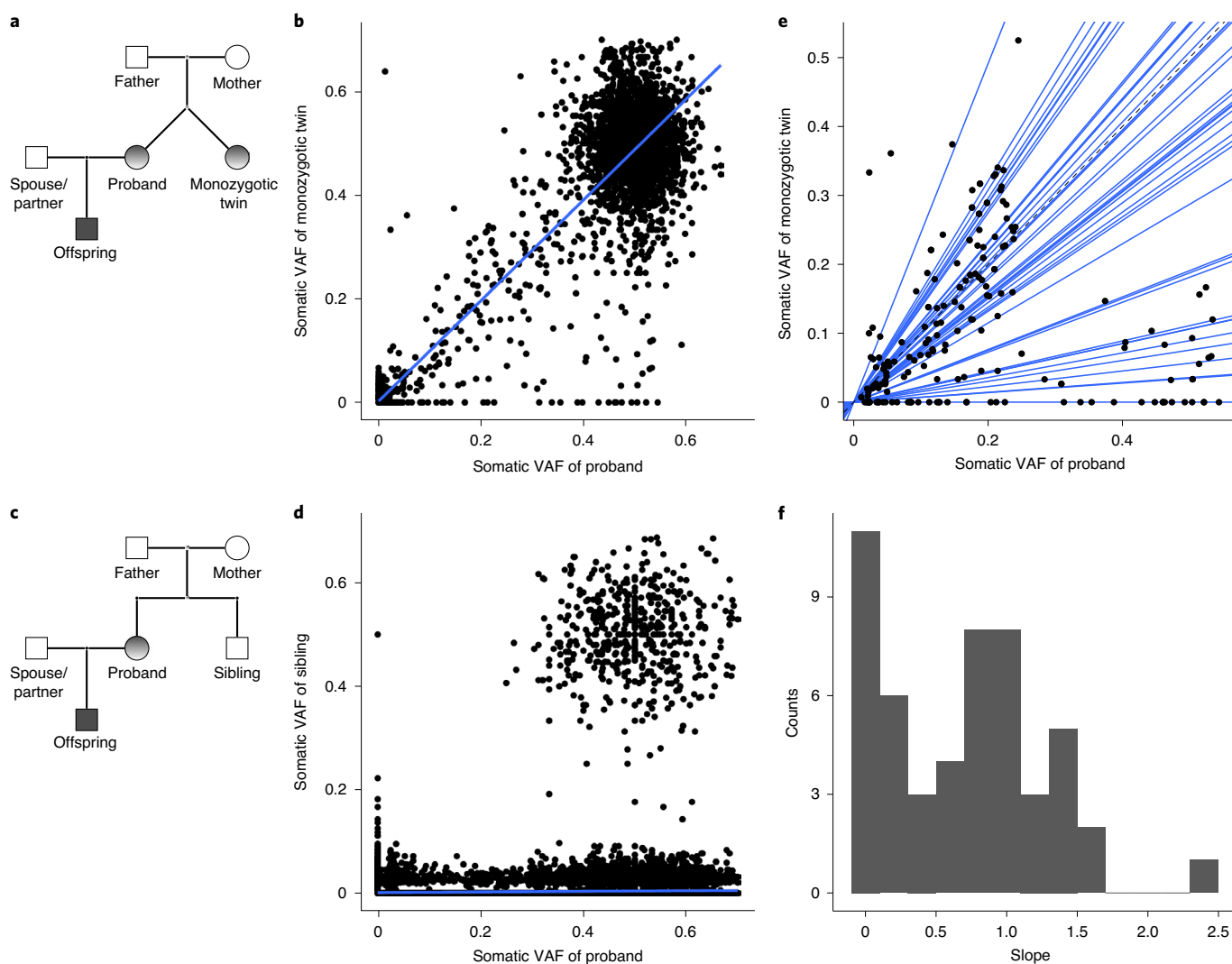


Fig. 5 | The three-generation approach. **a**, Extraction of pre-PGCS mutations using the spouse/partner, father and mother of the proband. Note that the mutation can be mosaic in the proband and monozygotic twin of the proband (<25% VAF or significant VAF difference between twins, $P < 0.001$). **b**, Somatic VAF of the proband and monozygotic twin. **c**, Same as **a**, except a nontwin sibling was used for comparison. **d**, Somatic VAF of the proband and sibling. Note that the points in 0.5, 0.5 consist of mutations that are shared by siblings but absent from their parents, indicating germline mosaicism or misgenotyped noncarrier calls for the parents. **e**, The somatic VAF of the proband and the monozygotic twin, restricting to pre-PGCS mutations. **b, e**, The blue lines are regression lines. The lines are derived from the regression per proband and spouse/partner pair; the intercept is fixed to the origin. **f**, Histogram of the slopes depicted in **e**.

that these mutations are truly pre-twinning mutations rather than somatic chimerism. Note that in this subset both twins were considered probands. For quality control, we performed the same analysis using 1,395 sibling pairs. As expected, we did not detect a population of shared mutations in these pairs because siblings were rooted in different zygotes (Fig. 5c,d).

We assessed the VAF relationship of pre-PGCS mutations in each twin pair by using regression analysis. The distribution of regression slopes suggests that there is considerable range in VAF dependency between twin pairs (Fig. 5e,f; see Extended Data Fig. 5 for the reciprocal regressions). At one extreme, the somatic VAFs of mutations are similar in the twins (regression slope > 0.5 , Fig. 5e,f; 31 out of 51 twin pairs with a pre-PGCS mutation). At the other extreme, mutations are present in probands at a high VAF and absent from their twins or with substantially lower VAF (regression slope < 0.5). Twin pairs with similar somatic VAFs are likely to have developed from the same cell lineages of the pre-twinning cell population. In contrast, probands with a nearly constitutional mutation that is absent

from the monozygotic sibling are perhaps the consequence of the expansion of a single clone from the pre-twinning cell population rooted in a cell where the mutation occurred, whereas their twin was formed from other cells. Another plausible explanation is drastic cell death before or after twinning, which could have reduced the cellular diversity of the proband to a single clone.

Mutation classes during early embryonic development. To compare the mutational processes operating in early development to those specific to germ cells, we tabulated the number of pre-PGCS and post-PGCS mutations in each mutation class. We used the mutations discovered by comparing the genomes of children and their parents (trio approach) as a surrogate for post-PGCS mutations. We analyzed the set of 705 pre-PGCS mutations identified in the trio and three-generation approach and compared it to a set of 321,106 mutations from 5,515 trios where family members did not include a monozygotic twin. Note that pre-PGCS mutations with no read support for the alternative allele in the transmitting parent will be

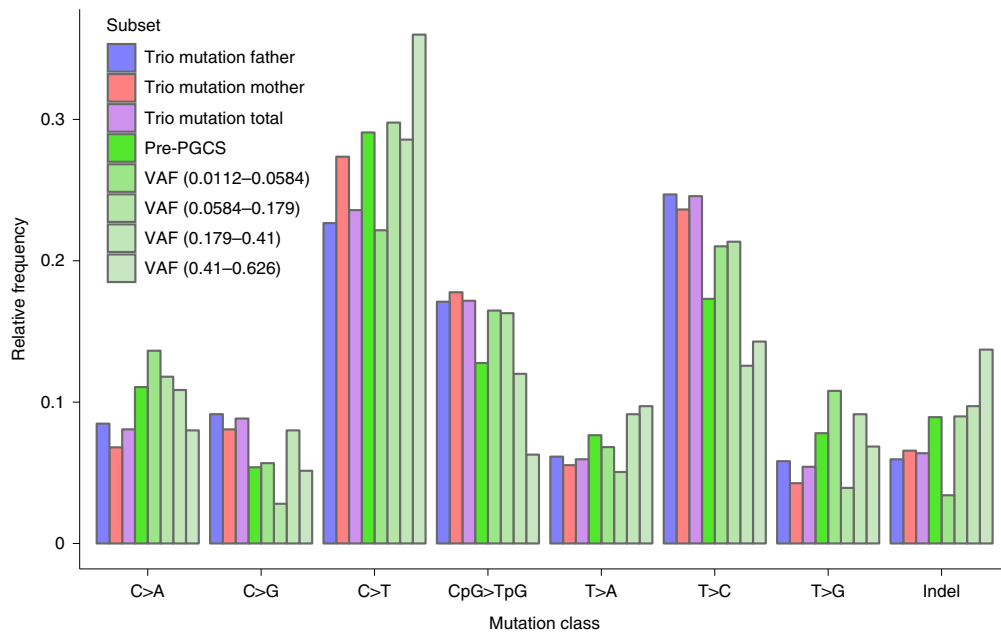


Fig. 6 | Mutation classes of pre-PGCS versus trio mutations. All trio mutations were considered; then, we restricted our analysis to mutations determined to be of paternal and maternal origin. We show the mutation class spectrum of pre-PGCS mutations falling within a VAF interval with different shades of green. The C>T mutation class does not include C>T mutations in a CpG context.

misclassified as post-PGCS mutations. Furthermore, postzygotic mutations with high VAF occurring in the offspring that are part of a trio will be misclassified as transmitted mutations. Therefore, our approach provides a conservative assessment of mutational spectrum differences between pre-PGCS and post-PGCS mutations. We found that C>A, C>T (non-CpG context), T>G and indel mutations were more frequent among pre-PGCS mutations than trio mutations (Fig. 6, Supplementary Table 5 and contingency table test). We did not replicate a previous report of a greater number of T>A mutations among pre-PGCS mutations than trio mutations¹³. The fraction of C>T and indel mutations increased with a higher VAF in the proband while the fraction of C>G, CpG>TpG and T>C mutations decreased. The later a mutation occurs during embryonic development, the lower its VAF. Therefore, these results indicate that early embryonic development contributes more C>T and indel mutations compared to subsequent development (Fig. 6 and Supplementary Table 6; mutational class fraction regressed on VAF). To test whether enrichment of a single mutational class was driving these results, we calculated the odds ratios conditional on omitting one mutational class at a time (Supplementary Tables 7 and 8). The results were robust to the exclusion of a single mutational class. T>C mutations often show strand bias at transcribed genes in cancer²⁰ and germline mutational profiles^{21,22}; hence, they have been associated with damage missed by transcription-coupled DNA repair. The decrease in fraction of T>C mutations with higher VAF could indicate that transcription-coupled DNA repair is less active during early development; however, we had limited data to test for strand bias among pre-PGCS T>C mutations. CpG>TpG mutations are generally thought to be the consequence of deamination of methylated cytosines^{23,24}. Thus, the lower contribution of CpG>TpG mutations in the first divisions might reflect the overall low CpG methylation status of the genome during early development^{25–27}. This would suggest that the genomes of cells that give rise to the germline and soma are being gradually methylated before PGCS. Direct estimation of the methylation profile at these first stages is difficult given the experimental and ethical constraints. However, our results suggest that we can retrospectively infer global methylation status of these key developmental stages by mutational spectrum shifts in monozygotic twins.

Allocation of cells during early embryonic development.

Developmental mutations in monozygotic twins can be used to trace the allocation of cells throughout development. For example, mosaicism in a proband's twin is informative about the pre-twinning cell population. For the family portrayed in Fig. 7a, a near-constitutional mutation in the proband was present in 11.2% of the twin's cells (VAF = 5.6%, 95% CI = 2.5–32.5%). The VAF difference of this pre-twinning mutation and in general the diverse VAF relationships across twins suggest that there is considerable stochasticity in the allocation of cell lineages at twinning. We simulated cell proliferation in early development under different twinning scenarios to assess this (Supplementary Note and Supplementary Table 10). Several scenarios were compatible with the data; however, considerable sampling in the allocation of cell lineages is needed to create these differences in VAF. Furthermore, if probands with near-constitutional pre-PGCS mutations were the consequence of a drop in clonal diversity specific to the proband, such as a single cell splitting from a cell mass, then the twin would lack near-constitutional mutations (Supplementary Note). However, the observation of a near-constitutional pre-PGCS mutation in one twin predicts a near-constitutional pre-PGCS mutation in the other twin. This suggests a drop in cellular diversity shared by the twins. We observed this dependency in our simulations by varying twinning generation. However, we could not discern between late twinning of related cells and early twinning of distantly related cells.

To explore cell allocation during early development, we also used monozygotic triplets. We searched for pre-PGCS mutations in a single family consisting of monozygotic triplets and their offspring (not a part of the three-generation dataset described above). Two of the triplets shared 2 mutations at a VAF of 9–19% and 27–39%, whereas the third triplet did not (Fig. 7b). Furthermore, we found a single pre-PGCS mutation at a VAF of 39% in the third triplet that was absent from the other two. Thus, the cell lineages in two of the triplets were closer to each other than either was to that of the third triplet. Furthermore, their mutation sharing suggests that two of the triplets were formed from the descendants of the same cell whereas the third was formed from a different set of cells. Overall,

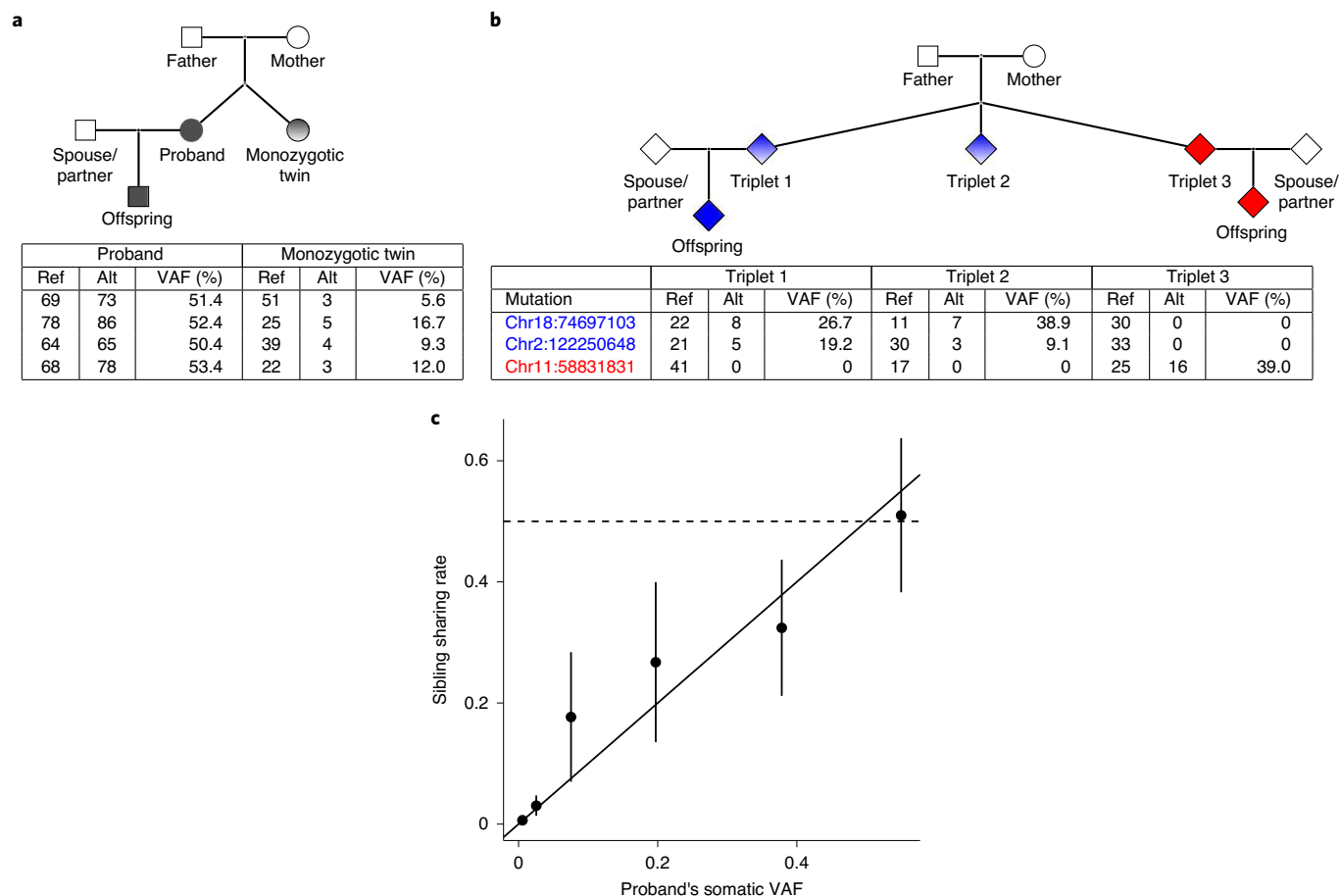


Fig. 7 | Cell allocation in early human development. **a**, Example of a family with pre-twinning mutations. **b**, Pre-PGCS mutations detected in a family of monozygotic triplets. We have omitted the sex of the monozygotic triplets and their offspring because of privacy concerns. The average genome-wide coverage of the monozygotic triplets was 31.6. **c**, Sibling sharing rate as a function of the proband's VAF. We binned the pre-PGCS mutations on the basis of the quintiles of the somatic VAF of the probands and added an extra bin for the post-PGCS mutations (VAF = 0%). The centers are the means. The error bars are the 95% CIs based on jackknifing over the proband and spouse/partner combinations (Methods). The unique proband and spouse/partner combinations in the VAF bins are 135, 130, 36, 26, 29 and 21.

our results demonstrate that there is a substantial stochastic component in cell allocation throughout early human development.

The data from the twins can also be used to evaluate the allocation of cells to the germline by contrasting transmission rates and somatic VAFs of pre-PGCS mutations. Previously, we showed that the rate of transmission of a mutation to more than one offspring is a function of its somatic VAF in the parent¹⁷. However, in our previous study we could not distinguish between pre-PGCS mutation with high somatic VAF and germline variants inherited from the proband's parents. In this study, we used the absence of a mutation from a monozygotic twin to determine that a mutation with a high somatic VAF occurred during the proband's development rather than being transmitted as a *de novo* mutation from a parent, even in the case of constitutional mutations. We found that the rate of a second transmission of a pre-PGCS mutation was proportional to the VAF of the mutation in the proband (Fig. 7c; slope = 0.95, 95% CI = 0.77–1.12; block jackknife, $P = 9.3 \times 10^{-27}$), which provides an insight into the recurrence risk of pediatric diseases caused by mutations. This relationship is robust to the removal of post-PGCS and near-constitutional mutations (Supplementary Table 9). This relationship shows that germline development is not dependent on a specific cell lineage, but rather it suggests that the same founder cells give rise to both germ and somatic cells.

Discussion

Phenotypic discordance between monozygotic twins has generally been attributed to the environment. This assumes that the contribution of mutations that separate monozygotic twins is negligible; however, for some diseases such as autism and other developmental disorders, a substantial component is due to *de novo* mutations³⁸. Our analysis demonstrates that in 15% of monozygotic twins a substantial number of mutations are specific to one twin but not the other. This discordance suggests that in most heritability models the contribution of sequence variation to the pathogenesis of diseases with an appreciable mutational component is underestimated.

Formation of more than one individual from the same zygote is a unique window into early embryonic development. The inner cell mass that gives rise to the embryo is formed at 4 d postfertilization concurrently with the trophoblast, the cell layer that gives rise to the chorion (the fetal part of the placenta). It has been suggested that the timing of monozygotic twinning may be deduced from the number of chorions and amniotic sacs⁷. According to this model, the trophoblast can be used to time the divergence of twin pairs, such that shared placentas (70–75% of twins) indicate twinning 4–7 d postfertilization, whereas separate placentas point to twinning 0–3 d postfertilization⁷. While this is the standard model, it is based on limited direct evidence^{8,9}. Unfortunately, the number of chorions and amniotic sacs was not documented for the twins

presented in this study; thus, we could not correlate the number of chorions with the pattern of sharing of mutations. If sharing of placentas dictated sharing of pre-PGCS mutations, then we would expect the fraction to be the same or similar. Fifty percent of twin pairs share a pre-PGCS mutation, which is less than the 70–75% fraction of monozygotic twins reported in epidemiological studies⁷; however, this is conditioned on observing a pre-PGCS mutation. Future studies could assess this by comparing the chorionicity of the twin pair with sharing of pre-PGCS mutations. The VAF range among the pre-PGCS mutations provides a unique view of twin development and suggests that there is a considerable range in the number of cells allocated to each twin during the twinning event. Models of early embryonic development should incorporate this stochastic sampling of early cell lineages.

Embryonic development can be thought of as a series of cell allocations to form different cell types. According to this line of thought, twinning is a consequence of allocation of totipotent cells. In a subset of twins, a single cell lineage appears to have given rise to the proband whereas the same cell lineage is absent from the proband's twin. An unequal contribution of cell lineages to the inner cell mass^{29,30} could create a VAF difference between cell populations contributing to twins. The VAF difference between twins could also be due to the death of cell lineages. In this study, we used sequencing of a single somatic sample per proband to detect pre-PGCS mutations, but we acknowledge that sequencing multiple tissues per proband would probably enable the detection of more pre-PGCS mutations¹⁸.

The accumulation of mutations is a function of the number of cell divisions, regardless of whether the mutation is induced by cell division or DNA damage before cell division. Probands with a nearly constitutional mutation transmitted 3.5 pre-PGCS mutations compared to 0.9 pre-PGCS mutations transmitted by those without a constitutional mutation. If the excess of 2.6 mutations is due to pre-twinning mutations, we estimate 5.2 more pre-twinning mutations in probands with a nearly constitutional mutation than in those without after accounting for the nontransmitted half of the proband's genome. This difference indicates that all cells from probands with nearly constitutional mutation derive from a single cell lineage formed after 5–6 mitoses from the initial zygote formation assuming roughly one mutation per mitosis. Formation of the cell lineage could be the twinning event, allocation of related cells, cell death or a combination thereof. The mutation rate could be variable in these first divisions³¹, thereby confounding the interpretation of the formation of the cell lineage. Whatever the underlying process, our results suggest that there is a considerable cellular diversity lost in the early development of these probands.

The proportional relationship between transmission rate and somatic VAF shows that allocation of cell lineages to the germline is mainly driven by their frequencies in the developing embryo. Furthermore, this indicates that the VAF of pre-PGCS mutations is similar across all human tissues and is mainly influenced by sampling variation. Therefore, our results are consistent with the notion that sampling of cell lineages in early human development is a major contributor to genomic differences between pairs of twins.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41588-020-00755-1>.

Received: 7 October 2019; Accepted: 20 November 2020;
Published online: 7 January 2021

References

1. Van Dongen, J., Slagboom, P. E., Draisma, H. H. M., Martin, N. G. & Boomsma, D. I. The continuing value of twin studies in the omics era. *Nat. Rev. Genet.* **13**, 640–653 (2012).
2. Vadlamudi, L. et al. Timing of de novo mutagenesis: a twin study of sodium-channel mutations. *N. Engl. J. Med.* **363**, 1335–1340 (2010).
3. Ehli, E. A. et al. De novo and inherited CNVs in MZ twin pairs selected for discordance and concordance on attention problems. *Eur. J. Hum. Genet.* **20**, 1037–1043 (2012).
4. Baranzini, S. E. et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature* **464**, 1351–1356 (2010).
5. Dal, G. M. et al. Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *J. Med. Genet.* **51**, 455–459 (2014).
6. Zink, F. Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017).
7. Hall, J. G. Twinning. *Lancet* **362**, 735–743 (2003).
8. Herranz, G. The timing of monozygotic twinning: a criticism of the common model. *Zygote* **23**, 27–40 (2015).
9. McNamara, H. C., Kane, S. C., Craig, J. M., Short, R. V. & Umstad, M. P. A review of the mechanisms and evidence for typical and atypical twinning. *Am. J. Obstet. Gynecol.* **214**, 172–191 (2016).
10. Tang, W. W. C., Kobayashi, T., Irie, N., Dietmann, S. & Surani, M. A. Specification and epigenetic programming of the human germ line. *Nat. Rev. Genet.* **17**, 585–600 (2016).
11. D'Gama, A. M. & Walsh, C. A. Somatic mosaicism and neurodevelopmental disease. *Nat. Neurosci.* **21**, 1504–1514 (2018).
12. Dou, Y., Gold, H. D., Luquette, L. J. & Park, P. J. Detecting somatic mutations in normal cells. *Trends Genet.* **34**, 545–557 (2018).
13. Sasani, T. A. et al. Large, three-generation families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *elife* **8**, e46922 (2019).
14. Campbell, I. M. et al. Parental somatic mosaicism is underrecognized and influences recurrence risk of genomic disorders. *Am. J. Hum. Genet.* **95**, 173–182 (2014).
15. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
16. Scally, A. Mutation rates and the evolution of germline structure. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **371**, 20150137 (2016).
17. Jónsson, H. et al. Multiple transmissions of de novo mutations in families. *Nat. Genet.* **50**, 1674–1680 (2018).
18. Lindsay, S. J., Rahbari, R., Kaplanis, J., Keane, T. & Hurler, M. E. Similarities and differences in patterns of germline mutation between mice and humans. *Nat. Commun.* **10**, 4053 (2019).
19. Harland, C. et al. Frequency of mosaicism points towards mutation-prone early cleavage cell divisions in cattle. Preprint at *bioRxiv* <https://doi.org/10.1101/079863> (2016).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
21. Seplyarskiy, V. B. et al. Population sequencing data reveal a compendium of mutational processes in human germline. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.01.10.893024> (2020).
22. Xia, B. et al. Widespread transcriptional scanning in the testis modulates gene evolution rates. *Cell* **180**, 248–262.e21 (2020).
23. Moorjani, P., Amorim, C. E. G., Arndt, P. F. & Przeworski, M. Variation in the molecular clock of primates. *Proc. Natl Acad. Sci. USA* **113**, 10607–10612 (2016).
24. Gao, Z. et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl Acad. Sci. USA* **116**, 9491–9500 (2019).
25. Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
26. Greenberg, M. V. C. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nat. Rev. Mol. Cell Biol.* **20**, 590–607 (2019).
27. Reik, W., Dean, W. & Walter, J. Epigenetic reprogramming in mammalian development. *Science* **293**, 1089–1093 (2001).
28. Coe, B. P. et al. Neurodevelopmental disease genes implicated by de novo mutation and copy number variation morbidity. *Nat. Genet.* **51**, 106–116 (2019).
29. Hardy, K., Handyside, A. H. & Winston, R. M. The human blastocyst: cell number, death and allocation during late preimplantation development in vitro. *Development* **107**, 597–604 (1989).
30. Tabansky, I. et al. Developmental bias in cleavage-stage mouse blastomeres. *Curr. Biol.* **23**, 21–31 (2013).
31. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2021

Methods

We sequenced 49,962 individuals by using Genome Analyzer_{II}, HiSeq, HiSeq X and NovaSeq Illumina systems. The target coverage per individual was at least 30×. Our sequencing effort was mainly focused on understanding the interplay between sequence variation and phenotypes. For this study, we augmented this strategy by enriching for the family setup portrayed in Fig. 3d, that is a proband, their spouse/partner, monozygotic twin and offspring, regardless of the phenotype of the proband. To assess the somatic VAF of pre-PGCS mutations accurately, we sequenced a subset of the twins to an average coverage of 152× (Extended Data Figs. 1 and 2).

Ethics statement. The National Bioethics Committee of Iceland and the Icelandic Data Protection Authority approved this study. Blood or buccal samples were taken from individuals participating in the various studies after informed consent was obtained from participants or their guardians.

DNA WGS. Extraction of DNA and subsequent library preparation is described in the Supplementary Note sections. Sequencing libraries were hybridized to the surface of paired-end flow cells using either the Illumina cBot or via on-board clustering (NovaSeq 6000). Paired-end sequencing by synthesis was performed on Illumina sequencers, including Genome Analyzer_{II}, HiSeq 2000/2500, HiSeq X and NovaSeq 6000 systems, respectively. Read lengths depended on the instrument and/or sequencing kit employed and varied from 2×76 cycles to 2×150 cycles of incorporation and imaging. Real-time analysis involved conversion of image data to base calling in real time. We monitored run performance by assessing base quality scores and clusters pass filtering. All sequencing runs were required to pass internal quality thresholds, where each sequencing lane was required to have >80% of sequenced bases with a Phred base quality score (*Q* score) >30. All samples from monozygotic twins were sequenced on either HiSeq X or NovaSeq 6000 systems, with read lengths of 2×150 base pairs (Supplementary Table 1); the year of sequencing of the read groups present in the final BAM files is in Supplementary Table 2.

Alignment of sequenced data. The alignment and postprocessing of the sequenced data was carried out as described in Jónsson et al.³².

Detection of postzygotic mutations. We applied two different approaches to call sequence variants, one for postzygotic mutations that are present in a subset of the cells and another optimized for germline mutations (described below). Briefly we extracted postzygotic mutation candidates with Strelka2 v.2.9.4 (ref.³³) using the monozygotic twins as paired samples (Supplementary Note), then we filtered those candidates by using the frequency of the mutation in the cohort and alignment statistics (Supplementary Note). For the filtered set, we assessed the quality of the postzygotic mutations (Supplementary Note) by read-tracing to nearby germline variants and by resequencing 46 variants to estimate the false positive rate of our filtered postzygotic mutation calls.

Germline variant calling and genotyping. We called variants and genotyped them with GraphTyper³⁴ v.1.4 using all 49,962 sequenced individuals. This resulted in 74,239,180 sequence variants, which were scanned for de novo mutations.

Detection of transmitted mutations. We used three different methods to extract transmitted mutation candidates. This is explained in detail in the Supplementary Note. All the methods compare the sequence of the child to its close relatives, the main difference between the methods being with whom the offspring is compared. Briefly, we searched for sequence variants present in the offspring but absent from the following sets: twin and spouse/partner of the proband (monozygotic method); proband and their spouse/partner (conventional trio method); and parents and spouse/partner of the proband (three-generation method). The transmitted mutation candidates were then filtered (Supplementary Note) based on alignment statistics supporting the mutation candidate in the offspring and the frequency of the mutation candidate outside the family. For the transmitted mutation passing the filter, we categorized the mutation depending on the presence in the somatic tissue of the proband (Supplementary Note). For all the transmitted mutations, we tried to determine the phase of the transmitted mutations by using inheritance patterns and physical phasing (Supplementary Note).

Assessment of germline frequency of mutations in the proband. We used the transmission from the proband to the younger siblings of the offspring to assess the frequency of pre-PGCS mutations in the germline of the proband. This was done because we did not have access to germ cell samples from the proband. We analyzed multiple offspring of the proband by splitting the offspring into pairs of siblings. To account for intrafamily correlation^{17,35}, we assessed the standard errors and significance of our estimates with a block jackknife³⁶ using the parent pair of the offspring (proband and spouse/partner) as the block.

Expected number of transmitted pre-PGCS mutations to the offspring of the monozygotic twin. Pre-twinning mutations present in the germline in both twins are expected to be transmitted to the offspring of the proband and monozygotic twin. Having offspring from the monozygotic twin allowed us to assess the

expected transmission of pre-PGCS mutations (detected in the proband) from the monozygotic twin to the offspring of the monozygotic twin. We did this by going through the pre-PGCS mutations identified in the proband and summing the VAFs of the twin (f_i); f_i is the product of the frequency of the cell lineage in the somatic tissue and 0.5. If the frequency of the cell lineage is proportional between the somatic tissue and germline, then the mutation defining the cell lineage will be transmitted at an f_i rate (since the gamete cell is haploid). Then, the expected number of transmitted pre-PGCS mutations to the offspring of the twin is $\sum_i f_i$, where the sum ranges over the pre-PGCS mutations identified in the proband. The resulting sum is 21 (if conditioned on the pre-PGCS mutations of the proband, we expect 21 transmissions to the offspring of the monozygotic twin). Note that in this analysis both twins were considered as probands.

Analysis of mutation classes. We classified the mutations and their complements into 8 mutation classes (C>A, C>G, C>T, CpG>TpG, T>A, T>C, T>G and indels). Subsequently, we calculated the odds ratio per mutation class among pre-PGCS de novo mutations against the late de novo mutations identified in the nonhomozygous trios (contingency table test). To assess the significance of the odds ratio difference from 1, we block jackknifed the odds ratio on a log scale using the proband and spouse/partner as a block. The results from this procedure are reported in Supplementary Table 5. We checked whether there was a difference between the relative frequencies of the mutational classes based on the sex of the proband. We calculated the chi-squared value for the entire table and transformed it using the Wilson–Hilferty standardization and then block jackknifed the transformed values and compared them to a normal distribution. We did not find a statistically significant difference ($P=0.06$).

We further assessed the dependency of the mutational class on the VAF of the proband by modeling the fraction of a mutational class as the function of the proband's VAF in a linear model (mutational class fraction regressed on the VAF). Then, we tested the null hypothesis that the slope is zero for each mutation class by block jackknifing the slope estimate using the proband and spouse/partner as a block (Supplementary Table 6). For computational reasons, we kept the late de novo mutation dataset fixed while jackknifing over the proband and spouse/partner pairs from the monozygotic trio set. To assess whether the mutational class enrichment results were induced by a single mutational class, we carried out the enrichment analysis by leaving one mutational class out at a time (Supplementary Tables 7 and 8).

Monozygotic triplets. The offspring of the monozygotic triplets presented in Fig. 7a were not present in the set of 49,962 sequenced individuals described above. To find pre-PGCS mutations in the triplets, we applied a different procedure from that in the rest of the dataset. We scanned the de novo mutation candidates present in the triplets, that is, using the absence from the parents and presence in the triplets. Then, we used the alignment model from the section 'Filtering of postzygotic mutation candidates' in the Supplementary Note to restrict to a prediction score >1%. This model does not take the VAF of the mutation candidate into account, thus allowing us to find nonconstitutional mutations in the triplets. Then, we genotyped the offspring at these sites and required the mutation to be transmitted to one of them. More specifically, we required the presence of these mutations in the offspring of the triplets with a VAF > 30% and at least 4 reads supporting the allele in the offspring. Finally, we restricted to the transmitted mutations where we could reject the VAF if it was 50% in one of the triplets by using a binomial test ($P=0.001$).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Access to these data is controlled; the sequence data cannot be made publicly available because Icelandic law and the regulations of the Icelandic Data Protection Authority prohibit the release of individual-level and personally identifying data. Data access can be granted only at the facilities of deCODE genetics in Iceland, subject to Icelandic law regarding data usage. Anyone wanting to gain access to the data should contact Kári Stefánsson (kstefans@decode.is). Data access consists of the lists of mutations identified in monozygotic twins with numbered proband identifiers. The lists of mutations are provided in Supplementary Data 1–3.

Code availability

The major components in our sequence data processing pipeline consist of publicly available software, notably Burrows–Wheeler Aligner–MEM for the alignment (<https://github.com/lh3/bwa>), Samtools for the processing of BAM files (<http://samtools.github.io/>), Picard for PCR duplication marking (<https://broadinstitute.github.io/picard/>) and GraphTyper for sequence variant calling (<https://github.com/DecodeGenetics/graph typer>). The implementation of the phasing and imputation of sequence variants is described in the data descriptor³².

References

- Jónsson, H. et al. Whole genome characterization of sequence diversity of 15,220 Icelanders. *Sci. Data* **4**, 170115 (2017).

33. Kim, S. et al. Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* **15**, 591–594 (2018).
34. Eggertsson, H. P. et al. Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).
35. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombinations through a sequence level genetic map. *Science* **363**, eaau1043 (2019).
36. Busing, F. M. T. A., Meijer, E. & Van Der Leeden, R. Delete-m jackknife for unequal m. *Stat. Comput.* **9**, 3–8 (1999).

Acknowledgements

We thank everyone who participated in our studies.

Author contributions

H.J., U.T., D.F.G. and K.S. wrote the manuscript with input from E.M., T.S., H.P.E., O.A.S., O.E., G.A.A., F.Z., E.A.H., I.J., A.G., Adalbjorg Jonasdottir, Aslaug Jonasdottir, D.B., G.L.N., O.T.M., G.M., B.V.H., A.H. and P.S. H.J., O.E. and E.A.H. analyzed the data. H.J., H.P.E., O.E., F.Z., E.A.H., A.G., G.M. and P.S. developed the methods. H.J., Adalbjorg Jonasdottir, Aslaug Jonasdottir, O.T.M. and G.L.N. performed the

experiments. G.A.A. and I.J. provided samples and information. H.J., P.S., D.F.G. and K.S. designed the study.

Competing interests

H.J., H.P.E., O.A.S., O.E., G.A.A., F.Z., E.A.H., I.J., A.G., Adalbjorg Jonasdottir, Aslaug Jonasdottir, D.B., G.L.N., O.T.M., G.M., B.V.H., U.T., A.H., P.S., D.F.G. and K.S. are employed by deCODE genetics/Amgen.

Additional information

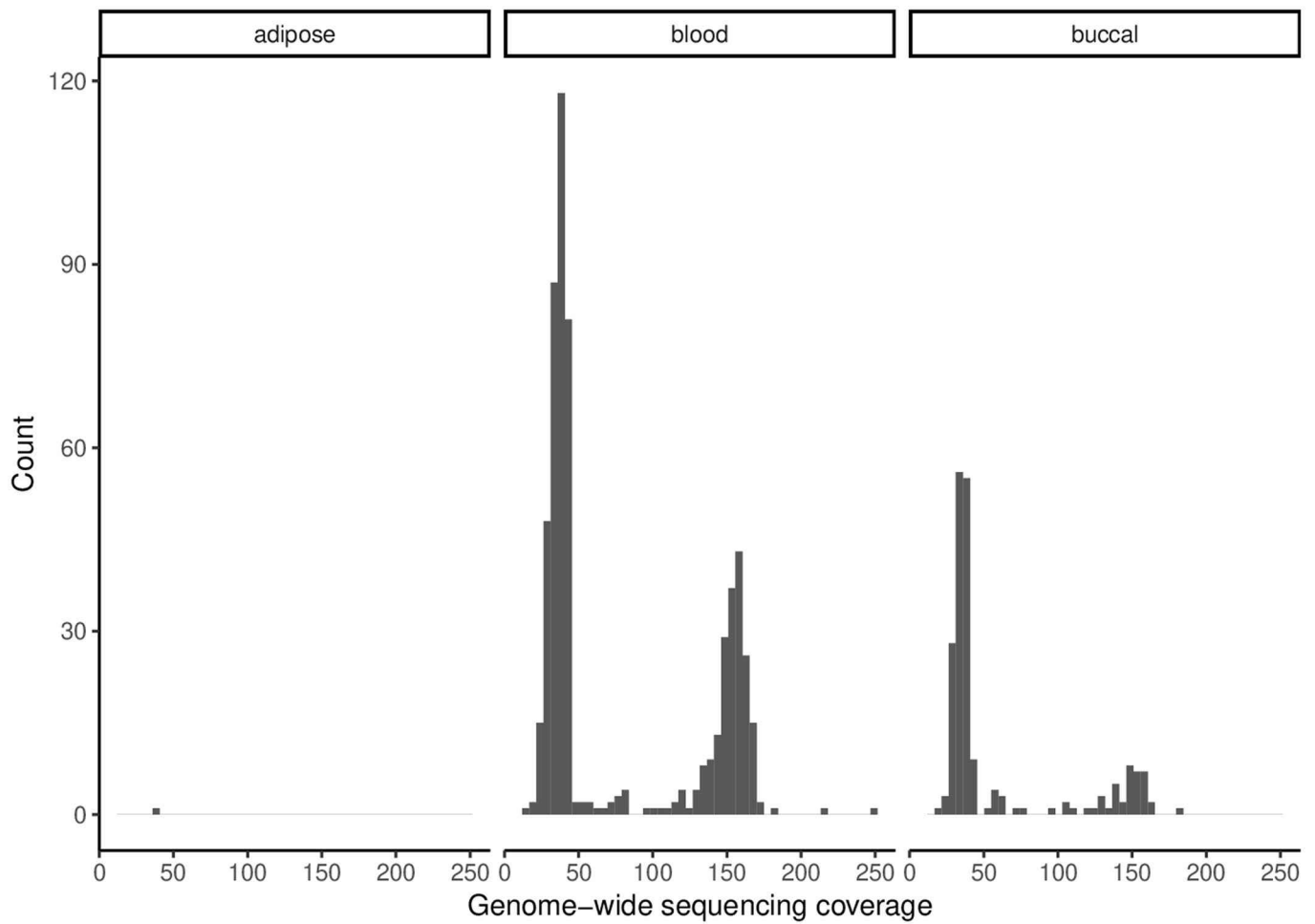
Extended data is available for this paper at <https://doi.org/10.1038/s41588-020-00755-1>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-020-00755-1>.

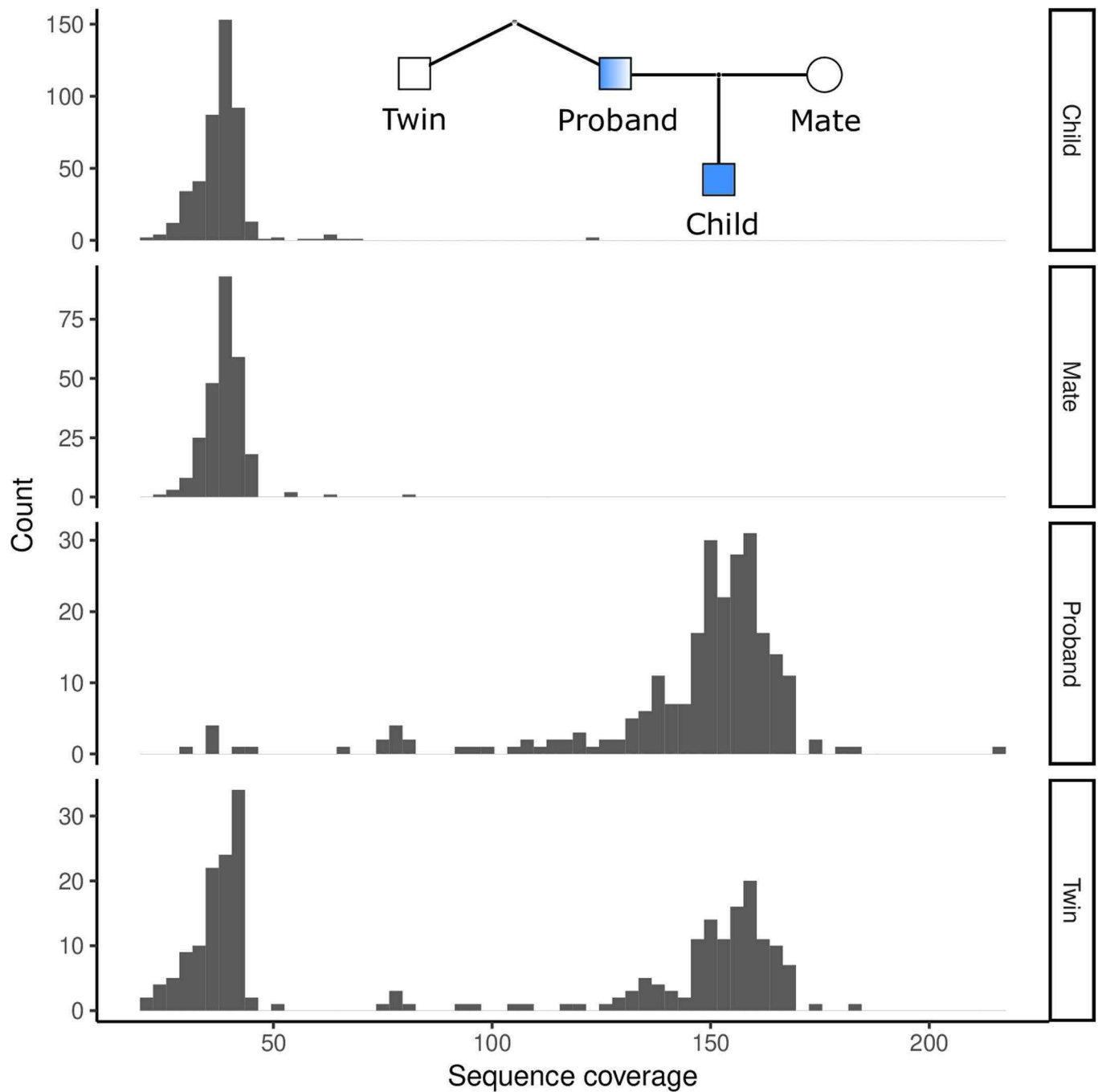
Correspondence and requests for materials should be addressed to H.J., D.F.G. or K.S.

Peer review information *Nature Genetics* thanks Jeffrey Beck, Dorret Boomsma, Ziyue Gao, Brandon Johnson, and Amy Williams for their contribution to the peer review of this work.

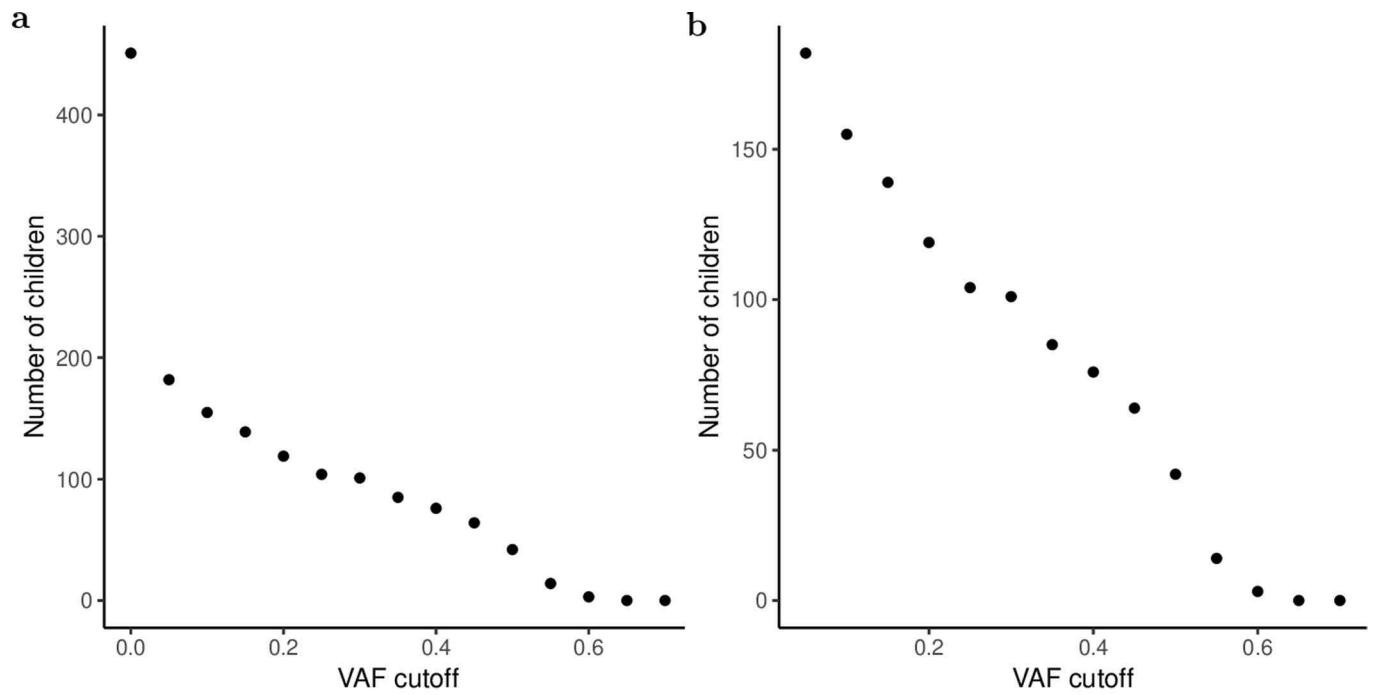
Reprints and permissions information is available at www.nature.com/reprints.



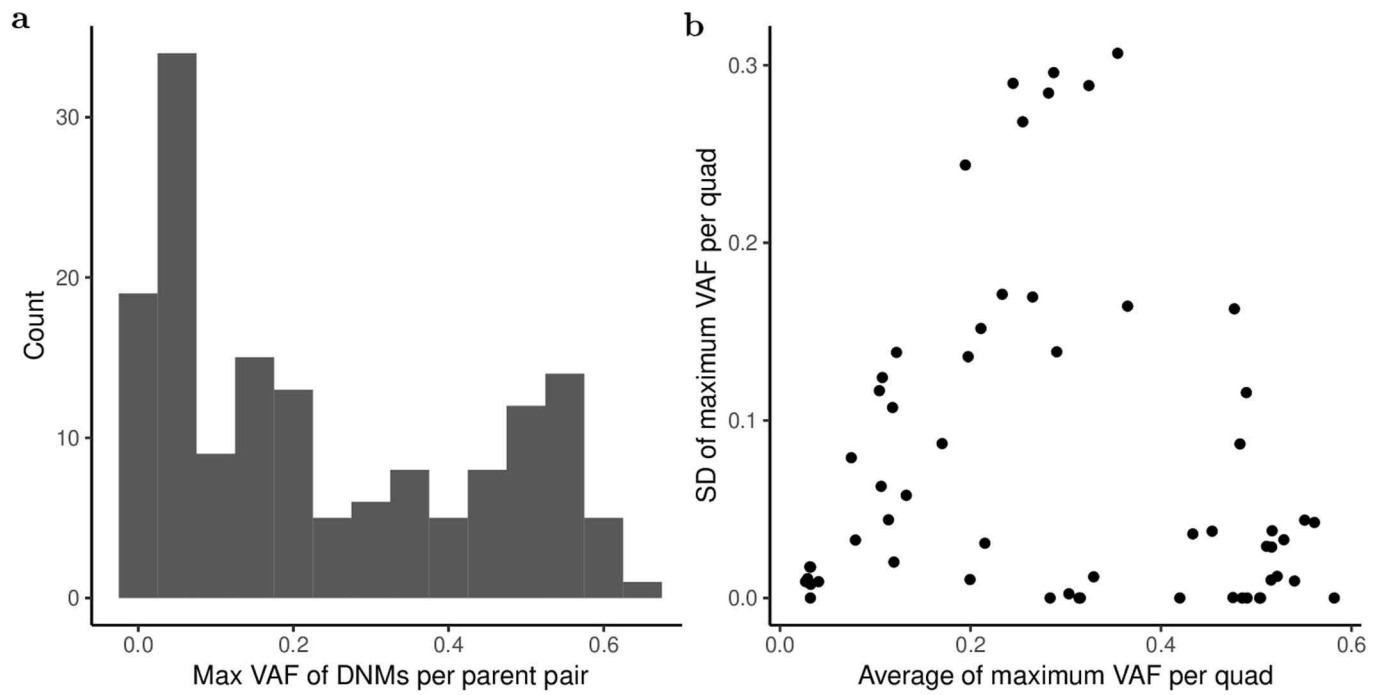
Extended Data Fig. 1 | Histogram of the genome-wide sequence coverage of the twins. Histogram of the genome-wide sequence, coverage of the twins. Note that the sequence coverage for the monozygotic twins was aggregated across several sequencing runs, and the aggregated sequence data were used for the subsequent analysis.



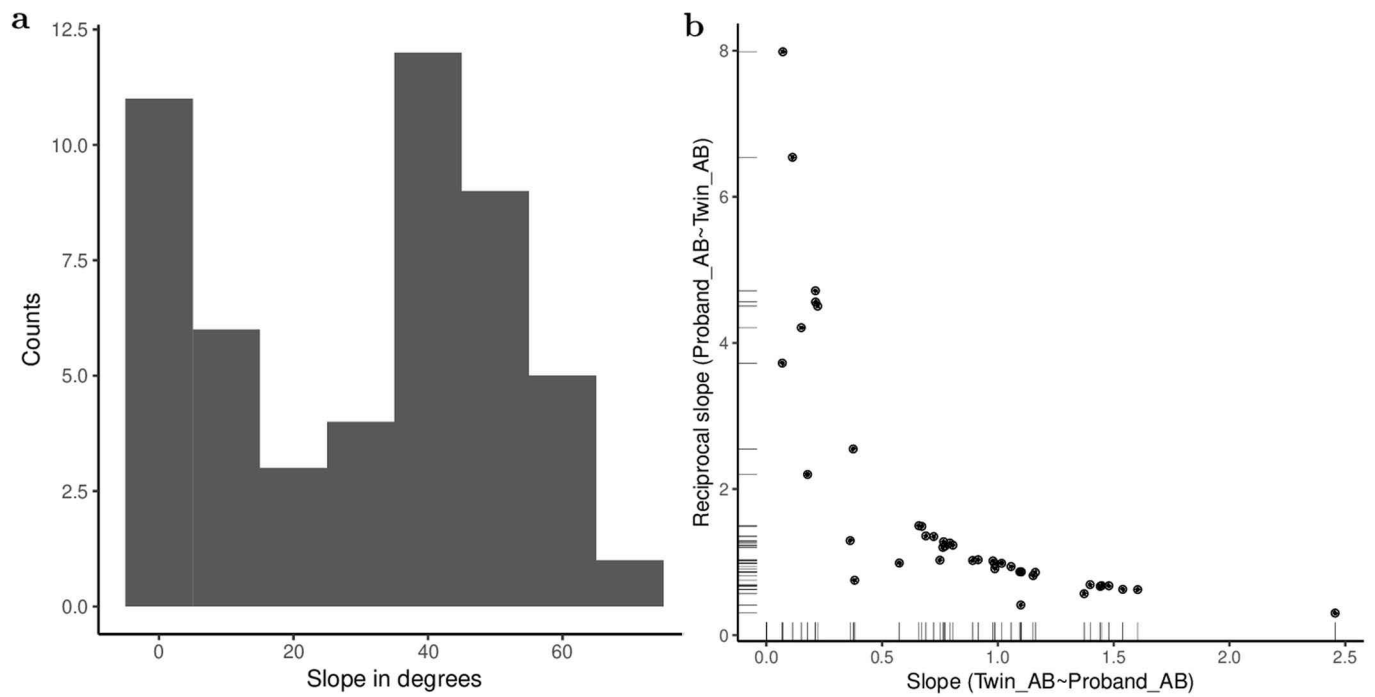
Extended Data Fig. 2 | The genome-wide sequence coverage of the probands' family members. The genome-wide sequence coverage of the probands' family members. The family members of the probands were used to detect pre-PGCS mutations. Note, that if both twins of a pair have sequenced children then they will appear as 'Proband' and as 'Twin'.



Extended Data Fig. 3 | Number of children with a pre-PGCS mutation. Number of children with a pre-PGCS mutation. a, We counted how many children have a pre-PGCS mutation with VAF higher than a cutoff. b, We restricted to children where at least one pre-PGCS mutation was detected.



Extended Data Fig. 4 | The maximum VAF of pre-PGCS mutations per proband/mate pair. The maximum VAF of pre-PGCS mutations per proband/mate pair. **a**, The maximum VAF of pre-PGCS mutations per proband/mate pair. **b**, The standard deviation of the maximum VAF per proband/mate pair against the average of the maximum VAF.



Extended Data Fig. 5 | Alternative calculations of the slopes from the three-generation approach. Alternative calculations of the slopes from the three-generation approach. a, Histogram of the slopes as Fig. 5e, except the slopes are transformed with atan. b, The slopes in three generation approach with swapped roles. Note that the reciprocal slopes are not defined for near constitutional probands due to zero sample variance.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Description of the code used for the sequencing, alignment and imputation setup is described in the Data Descriptor "Whole genome characterization of sequence diversity of 15,220 Icelanders."

Data analysis

The major components in our sequence data processing pipeline are composed of publicly available software, notably BWA mem for the alignment (version 0.7.10-r789; <https://github.com/lh3/bwa>), Samtools for processing of bam files (version 1.9; <http://samtools.github.io/>), Picard tools for PCR duplication marking (version 1.117; <https://broadinstitute.github.io/picard/>), and Gruptyper for sequence variant calling (version 1.4; <https://github.com/DecodeGenetics/graph typer>). The implementation of the phasing and imputation of sequence variants is described in the Data descriptor "Whole genome characterization of sequence diversity of 15,220 Icelanders." .

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Access to these data is controlled, the data access consists of lists of mutations identified in the offspring of the monozygotic twins with enumerated proband identifiers. The lists of mutations are provided as Supplementary Data Sets.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample size calculation was performed. To accurately characterize the somatic VAF of the proband we aimed to sequence the proband using 4 lanes of Hiseq X, resulting in a target coverage of 120X. We prioritize probands with sequenced family members and multiple offspring, using around 1,400 lanes of Hiseq X sequencing. The set was enriched for monozygotic twins and their offspring. Our sample size is adequate as we detect a large number of post-zygotic mutations in the monozygotic twins.
Data exclusions	No data were excluded from the analyses, we included all whole genome sequenced quads for the analysis. More specifically, the quad consist of a proband, monozygotic twin of the proband, mate of the proband and finally child of the proband.
Replication	We randomly selected 46 post-zygotic mutations from the somatic approach for targeted resequencing, 43 had sufficient coverage and of those 31 were validated, resulting in false positive rate of 28% (95%-CI:15-44%). Further, the segregation of mutations in three generation families provides an excellent control of the false positive rate. In addition the absence of the mutations from the monozygotic twins verifies directly the post-zygotic nature of the mutations. None of the mosaicism patterns were observed when replacing the monozygotic twins with siblings, as expected as they are derived from different zygotes.
Randomization	The canonical experimental group in this study is a family consisting of a proband, his or her mate and their offspring. Participants were not allocated randomly to experimental groups as the experimental group is the family of the participants. All reported estimates of standard deviation in the study measure the Proband-Mate or inter-family standard deviation, hence intra-family covariates that differ between the families will increase the inter-family standard deviation.
Blinding	It is not possible to be agnostic to the group allocation of the samples in this study, as the experimental group is the family and we are looking for post-zygotic mutations within a family.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern		

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Supplementary Tables 3-4 summarizes the families used in the study, e.g. the number of the quads and whether the DNA of the proband was derived from a blood or buccal sample.
Recruitment	Participants were recruited through various projects from the beginning of deCODE genetics, most of them are focused on understanding the interplay between genetics and phenotypes. The monozygotic twins and their family members analyzed here were recruited through these studies. There was no specific recruitment for this study. The individuals that were sequenced specifically for this study were not selected based on their phenotypes. We aimed for sequencing monozygotic twins with at least one offspring, this could lead to enrichment of older twins which are fertile.
Ethics oversight	The National Bioethics Committee of Iceland.

Note that full information on the approval of the study protocol must also be provided in the manuscript.