



# HHS Public Access

Author manuscript

*Cancer Discov.* Author manuscript; available in PMC 2022 October 01.

Published in final edited form as:

*Cancer Discov.* 2022 April 01; 12(4): 949–957. doi:10.1158/2159-8290.CD-21-1110.

## Cancer Causative Mutations Occurring in Early Embryogenesis

Fresia Pareja<sup>1,\*</sup>, Ryan N. Ptashkin<sup>1,\*</sup>, David N. Brown<sup>1,\*</sup>, Fatemeh Derakhshan<sup>1</sup>, Pier Selenica<sup>1</sup>, Edaise M. da Silva<sup>1</sup>, Andrea M. Gazzo<sup>1</sup>, Arnaud Da Cruz Paula<sup>4</sup>, Kelsey Breen<sup>2</sup>, Ronglai Shen<sup>3</sup>, Antonio Marra<sup>1</sup>, Ahmet Zehir<sup>1</sup>, Ryma Benayed<sup>1</sup>, Michael F. Berger<sup>1</sup>, Ozge Ceyhan-Birsoy<sup>1</sup>, Sowmya Jairam<sup>1</sup>, Margaret Sheehan<sup>2</sup>, Utsav Patel<sup>1</sup>, Yelena Kemel<sup>5</sup>, Jacklyn Casanova-Murphy<sup>1</sup>, Christopher J. Schwartz<sup>1</sup>, Mahsa Vahdatinia<sup>1</sup>, Elizabeth Comen<sup>6</sup>, Laetitia Borsu<sup>1</sup>, Xin Pei<sup>7</sup>, Nadeem Riaz<sup>7</sup>, David H Abramson<sup>8</sup>, Britta Weigelt<sup>1</sup>, Michael F. Walsh<sup>2,8</sup>, Anna-Katerina Hadjantonakis<sup>9</sup>, Marc Ladanyi<sup>1</sup>, Kenneth Offit<sup>2,5</sup>, Zsofia K. Stadler<sup>2,5</sup>, Mark E. Robson<sup>2,5,6</sup>, Jorge S. Reis-Filho<sup>1</sup>, Diana Mandelker<sup>1</sup>

<sup>1</sup>Department of Pathology, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>2</sup>Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>3</sup>Department of Epidemiology & Biostatistics, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>4</sup>Department of Surgery, Memorial Sloan Kettering Cancer Center; New York City, New York, USA

<sup>5</sup>Niehaus Center for Inherited Cancer Genomics, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>6</sup>Breast Medicine Service, Department of Medicine, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>7</sup>Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY, USA

<sup>8</sup>Department of Ophthalmology, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

<sup>9</sup>Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center; New York City, New York, USA.

### Abstract

Mosaic mutations in normal tissues can occur early in embryogenesis and be associated with hereditary cancer syndromes when affecting cancer susceptibility genes (CSGs). Their contribution to apparently sporadic cancers is currently unknown. Analysis of paired tumor/blood sequencing data of 35,310 cancer patients revealed 36 pathogenic mosaic variants affecting CSGs,

---

**Correspondence to:** Dr. Diana Mandelker, MD PhD, Department of Pathology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA. Phone: +1 212 639 4820. mandelkd@mskcc.org, Dr. Jorge S Reis-Filho, MD PhD FRCPath, Department of Pathology, Memorial Sloan Kettering Cancer Center, 1275 York Avenue, New York, NY 10065, USA. Phone: +1 212 639 8054. reisfilj@mskcc.org.

\*Equal contribution

most of which were not detected by prior clinical genetic testing. These CSG mosaic variants were consistently detected at varying variant allelic fractions in microdissected normal tissues (n=48) from distinct embryonic lineages in all individuals tested, indicating their early embryonic origin, likely prior to gastrulation, and likely asymmetrical propagation. Tumor-specific biallelic inactivation of the CSG affected by a mosaic variant was observed in 91.7% (33/36) of cases and tumors displayed the hallmark pathologic and/or genomic features of inactivation of the respective CSGs, establishing a causal link between CSG mosaic variants arising in early embryogenesis and the development of apparently sporadic cancers.

## Keywords

Embryogenesis; mutations; mosaicism; cancer

---

## INTRODUCTION

Normal cells undergo mutagenesis throughout an individual's lifespan (1), resulting in somatic mosaicism, a phenomenon by which an individual has two or more cell populations with different genotypes (2). Various normal tissues, such as skin, bladder and esophagus have been shown to comprise mosaics of evolving clones harboring a panoply of somatic mutations (3–5). RNA sequencing analyses of 29 different tissue types from over 400 individuals have revealed somatic clonal expansions in normal tissues (6). Notably, mosaic variants acquired as early as in embryogenesis may be present in adult normal cells, as evidenced by whole-genome sequencing analyses of normal blood from 241 adults where early embryonic somatic mutations were identified (7).

Mosaicism has been reported to affect cancer susceptibility genes (CSGs) in individuals meeting clinical criteria for hereditary predisposition syndromes, such as tuberous sclerosis and neurofibromatosis (8,9). The contribution of mosaic variants to the development of apparently sporadic cancers, however, has yet to be determined. In addition to the potential cancer predisposition risk conferred by mosaicism affecting CSGs, these variants may be transmitted to offspring if present in gonadal tissue. In addition, some of the CSGs affected by mosaic variants constitute indications for targeted therapies. Hence, their detection and the definition of their role in cancer predisposition are critical for genetic counselling and clinical management. The low variant allelic fraction (VAF) in blood of mosaic variants poses challenges to their detection using current clinical genetic testing methodologies (10). In the absence of matched paired tumor/normal samples, their distinction from sequencing artifacts, clonal hematopoiesis (CH) and circulating tumor cells (CTCs) is challenging.

Here, through the analysis of >35,000 unselected cancer patients undergoing tumor/normal sequencing, we sought to determine whether mosaic variants affecting CSGs arising in embryogenesis might contribute to the development of seemingly sporadic cancers. Our analyses revealed that mosaic variants affecting CSGs can occur early in embryogenesis, likely contribute to tumorigenesis, and result in increased risk of cancer development.

## RESULTS

### Detection of cancer causative mosaic variants in cancer susceptibility genes

To detect candidate patients harboring mosaic variants affecting CSGs, following Institutional Review Board approval from Memorial Sloan Kettering Cancer Center (MSKCC), we applied a set of filtering criteria to tumor/blood sequencing data from 35,310 unselected cancer patients using the FDA-cleared MSK Integrated Mutation Profiling of Actionable Targets assay (MSK-IMPACT (11) (Fig. 1A and Supplementary Table S1). We sought to identify mosaic pathogenic/likely pathogenic (P/LP) variants in the 61 CSGs included in the MSK-IMPACT panel (Supplementary Table S2). 3,113,229 variants affecting any of the 61 CSGs in the MSK-IMPACT panel were detected in the blood at any VAF in the 35,310 patients included in our study (Fig. 1A, Supplementary Tables S1–S2). To select mosaic variants that would be causative of cancer, we applied a set of filters to identify candidate variants with a VAF of 1.5%–30% in blood, 10% VAF in the tumor with a VAF at least 1.5x higher in the tumor as compared to the matched blood, total depth of sequencing coverage  $\geq 50$  for both tumor and blood,  $\geq 6$  reads of the alternate allele in blood, and an allele frequency  $< 1\%$  in any Genome Aggregation Database (gnomAD; v2.1.1; Fig. 1A; Supplementary Figs. S1A–S1E, S2A–S2D and S3A–S3B; see Methods and Supplementary Methods). To distinguish true mosaic variants from CTC variants and CH, we excluded patients with  $> 1$  variant meeting these criteria and applied established CH criteria (12), respectively (Fig. 1A, Supplementary Methods). 123 CTC variants, including 48 affecting CSGs, were identified in 25 individuals, with an average of 4.9 (range, 2–24) CTC variants/case which largely (71.4%) displayed mutational signatures matching those of their respective tumors; only 23.6% of CTC variants were *bona fide* loss-of-function (LOF) mutations (Supplementary Tables S3–S4).

This analysis revealed 53 candidate mosaic variants affecting CSGs including 36 P/LP variants and 17 variants of uncertain significance (VUS) in 36 and 17 individuals, respectively (Fig. 1A, Supplementary Table S5). The median VAFs of the P/LP mosaic variants detected in blood and tumors were 7.8% (range, 1.7–28.3%) and 49.4% (range, 22.8–79.7%), respectively, and the median blood and tumor sequencing depths at the mosaic variant sites were 474 (range, 256–923) and 629 (range, 192–1520), respectively (Fig. 1B and Supplementary Table S5). Upon normalization by sequencing depth, the VAF remained increased in tumors relative to blood (Supplementary Fig. S4A–S4E). When taking tumor purity and ploidy into consideration, all mosaic variants were found to be clonal in the tumors analyzed (Supplementary Table S5). Upon applying the same algorithm to non-CSG included in the MSK-IMPACT panel ( $n=280$ , see Methods), we detected 79 mosaic variants meeting these criteria, all of which would be classified as VUS for hereditary cancer susceptibility (Supplementary Fig. S5). These findings demonstrate a significant enrichment for P/LP variants in CSGs as compared to non-CSGs taking into account their corresponding genomic footprints in the MSK-IMPACT panel ( $P=5 \times 10^{-31}$ ; two-tailed Fisher's exact test), supporting the adequacy of the approach employed for the detection of cancer causative mosaic variants.

## Clinical and genetic characterization of cancer causative mosaic variants

The CSGs most frequently affected by P/LP mosaic variants were *TP53* (n=16) and *RBI* (n=5; Fig. 1C and Supplementary Table S5). Most (27/36; 75%) P/LP mosaic variants were *bona fide* LOF mutations (i.e., truncating, frameshifting and splice-site single nucleotide variants (SNVs); Fig. 1B–1C; Supplementary Table S5). Tumor-specific biallelic inactivation of the CSGs affected by the mosaic variants was observed in 91.7% (33/36) of cases, as loss of heterozygosity the wild-type allele (28/36, 77.8%) or as a second inactivating somatic mutation (5/36, 13.9%; Fig. 1C and Supplementary Table S5). Tumors from patients with *MSH2* (MOS14 and MOS36) and *MSH6* (MOS9 and MOS38) mosaic variants displayed loss of MSH2 and MSH6 protein expression, respectively, as well as a dominant microsatellite instability (MSI) mutational signature, high tumor mutation burden, enrichment for short indels, and/or were MSI-high by PCR analysis (Fig. 2A–2D, Supplementary Table S6, Supplementary Figs. S6A–S6B and S7A–S7D). The ovarian (MOS1) and breast (MOS8) cancers in *BRCA2* P/LP mosaic variant carriers harbored genomic features indicative of homologous recombination DNA repair deficiency (HRD), such as dominant HRD mutational signature, genomic instability, high fraction of genome altered, increased number of large-scale state transitions, and/or increased indel length (Fig. 3A–3C, Supplementary Table S7). These findings support the notion that the mosaic variants here identified likely played an etiologic role in cancer development in these patients.

Tumors harboring candidate CSG mosaic variants were of various histologic types, the most frequent being breast cancer and sarcoma (6/36, each; Fig. 1C, Supplementary Fig. S8 and Supplementary Table S5), and stemmed from the different germ layers, including 13 tumors derived from endoderm and mesoderm, each, and 10 tumors from ectoderm (Fig. 1C and Supplementary Table S5). The CSGs affected by the mosaic variants were detected in cancer types expected in the syndromes caused by germline P/LP variants of the same genes in 80.6% (29/36) of cases. For instance, mosaic variants affecting *BRCA2* were identified in breast or ovarian cancers, whereas those affecting *APC* were found in colorectal carcinoma or its precursor, tubular adenoma. Notably, in 5/7 cases with a non-classic tumor type-gene association, the tumor type has been reported in association with the germline variants in the genes affected, albeit at a lower frequency, such as prostate and gastric cancer in *TP53* germline carriers(13), and sarcomas in Lynch syndrome(14). Our analysis also revealed the novel observation of *SMARCA4* mosaic variants in two patients with small cell carcinomas of the ovary, hypercalcemic type (SCCOHT, Fig. 1C and Supplementary Table S8), indicating that in addition to somatic or germline *SMARCA4* variants, SCCOHTs may also be underpinned by P/LP mosaic variants affecting *SMARCA4*. Akin to patients with germline variants in CSGs (15,16), 27.8% (10/36) of individuals carrying a CSG mosaic variant had multiple tumor types (median=2; range, 2–3; Supplementary Table S8) and an age of onset intermediate between sporadic and germline cases (Supplementary Figs. S9–S10).

Out of the 24 patients who had previous germline genetic testing, including next generation sequencing (n=23) and Sanger sequencing (n=1), only 6 (25%) had been reported as mosaic, whereas the remaining 18 (75%) were reported as negative for germline genetic testing. In the context of tumor/normal sequencing, 10/18 were mis-reported as tumor-derived

mutations, whereas 8/18 were not reported at all due to not meeting filtering criteria for somatic variant detection (Fig. 1C and Supplementary Tables S8–S9), highlighting the need for the development of a systematic methodology for the detection of mosaic variants. Only approximately half (51.4%, 18/35) of the patients with evaluable medical history met the clinical criteria for germline genetic testing for the gene affected by the mosaic variant based on personal history, whereas only 2.9% (1/35) of cases met those criteria based on family history (Supplementary Tables S8–S9). Hence, when a germline susceptibility is suspected, yet routine germline clinical genetic testing yields a negative result, assessment for mosaicism may be a reasonable approach, given that detection of these important variants would allow gene-specific cancer screening and prophylactic measures, screening of the patient's offspring, and even the potential use of specific therapies (e.g. PARP inhibitors in the case of *BRCA2* mosaic P/LP variants or immune checkpoint inhibitors in the case of mosaic P/LP variants affecting mismatch repair genes).

### Validation of mosaic variants

To validate the mosaic nature of the candidate variants identified, we interrogated their presence in normal tissues deriving from different embryonic lineages in 10 patients with available formalin-fixed paraffin-embedded (FFPE) material (Fig. 2A). To ensure the adequate purity of the different tissues, following histopathologic evaluation and assessment of leukocyte infiltration, we conducted laser capture microdissection (LCM) of normal (n=48) and tumor (n=15) FFPE tissues (Fig. 2A and Supplementary Table S10, Supplementary Figs. S11A–S11L and S12), and subjected them to targeted sequencing using MSK-IMPACT. A median of 4 (range, 3–9) different normal tissues were analyzed per case (Supplementary Table S10). We identified the CSG mosaic variants at varying VAFs (median, 7.6%; range, 1.5%–25.8%) in all normal samples interrogated, which included normal tissues of mesodermal, endodermal and/or ectodermal lineages (Fig. 2B and Supplementary Table S10). The CSG mosaic variants were found to be enriched in the respective tumor tissues with a median VAF of 53.4% (range, 25.6%–92.9%;  $P < 0.001$ , Mann Whitney *U*-test; Figs. 2B–2D, 3A–3B and Supplementary Fig. S13A–S13F and Supplementary Table S10). The tumors of 8/10 cases analyzed harbored additional non-synonymous somatic mutations (n=76; median number of non-synonymous mutations=4.5/case; range, 1–34), none of which were detected in the normal tissues interrogated (Figs. 2B–2D, 3A–3B, Supplementary Fig. S13C–S13F and Supplementary Table S10). Moreover, we conducted an additional orthogonal validation of the mosaic variants in 39 normal and 13 tumor tissues, respectively, of individuals with available FFPE material (n=10) using amplicon sequencing. Our analyses validated the mosaic nature of the variants showing a strong positive correlation ( $r=0.93$ ;  $P=2.2 \times 10^{-16}$ ) in the VAF detected by MSK-IMPACT and amplicon sequencing (Supplementary Fig. S14A). In all cases analyzed, the detection of the candidate mosaic variants in multiple tissue lineages in the absence of tumor-derived somatic mutations provides further evidence that the candidate variants interrogated were mosaic in nature.

### Cancer causative mosaic variants arise early in embryogenesis

Next, we sought to determine the time during development when mosaic variants arose. Assuming i) a symmetrical cell contribution model, in which the genetic material of two

daughter cells of a dividing progenitor cell contribute equally to adult tissues, and ii) that these mosaic variants are heterozygous, variants occurring in the first five cell divisions would be expected to have a VAF in normal tissues ranging from 25% (first division) to 1.6% (fifth division, Fig. 4A) (7). Given that ancestral clones of blood emerge early in embryogenesis, before gastrulation (17,18), and the limits of detection of the tumor/normal sequencing assay employed, mosaic variants arising after the 5<sup>th</sup> division, which would have an expected VAF lower than 1%, are unlikely to be detectable. We observed that although the VAFs of the 36 mosaic variants in blood fell within this range, their VAF distribution did not show peaks at the values expected for each of the first five cell divisions (Fig. 4A–4B), suggesting an asymmetry in cell contribution, in agreement with previous studies (7,18,19). Using a Log likelihood model (7) (Supplementary Methods), we sought to investigate the cell contribution asymmetry by introducing an asymmetry factor. We opted for a heuristic approach in which we modified the expected VAF of two cell divisions at a time whilst maintaining other cell divisions symmetric, as previously described (7). Our analyses revealed that the best fitting asymmetric cell contribution model was a better representation of the data compared to the symmetric cell contribution model for both, cell divisions 1 and 2 ( $P=5.1 \times 10^{-6}$ , likelihood ratio test) and cell divisions 3 and 4 ( $P=6.5 \times 10^{-6}$ , likelihood ratio test; Fig. 4C). These findings indicate that cell contribution during early embryogenesis is likely asymmetrical, as previously reported (7,18,19).

In agreement with these findings, we observed that although the VAF of mosaic variants in blood and in other normal tissues was similar when aggregated across patients ( $P>0.05$ , Mann Whitney *U*-test), the VAF of mosaic variants showed differences across normal tissues of the same individual (Figs. 2C–2D, 3A–3B, Supplementary Figs. S13A–S13F and S14B). In some individuals (MOS1 and MOS6), the VAF of the mosaic variant was relatively similar in the different normal tissues interrogated, whereas in others, such as MOS9, a wider VAF distribution across different tissues was observed, even across tissues of the same germ layer (Supplementary Fig. S15A). These findings provide further support to the notion that the daughter cells of a given cell division might contribute to adult tissues in an asymmetrical manner, and suggest that the degree of cell contribution asymmetry has inter-individual variability, consistent with previous reports(18,19).

Using a clustering model based on Beta-Binomial mixture distributions, without considering expected VAFs from the different specific cell divisions, we sought to determine mosaic variant clusters based on their VAF in blood. The number of mixture distributions was determined by bootstrapping whereby four mosaic variant VAF clusters was the simplest model that maximized the log-likelihood of the model (Figs. 4B, 4D and Supplementary Fig. S15B). Based on the posterior probability distribution, we assigned the mosaic variants to one of the four VAF clusters and observed that the majority (72.2%; 26/36) belonged to the third and fourth clusters (Fig. 4E), which would suggest that most of these variants were acquired during the 3<sup>rd</sup> or 4<sup>th</sup> cell divisions of embryogenesis. Nonetheless, asymmetry in cell contribution was observed, which showed marked interindividual variability in agreement with prior studies (18,19). Due to the FFPE nature of our samples, single cell sequencing which would allow individual phylogenetic inference per patient and incorporation of mutation rate, ideal for the assessment of mutation timing, could not be

conducted. Therefore, we can conclude that the variants developed within the first five cell divisions, however the exact timing could not be fully ascertained.

To define the biological processes involved in their genesis, we explored the mutational spectra and mutational signatures that shaped the mosaic variants in our cohort. The 13 mosaic indels identified had heterogeneous profiles, whereas the 23 mosaic SNVs were frequently C>T substitutions predominantly at CpG sites, consistent with a dominant clock-like/aging mutational signature (20) (Fig. 4F), akin to what was reported by Ju et al. for early embryonic mutations (7), and as observed in the context of germline variants (21).

## DISCUSSION

Here, we detected P/LP mosaic variants affecting CSGs in patients with apparently sporadic cancers subjected to tumor/normal sequencing using an FDA-cleared assay. These mosaic variants were present in tumors whose phenotypes are typical of the cognate syndromes caused by germline variants affecting the respective genes. Analysis of the tumor samples of the 36 patients with detectable P/LP mosaic variants targeting CSGs revealed biallelic inactivation of the respective CSG in 91.7% of cases, and that the tumors displayed pathologic and/or genomic features consistent with inactivation of the respective CSG. Given the identification of the P/LP mosaic variants in tissues derived from mesoderm, endoderm and ectoderm, the most parsimonious explanation is that the mutational process resulting in mosaicism occurred early in embryogenesis, before gastrulation, when the different primary germ layers are established (22). This finding is consistent with their detection in our initial screen, which required the mosaic variants to be present in both blood (mesoderm-derived) and in tumors (mesoderm-, ectoderm- or endoderm-derived). Our findings not only confirm previous reports of early mosaic variants in healthy human and mouse tissues (7,23,24), including those affecting cancer genes (3,4,6), but also provide a causative link between mosaic variants arising in early embryogenesis and seemingly sporadic cancers.

Through analyses of normal tissues of different lineages, we observed a likely asymmetry in the contribution of mosaic variants to different tissues, in agreement with studies reporting that cell contribution to postnatal tissues may not be symmetrical (7,23). These differences could stem from evolutionary bottlenecks, distinct rates of proliferation and involution taking place in embryogenesis and postnatally, and a potential lineage-dependent positive or negative selection effect of mosaic variants affecting CSGs.

Our study has important limitations. The approach we employed allowed only for the detection of patients with CSG mosaic variants present in blood, a tissue that differentiates early during embryogenesis, and our analysis was restricted to CSG included in the MSK-IMPACT panel. Hence, the approximate 1/1,000 frequency of CSG P/LP mosaic variants in cancer patients we identified, likely constitutes a conservative estimate of their impact on cancer predisposition. Furthermore, the validation of mosaic variants was restricted to the samples available for each case. The study of a wider spectrum of tissues would require post-mortem analyses with *a priori* identification of patients with somatic mosaic variants. Finally, owing to the FFPE nature of the samples, single cell sequencing, which

would constitute an orthogonal validation of the observations made, and would allow the individual phylogenetic inference per patient, ideal for the assessment of mutation timing given the vast inter-individual variability in terms of asymmetry we observed, and reported by others(18,19) could not be conducted. Despite these limitations, our study demonstrates that mosaic variants affecting CSGs can be detected using a clinical grade tumor/normal targeted sequencing assay, which is an additional advantage of this sequencing methodology in the clinical setting. Our study also provides a comprehensive analysis of mosaicism affecting CSGs in a large population of cancer patients and demonstrates that some of these individuals harbor mosaic variants in CSGs occurring in early embryogenesis that likely contributed to cancer development.

## METHODS

### Subjects and samples

This study was approved by MSKCC (NY, USA) Institutional Review Board (IRB), protocols 12–245 (Genomic profiling in cancer patients) and 19–154 (Prevalence of somatic mosaicism in advanced cancer patients). Written informed consent was obtained according to IRB protocols. De-identified tumor/blood MSK-IMPACT sequencing data of 35,310 cancer patients enrolled on the institutional protocol 12–245 (NCT01775072) who underwent sequencing between January 2014 and December 2019 were retrieved.

### Filtering criteria for detection of cancer causative mosaic variants

61/341 genes present across MSK-IMPACT versions were determined to be associated with increased cancer susceptibility and a dominant mode of inheritance. These 61 CSGs were analyzed for candidate mosaic variants, and the variant pathogenicity was reviewed by a board-certified molecular pathologist (DM) according to the American College of Medical Genetics and Genomics (ACMG) criteria (25).

To define candidate mosaic variants affecting CSGs (Supplementary Methods), given that mosaic variants are expected to have a VAF <50%, seen for heterozygous variants, and that the MSK-IMPACT sequencing resolution is limited at VAFs <1.5%, we selected variants with a blood VAF 1.5%–30%. To detect cancer causative mosaic variants that would therefore be enriched in tumors, we selected variants with a tumor VAF 10% and a tumor/blood VAF ratio 1.5, given that the individual is mosaic, but the cell giving rise to the tumor was heterozygote for the variant at the outset. To minimize artifacts, only reads with a mapping quality >20 and cases with a sequencing depth 50 for both tumor and blood samples with 6 reads of the alternate allele in blood were included. Detected variant had to have an allele frequency <1% in the Genome Aggregation Database (gnomAD; v2.1.1). In addition, we excluded variants in highly repetitive regions and variant calls with strand bias, as described (12). To distinguish mosaic variants from variants detected in CTCs present in the blood samples, cases with 1 variant in blood targeting CSGs or non-CSGs meeting the above criteria were excluded. To exclude circulating malignancies, variants from patients with hematologic malignancies were removed. To distinguish mosaic variants from CH, established CH criteria were applied (12).



The pathogenicity of mosaic variants targeting non-CSGs (n=280) with known gene-disease associations were determined according to the ACMG criteria.

### Immunohistochemistry

Expression of MSH2, MSH6 and CD45 was assessed by immunohistochemistry on a Bond-3 automated stainer platform (Leica Biosystems, Wetzlar, Germany; Supplementary Methods).

### Statistical analysis

Statistical analyses were conducted using Rv3.1.2. VAF 95% confidence intervals were calculated using the Wilson procedure. Mann-Whitney *U*-test and Fisher's exact test were used for continuous and categorical variables, respectively. *P*-values were two-sided and adjusted for multiple testing wherever appropriate. *P*-values<0.05 were considered significant.

### Data availability

Somatic mutations and copy number alterations for the 36 mosaic cases identified in this study are available on cBioPortal ([https://www.cbioportal.org/study/summary?id=mixed\\_mos\\_msk\\_2021](https://www.cbioportal.org/study/summary?id=mixed_mos_msk_2021)). Targeted sequencing data (aligned BAM files) of tumor and normal tissues included in the validation cohort supporting the findings of this study have been deposited in the Sequence Read Archive (SRA) under accession number SUB10801254.

### Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

### Acknowledgments:

This study was partially funded by the Breast Cancer Research Foundation and by the Sarah Leigh Fund. Research reported in this publication was partly funded by a Cancer Center Support Grant of the National Institutes of Health (NIH)/ National Cancer Institute (grant No P30CA008748). F.P. is partially funded by an NIH K12 CA184746 grant, and BW is partially funded by a Cycle for Survival grant. FP, RS, NR, BW and JSR-F are funded in part by the NIH/NCI P50 CA247749 01 grant. YK, KW, and ZS are partially supported by The Robert and Kate Niehaus Center for Inherited Cancer Genomics and the Sharon Corzine Research Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### Conflict of interest disclosures:

JSR-F reports receiving personal/consultancy fees from Goldman Sachs, REPARE Therapeutics and Paige.AI, membership of the scientific advisory boards of VolitionRx, REPARE Therapeutics, Paige.AI and Personalis, membership of the Board of Directors of Grupo Oncoclinicas, and ad hoc membership of the scientific advisory boards of Roche Tissue Diagnostics, Ventana Medical Systems, Novartis, Genentech and InVivo, outside the scope of this study. BW reports ad hoc membership of the scientific advisory board of REPARE Therapeutics, outside the scope of the submitted work. M.L. has received advisory board compensation from Merck, Lilly Oncology, Bristol-Myers Squibb, Takeda, Blueprint Medicines, Bayer, Janssen Pharmaceuticals, and Paige.AI, and research support from LOXO Oncology, Helsinn Healthcare, Elevation Oncology, and Merus. Z.K.S. reports that an immediate family member holds consulting/advisory roles in Ophthalmology with Allergan, Adverum, Genentech/Roche, Novartis, Optos, Regeneron, Regenxbio, Neurogene, Gyroscope. MR has received honoraria from Intellisphere, Physicians' Education Resource, and Research to Practice, consulted for or served on advisory boards for AstraZeneca (uncompensated), Change Healthcare, Daiichi-Sankyo (uncompensated), Epic Sciences (uncompensated), Merck (uncompensated), and Pfizer (uncompensated).has received research funding from AbbVie

(institution), AstraZeneca (institution), Invitae (institution), Merck (institution), and Pfizer (institution), travel reimbursement from Merck, and other transfers of value (editorial services) from AstraZeneca and Pfizer. All other authors declare no conflicts of interest.

## REFERENCES

1. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr., Kinzler KW. Cancer genome landscapes. *Science* 2013;339(6127):1546–58 doi 10.1126/science.1235122. [PubMed: 23539594]
2. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nat Rev Genet* 2013;14(5):307–20 doi 10.1038/nrg3424. [PubMed: 23594909]
3. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 2015;348(6237):880–6 doi 10.1126/science.aaa6806. [PubMed: 25999502]
4. Lawson ARJ, Abascal F, Coorens THH, Hooks Y, O'Neill L, Latimer C, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* 2020;370(6512):75–82 doi 10.1126/science.aba8347. [PubMed: 33004514]
5. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. *Science* 2018;362(6417):911–7 doi 10.1126/science.aau3879. [PubMed: 30337457]
6. Yizhak K, Aguet F, Kim J, Hess JM, Kubler K, Grimsby J, et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* 2019;364(6444) doi 10.1126/science.aaw0726.
7. Ju YS, Martincorena I, Gerstung M, Petljak M, Alexandrov LB, Rahbari R, et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* 2017;543(7647):714–8 doi 10.1038/nature21703. [PubMed: 28329761]
8. Verhoef S, Bakker L, Tempelaars AM, Hesselting-Janssen AL, Mazurczak T, Jozwiak S, et al. High rate of mosaicism in tuberous sclerosis complex. *Am J Hum Genet* 1999;64(6):1632–7 doi 10.1086/302412. [PubMed: 10330349]
9. Kehrer-Sawatzki H, Cooper DN. Mosaicism in sporadic neurofibromatosis type 1: variations on a theme common to other hereditary cancer syndromes? *J Med Genet* 2008;45(10):622–31 doi 10.1136/jmg.2008.059329. [PubMed: 18511569]
10. Mandelker D, Ceyhan-Birsoy O. Evolving Significance of Tumor-Normal Sequencing in Cancer Care. *Trends Cancer* 2020;6(1):31–9 doi 10.1016/j.trecan.2019.11.006. [PubMed: 31952779]
11. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, et al. Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. *J Mol Diagn* 2015;17(3):251–64 doi 10.1016/j.jmoldx.2014.12.006. [PubMed: 25801821]
12. Bolton KL, Ptashkin RN, Gao T, Braunstein L, Devlin SM, Kelly D, et al. Cancer therapy shapes the fitness landscape of clonal hematopoiesis. *Nat Genet* 2020;52(11):1219–26 doi 10.1038/s41588-020-00710-0. [PubMed: 33106634]
13. Kratz CP, Freycon C, Maxwell KN, Nichols KE, Schiffman JD, Evans DG, et al. Analysis of the Li-Fraumeni Spectrum Based on an International Germline TP53 Variant Data Set: An International Agency for Research on Cancer TP53 Database Analysis. *JAMA Oncol* 2021 doi 10.1001/jamaoncol.2021.4398.
14. Nilbert M, Therkildsen C, Nissen A, Akerman M, Bernstein I. Sarcomas associated with hereditary nonpolyposis colorectal cancer: broad anatomical and morphological spectrum. *Fam Cancer* 2009;8(3):209–13 doi 10.1007/s10689-008-9230-8. [PubMed: 19130300]
15. Vogt A, Schmid S, Heinimann K, Frick H, Herrmann C, Cerny T, et al. Multiple primary tumours: challenges and approaches, a review. *ESMO Open* 2017;2(2):e000172 doi 10.1136/esmoopen-2017-000172. [PubMed: 28761745]
16. Cybulski C, Nazarali S, Narod SA. Multiple primary cancers as a guide to heritability. *Int J Cancer* 2014;135(8):1756–63 doi 10.1002/ijc.28988. [PubMed: 24945890]
17. Lee-Six H, Obro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* 2018;561(7724):473–8 doi 10.1038/s41586-018-0497-0. [PubMed: 30185910]

18. Park S, Mali NM, Kim R, Choi JW, Lee J, Lim J, et al. Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* 2021;597(7876):393–7 doi 10.1038/s41586-021-03786-8. [PubMed: 34433967]
19. Coorens THH, Moore L, Robinson PS, Sanghvi R, Christopher J, Hewinson J, et al. Extensive phylogenies of human development inferred from somatic mutations. *Nature* 2021;597(7876):387–92 doi 10.1038/s41586-021-03790-y. [PubMed: 34433963]
20. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. *Nature* 2020;578(7793):94–101 doi 10.1038/s41586-020-1943-3. [PubMed: 32025018]
21. Rahbari R, Wuster A, Lindsay SJ, Hardwick RJ, Alexandrov LB, Turki SA, et al. Timing, rates and spectra of human germline mutation. *Nat Genet* 2016;48(2):126–33 doi 10.1038/ng.3469. [PubMed: 26656846]
22. Ghimire S, Mantziou V, Moris N, Martinez Arias A. Human gastrulation: The embryo and its models. *Dev Biol* 2021;474:100–8 doi 10.1016/j.ydbio.2021.01.006. [PubMed: 33484705]
23. Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, et al. Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 2014;513(7518):422–5 doi 10.1038/nature13448. [PubMed: 25043003]
24. Bae T, Tomasini L, Mariani J, Zhou B, Roychowdhury T, Franjic D, et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* 2018;359(6375):550–5 doi 10.1126/science.aan8690. [PubMed: 29217587]
25. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015;17(5):405–24 doi 10.1038/gim.2015.30. [PubMed: 25741868]

**STATEMENT OF SIGNIFICANCE**

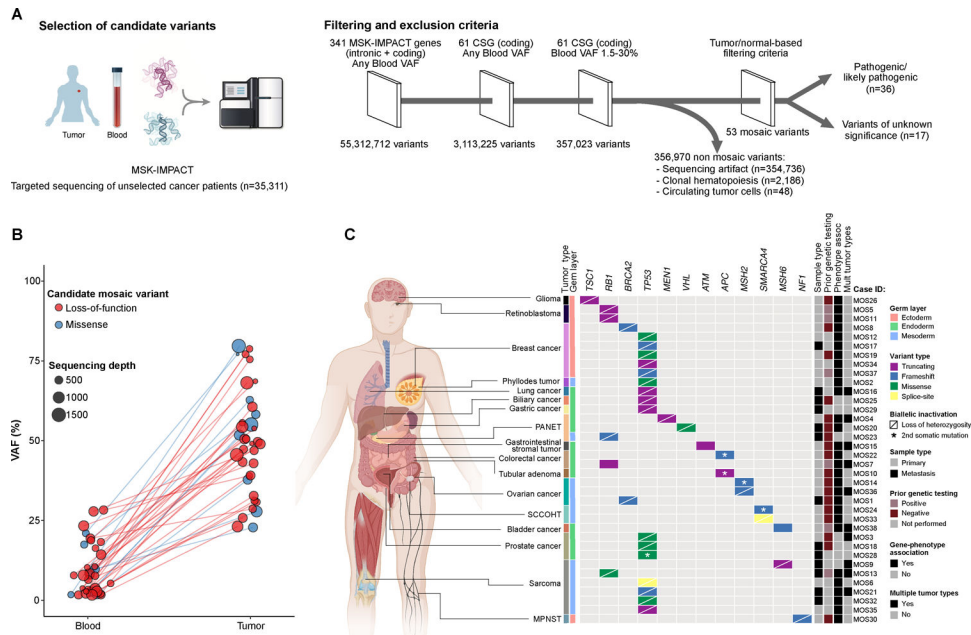
Here we demonstrate that mosaic variants in CSGs arising in early embryogenesis contribute to the oncogenesis of seemingly sporadic cancers. These variants can be systematically detected through the analysis of tumor/normal sequencing data, and their detection may impact therapeutic decisions as well as prophylactic measures for patients and their offspring.

Author Manuscript

Author Manuscript

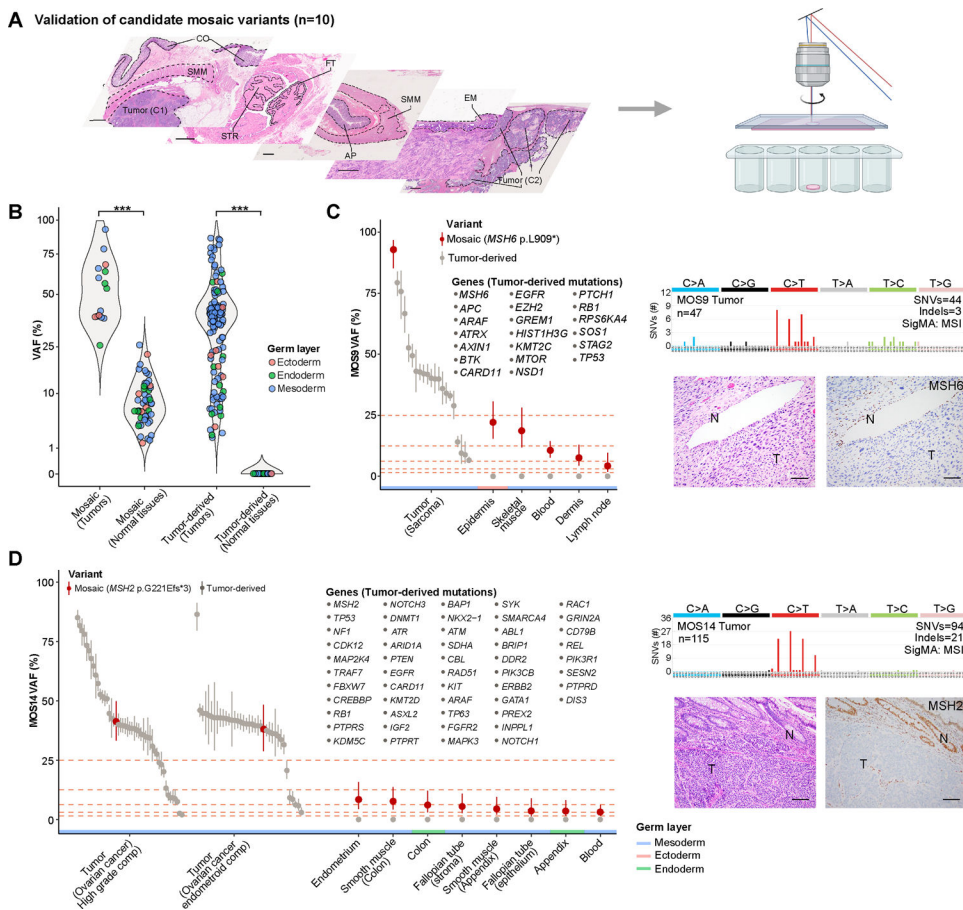
Author Manuscript

Author Manuscript



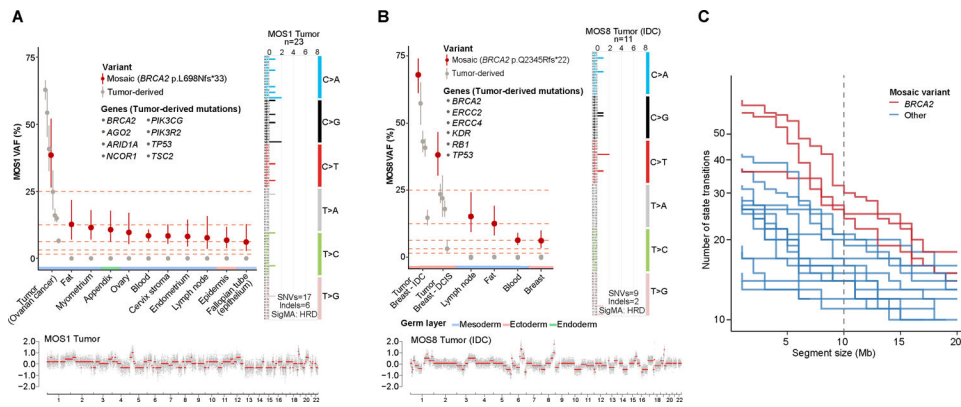
**Figure 1. Identification of cancer patients with candidate early mosaic variants.**

**A**, Schematic representation of the methodology for selection of patients with candidate mosaic variants in CSGs, sequencing methods, selection algorithm, filtering and exclusion criteria and selected variants. **B**, Variant allele fraction (VAF), sequencing depth and variant type of the candidate pathogenic/likely pathogenic (P/LP) mosaic variants identified by the set of filters in blood and tumor. **C**, Tumor type and germ layer derivation, CSG affected, variant type, biallelic inactivation of P/LP candidate mosaic variants (n=36). Phenobars (right) depict clinical characteristics. CSG, cancer susceptibility gene, PANET, pancreatic neuroendocrine tumor; MPNST, malignant peripheral nerve sheath tumor; SCCOHT, small cell carcinoma of the ovary hypercalcemic type.



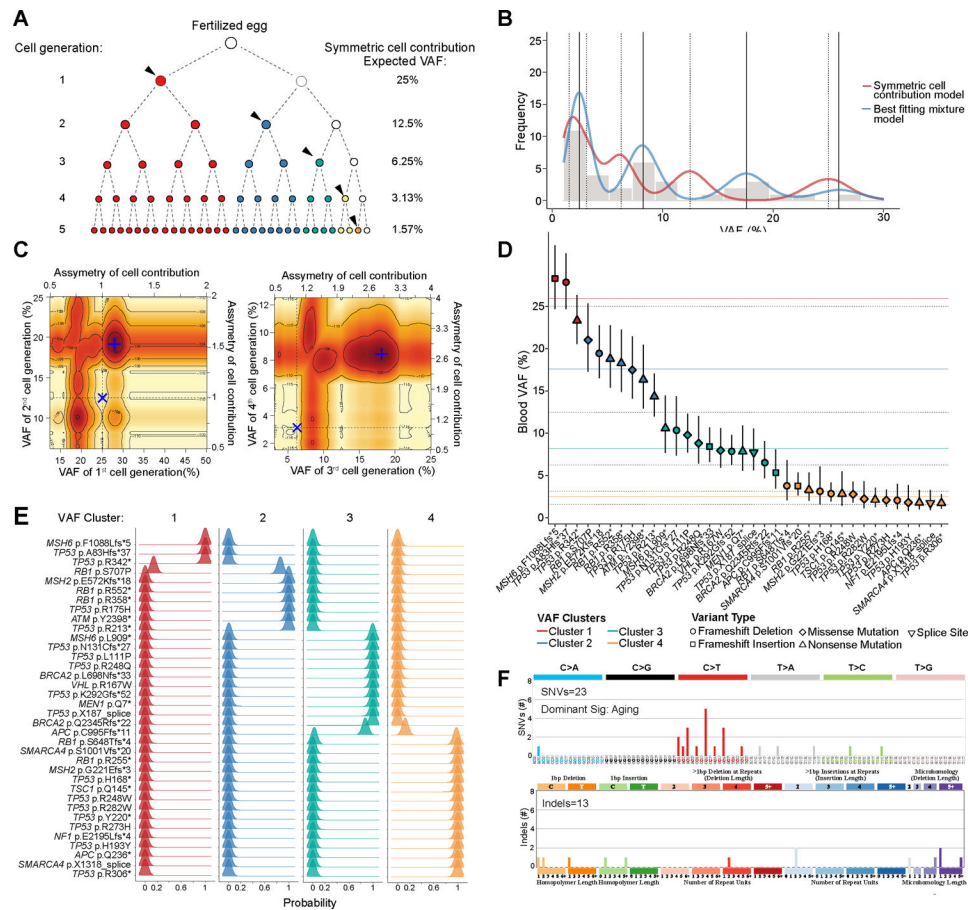
**Figure 2. Validation of candidate mosaic variants affecting mismatch repair genes by targeted sequencing.**

**A**, Schematic representation of the validation method of candidate mosaic variants and representative micrographs of hematoxylin and eosin (H&E)-stained tumors and normal tissues of case MOS14. Scale bars, 1 mm. **B**, Variant allele fraction (VAF) of the mosaic variants and tumor-derived non-synonymous somatic mutations in microdissected tumors and normal tissues according to germ layer. **C-D**, VAF of the mosaic variants affecting mismatch repair genes (red) and of tumor-derived non-synonymous somatic mutations (gray) in tumor and normal tissues and tumor mutational signatures. Error bars, 95% CI. Representative photomicrographs of H&E-stained slides and immunohistochemistry analysis of mismatch repair proteins are depicted. Scale bars, 50  $\mu$ m. AP, appendix; C1, histologic component 1; C2, histologic component 2; CO, colon; EM, endometrium; FT, fallopian tube epithelium; MSI, microsatellite instability; N, normal; SMM, smooth muscle; SNV, single nucleotide variation; T, tumor; VAF, variant allele fraction.



**Figure 3. Validation of candidate mosaic variants affecting homologous recombination deficiency genes by targeted sequencing.**

**A-B**, Variant allele fraction (VAF) of the mosaic variants affecting *BRCA2* (red) and of tumor-derived non-synonymous somatic mutations (gray) in tumor and normal tissues and tumor mutational signatures. Copy number plots depicting segmented  $\text{Log}_2$  ratios (y-axis) according to genomic position (x-axis). **C**, Number of state transitions according to segment size. Each line corresponds to a tumor (n=15) of the 10 individuals in the validation cohort. DCIS, breast ductal carcinoma in situ; HRD, homologous recombination deficiency; IDC, breast invasive ductal carcinoma; SNV, single nucleotide variation; VAF, variant allele fraction.



**Figure 4. Timing of occurrence and mutational processes of mosaic variants.**

**A**, Schematic representation of the first cell divisions of embryogenesis along with the expected variant allele fraction (VAF) of mosaic variants per cell generation in adult normal tissues assuming a symmetrical cell contribution. **B**, VAF distribution of the 36 CSG mosaic variants. Expected VAF distribution from symmetric cell contribution (red line) and best fitting cell contribution mixture model (blue line). **C**, Contour plots depicting the Log likelihoods of symmetric and asymmetric cell contributions. The x axis and y axis display the expected VAFs of the first and second cell divisions, respectively (left), and of the third and fourth cell divisions, respectively (right), given different cell contribution asymmetry levels (right x and top y axes). The dotted lines represent the expected VAFs of the respective cell divisions as per a symmetric cell contribution model. (X), symmetric model; (+), best fitting asymmetric model. **D**, Assignment of the 36 mosaic variants to the VAF clusters of the best fitting mixture model. The means of four VAF clusters are shown in red (cluster 1), blue (cluster 2), green (cluster 3) and orange (cluster 4). The expected VAF of mosaic variants occurring in the first five cell generations assuming a symmetric cell contribution are shown as black dotted lines. Error bars, 95 CI. **E**, Posterior probabilities of the 36 mosaic variants to belong to each of the four VAF clusters identified using the Beta-Binomial mixture model. **F**, Single nucleotide variant (SNV) and indels mutational signatures of the 36 mosaic variants identified.