TITLE

Extended family with germline pathogenic variant in polymerase delta provides strong evidence for recessive effect of proofreading inactivation

AUTHORS

Maria A. Andrianova[1,2†], Vladimir B. Seplyarskiy[3,4†], Mariona Terradas[5], Ana Beatriz Sánchez-Heras[6,7], Pilar Mur[5], José Luis Soto[6,8], Gemma Aiza[5], Fyodor A. Kondrashov[9,10], Alexey S. Kondrashov[11], Georgii A. Bazykin[1,2], Laura Valle[5,12*]

[†]These authors contributed equally

[*]Corresponding authors

AFFILIATIONS

[1]Skolkovo Institute of Science and Technology, Skolkovo, Russia

[2]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevitch Institute), Moscow, Russia

[3]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

[4]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[5]Hereditary Cancer Program, Catalan Institute of Oncology; Oncobell Program, IDIBELL, Hospitalet de Llobregat, Barcelona, Spain

[6]Foundation for the Promotion of Health and Biomedical Research of Valencia Region (FISABIO), Elche Health Department, Elche, Spain

[7]Medical Oncology Department, Cancer Genetic Counseling Unit. Elche University Hospital, Elche, Spain

[8]Molecular Genetics Unit, Elche University Hospital, Elche, Spain

[9]Institute of Science and Technology Austria, 1 Am Campus, 3400, Klosterneuburg, Austria

[10]Evolutionary and Synthetic Biology Unit, Okinawa Institute of Science and Technology

Graduate University, Japan

[11]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA

[12]Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain

ABSTRACT

Mutational processes in germline and in somatic cells are vastly different, and it remains unclear how the same genetic background affects somatic and transmissible mutations. Here, we estimate the impact of a germline pathogenic variant in the exonuclease domain of polymerase delta (Polδ) on somatic and germline mutational processes and cancer development. In germline cells and in non-cancer somatic cells, the *POLD1* L474P variant only slightly increases the mutation burden, contributing to ~11.8% and ~14.7% of mutations respectively, although it strongly distorts the mutational spectra. By contrast, tumors developed by carriers of germline pathogenic variants in *POLD1* harbor a DNA rearrangement that results in a homozygous state of the pathogenic variant, leading to an extremely high mutation rate. Thus, Polδ proofreading dysfunction has a recessive effect on mutation rate, with mutations in both *POLD1* alleles leading to a dramatic rate of mutation accumulation and cancer development. These results clarify the link between the effect of *POLD1* mutator variants on germline and somatic replication, and, together with previous findings, illustrate the important differences in disruption of replication fidelity caused by mutations in main replicative polymerases.

INTRODUCTION

Changes in DNA sequence can arise due to errors during replication or due to DNA damage caused by intracellular or environmental factors. Mutation rate and patterns are drastically different between somatic and germline cells[1–3], likely due to a different exposure to external and internal mutagens or possible differencies in proficiency of DNA repair. The relative contribution of replicative and non-replicative errors is a matter of debate for both somatic and germline mutations[4]. In germline cells, the number of germline mutations inherited from the father is correlated with the father's age, which has been interpreted as a high impact of replicative errors[5–8], although this interpretation has been questioned recently[9]. In cancerous cells, most mutations result from exposure to exogeneous and endogeneous factors[2,10], and a significant contribution of replication-associated mutations is only observed in cells with deficiency in proofreading of replicative polymerases or replication-associated mismatch repair[11–15].

Cells with deficiencies in systems controlling replication fidelity have a high potential for the study of the impact of replicative mutations. First, they are characterized by pronounced and characteristic mutational spectra[15]. Second, the accumulated mutations are believed to depend strongly on the number of cell divisions. The strong conservation of eukaryotic replication mechanisms[16,17] suggests that such deficiencies likely affect both germline and soma in a similar way. To understand the impact of replication errors and the effect of a particular germline variant that causes decreased replication fidelity on inheritable and somatic mutations simultaneously, we studied the mutation rate and spectra in an extended family harboring the germline *POLD1* c.1421T>C p.(Leu474Pro) pathogenic variant (from herein on, L474P) which affects the proofreading function of Polδ polymerase. We showed that L474P contributes a similar proportion of mutations to germline cells and fibroblasts, slightly increasing the overall mutation burden but significantly disturbing mutational spectra.

Germline pathogenic mutations affecting the proofreading activity of Polδ cause a multi-organ tumor predisposition syndrome, where colorectal adenomatous polyposis, colorectal and endometrial cancers are the most prevalent phenotypic features[18]. Cancer development in the carriers is thought to be associated with an increased somatic mutation rate. However, we and others have shown that heterozygous inactivation of *POLD1* leads to just a modest increase in mutation rate[19]. A deeper investigation of the cancer samples developed by patients harboring

germline *POLD1* cancer-predisposing variants helped us solve this paradox.

RESULTS

Extended family with the germline pathogenic variant in *POLD1*

In the family investigated in this study, *POLD1* L474P was identified in eight members, three of whom had been diagnosed with colorectal cancer (ages at cancer diagnosis: 23-50), and one with endometrial cancer at age 58. Most L474P carriers were diagnosed with gastrointestinal polyps, some of them with a clear attenuated adenomatous polyposis phenotype (10-100 adenomas) (Fig. 1a, Supplementary Table 1). It is intuitive to assume that the *POLD1* L474P genotype increases the risk of cancer due to a higher mutation rate in somatic tissues caused by the polymerase proofreading defect. However, the mutagenic effect of this variant in heterozygotic state in human normal tissues has been recently shown to be modest[19]. Nevertheless, a mutator phenotype was observed when the homologous residue was substituted in *Saccharomyces cerevisiae*[20].

Influence of germline *POLD1* L474P on somatic mutagenesis

To estimate the influence of *POLD1* L474P on somatic mutagenesis, we sequenced single-cell derived colonies obtained from immortalized fibroblasts (origin: skin punches) from eight family members (six carriers of *POLD1* L474P and 2 non-carriers) (Fig. 1a). The mutations accumulated in the fibroblasts of the individuals over their lifetime had at average of 9705 mutations caused by ultraviolet (UV) radiation, represented by the COSMIC single base substitution signature 7 (SBS7) and a high proportion of CC>TT double substitutions (Fig. 2a, Supplementary Fig. 1). Interestingly, the amount of UV-induced mutations identified was higher than previously reported for skin fibroblasts[21] and more in line with numbers observed for melanocytes[22]. In three out of six carriers of *POLD1* L474P, we observed the SBS10c signature corresponding to Polδ proofreading deficiency; the remaining three samples had a high load of UV-induced mutations in fibroblasts, which probably masked the signal of Polδ deficiency.

To study the effect of *POLD1 L474P* on the somatic mutation rate directly, we designed a controlled experiment for six out of the eight fibroblast colonies (for individuals III.2, IV.2, IV.3, III.4, IV.4 and IV.5, Fig. 1a). The cell lines were grown for ~40 passages after single cell isolation (Supplementary Table 2), and genome sequencing was performed on the DNA obtained from the cultured cells at the endpoints of the experiment (Fig.1b, Methods). The detection of mutations accumulated during the passages was performed using a standard procedure for somatic single nucleotide calling (Fig. 1c, Methods). While the number of mutations accumulated during the experiment in the fibroblasts harboring *POLD1* L474P was higher than that in the wild-type fibroblasts, for most samples this increase was relatively minor (Fig. 2b). However, the *POLD1* L474P status was associated with a radical shift in the mutational spectrum. *De novo* extraction and decomposition of mutational signatures in the endpoint fibroblasts from the four carriers and the two non-carriers of the *POLD1* L474P heterozygous pathogenic variant yielded six COSMIC signatures. In all four *POLD1* L474P carriers, we observed a strong enrichment in mutations attributed to SBS10c (Fig. 2c), a signature associated with Polδ proofreading deficiency[19]. In fact, the excess of mutations in *POLD1* L474P fibroblasts could be mostly attributed to SBS10c (Fig. 2d). This signature was not detected in the cultured fibroblast from the two non-carrier family members (Fig. 2c).

To orthogonally assess the effect of *POLD1* L474P on mutational spectrum, we calculated the 96-context mutational spectra for the fibroblasts with and without *POLD1* L474P, and applied principal component analysis (PCA) to these spectra (Fig. 2f). PCA clearly separated the carriers of *POLD1* L474P from non-carriers: PC1 explained ~48% of the variance in the mutational spectrum across samples, and reflected the experimentally obtained difference in spectra between

POLD1 mutated and non-mutated individuals (Fig. 2e). Comparison of principal components with COSMIC mutational signatures showed that the PC1 had the strongest correlation with SBS10c (cosine similarity = 0.52). PC1 also separated the previously published samples of somatic tissues from individuals with pathogenic germline variants in *POLD1* (Fig. 2g). Moreover, the values of PC1 differed between samples with different pathogenic variants in *POLD1:* S478N showed the highest PC1 values, and D316N, the lowest. This is consistent with previous findings that showed that S478N had the strongest effect on *POLD1* proofreading activity and caused the highest mutational burden, while this effect was lower for D316N[19].

Influence of germline *POLD1* L474P on germline mutagenesis

Next, we studied the impact of *POLD1* L474P on mutation patterns in the germline. For this, we sequenced four sub-families from the studied family, each including two parents and one to three children, comprising a total of eight offspring (Fig. 1a). In two sub-families *POLD1* L474P was carried by the father; in one sub-family, by the mother; and in the other sub-family, both parents had wild-type *POLD1*. *De novo* mutations in each offspring were called using the standard GATK4 pipeline with additional filters (Methods), focused on identifying mutations present in the offspring but absent in both parents (Fig. 1d). The number of called *de novo* mutations per individual varied between 24 and 88 (Fig. 3a).

Theoretically, *POLD1* L474P should affect the number of replication errors. However, based on the obtained fibroblast data, we expected this effect to be minor, with a no more than ~20% increase in the overall number of mutations in the offspring due to the *POLD1* L474P genotype. Because of the ~10-fold higher number of cell divisions in spermatocytes compared to oocytes[23], one would also expect a larger increase in the mutation rate in male gametes. Our limited data showed no significant effect of *POLD1* L474P on the burden of *de novo* mutations, when either the father or the mother carried the variant (Fig. 3a). We next performed an analysis of mutational spectra, which would be better powered, because *POLD1* L474P produces mutations in very specific contexts and its spectrum is well described by four main mutation types: Cp<u>C</u>pT>A, Tp<u>C</u>pT>A, Ap<u>T</u>pT>A and Cp<u>T</u>pT>G (Fig. 2e). The proportion of *de novo* mutations in these contexts among all 96 contexts estimated in previously published trios[24,25] was ~2.4%. Given this and the total number of observed *de novo* mutations per offspring of fathers with *POLD1* L474P variant, we expected a total of 5.7 mutations in these contexts for 4 offspring. Instead, we observed 28 such mutations (11.8% of all mutations), suggesting a ~4.9 fold increase (rate ratio test p-value = 6.619e-05, Fig. 3b). These observations indicate that in spermatocytes, the fraction of mutations caused by replicative errors introduced by *POLD1* L474P is comparable to that observed in fibroblasts (14.7% in soma and 11.8% in germline). By contrast, in the offspring of *POLD1* L474P mothers, the fraction of mutations in the four contexts mostly affected by *POLD1* L474P was practically the same as in the wild-type trios (Fig. 3b, Supplementary Fig. 2), indicating that this genotype is responsible for just 3.1% of all mutations in oocytes of those mothers.

Consistently, the projection of mutational spectrum of *de novo* mutations in the offspring of *POLD1* L474P fathers to -PC1 showed a significantly higher value compared to the offspring of mothers with *POLD1* L474P and to the offspring of wild-type parents (KS test p-value = 0.0019) (Fig. 3c and 3e). Altogether, our analyses demonstrate that the presence of a mutator allele could be more easily detectable from changes in the mutational spectra than from an increase in the overall mutation burden (Fig. 3d-f). In particular, and based on our limited data, the offspring of fathers harboring *POLD1* L474P could not be distinguished from the offspring of wild-type parents based on the number of *de novo* mutations (Fig. 3d). However, *de novo* mutations in *POLD1*-specific mutational contexts were clearly enriched in the offspring of *POLD1* L474P fathers compared to the offspring of wild-type parents (Fig. 3e-f).

Recently, pathogenic variants in known DNA repair genes were shown to contribute to germline hypermutation in sequenced human trios[26]. Could the offspring of parents with disrupted Polδ proofreading activity contribute to the public datasets of sequenced families? To study this, we mapped onto our PCA coordinates the mutational spectra of 6,233 offspring belonging to 4,638 families with assumedly wild-type *POLD1* status. Although the mean value of -PC1 in this sample was significantly lower than that in the offspring of *POLD1* L474P fathers, we found that 138 (2.2%) of the offspring had –PC1 above this value (Fig. 3c). The observed proportion of samples with a high –PC1 value could indicate the presence of undetected Polδ deficiency in parents of these trios; alternatively, it could arise for stochastic reasons. To distinguish between these alternatives, we generated an artificial dataset of trios by randomly sampling mutations according to their fractions in the spectra of *de novo* mutations in offspring of wild-type parents. Among these simulated trios, 2.0% of the offspring had values of -PC1 above the mean value observed in the five offspring of the two *POLD1* L474P fathers studied here. The presence of such cases in our simulated dataset indicates that a high value of -PC1 in some of the trios could be obtained by chance, and a comparable proportion of such cases in the data and in the simulation suggests either that *POLD1* mutated fathers are absent in the dataset or their fraction is smaller than the resolution of this method. To estimate the fraction of fathers with *POLD1* L474P that would detectably change the distribution of -PC1 values in the sample, we generated additional synthetic datasets by mixing simulated offspring with mutations sampled from the *de novo* spectrum of wild-type parents with small fractions of simulated offspring with mutations sampled from the spectra of offspring of *POLD1* L474P fathers. For the 1% admixture, the proportion of samples with -PC1 above the mean in offspring of *POLD1* L474P fathers was 2.5%, similar to simulation with wild-type parents only. For the 5% admixture, the distribution had a pronounced shift to higher -PC1 values, and the corresponding proportion was 3.98% (a 1.8-fold increase compared to the wild-type simulation) (Fig. 3h). We can thus conclude that the presence of a mutation rate modifier is detectable in a sample of trios based on the heterogeneity in the mutational spectrum in the offspring if the fraction of parents carrying the modifying mutation is substantial (e.g. ~5% for *POLD1* L474P).

Sequencing more than one child in a family should provide more power to detect families with disrupted Polδ, as it is unlikely to obtain high values of -PC1 in multiple offspring by chance alone. By simulating multiple offspring per family with wild-type parents, we estimated that only 0.16% of the quartets with two sequenced siblings, and 0.04% of the quintets with three sequenced siblings, are expected to have the -PC1 value averaged between siblings higher than the mean -PC1 value in the offspring of fathers harboring *POLD1* L474P (Supplementary Fig. 3). This confirms that sequencing more than one child per family decreases the rate of false positive prediction of a mutation rate modifier presence in the father, thus increasing the power of this approach. Among the 4,638 analyzed families, 1,555 were quartets with two sequenced siblings. Among them, the observed fraction of families with increased -PC1 value averaged between siblings (0.19%) was similar to the 0.16% estimated from the simulation, again arguing against a detectable contribution of families with Polδ deficiency to the studied dataset.

Inactivation of *POLD1* proofreading in both alleles is essential for cancer development

Inherited heterozygous mutations in the exonuclease domain of *POLE* or *POLD1* cause a highly penetrant syndrome known as polymerase proofreading-associated polyposis (PPAP) characterized by numerous polyps in the colon and an increased risk of colorectal cancer, endometrial cancer, and other tumor types[27]. *POLD1* L474P is classified as pathogenic following the ACMG/AMP variant classification guidelines for PPAP[28], and, as expected, co-segregates with colorectal cancer, endometrial cancer and gastrointestinal benign tumors in the studied family. The proofreading function of *POLE* and *POLD1* was previously proposed to be haploinsufficient[18].

However, unlike *POLE* pathogenic variants which strongly affect germline and somatic mutation rates even in heterozygous state[19], we and others[19] found only a minor effect of heterozygous *POLD1* L474P on the mutation rate in non-tumoral tissues. To better understand how such a weak mutation rate modifier can drive a highly penetrant cancer phenotype[28–30], we sequenced a tumor sample from the same family (Fig. 1a; colorectal cancer developed by individual IV.1). In contrast to cultured fibroblasts, the sequenced tumor was characterized by an ultrahypermutable phenotype, with 226 mutations per megabase (Fig. 4a). The mutational spectrum of the tumor also differed from the spectrum identified in the cultured fibroblasts of the *POLD1* L474P carriers, and in phenotypically normal crypts from carriers of different germline *POLD1* pathogenic variants sequenced in another study[19]. Specifically, it was enriched in C>A mutations (~54% of observed mutations), particularly in TpCpA and TpCpT contexts (Fig.4b). This mutational spectrum perfectly matched SBS10d (cosine similarity = 0.96) found in hypermutable polyps from germline carriers of *POLD1* S478N variant[19]. We also sequenced and analyzed a colorectal cancer developed by a carrier of the pathogenic variant *POLD1* D316H, which affects a catalytic site of the Polδ exonuclease. This tumor was also characterized by a high mutation rate (96.4 mutations/Mb) and prevalence of SBS10d (Fig. 4b). Thus, all three of these samples with an extremely high mutation rate were characterized by a similar shift in the mutational spectrum, suggesting a mutational process different from the mild mutagenesis observed in normal tissues of individuals with heterozygous *POLD1* mutations. The two analyzed tumors were microsatellite stable and had normal expression of mismatch repair (MMR) proteins MLH1, MSH2, MSH6 and PMS2, excluding inactivation of this DNA repair system. Tumors with dysfunctional MMR and an error-prone version of polymerase delta usually have spectra that correspond to mutational signature SBS20[13,31], which was not observed here.

To explain this discrepancy, we considered the possibility of somatic deactivation of the exonuclease function in the second copy of *POLD1* in our samples. Indeed, in all three studied samples with elevated mutation rate, we detected a copy neutral loss of heterozygosity (cnLOH), whereby the loss of the wild-type *POLD1* allele resulted in homozygosity of the pathogenic *POLD1* variant in tumor cells (Fig. 4d, Supplementary Fig. 4). These observations suggest a high level of haplosufficiency of *POLD1* in human cells. Our findings show that the presence of a single copy of wild-type *POLD1* can prevent a strong increase in the mutation rate. In line with our results, haplosufficiency of *POLD1* was recently shown in yeast experiments[32]. The proposed explanation for haplosufficiency is the ability of wild-type Polδ to proofread mismatches extrinsically, i.e., those produced by other mutant enzymes, thus preventing hypermutability. Our observations in normal and cancer tissues from heterozygous germline *POLD1* variant carriers are concordant with observations in yeasts and suggest a capability of Polδ to proofread mismatches extrinsically in human cells.

To identify and characterize the epigenetic covariates for the extrinsic proofreading effect of Polδ, we compared the relationship between replication timing and mutation rate in isolated colon crypts with homozygous and heterozygous *POLD1* S478N (data obtained from published dataset[19]). Both in hetero- and homozygous *POLD1*-mutated samples, mutation rates strongly depend on replication timing (Supplementary Fig. 5). However, in heterozygous samples, this dependence is more pronounced, suggesting that the extrinsic proofreading effect of *POLD1* is stronger in early replicating regions.

DISCUSSION

Taken together, our findings indicate that germline heterozygous mutations in the exonuclease domain of *POLD1* have a mild effect on somatic and germline mutation rates, although its carriers are predisposed to develop highly mutable cancers. This suggests that the capability of wild-type

Polδ (produced by the wild-type allele) to proofread the mismatches produced by mutated Polδ molecules (produced by the mutated allele) prevents hypermutability in cells with heterozygous *POLD1* pathogenic variants. In turn, the absence of a wild-type copy of the exonuclease of *POLD1* prevents both intrinsic and extrinsic proofreading of mismatches, drastically increasing the mutation rate (Fig. 5).

Our observations are well suited to explain the prevalence of different repair deficiencies in cancer. Intuitively, a high mutation rate is a selected phenotype during cancer development. Here, we found that a heterozygous mutation in the exonuclease domain of *POLD1* leads to just a minor increase in mutation rate if other repair systems are intact. Thus, heterozygous *POLD1* mutations should be almost neutral on a normal genetic background. Indeed, to our knowledge, there are no reported tumors where *POLD1* is somatically mutated but MMR is intact. Meanwhile, somatic inactivation of the second copy of *POLD1* on the background of a heterozygous germline mutation is under strong positive selection and is a major avenue for cancer development. We could expect somatic MMR inactivation as an alternative mechanism; however, MMRd is rarely found in cancers from individuals with inherited heterozygous *POLD1* mutations (Supplementary Table 3). This probably reflects the higher odds to mutate the second copy of *POLD1* compared to achieving MMRd. Thus, while most tumors developed by carriers of germline *POLD1* pathogenic variants have intact MMR, tumors with somatic pathogenic mutations in *POLD1* seem to be always accompanied by MMR deficiency (MMRd). In these tumors, the presence of a mutational signature of MMRd in addition to the mutational signature of simultaneous inactivation of *POLD1* proofreading and MMR indirectly suggests that MMR inactivation preceded the acquisition of *POLD1* mutation (Supplementary Fig. 6). MMRd alone increases the mutation rate and represents an advantageous genotype; however, after a clonal expansion of MMRd, *POLD1* mutations will further increase the mutation rate by an order of magnitude (Fig. 4f) and thus should also be under positive selection. These two examples demonstrate non-additive effects of repair-disrupting variants on mutation rate and could explain why there are no tumors with only a heterozygous *POLD1* mutation, but mutation of the second *POLD1* allele or an MMRd background[15] are required for positive selection.

Overall, in this study, we showed that heterozygous germline *POLD1* L474P has a minor effect on mutation burden in germline and soma but leads to a prominent change in mutational spectra. Sequencing a large number of trios with mutations in genes involved in replication could help better understand the mutational mechanisms in the germline and identify the relative roles of replicative errors and DNA damage.

We also uncovered a recessive effect of polymerase delta proofreading deficiency on mutation rate in the context of cancer, suggesting that a germline heterozygous *POLD1* exonuclease mutation may lead to cancer through a double hit mechanism, similarly to somatic inactivation of the second MMR gene copy in Lynch syndrome patients[33–35]. This and other findings suggest that the non-additive effect of polymerase proofreading and repair genes inactivation on mutation rate is translated into recessivity and/or epistatic interactions in cancer.

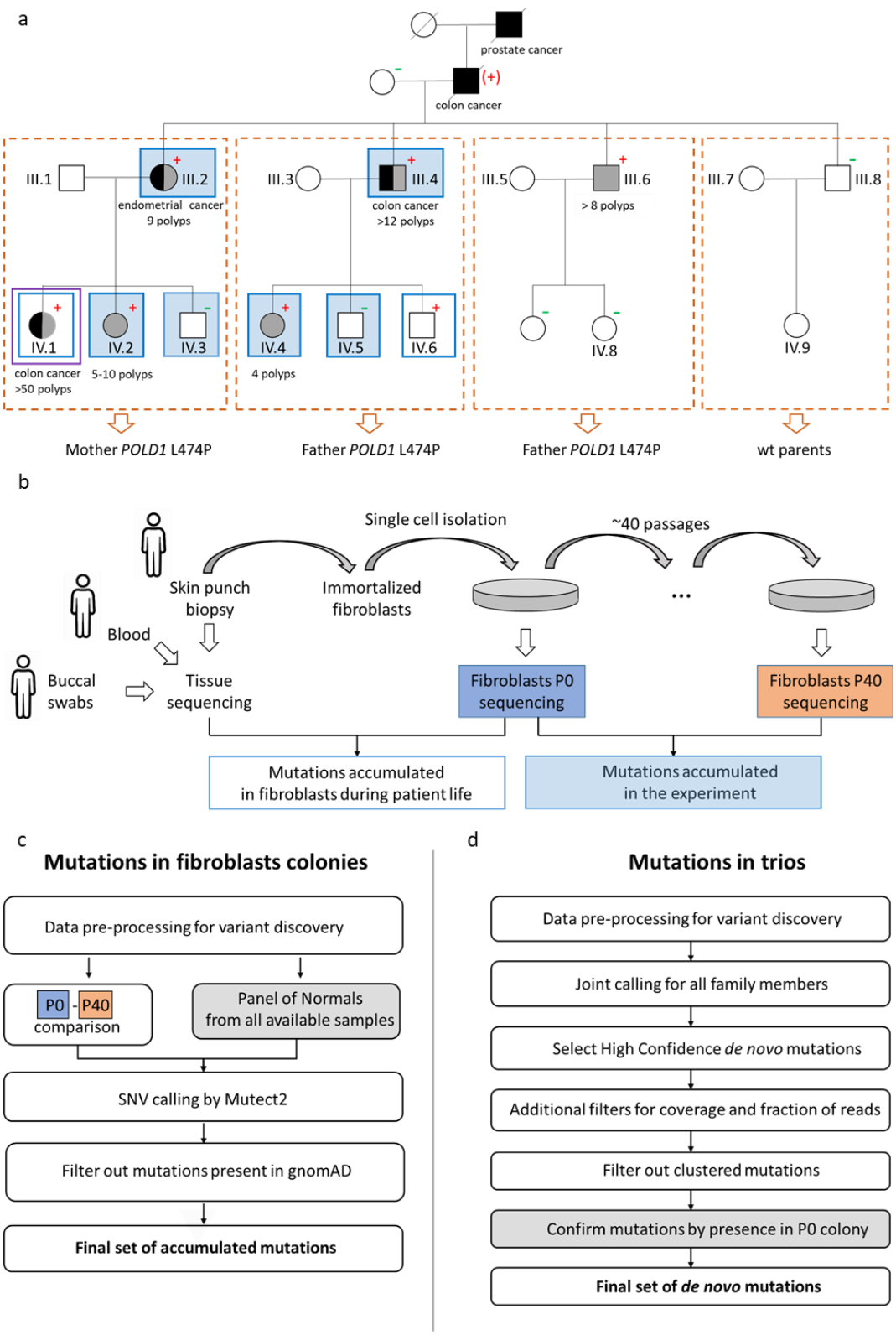

Fig. 1: Family, sample description, and working pipeline.

a, Family pedigree. Individuals diagnosed with cancer are marked in black, and with polyps, in gray. The tumor types and number of polyps are indicated below the corresponding individual symbol. Detailed tumor phenotypes and ages at diagnosis are shown in Supplementary Table 1. Blue squares mark the individuals from whom skin biopsies for single-cell fibroblast colonies were obtained. A violet square depicts the individual with a sequenced tumor sample. Sequenced sub-families are marked in orange with individual IDs shown for the members that were sequenced. Plus and minus signs mark the carrier status for *POLD1* L474P, and a plus sign between parentheses (+) indicates an obligate carrier. **b,** Scheme of the experiment performed on immortalized fibroblasts from family members for the assessment of somatic mutation accumulation. **c,** Algorithm used for calling somatic mutations accumulated during the growth of the cell lines. **d,** Algorithm used for calling germline *de novo* mutations.

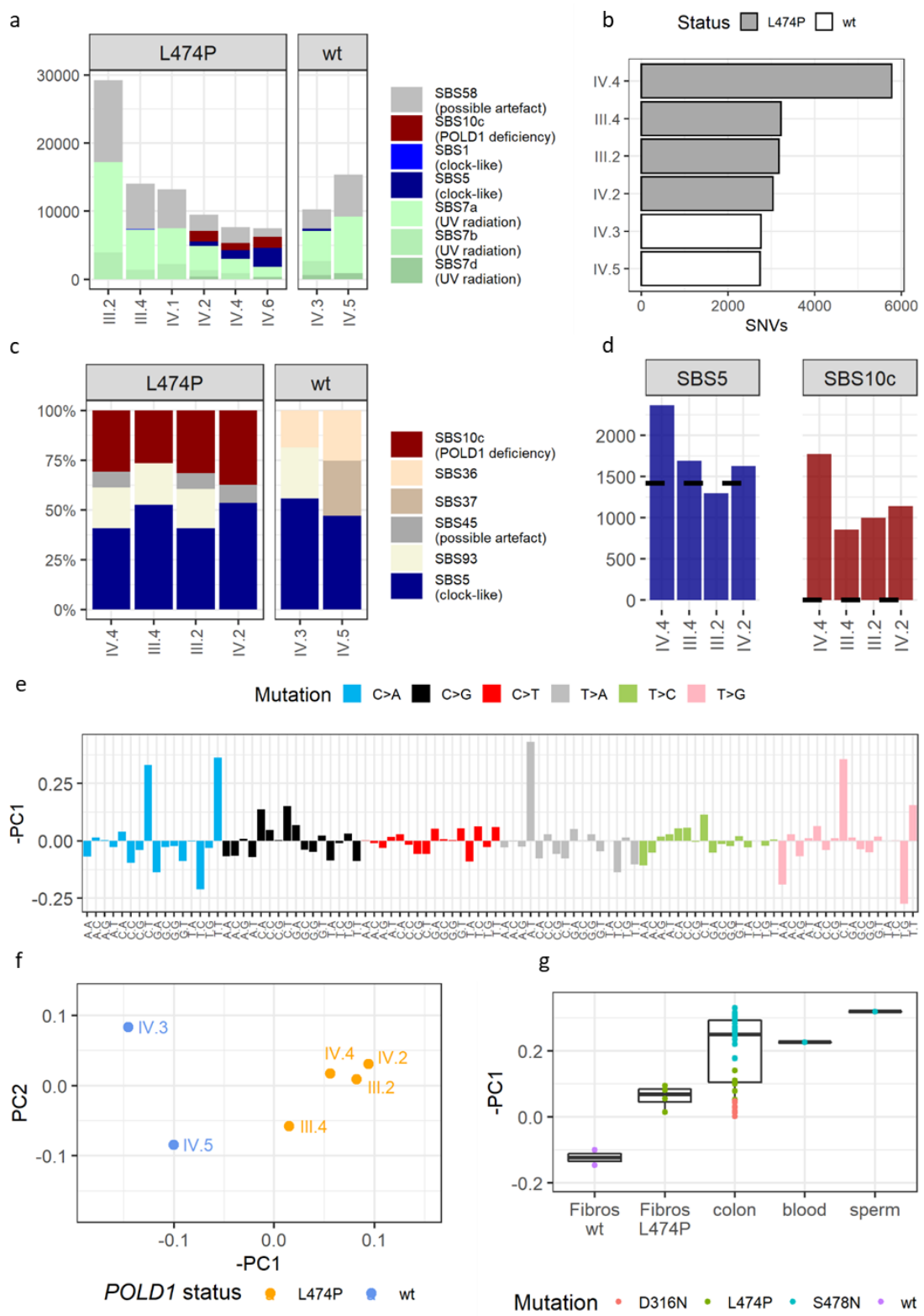

Fig. 2: Somatic mutation burden and spectrum. a, Number of mutations attributed to different COSMIC signatures observed in the sequenced fibroblasts. *De novo* signatures were extracted and decomposed to COSMIC signatures using SigProfilerExtractor. **b,** Number of SNVs accumulated during the experiment in the cultured fibroblasts of carriers and non-carriers of *POLD1* L474P. **c,** Proportion of mutations attributed to the different COSMIC signatures accumulated in the cultured fibroblasts. *De novo* signatures were extracted and decomposed to COSMIC signatures using SigProfilerExtractor. **d,** Excess of mutations corresponding to the different mutational signatures in the cultured fibroblasts of *POLD1* L474P heterozygous carriers. The black dashed lines mark the mean number of mutations attributable to the corresponding signature in wild-type fibroblasts. **e,** Loadings of mutational contexts in –PC1. **f,** PCA analysis of 96-context mutational spectra of mutations accumulated in the cultured fibroblasts. **g,** Values of the PC1 component in wt and L474P fibroblasts and in other normal tissues obtained from heterozygous carriers of *POLD1* pathogenic variants.[19] The different colors correspond to different *POLD1* pathogenic variants.

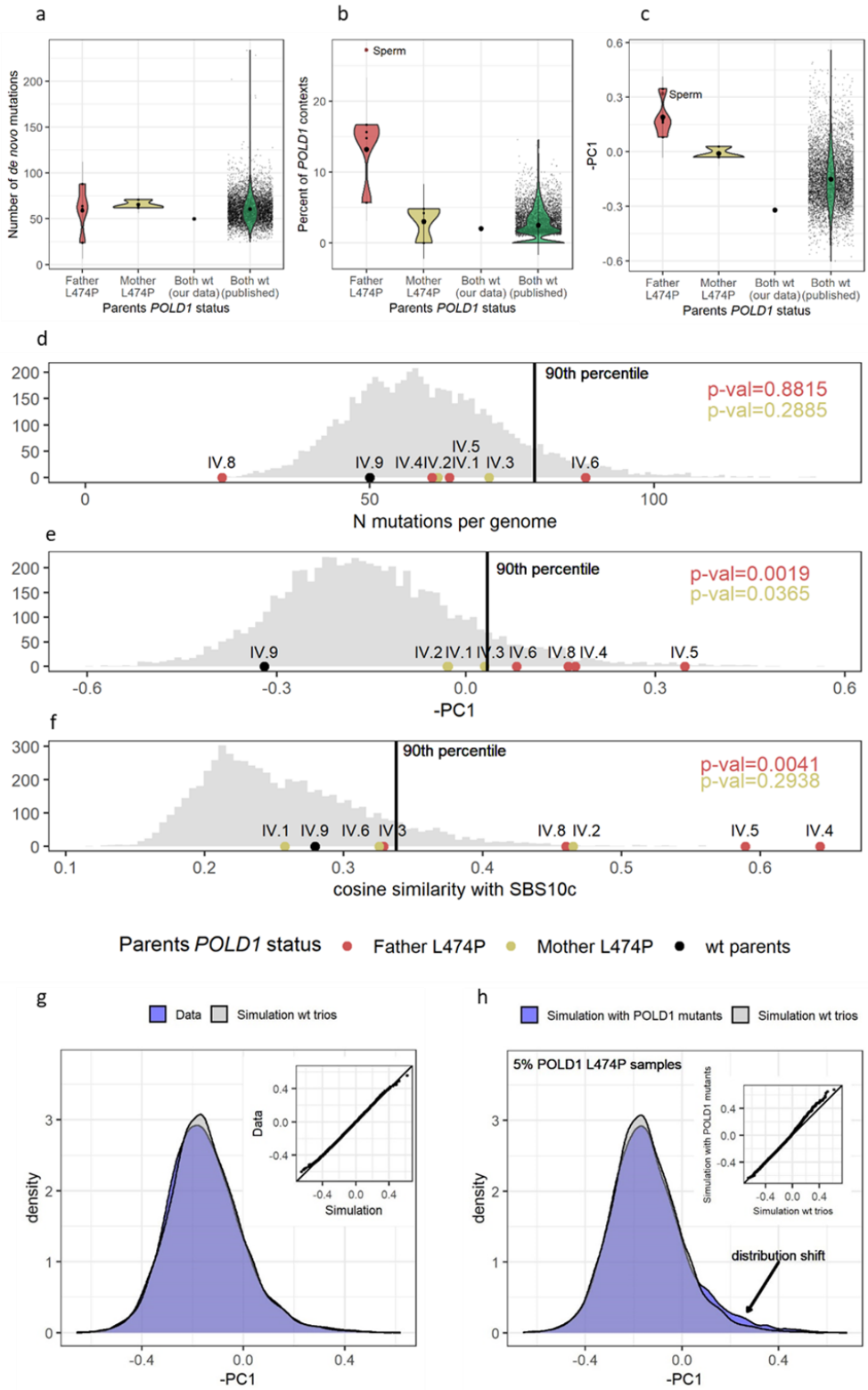

Fig. 3: Germline mutation burden and spectrum. a, Number of *de novo* mutations in offspring of parents with wild-type or mutated *POLD1.* **b,** Proportion of mutated *POLD1*-specific tri-nucleotide contexts in *de novo* mutations in offspring of parents with wild-type or mutated *POLD1.* **c,** Value of -PC1 in *de novo* mutations in offspring of parents with wild-type or mutated *POLD1.* In **b** and **c**, the red dot (sperm) corresponds to the sperm sample from an individual with a germline *POLD1* S478N variant previously published[19]. **d-f,** Representation of the number of mutations (**d**), -PC1 values (**e**), and cosine similarity with SBS10c signature (**f**) in sequenced trios from the family compared to the distribution of the corresponding parameter in publicly available trios[24,25]. The p-value for Kolmogorov-Smirnov test of public trios against fathers-carriers of *POLD1* variant is shown in red, and against mothers-carriers of *POLD1* variant is shown in khaki. **g,** Distribution of the -PC1 values in published and simulated *de novo* mutation spectra. **h,** Distribution of -PC1 values in simulations of *de novo* mutations with a mixture of offspring of *POLD1* L474P fathers.

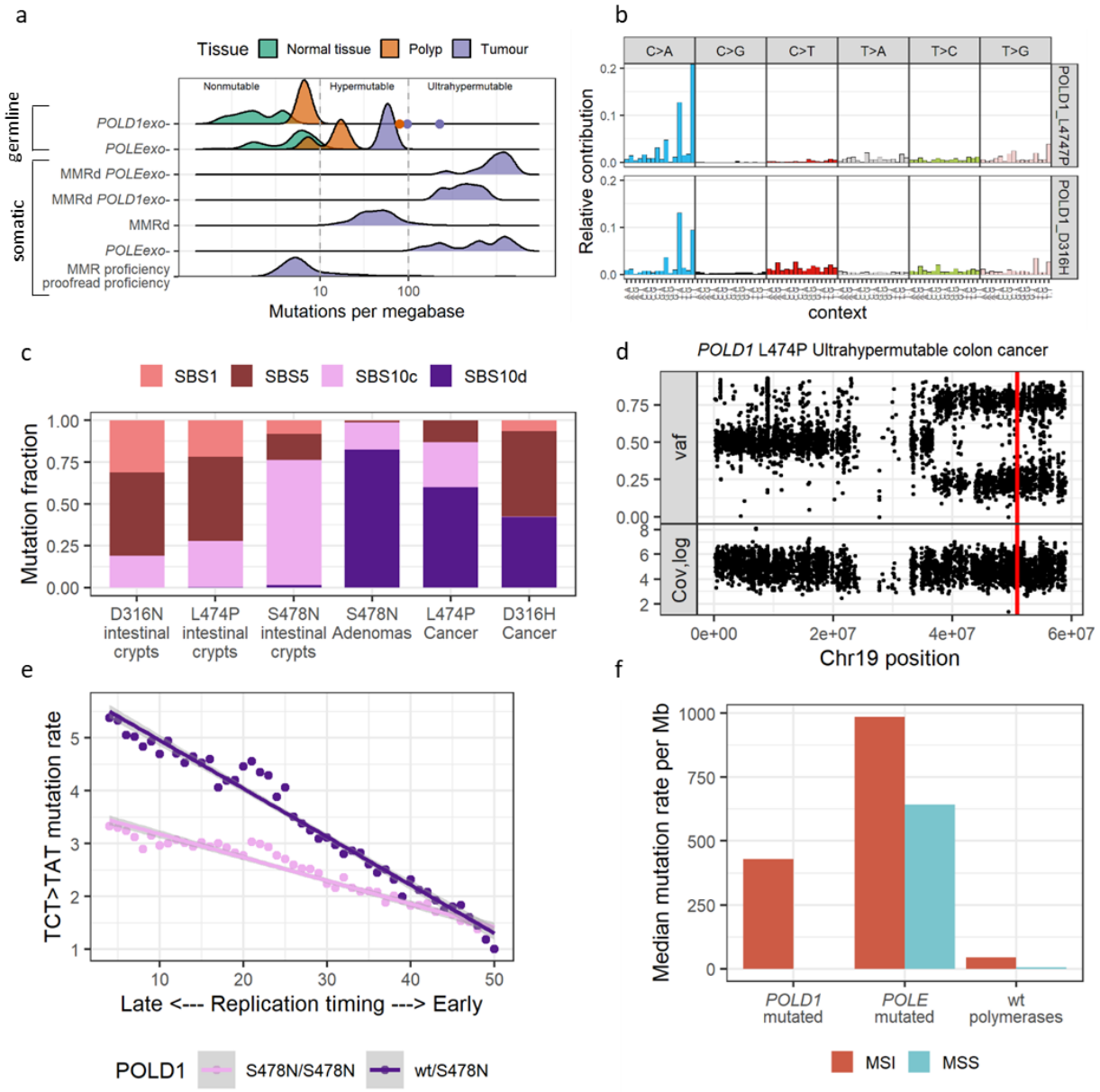

Fig. 4: Hypermutability of germline *POLD1* variants. a, Mutability of normal and cancer colon samples with germline mutations in the exonuclease domain of *POLE* or *POLD1* and their formal classification based on the number of mutations per megabase. Dots mark hypermutable and ultrahypermutable samples with germline pathogenic mutations in *POLD1*. Mutation rates in uterine corpus endometrial carcinoma samples from The Cancer Genome Atlas (TCGA) with somatic inactivation of mismatch repair system (MMRd, MMR deficiency)) and/or proofreading activity of replicative polymerases (*POLE*exo- or *POLD1*exo-) are shown for comparison. **b,** 96-nucleotide context mutational spectrum of the ultrahypermutable cancer sample from the carrier of *POLD1* L474P and from an hypermutable cancer sample from a carrier of *POLD1* D316H. **c,** Fraction of mutations attributed to SBS10c and SBS10d mutational signatures in normal crypts, polyps[19] and cancer samples from germline carriers of *POLD1* variants. **d,** Variant allele frequency (top) and total coverage (bottom) along chromosome 19 in the tumor of a *POLD1* L474P carrier, showing cnLOH of the genomic region. The red line indicates the position of the POLD1 L474P mutation. **e,** Fold change in mutation rate in bins of different replication timing compared to the bin with the earliest replication timing. **f,** Mutation rate in cancer exomes with somatic mutations in *POLE* or *POLD1* and varying status of the MMR system calculated for all available TCGA UCEC samples.

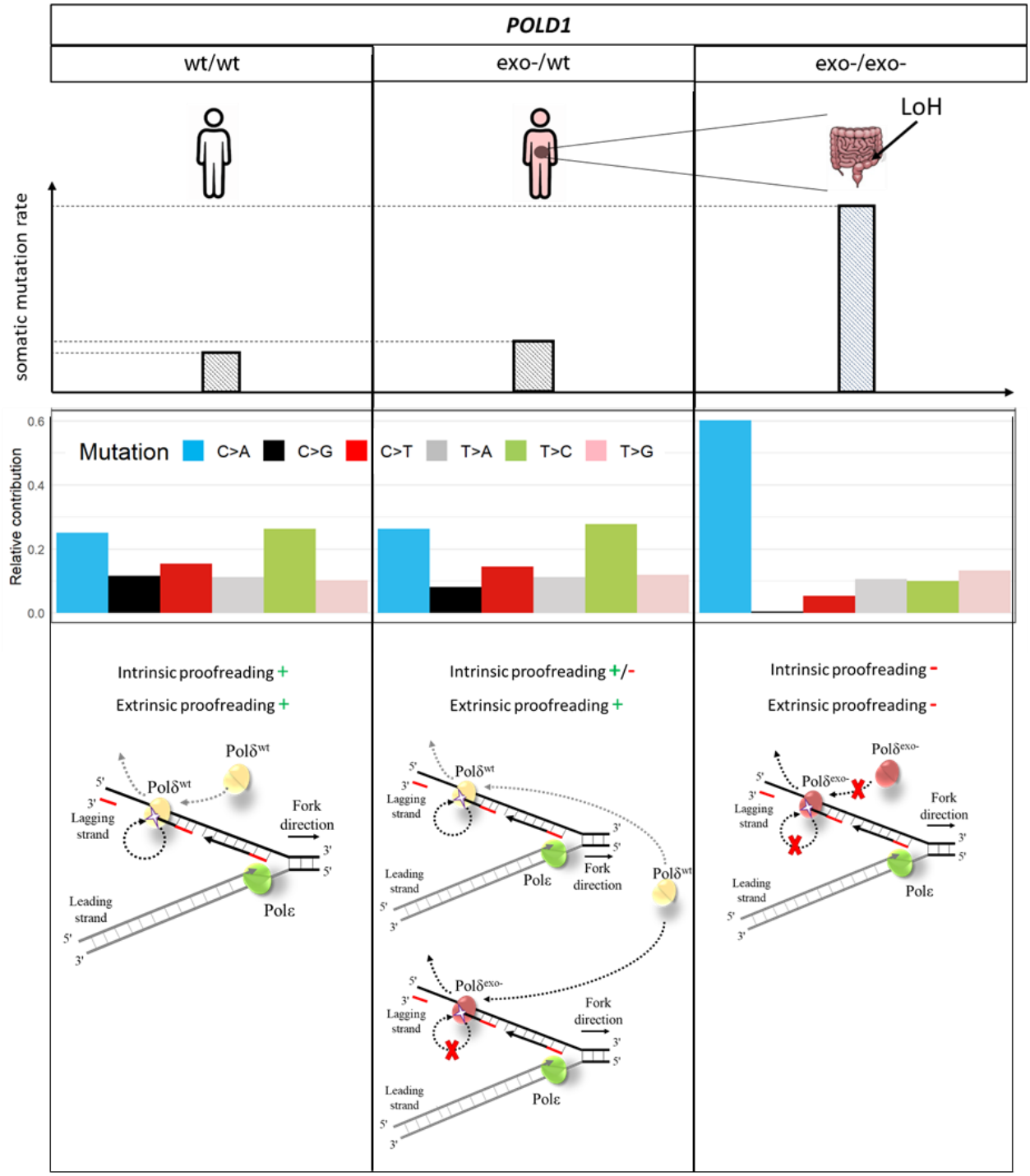

Fig. 5: Schematic representation of the possible underlying mechanisms for monoallelic and biallelic inactivation of POLD1 proofreading activity.

REFERENCES

1. Heredia-Genestar, J. M., Marquès-Bonet, T., Juan, D. & Navarro, A. Extreme differences between human germline and tumor mutation densities are driven by ancestral human-specific deviations. *Nat Commun* **11**, 2512 (2020).
2. Moore, L. *et al. The mutational landscape of human somatic and germline cells*. http://biorxiv.org/lookup/doi/10.1101/2020.11.25.398172 (2020) doi:10.1101/2020.11.25.398172.
3. Milholland, B. *et al.* Differences between germline and somatic mutation rates in humans and mice. *Nat Commun* **8**, 15183 (2017).
4. Seplyarskiy, V. B. & Sunyaev, S. The origin of human mutation in light of genomic data. *Nat Rev Genet* **22**, 672–686 (2021).
5. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
6. Besenbacher, S. *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* **6**, 5969 (2015).
7. Genome of the Netherlands Consortium *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47**, 822–826 (2015).
8. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
9. de Manuel, M., Wu, F. L. & Przeworski, M. *A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions*. http://biorxiv.org/lookup/doi/10.1101/2022.02.07.479417 (2022) doi:10.1101/2022.02.07.479417.
10. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
11. Meier, B. *et al.* Mutational signatures of DNA mismatch repair deficiency in *C. elegans* and human cancers. *Genome Res.* **28**, 666–675 (2018).
12. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science* **376**, abl9283 (2022).

13. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).

14. Haradhvala, N. J. *et al.* Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* **164**, 538–549 (2016).

15. Haradhvala, N. J. *et al.* Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. *Nat Commun* **9**, 1746 (2018).

16. Fukui, K. DNA Mismatch Repair in Eukaryotes and Bacteria. *Journal of Nucleic Acids* **2010**, 1–16 (2010).

17. Kunkel, T. A. Evolving Views of DNA Replication (In)Fidelity. *Cold Spring Harbor Symposia on Quantitative Biology* **74**, 91–101 (2009).

18. The CORGI Consortium *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet* **45**, 136–144 (2013).

19. Robinson, P. S. *et al.* Increased somatic mutation burdens in normal human cells due to defective DNA polymerases. *Nat Genet* **53**, 1434–1442 (2021).

20. Murphy, K., Darmawan, H., Schultz, A., da Silva, E. F. & Reha-Krantz, L. J. A method to select for mutator DNA polymerase δs in *Saccharomyces cerevisiae*. *Genome* **49**, 403–410 (2006).

21. Saini, N. *et al.* UV-exposure, endogenous DNA damage, and DNA replication errors shape the spectra of genome changes in human skin. *PLoS Genet* **17**, e1009302 (2021).

22. Tang, J. *et al.* The genomic landscapes of individual melanocytes from human skin. *Nature* **586**, 600–605 (2020).

23. Gao, Z. *et al.* Overlooked roles of DNA damage and maternal age in generating human germline mutations. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 9491–9500 (2019).

24. Halldorsson, B. V. *et al.* Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science* **363**, eaau1043 (2019).

25. An, J.-Y. *et al.* Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science* **362**, eaat6576 (2018).

26. Kaplanis, J. *et al. Genetic and chemotherapeutic causes of germline hypermutation*. http://biorxiv.org/lookup/doi/10.1101/2021.06.01.446180 (2021) doi:10.1101/2021.06.01.446180.
27. Church, J. M. Polymerase Proofreading-Associated Polyposis: A New, Dominantly Inherited Syndrome of Hereditary Colorectal Cancer Predisposition. *Diseases of the Colon & Rectum* **57**, 396–397 (2014).
28. Mur, P. *et al.* Role of POLE and POLD1 in familial cancer. *Genetics in Medicine* **22**, 2089–2100 (2020).
29. Valle, L. *et al.* New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. *Human Molecular Genetics* **23**, 3506–3512 (2014).
30. Bellido, F. *et al.* POLE and POLD1 mutations in 529 kindred with familial colorectal cancer and/or polyposis: review of reported cases and recommendations for genetic testing and surveillance. *Genetics in Medicine* **18**, 325–332 (2016).
31. Hao, D., Wang, L. & Di, L. Distinct mutation accumulation rates among tissues determine the variation in cancer risk. *Sci Rep* **6**, 19458 (2016).
32. Zhou, Z.-X. *et al.* How asymmetric DNA replication achieves symmetrical fidelity. *Nat Struct Mol Biol* **28**, 1020–1028 (2021).
33. Porkka, N. *et al.* Sequencing of Lynch syndrome tumors reveals the importance of epigenetic alterations. *Oncotarget* **8**, 108020–108030 (2017).
34. Hemminki, A. *et al.* Loss of the wild type MLH1 gene is a feature of hereditary nonpolyposis colorectal cancer. *Nat Genet* **8**, 405–410 (1994).
35. Sanchez de Abajo, A. *et al.* Dual role of LOH at MMR loci in hereditary non-polyposis colorectal cancer? *Oncogene* **25**, 2124–2130 (2006).

METHODS

Study participants and Ethical approval

The family included in the study was recruited through the Hereditary Cancer Genetic Counseling and Molecular Genetics Lab at the University Hospital of Elche (Spain), where the clinical information and blood and skin punches were obtained. FFPE tumor material was obtained through the Valencian Biobank Network.

The study received the approval of the Ethics Committees of IDIBELL (PR235/16) and of the University Hospital of Elche (PI37/2018).

Experimental procedure to assess somatic mutation accumulation in fibroblasts

Fibroblast obtention and immortalization

A skin punch biopsy was obtained from individuals III.2, III.4, IV.1, IV.2, IV.3, IV.4, IV.5, and IV.6 (Figure 1; Supplementary Table 1). The obtained biopsies were kept on culture media [DMEM with 10% fetal bovine serum and 100U/ml penicillin-streptomycin (Gibco, Thermo Fisher Scientific, Walthman, MA] at 4ºC until processing. After a 1X PBS wash, the sample was incubated ON at 37ºC with digestion media (DMEM supplemented with 160 U/ml collagenase and 1.25 U/ml dispase) and then disaggregated by pipetting. The sample was washed with culture media and seeded in one well of a 12-multiwell plate and maintained at 37ºC in a 5% CO_2 atmosphere. Cells were expanded for two passages before immortalization with lentiviral transduction with hTERT.

The lentiviral plasmid carrying the catalytic subunit of human telomerase protein (pLVX-IRES-hTERT-tdTOMATO) as well as envelope and packaging plasmids (pPAX2 and pMD2G) were kindly provided by Dr. Manel Esteller. Lentiviruses were produced transfecting HEK293 cells growing on a T75 flask with 20ug of total DNA and Lipofectamine 2000 (Thermo Fisher Scientific) for 16h. The culture medium was changed, collected at 72h and filtered using a 0.45 µm filter. The lentivirus enriched-medium was immediately used to transduce the human fibroblasts growing on a T25 flask in the presence of 8µg/ml polybrene (Sigma-Aldrich, San Luis, MO). Infected fibroblasts were maintained in culture until having enough cells for cell sorting. Cells expressing hTERT-tdTOMATO were enriched by fluorescent activated cell sorting (FACS). Cells were analyzed with the cell sorter MoFlo Astrios (Beckamn Coulter, Brea, CA) using a 561nm laser. Immortalized fibroblasts were maintained in culture media at 37ºC in a 5% CO_2 atmosphere and split at 1:3 ratio.

Single cell isolation and clonal expansion

To generate single-cell clones, fibroblasts enriched in hTERT-tdTOMATO were single-cell sorted using the MoFlo Astrios and the 561nm laser. Each single cell was automatically plated in a well of a 96-well plate in the presence of 200 µl of culture medium. The individual cultures were followed up to ensure clonal expansion. Each pool of cells was passed to growing sizes of culture plates up to T25 flasks. The first confluent T25 flask was considered the starting passage (p0). Two clones per individual were maintained in culture for 30-45 additional passages (Supplementary table 2) as described above. One clone per individual was sequenced.

DNA extractions

Peripheral blood DNA was extracted using the FlexiGene DNA kit (Qiagen, Valencia, CA). DNA from fibroblasts was obtained with the Quick-DNA Miniprep Plus kit (Zymo Research, Orange, CA). DNA from buccal swabs obtained with Isohelix swab packs and maintained in BuccalFix

tubes (Isohelix, Cell Projects Ltd, UK), was extracted using a standard phenol-chloroform protocol. DNA from formalin-fixed paraffin-embedded (FFPE) tumor samples was isolated with the kit QIAamp DNA FFPE tissue Kit (Qiagen). All extractions were carried out following the manufacturers' instructions.

Whole-genome sequencing (WGS)

DNA preparation for WGS was performed with the TruSeq Nano DNA Library, and sequencing was carried out in a NovaSeq 6000 150 PE (2x150 bp). Sequencing was performed at a minimum coverage of 90 Gb (30x) for the germline studies in the family members, and at a minimum coverage of 150 Gb (50x) for the experiments with fibroblasts (P0 – ~P40 passages of the fibloblasts' cultures). Sequencing was performed at Macrogen (Macrogen Inc, Seoul, South Korea).

Whole-exome sequencing

Exome sequencing was performed in FFPE tumor DNA of individual IV.6. Exome capture was performed with Kappa HyperExome Probes (Roche) and sequenced in a NovaSeq 6000 S1 (2x100bp). Sequencing was performed at Centro Nacional de Análisis Genómico (CNAG, Barcelona, Spain).

Analysis of MMR status in tumors

MMR status in tumor tissue was assessed by standard methods, using immunohistochemistry of MMR proteins MLH1, MSH2, MSH6 or PMS2, and/or by analysis of microsatellite instability (MSI) by PCR-based analysis of microsatellite markers.

Variant calling in fibroblasts and post-processing filters

Sequenced reads were aligned to hg19 reference human genome downloaded from USCS (https://hgdownload.soe.ucsc.edu/downloads.html#human) using Burrows–Wheeler alignment (BWA-MEM[1]).

Somatic mutations in single-cell derived colonies were called using Mutect2. DNA genome sequencing data available from normal tissue (blood, buccal swab or fibroblasts) from the corresponding individual was used as matched normal DNA. Panel of normals (--panel-of-normals Mutect2 argument) was created from all available sequenced blood samples. The population allele frequencies in gnomAD were used as a prior for germline variant detection (--germline-resource Mutect2 argument). Additional filter for the variant allele frequency observed in sequenced data was applied: only mutations with vafs between 0.25 and 0.75 were selected for the analysis.

Somatic mutations accumulated during the experiment were called using Mutect2. Mutations present in the end point (P40) but absent in the start (P0) were selected. To filter out recurrent technical artifacts all available sequencing data from samples that did not correspond to the mutation accumulation experiment were used to create a panel of normals (--panel-of-normals Mutect2 argument). The population allele frequencies in gnomAD were used as a prior for germline variant detection (--germline-resource Mutect2 argument).

Mutational spectra and PCA analysis

The 96-mutational spectrum was calculated dividing the number of mutations of a particular type in fixed 3-nucleotide contexts by the total number of mutations of that type. Principal component analysis (PCA) was used on these 96-dimensional vectors for mutations obtained from fibroblast colonies. Mutational spectra obtained by the same procedure across different datasets (Fig. 2f, Fig.

3c,e, Fig.3g,h) were projected in the obtained PC space.

Extraction of *de novo* signatures

Mutational signatures in fibroblasts were extracted *de novo* and then decomposed to COSMIC signatures using SigProfilerExtractor[2]. The number of mutations attributed to each COSMIC signature was obtained as output.

For mutations accumulated in fibroblast colonies during the experiment, SigProfilerExtractor was run with the following parameters: minimum_signatures=1, maximum_signatures=15, nmf_replicates=300. The solution with the most stable signatures (n_signatures = 3) was selected.

'sigfit' R package[3] was used to estimate exposure to mutation signatures in intestinal crypts, adenomas and cancer samples with germline *POLD1* mutations. List of signatures to fit was limited to SBS1, SBS5, SBS10c and SBS10d.

Calling of germline *de novo* mutations and post-processing filters

Sequenced reads were aligned to the hg19 reference human genome downloaded from UCSC using Burrows–Wheeler alignment (BWA-MEM). *De novo* mutations were called using standard GATK4 best practices pipeline. All samples were jointly called, but only high-confidence mendelian violation sites with minimum GQ=20 for each trio member were selected. Subsequently, additional filters were applied: i) coverage of each trio member ≥10; ii) absence of reads confirming alternative allele in parents; iii) number of reads confirming reference and alternative allele in proband ≥5; iv) allele frequency of the alternative allele in the proband ≥0.3. Additionally, clustered mutations (distance between mutations ≤1000bp) were filtered out, as potential false positives. Mutations present in any other family member were also excluded. To eliminate false positive mutations coming from the sequencing of proband fibroblasts, we used other fibroblasts from the same individual (used in MA experiments) as additional confirmation of mutation presence: true *de novo* mutations have to be present in all tissues of the proband, including the other fibroblast sample. This additional filter mainly removed variants from the left tail of the variant allele frequency distribution of potential *de novo* mutations, thus keeping the distribution more symmetric around 0.5 (Supplementary Fig. 7). To evaluate the performance of the filter we took the data from our preliminary sequencing of blood samples for two trios from the family (IV.5 and IV.6 probands) and compared the *de novo* mutations obtained in the two attempts. We found out that for IV.6 offspring this additional filter removed 1257 out of 1360 variants. Only 6 of them had been proposed as *de novo* mutations in the first round of trio sequencing. For the IV.5 individual, the filter removed 56 out of 121 candidates and none of them had been called as *de novo* mutations in the preliminary sequencing (Supplementary Fig. 8). Blood samples for each individual in the trios had been preliminary sequenced with 20X coverage, *de novo* mutations had been called using PhaseByTransmission GATK3 tool (--prior = 1e-4) and sites with violation from mendelian inheritance were selected.

Enrichment of *de novo* mutations in *POLD1* contexts

According to analysis of mutations in fibroblasts, four contexts were enriched in mutations in carriers of *POLD1* L474P variant: Cp<u>C</u>pT>A, Tp<u>C</u>pT>A, Ap<u>T</u>pT>A and Cp<u>T</u>pT>G.

We calculated a proportion of mutations in these contexts in *de novo* mutations of published trios equal to 2.4%. This is thus the expected proportion for such mutations in offspring of wild type *POLD1* parents. We then calculated the expected number of such mutations in each offspring in our experiment by multiplying the expected proportion by the total number of observed mutations in offspring. The sum of expected numbers calculated by all offspring of father-carriers of *POLD1*

L474P variant was compared to the observed sum to estimate the enrichment. Rate-ratio test was used for comparison of observed vs. expected data.

Simulation of *de novo* mutations

To test how many offspring of wild-type parents in the sample have –PC1 values higher than the mean value of –PC1 in the offspring of fathers harboring *POLD1* L474P by chance, we generated an artificial dataset of trios by randomly sampling *de novo* mutations according to their fractions in homogeneous underlying spectra. To obtain the underlying spectra we used data from 6233 offspring from two publicly available datasets[4,5], aggregated all *de novo* mutation together and calculated the proportion of each mutation type in each possible 3-nucleotide context. For each offspring in the dataset, we counted the observed number of *de novo* mutations, and, sampled the same number of mutations from the obtained spectrum. For each simulated set of *de novo* mutations, we created the vector of mutational probabilities and estimated the value of the -PC1 component by projecting in the PC space obtained previously from the analysis of fibroblasts. We repeated the same procedure generating two or more offspring for each family and averaging the –PC1 among siblings.

Similarly, a sample of trios with a mixture of offspring of *POLD1* L474P carriers was simulated. For 95% of the offspring we sampled mutations according to their probability in the wild-type trios and for the remaining 5% of samples, we sampled mutations based on their probability of occurring in the offspring of *POLD1* L474P fathers in the studied family. We then calculated the mutational spectrum per simulated individual and the -PC1 values. The same procedure was repeated for 1% admixture of offspring of *POLD1* L474P fathers.

Calling of mutations in tumors with *POLD1* variant

Sequenced reads were aligned to the hg19 reference human genome downloaded from USCS (https://hgdownload.soe.ucsc.edu/downloads.html#human) using Burrows-Wheeler alignment (BWA-MEM). Somatic mutations in tumor samples were called against normal samples from the same individual using standard GATK4 best practices pipeline. The Learn Orientation Bias Artifacts tool was used to control for orientation bias, which is critically important for FFPE samples. Mutations with variant allele frequency <0.15 were filtered out, as low-frequency variants are known to be enriched in formalin fixation artifacts in FFPE samples[6].

Epigenetic covariates for the extrinsic proofreading effect of Polδ

Genome was splitted in 100-kb non-overlapping windows. The number of Tp<u>C</u>pT>A/ApGpA>T mutations and TpCpT/ApGpA sites was calculated in each window. Mean replication timing for each window was obtained using Wavelet-smoothed Signal for Repli-seq data for HeLa-S3 cell line (GSE34399). Genome was separated into 50 equal bins according to the mean replication timing value. Mutation rate in each bin was calculated as a sum of all mutations in windows in this bin divided by the sum of target sites. The obtained mutation rate in each bin was normalized for the mutation rate in the bin of the earliest replication time.

SUPPLEMENTARY REFERENCES

1. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013) doi:10.48550/ARXIV.1303.3997.
2. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. http://biorxiv.org/lookup/doi/10.1101/2020.12.13.422570 (2020) doi:10.1101/2020.12.13.422570.
3. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures.

http://biorxiv.org/lookup/doi/10.1101/372896 (2018) doi:10.1101/372896.
4. Halldorsson, B. V. et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science 363, eaau1043 (2019).
5. An, J.-Y. et al. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. Science 362, eaat6576 (2018).
6. Bhagwate, A. V. et al. Bioinformatics and DNA-extraction strategies to reliably detect genetic variants from FFPE breast tissue samples. BMC Genomics 20, 689 (2019).

DATA AVAILABILITY

DNA sequencing data are deposited in the European Genome-Phenome Archive (EGA) with accession code EGAS00001006434. Called de novo and somatic mutations are available online (https://github.com/andrianovam/POLD1-project_scripts). All other data are available from the authors upon request.

CODE AVAILABILITY

Code required to reproduce the analyses in this paper is available online. All custom code used in this study is available online (https://github.com/andrianovam/POLD1-project_scripts).

ACKNOWLEDGMENTS

We would like to thank the members of the family under study for their selfless participation, and Dr. Anouk Jaen Larrieu of the Dermatology Department of Elche University Hospital, who performed the skin biopsies for the obtention of the fibroblasts.

FUNDING

This study was funded by the Spanish Ministry of Science and Innovation (Agencia Estatal de Investigación), co-funded by FEDER funds a way to build Europe [PID2020-112595RB-I00 (LV)], Instituto de Salud Carlos III [CIBERONC CB16/12/00234 (LV); ISCIII-AES-2017 PI17/01082 (JLS)], Government of Catalonia [AGAUR 2017SGR1282, CERCA Program for institutional support (LV)], Scientific Foundation *Asociación Española Contra el Cáncer* [AECC Investigador (MT)], Austrian Science Fund FWF [Grant Agreement # I5127-B (FK)], German Research Foundation DFG [Grant Agreement # 429960716 (FK)], and ERC Consolidator [Grant Agreement # 771209 ChrFL (FK)].

CONTRIBUTIONS

V.B.S., F.A.K., A.S.K., G.A.B. and L.V. conceived the project and designed the experiments. M.A.A. and V.B.S. conducted bioinformatic and statistical analyses. M.T., P.M., and G.A. conducted the experiments in the fibroblasts from family members (immortalization, cell culture and single cell isolation) under L.V.'s supervision, and performed DNA extractions from blood, buccal swabs and tumor samples. A.B.S-H. and J.L.S. obtained the clinical information, informed consents and samples from the family under study. M.A.A., V.B.S. and L.V. wrote the original manuscript. F.A.K., A.S.K. and G.A.B. consulted on the project and edited the manuscript.

COMPETING INTERESTS

No competing interests are declared by the authors of this study.