The PREGCARE study: Personalized recurrence risk assessment following the birth of a child with a pathogenic *de novo* mutation

Keywords: mosaicism, gonadal mosaicism, de novo mutation, WGS, trio sequencing Personalised Recurrence Risk, Precision genetic counselling.

Authors

Marie Bernkopf[1,21], Ummi B. Abdullah[1,21], Stephen J. Bush[1], Katherine Wood[1], Sahar Ghaffari[1], Eleni Giannoulatou[2], Nils Koelling[1], Geoffrey J. Maher[1], Loïc M. Thibault[3], Jonathan Williams[4], Edward M. Blair[5], Fiona Blanco Kelly[5], Angela Bloss[5], Emma Burkitt-Wright[6], Natalie Canham[7], Alexander T. Deng[8], Abhijit Dixit[9], Jacqueline Eason[9], Frances Elmslie[10], Alice Gardham[11], Eleanor Hay[12], Muriel Holder[8], Tessa Homfray[10], Jane A. Hurst[12], Diana Johnson[13], Wendy D. Jones[12], Usha Kini[5], Emma Kivuva[14], Ajith Kumar[12], Melissa M. Lees[12], Harry G. Leitch[11,15], Jenny E. V. Morton[16], Andrea H. Németh[5,17], Shwetha Ramachandrappa[10], Katherine Saunders[5], Deborah J. Shears[5], Lucy Side[18], Miranda Splitt[19], Alison Stewart[13], Helen Stewart[5], Mohnish Suri[9], Penny Clouston[4], Robert W. Davies[20], Andrew O. M. Wilkie[1,22], Anne Goriely[1,22*]

[1]Clinical Genetics Group, MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford; NIHR Oxford Biomedical Research Centre, Oxford, UK

[2]Victor Chang Cardiac Research Institute, Darlinghurst, NSW, Australia

[3] Centre for Population Genomics, Garvan Institute of Medical Research and School of Mathematics and Statistics, UNSW Sydney, Sydney, New South Wales; Centre for Population Genomics, Murdoch Children's Research Institute, Melbourne, Victoria, Australia

[4] Oxford Genetics Laboratories, Churchill Hospital, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

[5]Oxford Centre for Genomic Medicine, Nuffield Orthopaedic Centre, Oxford University Hospitals NHS Foundation Trust, Oxford, UK

[6]Manchester Centre for Genomic Medicine, Manchester University NHS Foundation Trust and Division of Evolution and Genomic Sciences, University of Manchester, Manchester, Greater Manchester, UK.

[7]Department of Clinical Genetics, Liverpool Women's NHS Foundation Trust, Liverpool, UK

[8]Clinical Genetics Department, Guy's Hospital, Guy's & St Thomas' NHS Foundation Trust, London, UK

[9]Nottingham Regional Genetics Service, City Hospital Campus, Nottingham University Hospitals NHS Trust, Nottingham, UK

[10]South West Thames Regional Genetics Service, St George's University Hospitals NHS Foundation Trust, London, UK

[11]North West Thames Regional Genetics Service, London North West University Healthcare NHS Trust, Northwick Park Hospital, Harrow, UK.

[12]North East Thames Regional Genetics Service, Great Ormond Street Hospital NHS Foundation Trust, London, UK

[13]Sheffield Clinical Genetics Service, Sheffield Children's NHS Foundation Trust, Sheffield, UK

[14]Clinical Genetics, Royal Devon University Healthcare NHS Foundation Trust, Exeter, UK

[15]MRC London Institute of Medical Sciences, Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK

[16]West Midlands Regional Clinical Genetics Service and Birmingham Health Partners, Birmingham Women's and Children's Hospitals NHS Foundation Trust, Birmingham, UK

[17]Nuffield Department of Clinical Neurosciences, University of Oxford, UK

[18]Wessex Clinical Genetics Service, University Hospital Southampton, Princess Anne Hospital, Southampton, UK

[19]Northern Genetics Service, The Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle, UK

[20]Department of Statistics, University of Oxford, Oxford, UK

[21]These authors contributed equally: M. Bernkopf, U. Abdullah

[22]These authors contributed equally: A.O.M. Wilkie, A. Goriely

*Correspondence to A.G. Email: anne.goriely@imm.ox.ac.uk

Abstract

Next-generation sequencing has led to a dramatic improvement in molecular diagnoses of serious pediatric disorders caused by apparently *de novo* mutations (DNMs); by contrast, clinicians' ability to counsel the parents about the risk of recurrence in a future child has lagged behind. Owing to the possibility that one of the parents could be mosaic in their germline, a recurrence risk of 1-2% is frequently quoted, but for any specific couple, this figure is usually incorrect. We present a systematic approach to providing individualized recurrence risk stratification, by combining deep-sequencing of multiple tissues in the mother-father-child trio with haplotyping to determine the parental origin of the DNM. In the first 58 couples analysed (total of 59 DNMs in 49 different genes), the risk for 35 (59%) DNMs was decreased below 0.1% but for 6 (10%) couples it was increased owing to parental mosaicism - that could be quantified in semen (recurrence risks of 5.6-12.1%) for the paternal cases. Deep-sequencing of the DNM efficiently identifies couples at greatest risk for recurrence and may qualify them for additional reproductive technologies. Haplotyping can further reassure many other couples that their recurrence risk is very low, but its implementation is more technically challenging and will require better understanding of how couples respond to information that reduces their risks.

Main text:

The birth of a child with a serious clinical disorder to a healthy couple with no previous family history is a life-changing event. Added to the challenges posed by caring for their child, is the anxiety that their future children could be similarly affected. Whilst robust frameworks for addressing this possibility are increasingly available for common chromosomal abnormalities and recessive monogenic diseases, no systematic approach has been developed for dominant disorders caused by apparently *de novo* mutations (DNMs). Such disorders are collectively common, estimated to affect at least 1 in 295 births[1], but extremely heterogeneous; for example, mutations in over 650 genes are currently recognised to cause developmental disorders through a dominant mechanism of action[1,2]. The need to address this issue has been made more pressing by the success over the past decade of next-generation sequencing (NGS) technologies in identifying DNMs, leading to a deluge of new causative genes and diagnoses.

The implementation of NGS technologies across large populations has contributed to a better understanding of the patterns of occurrence of DNMs. It is now well established that DNMs are rare events (spontaneous human mutation rate is ~1.2×10^{-8} per bp, per generation), mainly occurring as "one-off" copying errors during sperm production, or less frequently in oocytes[3,4]. While in these instances, the risk of recurrence for future siblings would be negligible, DNMs can also occur post-zygotically (either in one of the two clinically unaffected parents, or in the affected child) leading to a mosaic genotype that alters the recurrence risk. Mosaicism populating multiple germinal cells in the ovaries or testes (arising during one of the parent's own development), termed gonadal (or germline) mosaicism, may be associated with a substantial recurrence risk for further offspring, reaching up to 50% in some cases; by contrast, convincing demonstration of post-zygotic mosaicism in the offspring would eliminate the chance of sibling recurrence[5].

Although mosaicism has long been recognised as a source of DNMs, few studies have attempted (or had the power) to define its exact contribution to spontaneous disease. Overall, current NGS methods used to identify DNMs rely on mother-father-proband trio sequencing and are poorly suited for detection of mosaic cases - either for cases of low-level (parental) mosaics[6], or to distinguish high-level variant allele frequency (VAF) from constitutional (50%) presentation in post-zygotic (proband) cases[5,7]. For example, the limit of VAF sensitivity of WES/WGS trio sequencing, which is typically performed at a depth of 25-30x, is ~10-15%, similar to that of dideoxy-sequencing[6,8]. Moreover, routine genetic analysis relies on the interrogation of a single somatic tissue (blood or saliva), which is not adequate to identify mosaicism in parental gametes or variable VAF in a proband's tissues.

The recognition that the tissue distribution and VAF of a DNM are determined by the timing at which it first occurred, allows us to identify three key time points during development with different predicted presentations: (1) very early in development - before the segregation of germline and somatic lineages at ~day 14 of human embryogenesis, yielding cases of mixed (somatic and gonadal) mosaicism; (2) post-15 days of development in the germline, resulting in confined gonadal mosaicism; (3) or much later in the developing or adult gonad, yielding a "one-off" mutation (Supplementary Fig S1). Furthermore, by taking into account the individual in whom the DNM originated (mother, father, or affected child), it becomes possible to distinguish a total of seven scenarios whereby a DNM can occur (Fig. 1). The overall relative prevalence of these seven scenarios can be estimated quite accurately based on previous analyses of the parental origin of DNMs and the prevalence of

mosaicism from population studies (see Supplementary Note 1).

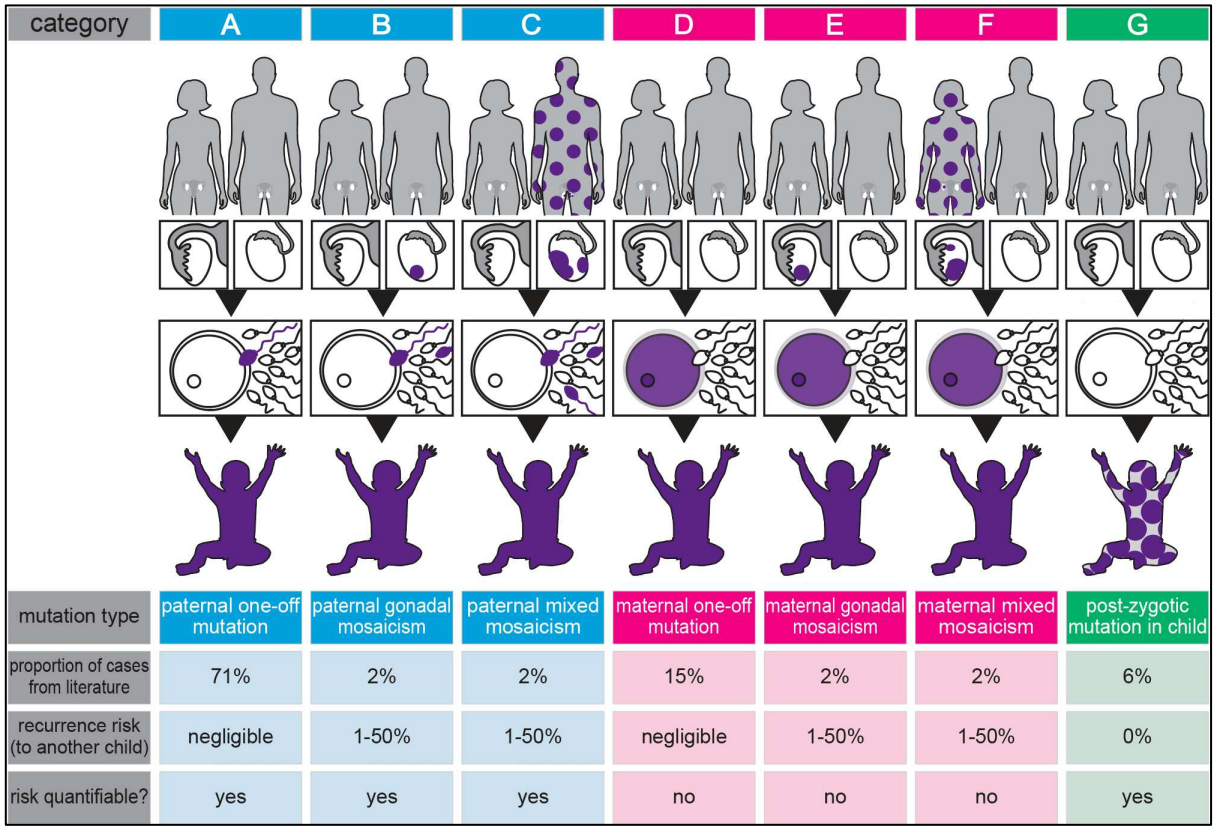


Fig. 1*: Stratification of DNMs into seven categories. Establishing the origin (paternal (blue), maternal (pink) or post-zygotic (proband, green), and timing of the mutational events (purple colour indicates mutant cells), yields widely different recurrence risks in different families. See main text, Supplementary Fig S1and Supplementary Note 1.*

Here, we have developed a systematic strategy to categorise pathogenic DNMs in a mixed clinical population of 60 couples who had one or more children with a serious developmental disorder caused by an apparent DNM, and were seeking individualized reproductive counselling about recurrence risk in a future pregnancy. By combining deep-sequencing of multiple tissues to detect occult mosaicism with haplotyping to determine parent-of-origin of the DNM, we show that we can reliably stratify individual couples into discrete categories that are associated with substantially different risks to the offspring. This personalised approach to recurrence risk assessment offered *prior* to a new pregnancy should provide reassurance to the majority of couples in whom the risk is very low or negligible and help to focus resources on the minority of families at increased recurrence risk.

Results

Population sampled

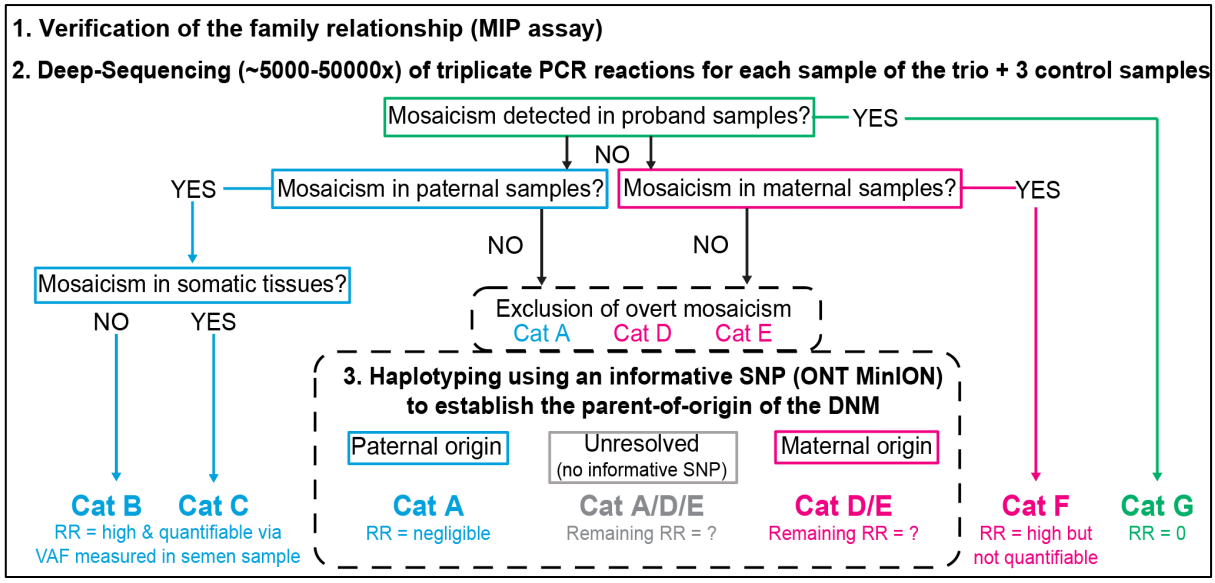
Following ethical approval we recruited, through the network of Clinical Genetics centres in England, 60 couples who had one or more children (or fetuses) affected by a serious clinical disorder caused by an identified DNM, which was not present in the parents' DNA on routine analysis (PREGCARE study; Online Methods). Two families (FAM17 and FAM60) had three affected siblings/pregnancies, indicating that one of the parents must be a gonadal mosaic, but routine diagnostic analysis performed on parental blood DNA had failed to identify the parent-of-origin. To eliminate ascertainment bias, these two families are excluded from the quantitative presentation of the data but included in the specific analysis of mosaicism. Hence, our primary cohort comprises data from 58 parent-child trios, including one trio with two different pathogenic DNMs (FAM12). These 59 DNMs comprised 40 single nucleotide substitutions, 14 small (1-2 nucleotides) indels and 5 larger (4-44 nucleotides) indels in 49 different genes, providing a broad and representative spectrum of pathogenic molecular lesions encountered in clinical practice (Supplementary Table S1).

Deep-sequencing of multiple tissues identifies mosaic cases

Four of the seven categories shown in Fig. 1 (B, C, F and G) involve mosaic states that can be directly identified without requiring invasive sample collection and distinguished by deep-NGS of tissues collected from the family trio. We therefore obtained up to 14 biological family samples (child: blood, buccal mucosa left + right; mother and father: blood, saliva, buccal mucosa left + right, urine; plus paternal semen) to seek evidence of mosaicism. Collection of parent samples was designed to include all three embryonic germ layers (ectoderm, buccal; mesoderm, blood; endoderm, urine), plus germline in the father (Supplementary Fig. S1).

169 The overall strategy deployed for the analysis is shown in Fig. 2.

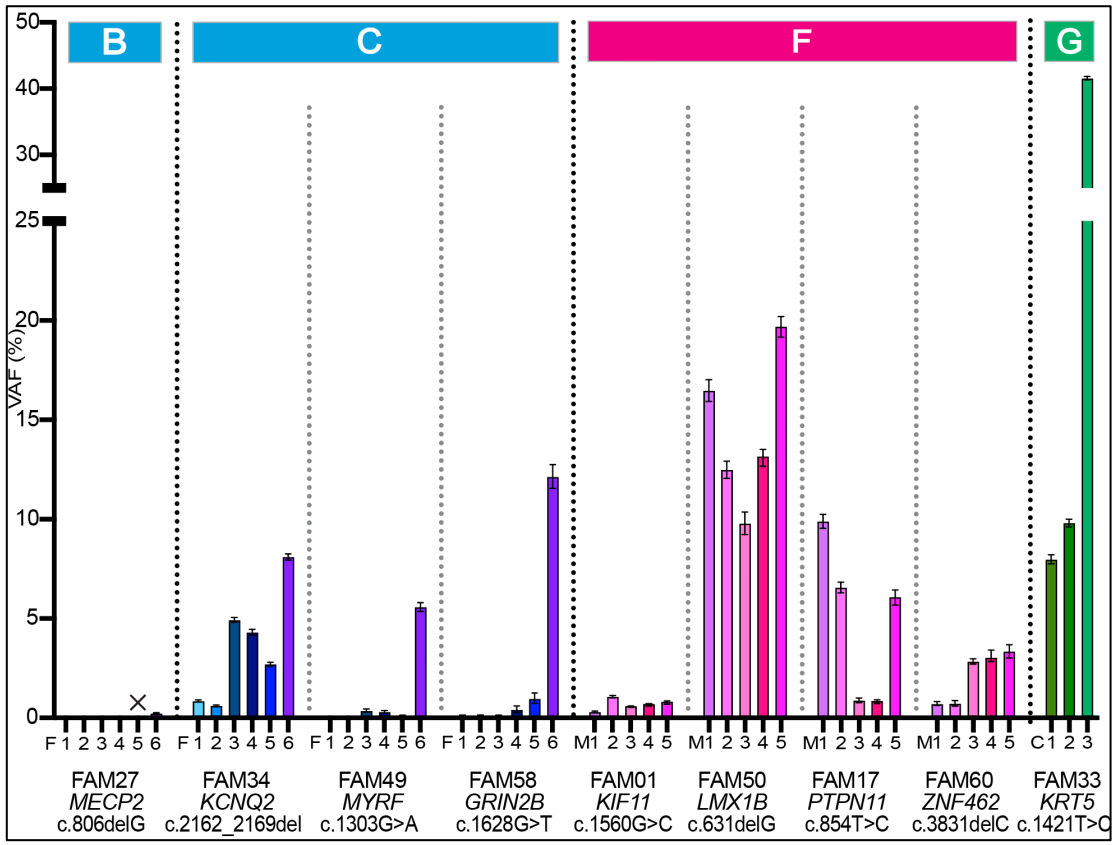
170



171

172 **Fig. 2: Flow chart describing the three-tier sample analysis in the PREGCARE study**
173 *Following collection of up to 14 different biological samples per family and verification of*
174 *the familial relationships between the 3 individuals of the trio, the DNM site was deep-*
175 *sequenced in all family samples (performed in triplicates together with 3 unrelated controls)*
176 *to detect low-levels of parental mosaicism or instances of post-zygotic mosaicism in the*
177 *proband. For those families without evidence of overt mosaicism, haplotyping was performed*
178 *to resolve the parental origin of the DNM and further stratify the recurrence risk (RR). Refer*
179 *to Fig. 1 for category (Cat) description.*
180
181
182 Following verification of sample relationships and parentage in each family using a panel of
183 bespoke multiplex inversion probes (MIP) targeting 168 common single nucleotide
184 polymorphisms (SNPs), we designed a custom PCR assay covering 65-224 bp around the
185 family-specific DNM site and performed triplicate reactions from each available tissue, and
186 three unrelated control DNAs, before undertaking deep-NGS (target depth 5,000-50,000x) on
187 the Illumina MiSeq platform. Reads were processed using amplimap[9] and VAF quantified at
188 the genomic position of the DNM (see Methods). NGS was poorly suited to analyse two
189 DNMs associated with the larger indels (a 35 bp deletion in FAM12b and a 44 bp duplication
190 in FAM54). Hence, to rule out the possibility of occult mosaicism in these samples, we

191 performed mutant allele-specific PCR on all available samples from the trio (Supplementary
192 note 2).
193 Overall, deep-NGS (and/or allele-specific PCR) identified 7/59 (11.9%) cases with strong
194 evidence of mosaicism in one family member (Fig. 3; Supplementary Fig. S2 & Table S2).
195 These comprised DNMs belonging to Categories B (paternal gonadal mosaicism; FAM27), C
196 (paternal mixed mosaicism; FAM34, FAM49, FAM58), F (maternal mixed mosaicism;
197 FAM01, FAM50) and G (post-zygotic mosaicism in proband; FAM33). Analysis of the two
198 additional families in which recurrence in siblings was already documented (FAM17,
199 FAM60) showed that both were attributable to maternal mixed mosaicism (Fig. 3).
200



201
202

203 **Fig. 3: Mutation levels observed in the families presenting with mosaicism.**
204 *Variant allele frequencies (VAF) in different samples from the family member in whom mosaicism was*
205 *detected by deep-NGS sequencing. Family number, gene, cDNA coordinates of the DNM and the origin of*
206 *the different samples are indicated on the x-axis in the same order for each family and distinguished by*
207 *colors for ease of visualization. The category classification is indicated at the top of the figure. X*
208 *represents a sample failure. Full data for the other family members and controls are presented in*
209 *Supplementary Fig S2 and Table S2. FAM17 and FAM60 are the two families with multiple affected*

pregnancies and belong to category F (maternal mixed mosaicism); note the low VAF in the maternal blood samples (M3) for both families. The corrected VAF (see Methods) is plotted and error bars represent the 95% binomial confidence intervals across the three technical replicates. Abbreviations: F=father; M=mother; C=child; 1=buccal swab (left); 2=buccal swab (right); 3=blood; 4=saliva; 5=urine; 6=sperm; 7=gDNA from original testing.

Identifying these mosaic families is particularly important, because whereas the recurrence risk associated with post-zygotic mosaicism (Category G) is effectively zero, the other three mosaic categories (B, C, F) are associated with increased recurrence risks. While the offspring risk is not directly quantifiable for the maternal mosaics because of the inaccessibility of ovarian tissue, it could be quantified in the paternal mosaic cases via the VAF measured in sperm and ranged from 0.23% (FAM27) to 12.1% (FAM58) (F6 bars in Fig. 3; Sup Fig S2). Importantly, in the three paternal cases of mixed mosaicism the level of the cognate DNM in sperm (5.6-12.1%) was substantially higher than in any of the other tissues sampled and variability in mutation levels was present between different somatic tissues, with no one tissue providing a reliable indicator of the level in sperm (Fig. 3).

In seven of the eight parental mosaic cases, where DNA derived from blood (the most widely used source of DNA for genetic analysis) was analysed, the level of mutation in the transmitting parent was below 5% (F3 and M3 values in Fig. 3; Supplementary Table S2). Such VAFs would be impossible to detect systematically using standard diagnostic NGS read depths (~25-30x) or dideoxy-sequencing, illustrating the importance of deep-sequencing (>5000x) and the value of collecting additional tissue samples to increase sensitivity for ascertaining occult mosaicism. In the single identified instance of paternal confined gonadal mosaicism, a relatively low level of sperm mutation was observed (0.23%), consistent with a slightly later timing of mutational origin (Supplementary Fig S1) and in line with empiric data on mutations in sperm[10-13].

Also of note, the VAFs for the proband samples from FAM33 with the post-zygotic mutation (C1, C2 and C7 values in Fig. 3) were markedly different across the tissues

238 analysed (blood 41.6%; buccal mucosa 8.0% and 9.8%), demonstrating the benefit of
239 analysing several tissue samples from an individual to distinguish post-zygotic mosaicism
240 associated with high VAF levels from constitutional (50%) presentation.
241
242 **Haplotype phasing enables determination of parental origin of DNM**
243 For the remaining 52 DNMs that did not classify into one of the four mosaic categories
244 described above, further stratification was attempted through haplotyping to determine the
245 parental origin of the mutation (Fig. 2). For these families, only one category (Category E,
246 maternal gonadal mosaicism) is associated with a recurrence risk to offspring (Fig. 1).
247 Although it is not possible to distinguish Categories D and E (because oocytes are not
248 accessible), most of the remaining DNMs (~88%) are predicted to belong to Category A
249 (paternal one-off), which is associated with a negligible risk to offspring.
250 Parental origin could be inferred for two families without performing haplotyping: FAM26
251 (mutation in the X-linked *MID1* gene in a male proband, implying a maternal origin) and
252 FAM54 (a 34 bp duplication in *MAGEL2*, a gene known to be maternally imprinted and for
253 which pathogenic mutations are exclusively paternal in origin[14]).
254 To perform haplotyping of the other 50 DNMs, we sought an informative SNP or other
255 variant in close proximity to the DNM, to enable phasing of the parental alleles. In the most
256 common informative scenario, the child is heterozygous for the SNP (genotype AB) whereas
257 one of the parents is homozygous (genotype AA or BB), making it possible for the inherited
258 parental chromosomes to be distinguished. In three cases an informative SNP was present in
259 the PCR product used for the deep-sequencing, enabling the parental origin to be determined
260 directly by examining the phase of the DNM on the Illumina reads. To haplotype the 47
261 remaining DNMs, we designed two long PCR products extending away on either side of the
262 DNM (total genomic region covered ~7-30kb), and sequenced the resulting fragments for the

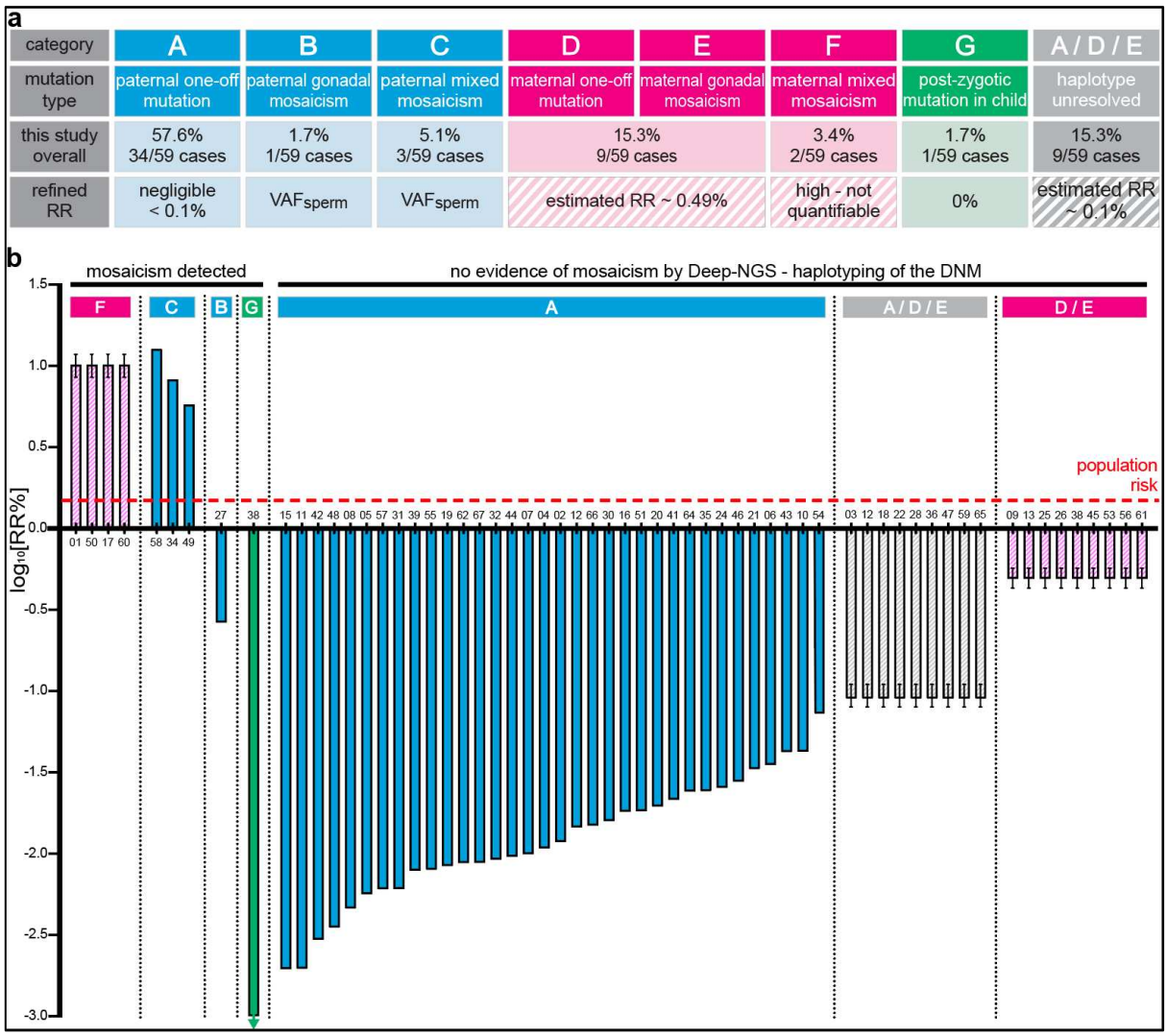
three family members using the MinION platform from Oxford Nanopore Technology (ONT). Reads for each trio were processed and analysed with an in-house custom pipeline combining Medaka and pile-up processing (see Methods and Supplementary Note 3). This haplotyping strategy was successful in the majority (38/47) of cases (Supplementary Table S3), including three families (FAM11, FAM38, FAM67) that required a more complex analysis involving two SNPs to distinguish the parental alleles. In one of these (FAM38), due to the local genomic context of the DNM (a single G-nucleotide deletion within a homopolymeric region), phasing by direct analysis of sequencing traces could not be resolved by ONT sequencing. Nevertheless, this approach identified an informative SNP in the proband which was used to design a bespoke allele-specific PCR and determine the DNM parental origin (Supplementary Note 4 & Table S3B).

Overall, parental origin could be established for 82.7% of DNM (43/52), which included 34 DNMs of paternal origin (79%) and 9 (21%) present on the maternally-derived allele (Fig. 4) - a result in line with the expected ~4:1 male to female ratio of mutational origin[3] (Supplementary Note 1).

Combining deep-sequencing and haplotyping allows category stratification and individualized recurrence risk

Having singled out the mosaic cases by deep-NGS of multiple familial tissues, the particular value of the combined approach of deep-sequencing of semen samples with haplotyping is to identify those families in which the DNM is paternal in origin (34/52), as they belong to Category A (Figs. 1 and 2). Deep-sequencing of sperm of these paternal mutations allows measurement of the VAF and derivation of the upper confidence limit for the level of DNM present in sperm (Supplementary Table S2). As NGS is subject to background sequencing errors, we corrected the raw VAF values using measurements from three unrelated control

288 samples. The corrected VAFs were estimated by numerically maximizing their marginal
289 likelihood and 95% confidence intervals were obtained by using profile likelihood (see
290 Methods for details). For Category A samples, the upper bound (95% CI) of the VAF
291 measured by deep-NGS in sperm was below 0.05% in all cases (Supplementary Table S2;
292 Fig. 4). These data point to DNMs in this category having originated as 'one-off' events
293 during late gonadal development or adult spermatogenesis.
294



295

296 ***Fig. 4: Overview of the results of the PREGCARE study showing refinement of individual recurrence risk for***
297 ***all families (a)*** *Summary table of the PREGCARE results for the 59 DNMs analyzed in this study and overview*
298 *of the refined recurrence risk (RR).* ***(b)*** *Personalized recurrence risk (RR%) estimates for each of the 60 families*
299 *(61 DNMs) enrolled in the PREGCARE study represented on a logarithmic scale. The red dotted line represents*
300 *the generic population RR given to couples who have had a child with a DNM (~1.5%). The RR can be*
301 *quantified (block colours) for paternal (Categories A-C, via semen analysis) and post-zygotic (G) DNMs; note*
302 *that to be conservative in our estimates of RR, we have plotted the upper 95% binomial CI from the corrected*

VAF_{Sperm} measured by deep-NGS for the paternally-derived DNMs (Supplementary Table S2). The RR can only be estimated for maternal (pink) or haplotype-unresolved cases (grey). These estimates are represented by stripes, with error bars representing the upper and lower 95%CI - see Supplementary Note 5. Note that the DNM of FAM54 was analyzed by allele-specific PCR (Supplementary Note 2) and that Category F includes the two additional families with multiple affected pregnancies. Individual family numbers are indicated on the x-axis.

In 9/52 (17.3%) families, the haplotyping revealed a maternal origin of the DNM. In these cases, the negative findings from deep-NGS of maternal somatic tissues (i.e. maternal origin of mutation but with no evidence of somatic mosaicism) do not allow categories D and E to be distinguished (Figs. 1 and 2; 4). The relative risk for a DNM belonging to Category E (maternal gonadal mosaicism) and representing a recurrence risk in a future pregnancy, rather than to Category D ('one-off' maternal event) can be estimated to occur in ~2:15 families (Fig. 1 and Supplementary Notes 1 & 5), meaning that on average 1 in every 8 or 9 maternal DNMs is anticipated to have originated in early developing germ cells. While the prevalence of mosaicism has not been directly quantified in ovaries owing to their experimental inaccessibility, it can be assumed that these early events are similar in magnitude for ovaries and testes because germline lineages are specified several weeks prior to sex determination.[4] To gain further insight into the remaining recurrence risk for DNMs of proven maternal origin (D/E), we used estimates of VAFs observed for paternal confined mosaics obtained from sperm WGS[11] (Supplementary Note 5) . We obtained a recurrence risk estimate for the combined maternal categories D-E of 0.49% (95%CI: 0.43% - 0.57%), a modest reduction over the population average (Fig. 1 and Fig. 4).

Evaluation of the remaining risk for cases with unresolved parental origin

Finally, in a further 9 families, despite sequencing a ~13-27 (mean 22.2) kb region around the DNM in the proband, no informative SNP could be identified in the MinION reads. Hence the parental origin of the DNM could not be assigned, and the mutation could belong to any of Categories A, D or E (Fig. 1). As the majority of DNMs are predicted to be sporadic

(Categories A or D), the remaining risk (associated with Category E) for these couples can be estimated by combining the relative proportion of Category E cases (2:86) and the average VAF observed for gonadal mosaicism[11]. As a result, for a DNM with an unresolved parent of origin, the recurrence risk is estimated to be 0.09% (95%CI: 0.08%-0.11%), a reduction of approximately 10-fold compared to the population risk baseline (Fig. 4; Supplementary Note 5).

Discussion

We have applied a general framework to analyse systematically and at scale, the origins of DNMs presenting in a clinical setting. The work addresses a stark unmet clinical need to improve genetic counselling for couples who have had a child affected by a disorder caused by a DNM – a situation faced by almost a million parents annually – in order to provide them with a personalised risk assessment *prior* to a new pregnancy. The current standard of care, which is to provide these couples with a recurrence risk of ~1-2% is unsatisfactory, both because this figure is nearly always wrong (as illustrated by Fig. 4), but also because of the uncertainty it raises for the complex decision process of whether to extend their family. It is well documented that couples' attitudes to reproductive risk vary widely[15]: some will view the 1-2% risk as small and others would not contemplate extending their family in the face of any risk. In addition, while in many healthcare settings there may be the option of a prenatal diagnostic procedure (chorionic villus biopsy or amniocentesis), this is associated with a small risk of miscarriage (currently estimated as ~0.2-0.5% for each procedure[16-18]) and may not be ethically acceptable to some couples. Owing to a combination of cost and technical challenges, prenatal procedures that are non-invasive (assay of free fetal DNA from maternal blood sample) or those which avoid the possibility of termination of pregnancy in the event of recurrence (preimplantation genetic testing for monogenic disorders [PGT-M]), are not available in most public healthcare settings. For example, in the UK the eligibility threshold

for PGT-M is a risk >10% of having a child with a serious genetic condition, which excludes the parents of children with DNMs even though some couples will have a risk higher than this.

Over recent years several pioneering studies on DNM origins have provided a solid framework to quantify the relative contribution of different mutational processes to DNMs (Fig. 1; Supplementary Notes 1 & 5)[3-5,19-21]. We designed the PREGCARE study based on this framework, with the dual aims to seek evidence for mosaicism in each member of the parent-child trio (deep-sequencing), and to stratify the risk based on the likely timing and parental origin of the DNM. Important aspects of the study design include the recognition that (1) clinically-relevant mosaicism is caused by early embryonic mutations, that present either in both soma and germline (mixed mosaicism) or the germline only and affect males and females equally - because they originate before sex determination; (2) sampling of multiple tissues of different embryonic origins increase the likelihood of detecting instances of mixed mosaicism in parents (or post-zygotic events in the proband); and (3) analysis of a paternal semen sample allows direct quantification of risk for paternally-derived DNMs, which are anticipated to represent ~3/4 of cases. Hence, although the female germline is not accessible to direct analysis, data about the prevalence and VAF anticipated for maternal mosaic cases can be inferred from sperm data[11]. Moreover, the relative risks of mixed vs. confined gonadal mosaic events can be estimated based on data from deep-sequencing of paired blood and sperm samples,[10,11] which have shown that the average VAF measured in sperm for cases associated with mixed mosaicism is ~9%, while 'sperm-only' average VAF are ~3%[11] (Supplementary Note 5). These data suggest that because mixed mosaicism is caused by very early mosaic events, they are likely to have a wide tissue distribution and be present at higher VAF (Supplementary Fig S1). Moreover, the rate of spontaneous mutations may be elevated during the first embryonic divisions[22,23]. Hence, mixed mosaic cases likely

383 contribute to most of the recurrence risk in the next generation and identifying them by deep-
384 sequencing of somatic tissues (and semen) represents an efficient way to single out the
385 couples at higher risk.

386 Here we show through systematic analysis of a clinical series of 59 DNMs a very good
387 correspondence between the distribution of DNMs across the 7 different categories for the
388 families analysed to that anticipated from previous work (Figs. 1 & 4A; Supplementary Note
389 1). In our cohort, which consists of clinically-ascertained cases, DNMs originated from occult
390 parental mosaicism in ~10% (6/59) of cases. For five families it was detectable in the
391 transmitting parent's somatic tissues - although present at low VAFs in blood, illustrating the
392 importance of deep-sequencing (>5000x) and the value of collecting additional tissue
393 samples to increase sensitivity for ascertaining occult mosaicism.

394 Fig. 4, which summarises our overall findings, shows that we achieved risk alteration
395 for individual couples over more than three $\log_{10}$ orders of magnitude: for 54/59 DNMs, the
396 risk was reduced compared to the population baseline risk, and for 5/59 (the mixed mosaics),
397 it was likely increased (but only quantifiable in the 3/5 paternal cases).

398 Encouraging though these data are, we acknowledge several barriers before considering
399 clinical translation of this work. The first hurdle relates to technical implementation of
400 individualised recurrence risk measurement in a clinical setting, which requires robust
401 laboratory methods and will be challenging as a DNM-specific custom assay will be required
402 for most families. In this study we used two methods, deep-sequencing and haplotyping, that
403 provide complementary information. Deep-NGS is highly effective in singling out couples at
404 high recurrence risk, whereas haplotyping is essential to generate most of the very low
405 recurrence risks and reassure the majority of couples that they belong to Category A. The
406 Illumina platform used for deep-sequencing is technically straightforward and the associated
407 calling pipelines are readily available in most diagnostic settings. Of note, for efficient

evaluation of reproductive risk, the source of tissue samples ought to be a major consideration. While semen provides the ideal tissue for determining reproductive risks directly, surprisingly at present the prevailing working practices of clinical genetics do not include routine semen analysis. Our view is that this work and that of others[11-13,24] provides clear evidence to promote much more widespread collection and analysis of this material (as is standard, for example, in fertility clinics). As there is no easy access to maternal gonadal tissue, it would be valuable to know whether there is a particular somatic tissue that provides a better surrogate for the germline. This can be addressed by studies of male samples, but although we observed substantial variation, it was not possible to identify a clear surrogate[12,13,25], consistent with the fact that during early embryogenesis, cell populations are subject to bottlenecks and differential lineage commitments leading to considerable variation and stochasticity in cellular representation across tissues[20,26]. Hence reliance on assessment of a single tissue (blood) risks missing some mixed mosaics harbouring low mutation levels (or high level post-zygotic mosaicism in the proband). Of the other tissues we sampled, we found that saliva tended to reflect the results from blood[6,27] but occasionally exhibited a higher background that can bias low VAF interpretation, likely reflecting the fact that ~70% of saliva DNA is derived from white blood cells, while the remaining fraction contains bacterial and/or other genomes (potentially including that of other family members, including the proband)[28]. Unlike urine, which often yielded poor amounts of DNA, buccal brushings (left and right sides of the cheek sampled independently), are easy to collect (including from children), and store, and contained cells of a different embryological origin to blood, which often yielded informative data.

Overall, we conclude that clinical implementation of deep-sequencing of a few key tissues from the trio (blood, buccal brushings and paternal sperm) alone should be easy to achieve and would identify most of the high-risk cases, therefore reducing the risk of mosaicism

433 presentation for the remaining couples (Categories A, D, and E) to ~0.1% (Supplementary
434 Note 5).
435 To further refine the remaining risk in non-mosaic families, we used the ONT platform as a
436 second method in this study. Although harder to scale and process than Illumina data, ONT
437 showed good potential for implementation in diagnostics, but is not currently approved for
438 use in most clinical settings. Independently of technical considerations, one major limitation
439 of this approach, which led to a substantial minority (15.3%) of unresolved cases in this
440 study, is the requirement for the presence of a heterozygous SNP in the vicinity of the DNM
441 to distinguish the two parental alleles in the proband. Implementation of novel ultra-long
442 read WGS methods will facilitate SNP identification and systematic parent-of-origin
443 assignment of DNMs, but are not currently available in most settings[29].
444 Another potential barrier to clinical implementation relates to how these refined risks are
445 viewed by couples and whether changes in risk actually result in altered decision-making.
446 Concerning the accuracy of our risk estimations, among the 61 DNMs analysed, in only 39
447 do we consider the risks to be reasonably accurate; these include the 38 DNMs shown to be
448 paternally originating, in which we could directly measure levels of mutation in sperm, and
449 the single post-zygotic case (Fig. 4). In 36/39 we reported a risk lower than baseline, while in
450 the three mixed mosaic cases it was increased (to 5.6%, 8.1% and 12.1%). By contrast, in the
451 remaining cases shown either to be of maternal origin (13/61) or unresolved (9/61), the risk
452 estimation has been refined but remains inaccurate and may be viewed differently by parents
453 and healthcare professionals. Even in proven cases of maternal mixed mosaicism, VAFs in
454 somatic tissues are poor predictors for the germline, as illustrated by the two families with
455 multiple recurrences in whom we detected relatively low VAF in maternal somatic tissues
456 (maximum of 3.3% and 9.9% in the samples analyzed), despite three affected pregnancies in
457 each sibship (Fig. 3). Nevertheless, detection of mixed mosaicism in maternal tissue will

458 warrant caution in future pregnancy and it should also be noted that some diagnostic options
459 may be more complicated for these families because of the unsuitability of non-invasive
460 prenatal testing via analysis of cell-free fetal DNA in maternal plasma[30].

461 In those cases where somatic mosaicism has been excluded but the DNM is proven or
462 possibly maternal in origin, the risk of maternal gonadal mosaicism (Category E, Fig. 1) may
463 remain an important factor in decision-making, despite the relative reduction in risk for these
464 subcategories (Category E corresponds to ~2:15 maternally-proven DNM and ~2:88
465 haplotype-unresolved DNM with an estimated average VAF of ~4%).

466 An interesting illustration from this work of the complexity of recurrence risk
467 counselling is provided by the case of paternal confined gonadal mosaicism (Category B) in
468 FAM27, in which the risk for the *MECP2* mutation was found to be 0.23% (95% CI: 0.19%-
469 0.26%), over 5-fold lower than the 1.2% baseline population risk. How to counsel a couple in
470 this situation, where stratification to the "at risk" Category B predicts increased caution,
471 remains difficult. This risk may also need to be leveraged against the current UK
472 recommendation for 'higher risk' in respect to the probability of carrying a fetus with Down
473 Syndrome for a 35-year old mother ~ 1/150 (0.66%) for which prenatal screening is routinely
474 offered[31].

475

476 Overall, we show that providing pre-conception recurrence risk assessment to couples
477 who have had a child with a DNM can be achieved and offers the prospect of driving a major
478 transformation in the practice of genetic counselling. Our data demonstrate that for all
479 couples, it is possible to refine the risk of having another affected child with the same DNM
480 and in the majority of cases (>64%) the risk is in fact very small, potentially reducing anxiety
481 and the need for expensive pre-implantation or prenatal diagnostic options. For couples in
482 whom we detected overt mosaicism, the risk is higher (and quantifiable through sperm

483 analysis for the paternal cases). Providing evidence-based estimation of the actual risk will
484 allow these couples to be singled out for further investigations and support, allowing them to
485 make informed choices (and for their clinicians to provide them with personalised advice and
486 risk assessment) about the different diagnostic options available to them.
487

488 **Acknowledgments**
489 This work was primarily supported by grants from NewLife (17/18/04), the Wellcome
490 (102731/Z/13/Z and 219476/Z/19/Z), and the National Institute for Health Research (NIHR)
491 Oxford Biomedical Research Centre Programme. We acknowledge WIMM core funding from
492 the Medical Research Council (MRC) through the WIMM Strategic Alliance (G0902418 and
493 MC_UU_12025) and the support of the NIHR UK Rare Genetic Disease Research Consortium
494 (Musketeers Memorandum). We thank Tim Rostron and the members of the Oxford Genomic
495 Medicine laboratory for technical support.
496 The funders had no role in study design, data collection and analysis, decision to publish, or
497 preparation of the manuscript.
498

499 **Authors contribution**
500 Conceived the study and designed the experiments: AOMW and AGo
501 Performed the experiments: MB, UBA, KW and SG
502 Performed data analysis: MB, UBA, SJB, EG, NK, GJM, LMT, JW, RWD, AOMW and
503 AGo
504 Recruited participants and/or provided samples: JW, EMB, AB, EB-W, FBK, NC, ATD, AD,
505 JE, FE, AGa, EH, MH, TH, JAH, DJ, WJ, UK, EK, AK, MML, HGL, JEVM, AHN, SR, KS,
506 DJS, LS, MS, AS, HS, MSu, PC
507 Wrote the manuscript: AOMW and AGo with input of MB, SJB, RWD

References

1. Deciphering Developmental Disorders, S. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433-438 (2017).
2. Kochinke, K.*, et al.* Systematic Phenomics Analysis Deconvolutes Genes Mutated in Intellectual Disability into Biologically Coherent Modules. *Am J Hum Genet* **98**, 149-164 (2016).
3. Rahbari, R.*, et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-133 (2016).
4. Sasani, T.A.*, et al.* Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *Elife* **8**, e46922 (2019).
5. Acuna-Hidalgo, R.*, et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *Am J Hum Genet* **97**, 67-74 (2015).
6. Gambin, T.*, et al.* Low-level parental somatic mosaic SNVs in exomes from a large cohort of trios with diverse suspected Mendelian conditions. *Genet Med* **22**, 1768-1776 (2020).
7. Krupp, D.R.*, et al.* Exonic Mosaic Mutations Contribute Risk for Autism Spectrum Disorder. *Am J Hum Genet* **101**, 369-390 (2017).
8. Cao, Y.*, et al.* A clinical survey of mosaic single nucleotide variants in disease-causing genes detected by exome sequencing. *Genome Med* **11**, 48 (2019).
9. Koelling, N.*, et al.* amplimap: a versatile tool to process and analyze targeted NGS data. *Bioinformatics* **35**, 5349-5350 (2019).
10. Breuss, M.W.*, et al.* Autism risk in offspring can be assessed through quantification of male sperm mosaicism. *Nat Med* **26**, 143-150 (2020).
11. Yang, X.*, et al.* Developmental and temporal characteristics of clonal sperm mosaicism. *Cell* **184**, 4772-4783 e4715 (2021).
12. Zhang, Q.*, et al.* Genomic mosaicism in the pathogenesis and inheritance of a Rett syndrome cohort. *Genet Med* **21**, 1330-1338 (2019).
13. Yang, X.*, et al.* Genomic mosaicism in paternal sperm and multiple parental tissues in a Dravet syndrome cohort. *Sci Rep* **7**, 15677 (2017).
14. Fountain, M.D.*, et al.* The phenotypic spectrum of Schaaf-Yang syndrome: 18 new affected individuals from 14 families. *Genet Med* **19**, 45-52 (2017).
15. Fumagalli, S.*, et al.* Prenatal diagnosis tests and women's risk perception: a cross-sectional study. *J Psychosom Obstet Gynaecol* **39**, 73-81 (2018).
16. Beta, J., Zhang, W., Geris, S., Kostiv, V. & Akolekar, R. Procedure-related risk of miscarriage following chorionic villus sampling and amniocentesis. *Ultrasound Obstet Gynecol* **54**, 452-457 (2019).
17. Salomon, L.J., Sotiriadis, A., Wulff, C.B., Odibo, A. & Akolekar, R. Risk of miscarriage following amniocentesis or chorionic villus sampling: systematic review of literature and updated meta-analysis. *Ultrasound Obstet Gynecol* **54**, 442-451 (2019).
18. Navaratnam, K., Alfirevic, Z., Royal College of, O. & Gynaecologists. Amniocentesis and chorionic villus sampling: Green-top Guideline No. 8 July 2021: Green-top Guideline No. 8. *BJOG* **129**, e1-e15 (2022).
19. Kong, A.*, et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-475 (2012).
20. Jonsson, H.*, et al.* Differences between germline genomes of monozygotic twins. *Nat Genet* **53**, 27-34 (2021).
21. Jonsson, H.*, et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).
22. Ju, Y.S.*, et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714-718 (2017).
23. Ye, A.Y.*, et al.* A model for postzygotic mosaicisms quantifies the allele fraction drift, mutation rate, and contribution to de novo mutations. *Genome Res* **28**, 943-951 (2018).

24. Bernkopf, M., *et al.* Quantification of transmission risk in a male patient with a FLNB mosaic mutation causing Larsen syndrome: Implications for genetic counseling in postzygotic mosaicism cases. *Hum Mutat* **38**, 1360-1364 (2017).
25. Huang, A.Y., *et al.* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell Res* **24**, 1311-1327 (2014).
26. Coorens, T.H.H., *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387-392 (2021).
27. Wright, C.F., *et al.* Clinically-relevant postzygotic mosaicism in parents and children with developmental disorders in trio exome sequencing data. *Nat Commun* **10**, 2985 (2019).
28. Samson, C.A., Whitford, W., Snell, R.G., Jacobsen, J.C. & Lehnert, K. Contaminating DNA in human saliva alters the detection of variants from whole genome sequencing. *Sci Rep* **10**, 19255 (2020).
29. Noyes, M.D., *et al.* Familial long-read sequencing increases yield of de novo mutations. *Am J Hum Genet* **109**, 631-646 (2022).
30. Scotchman, E., Shaw, J., Paternoster, B., Chandler, N. & Chitty, L.S. Non-invasive prenatal diagnosis and screening for monogenic disorders. *Eur J Obstet Gynecol Reprod Biol* **253**, 320-327 (2020).
31. https://view-health-screening-recommendations.service.gov.uk/downs-syndrome/

PREGCARE: Online Methods

Recruitment into the PREGCARE study

The PREGCARE (PREcision Genetic Counselling And REproduction) study was approved by the London - Queen Square Research Ethics Committee under the reference number 17/LO/1025 (IRAS reference: 225264). Couples with one (or multiple) children, stillbirths or terminated pregnancies affected by a likely pathogenic *de novo* mutation (DNM) and who were potentially interested in personalized transmission risk assessment for future pregnancies were invited to participate by healthcare professionals during routine clinical genetic consultation. A DNM was defined as a single-nucleotide or small insertion-deletion variant detected in the proband that was absent in the parents' DNA on routine diagnostic genetic analysis. DNMs occurring in one of six paternal age-effect genes (*FGFR2, FGFR3, HRAS, KRAS, PTPN11, RET*) were excluded, unless there were multiple affected pregnancies[1]. Couples where the mother was pregnant at the time of sample collection, those who were not both the biological parents of the affected child, or either the biological mother or father did not consent to participate, were also excluded.

Recruitment and sample collection took place at 13 of the 17 participating National Health Service (NHS) Trusts in England, UK.

Sample collection

Families interested in participating in the study were sent a box containing kits and instructions for collection at home of 2 ml saliva (Oragene DNA, OG-500, DNA-Genotek, Canada) and 50 ml morning midstream urine (Urine Collection And Preservation Tube, Norgen Biotek Corp., Canada) from both the mother and father, and an ejaculate of semen (following abstinence for three days before collection and stored at -20 °C) from the father. During the clinical visit, informed written consents were obtained and further samples were

collected from the three family members, including 5 ml peripheral blood (EDTA) from father and mother and buccal cells from the left and right inner cheek lining from mother, father, and the affected child using swabs (sterile PurFlock Ultra tip swab in dry transport tube, Puritan Medical Products, ME, USA). Samples and completed consent forms were sent at room temperature to the MRC Weatherall Institute of Molecular Medicine where they were witnessed-transferred and processed for extraction or long-term storage within 48 hours of collection. Overall, a total of 67 boxes were dispatched and 60 families completed collection and consents and were enrolled into the study.

In addition, the child's genomic DNA originally used for the molecular diagnosis was requested from the NHS genetic laboratory. This sample had usually been extracted from the proband blood or, occasionally, fetal tissues, amniocentesis or chorionic villus sample (CVS) (for details see Supplementary Table S1).

Sample processing and DNA extraction

Upon delivery of the box to the lab, the family samples were given a unique identifier and processed. The saliva samples were incubated at 50 °C for 60 min and then aliquoted. Blood samples were aliquoted as whole blood and isolated buffy coat. Urine samples were centrifuged at 2000 x *g* for 10 min and the cell pellet rinsed with 1 x phosphate-buffered saline (PBS) before storage. Semen samples were split into 50-100 µl volume aliquots which were rinsed with 1 x PBS (5000 x *g* for 5 min). Mouth swabs were kept frozen until extraction, when they were resuspended into 100 µl PBS.

Genomic DNA was extracted from all the collected family samples (2 saliva lysates, 2 whole blood samples, 2 urine cell pellets, 6 buccal swabs, 1 semen lysate) on the Maxwell RSC Instrument using the Maxwell RSC Blood DNA kit (both Promega, WI, USA) and following manufacturer's protocols. Aliquots of semen were pre-incubated in sperm lysis buffer (20

mM Tris HCl pH 8.0, 20 mM EDTA, 200 mM NaCl, 1% SDS) in the presence of proteinase K (250 µg/ml), dithiothreitol (DTT; 100 mM) and 0.6% SDS at 42°C for 4-12 hours. Concentrations of the final DNA eluates were assessed with standard fluorometric methods.

Genotyping assay for verification of familial relationship using molecular inversion probes (smMIP assay)

To confirm the familial relationships of each trio, we used an in-house custom single-molecule molecular inversion probes (smMIPs) genotyping assay to capture common single nucleotide polymorphisms (SNPs) across all chromosomes (total of 290 smMIP probes targeting 154 autosomal, 14 X-linked and 57 Y-linked markers; for SNP details and probe sequences, see Supplementary Tables S4A & S4B) following established smMIPs protocols[2] followed by sample barcoding, library preparation and 2 x 151 bp paired-end sequencing on a MiSeq instrument (Illumina, CA, USA). For each family, DNA from the proband sample obtained from the original diagnostic laboratory (or if unavailable, buccal swab DNA), the maternal blood sample and the paternal semen sample were analyzed. Sequencing data was processed using the 'pileups snps' tool in the amplimap v0.4.9[3] pipeline with default settings (alignment to GRCh38.p12 with BWA, variant calling with GATK) to generate counts for the reference (REF) and alternate (ALT) alleles at each locus. Subsequently, the autosomal and X-linked SNP genotype for each individual of the family trio was recorded as Homozygous REF (AA), Heterozygous (AB) or Homozygous ALT (BB). For genotyping, SNPs were considered informative when the parents were homozygous (AA or BB) and the proband exhibited the expected genotype such as when Parent1/Parent2/Proband were AA/AA/AA, BB/BB/BB, AA/BB/AB. Other SNPs were analyzed to ensure there was no genotype discordance across the 3 family members.

654 **Ultra-deep Illumina sequencing (Deep-NGS) of DNM sites**

655 Ultra-deep Illumina sequencing was performed in order to detect low levels of mosaicism in
656 parental samples or post-zygotic mosaicism in the child. For each family-specific DNM, a
657 pair of PCR primers tailed with generic CS1 (5'- ACACTGACGACATGGTTCTACA) and
658 CS2 (5'- TACGGTAGCAGAGACTTGGTCT) sequence tags was designed to amplify a
659 short genomic region (49-266 bp) around the DNM site; primer genomic locations (build
660 GRCh38.p12), are provided in Supplementary Table S1. Each primer set was tested on
661 control DNA with either High Fidelity Phusion or Q5 Polymerase (New England Biolabs,
662 MA, USA) and PCR amplification was performed following manufacturer's
663 recommendations using 30 ng of genomic DNA from triplicates of up to 14 biological
664 samples and three unrelated control DNAs in 10 µl PCR reactions, applying an initial
665 denaturation step for 30 s at 98 °C, followed by 30 cycles of 10 s at 98 °C, 30 s at 68 °C, and
666 30 s at 72 °C, and 8 min at 72 °C as final extension step. Successful amplification was
667 confirmed by running samples on an agarose gel. PCR-amplified fragments were diluted,
668 further PCR amplified using individually barcoded primers, pooled together to construct
669 libraries and ultra-deep sequenced, as previously described[4] on a MiSeq (Illumina) instrument
670 with 2 x 151 bp paired-end reads at an average depth of ~19,000 x for each sample.
671
672 **Deep-NGS data analysis and determination of the observed variant allele frequency**
673 **(VAF) at the DNM location.**
674 Illumina data were analyzed using amplimap[3], as above, to obtain both the variant allele
675 frequency (VAF) of each family-specific mutation and the total count of >Q30 bases at the
676 corresponding genomic position (GRCh38.p12) in each PCR replicate and sample. For each
677 family-specific dataset, DNM VAFs observed in each sample were corrected, to account for
678 the background alternate read counts observed in the control samples (false-positives) at the

DNM genomic location. Let *k1* and *k2* be the number of alternate reads observed in the control and case, and *n1* and *n2* be the total number of reads observed in the control and case, respectively. Let p denote the unobserved proportion of cells carrying a variant and let q be the false-positive rate of the sequencing and variant-calling procedure.

The joint likelihood of p and q is defined as follows

$$\mathcal{L}(p, q|k1, n1, k2, n2) = B(k1;\ n1, q) \cdot B(k2; n2, p + (1 - p) \cdot q)$$

where B denotes the binomial probability mass function and $B(k;\ n, p)$ is the probability of observing k successes in n trials with success probability p. The first term corresponds to the probability mass of observing k1 false-positives in the control, and the second term corresponds to the probability mass of observing *k2* alternate reads in the $p + (1 - p) \cdot q$ case. The rate in the second term corresponds to the fact that a read identified as carrying the variant in the case is either a true positive (i.e. actually carrying the variant) with probability p or a false positive (i.e. background noise, not carrying the variant but mistakenly identified as doing so) with probability $(1 - p) \cdot q$).

We treated q as a nuisance parameter and obtained the marginal likelihood of p by numerically integrating the joint likelihood over q using adaptive quadrature[5]. Finally, we obtained the maximum likelihood estimate of p by numerically maximizing the marginal likelihood and obtained 95% confidence intervals using profile likelihood[6]. Scripts describing this analysis are available at github.com/sjbush/pregcare.

Allele-specific PCRs

For two DNMs – a 44 bp deletion in *MECP2* in FAM12 and a 35 bp duplication in *MAGEL2* in FAM54 – the regions were successfully amplified as described above, but the deep-sequencing on the MiSeq platform did not lead to quantifiable results in the proband sample, making the assay unsuitable for mosaicism detection. Therefore, individual mutation-specific

PCR assays were designed and the resulting PCR products analyzed using gel electrophoresis. The individual assays' sensitivity was determined with dilution series experiments (Supplementary Note 2). Furthermore, an allele-specific PCR had to be designed for haplotyping the DNM of FAM38 in *AHDC1* due to a homopolymeric region around the mutation site for which the mutant and wildtype allele could not be phased with ONT sequencing (Supplementary Note 3 and Supplementary Table S3B).

Long-read haplotyping assay using Oxford Nanopore Technologies (ONT)

The MinION (Oxford Nanopore Technologies [ONT], UK) long-read sequencing technology was used to determine the parent-of-origin of the DNM in the proband. To do so, primers were designed to amplify two regions (~2-16 kb each, for locations of individual primer sequences, see Supplementary Table S1) on either side of the DNM. DNA from the two parental blood samples and the diagnostic genomic DNA from the proband were amplified using LongAmp Polymerase (New England Biolabs, UK) starting with 50 ng genomic DNA in a 20 µl reaction following manufacturer's recommendations and the cycling conditions: initial 2 min at 95 °C, 30 cycles of 30 s at 95 °C and 16 min at 65 °C, and a final extension at 65 °C for 20 min. PCR amplicons were checked on a 0.9% agarose gel and if amplification had been successful, regions 1 and 2 from one sample were pooled. For library preparation, the PCR barcoding amplicon protocol and 1D ligation kit and PCR expansion kit (all ONT, UK) were used to barcode individual samples in a 20 µl PCR reaction with LongAmp polymerase, 2 µM barcoding primers and 1:100 diluted target PCR with the cycling conditions as described above for 8 cycles. After adapter ligation, the pooled library was loaded onto a MinION SpotOn Mk I version R9 flowcell (ONT) for sequencing following the manufacturer's recommendations. For initial data processing (demultiplexing and basecalling) each set of fast5 files was processed using Guppy v4.5.4+66c1a77

(https://community.nanoporetech.com) with the parameter --config dna_r9.4.1_450bps_hac.cfg, producing one set of reads for each barcode/family member of the trio. Reads are deposited in the European Nucleotide Archive under BioProject accession number PRJEB53977 (http://www.ebi.ac.uk/ena/data/view/PRJEB53977).

Haplotype phasing of *de novo* mutations using Medaka and mpileup

ONT reads for each trio were aligned to the GRCh38.p12 primary assembly using minimap2 v2.18[7] with parameter -ax map-ont. Lower-quality (MAPQ < 20) and non-primary alignments were discarded using samtools view v1.12[8] with parameters -q 20 -F 256 -F 2048. For each target region (genomic coordinates are given in Supplementary Table S1), variants were called using the 'medaka_variant' workflow of Medaka v1.3.2 (https://github.com/nanoporetech/medaka, accessed 6th May 2021) with default parameters. The set of VCFs per region were then concatenated using BCFtools v1.12[8] to produce one VCF per BAM, subsequently annotated using dbSNP v153[9] (https://ftp.ncbi.nih.gov/snp/latest_release/VCF/GCF_000001405.38.gz, accessed 6th May 2021).

Where possible, Medaka uses the information contained within heterozygous SNPs to impute the haplotype of the aligned reads. In practice this means that a proportion of the calls in each VCF are phased, being assigned to a 'phase set' of SNPs on the same haplotype. Given that the sequencing data represent mother/father/proband trios, with each proband having a DNM, each VCF was parsed to determine whether Medaka had called and phased the DNM in the proband (but not in either parent, confirming their true "de novo" status). For each DNM called by Medaka, we obtained the associated phased set SNPs, retaining only those which had a total depth of coverage >10x. We cross-referenced the phased set SNPs with the VCFs from the mother and father and identified which calls (if any) had been made at those

positions. This produced a set of three haplotypes from which we used a custom script to classify the inheritance of the DNM as either maternal or paternal (the SNPs in phase with the DNM could only be derived from the chromosome inherited from the mother or father, respectively), else unresolved (Medaka either did not call the DNM in the child, called it but did not construct a phased set, or, if it did construct a phased set, either did not call its constituent SNPs in the parents or made identical calls for both of them).

DNMs not successfully phased using Medaka (Supplementary Note 3) were phased by programmatic and/or manual inspection of read pileups. A programmatic approach was implemented using a custom script which parsed read pileups (generated using samtools mpileup with parameters -aa --output-QNAME) to obtain a set of reads which contained both the ALT-allele for the DNM and a candidate phasing SNP (considered the closest one to it and for which there was a prior, namely inclusion in dbSNP). We then constructed a 2x2 count table (rows: number of reads calling REF/ALT at DNM position, columns: number of reads calling REF/ALT at phasing SNP positions) and resolved inheritance by identifying which of the two alleles for the phasing SNP, REF or ALT, were disproportionately found on the same read as the DNM ALT. Significance was assessed using Fisher's exact test.

Haplotypes flagged as not programmatically resolved by either Medaka or pileup were manually reviewed using IGV v2.11.2[10], with visual inspection also used to validate all the above calls. Full details, and all scripts used for this analysis, are available at github.com/sjbush/pregcare. For one family (FAM38) the phase could not be resolved with long read-sequencing due to a homopolymeric stretch around the mutation site. For this family, an allele-specific PCR was performed (Supplementary Note 4).

References for Online Methods

1. Wilkie, A.O.M. & Goriely, A. Gonadal mosaicism and non-invasive prenatal diagnosis for 'reassurance' in sporadic paternal age effect (PAE) disorders. *Prenat Diagn* **37**, 946-948 (2017).
2. Acuna-Hidalgo, R.*, et al.* Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am J Hum Genet* **101**, 50-64 (2017).
3. Koelling, N.*, et al.* amplimap: a versatile tool to process and analyze targeted NGS data. *Bioinformatics* **35**, 5349-5350 (2019).
4. Bernkopf, M.*, et al.* Quantification of transmission risk in a male patient with a FLNB mosaic mutation causing Larsen syndrome: Implications for genetic counseling in postzygotic mosaicism cases. *Hum Mutat* **38**, 1360-1364 (2017).
5. Berger, J.O., Liseo, B. & Wolpert, R.L. Integrated likelihood methods for eliminating nuisance parameters. *Stat Sci* **14**, 1-22 (1999).
6. Venzon, D.J. & Moolgavkar, S.H. A Method for Computing Profile-Likelihood-Based Confidence-Intervals. *J R Stat Soc C-Appl* **37**, 87-94 (1988).
7. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, bty191-bty191 (2018).
8. Li, H.*, et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079 (2009).
9. Smigielski, E.M., Sirotkin, K., Ward, M. & Sherry, S.T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research* **28**, 352-355 (2000).
10. Thorvaldsdóttir, H., Robinson, J.T. & Mesirov, J.P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**, 178-192 (2012).