Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PD, Bodian DL, Solomon BD, Veltman JA, Deeken JF, Gilissen C, Niederhuber JE.

Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence.

***Nature Genetics* 2018,**

https://doi.org/10.1038/s41588-018-0071-6

Germline *de novo* mutation clusters arise during oocyte aging in genomic regions with increased double-strand break incidence

Jakob M. Goldmann[1]*, Vladimir B. Seplyarskiy[2,3]*, Wendy S.W. Wong[4]*, Thierry Vilboux[4], Pieter B. Neerincx[5,6], Dale L. Bodian[4], Benjamin D. Solomon[7,8], Joris A. Veltman[9,10], John F. Deeken[4], Christian Gilissen[9#], John E. Niederhuber[4,11#]

[1]Department of Human Genetics, Radboud Institute for Molecular Life Sciences, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands

[2]Division of Genetics, Brigham & Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.

[3]Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Bolshoi Karetny pereulok 19, Moscow 127994, Russia

[4]Inova Translational Medicine Institute (ITMI), Inova Health Systems, Falls Church, VA, USA

[5]Department of Genetics, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[6]Genomics Coordination Center, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[7]Department of Pediatrics, Inova Children's Hospital, Inova Health System, Falls Church, VA, USA

[8]Department of Pediatrics, Virginia Commonwealth University School of Medicine, 1201 E Marshall St, Richmond, VA, USA

[9]Department of Human Genetics, Donders Centre for Neuroscience, Radboud University Medical Center , Geert Grooteplein 10, 6525 GA Nijmegen, the Netherlands

[10]Institute of Genetic Medicine, International Centre for Life, Newcastle University, Newcastle upon Tyne, United Kingdom

[11]Johns Hopkins University School of Medicine, 733 North Broadway Street, Baltimore, MD, USA

*These authors contributed equally.

[#] These authors jointly supervised this work.

To whom correspondence should be addressed: *Christian.gilissen@radboudumc.nl* and *John.Niederhuber@inova.org*

Clustering of mutations has been observed in cancer genomes as well as for germline *de novo* mutations (DNMs). We identified 1,796 clustered DNMs (cDNMs) within whole-genome sequencing data from 1,291 parent-offspring trios to investigate their patterns and inferred a mutational mechanism. We found that the number of clusters on the maternal allele was positively correlated with maternal age and that these consist of more individual mutations with larger inter-mutational distances compared to paternal clusters. More than 50% of maternal clusters were located on chromosomes 8, 9 and 16, in regions with an overall increased maternal mutation rate. Maternal clusters in these regions showed a distinct mutation signature characterized by C>G transversions. Finally, we found that maternal clusters associate with processes involving double-stranded-breaks (DSBs) such as meiotic gene conversions and *de novo* deletions events. This suggest accumulation of DSB-induced mutations throughout oocyte aging as an underlying mechanism for maternal mutation clusters.

De novo mutations (DNMs) arise spontaneously in parental gametes and result in approximately 50-100 germline mutations in their offspring[1-4]. As such, DNMs are both drivers of evolution as well as a common cause of sporadic disorders. The total number of DNMs is highly correlated with paternal age and, to a lesser extent, with maternal age[2-4]. The paternal age effect, giving rise to about one additional DNM in the offspring per year of life of the father before conception, is thought to be due to the higher number of cell divisions that spermatogonial cells of older men have undergone prior to this period[5,6]. The mechanisms underlying the maternal age effect, giving rise to about one additional DNM per 4 years of life of the mother, are still unknown. Approximately 2-3% of all DNMs in the offspring occur in close spatial proximities (below 20kb) as clustered mutations[4,7-11]. These clustered DNMs (cDNMs) have a distinct nucleotide substitution spectrum with an enrichment of C>G mutations, suggesting mutational mechanisms different from unclustered DNMs[4,9,10,12,13]. The precise composition of the mutation spectrum also varies with the inter-mutational distances of the clusters[10,14]. Contrary to unclustered DNMs, no paternal bias has been observed for the number of cDNMs[4,9,12]. Here, we investigated cDNMs, their potential contribution to the paternal and maternal age effect on the total number of DNMs, and the possible mechanisms underlying their occurrence.

Whole genomes of 1,291 parent-offspring trios from the Inova Translational Medicine Institute longitudinal childhood study cohort were sequenced using Illumina HiSeq2000 with average 40x coverage by Illumina services (La Jolla, USA; **Table 1, Supplementary Table 1**). This cohort represents a sample of the general population of average health giving birth at a single hospital[15]. After quality control, we identified 73,755 high-confidence DNMs using a random forest classifier (**Online Methods**, **Supplementary Table 2**). We defined cDNMs as DNMs within the same individual with all pair-wise inter-mutational distances smaller than 20kb. In total we identified 1,796 cDNMs (2.4% of all DNMs) distributed across 799 clusters, with 2-10 mutations per cluster, of which 678 clusters (85%) consisted of exactly two mutations (**Supplementary Tables 3-6**). 144 cDNMs in 72 clusters were directly adjacent. By performing read-phasing, we successfully identified the parent-of-origin for 700 cDNMs (39.0% of all cDNMs) across 400 clusters (**Table 1, Supplementary Table 7-8**). In 98.0% (204/208) of the fully phased clusters, all cDNMs arose on the same allele, which is in line with our simulations of the false detection rate of cluster detection (**Supplementary Table 9**). In contrast to unclustered DNMs, we did not observe an excess of cDNMs on the paternal allele (202 maternal clusters and 198 paternal clusters, chi-square goodness-of-fit p=0.84). In addition, we created a validation dataset based on four independent studies with phased DNMs from whole-genome sequencing (WGS)[4,9,10,12], resulting in a total of 1,643 cDNMs across 745 clusters, with limited information on parental ages (**Table 1**, **Supplementary Table 10**).

To investigate the contribution of cDNMs to the parental age effects, we used a linear regression model to correlate the age of each parent with the number of phased cDNMs in the offspring. Although the number of paternal cDNMs did not show a significant correlation with the paternal age (p=0.087), we found a highly significant correlation of maternal cDNMs with maternal age ($p<10^{-10}$). This effect was similar in our replication cohort (maternal p=0.00155 and paternal p=0.319, **Supplementary Figures 1,2**). In the primary cohort, the cDNMs accounted for 23% (95% c.i. 7-38%) of the maternal age effect (p-value for maternal age effect of unclustered DNMs $p=1.5x10^{-19}$). For the clusters where only a subset of cDNMs could be phased, we extrapolated the parent-of-origin. Based on this extrapolation, we also

observe a significant paternal age effect of a smaller amplitude than the maternal age effect (paternal effect size p=0.026/year, $p=8x10^{-7}$, maternal effect size 0.041/year, $p=3x10^{-11}$). While in the primary cohort, only 5% of the probands with the youngest mothers had one or more maternal cDNMs per genome, this was more than 5 times higher (risk ratio test, $p=1.4x10^{-11}$; c.i. 3.0-9.4) in probands from the oldest mothers (27% having a maternal cDNM, **Figure 1a**). This difference was not significant for the paternal cDNMs (13% vs 19%; risk ratio test p=0.08; risk ratio 1.42; 95% c.i. 0.95-2.12). In the replication cohort, the risk ratio was 3.02 for maternal cDNMS (c.i. 1.22-7.45; p=0.011, **Supplementary Figure 3**) and 0.60 (c.i. 0.30-1.22; p=0.15) for paternal cDNMs.

We found that this maternal age effect of clusters stems mostly from clusters with inter-mutational distances greater than 1kb (**Figures 1b,c**, **Supplementary Tables 11 and 12, Supplementary Figure 3**). Strikingly, the maximum number of DNMs in the phased clusters of an individual correlates positively with maternal age ($p<10^{-10}$, replication cohort $p<10^{-4}$), but is correlated only marginally significant with paternal age (p=0.050, replication cohort p=0.408, **Figure 1d,e, Supplementary Figure 3**). Clusters with more than two mutations were 4.2 times more likely to contain maternal cDNMs than paternal cDNMs (95% c.i. 2.5 – 7.6; $p=1.7x10^{-7}$). These results show that maternal clusters contain more cDNMs with larger inter-mutational distances.

We previously observed that maternal DNMs are enriched within specific genomic regions on chromosomes 8 and 16[4]. In this study, we found that 58.4% of maternal cDNMs localize to chromosomes 8, 9 and 16 ($p<10^{-16}$, replication cohort $p<10^{-16}$, Chi-square test; **Figure 2a, Supplementary Figures 4 and 5**). This in contrast to paternal cDNMs for which the number correlates with chromosome length ($R^2=0.72$, $p=6*10^{-7}$, replication cohort $R^2=0.43$, p=0.001). The maternal cDNMs on these three chromosomes occur specifically in regions that are also enriched for maternal unclustered DNMs (**Figure 2b, Supplementary Figures 6 and 7, Supplementary Note 1**) and their mutation spectrum is strongly enriched for C>G substitutions compared to other maternal cDNMs (**Figure 2c,d**, bootstrapping p=0.022). These observations are further supported by the patterns of clusters with more than two cDNMs, which are more likely to be on the maternal allele. These clusters are also enriched on the chromosomes 8, 9 and 16 (Chi-square test $p=3x10^{-09}$), and show an excess of C>G substitutions (Chi-square test $p=4.5x10^{-11}$). Taken together, this suggests a different mutational mechanism for maternal cDNMs in these regions compared to the rest of the genome.

To confirm these findings, we created a dataset of (unphased) clustered SNP variants based on publically available population-based genetic data[16] (**Online Methods**). This resulted in 1,146,891 clustered SNPs (cSNPs) across 522,487 clusters (**Supplementary Table 13**). We found that cSNPs on chromosomes that are enriched for maternal cDNMs are enriched for C>G substitutions (bootstrapping test, see **Online Methods,** p<0.001, **Figure 2e**). To further investigate this association, we calculated a genome-wide score for C>G cSNP enrichment (**Supplementary Methods**) and found that the number of maternal cDNMs in a region is significantly correlated with high C>G scores (Poisson regression $p<10^{-16}$ for maternal cDNMs, p=0.33 for paternal cDNMs, **Supplementary Figure 8**). Using this method we also identified an additional region on chromosome 2 that is enriched for maternal cDNMs (**Figure 2f**). This strong association between C>G scores of cSNPs with maternal cDNMs

highlights the maternal clusters' profound contribution to population polymorphisms in these regions.

The observed age-effect of maternal cDNMs suggests underlying mechanisms that are active during oocyte aging, a process that has been associated with the decreasing efficiency of double-strand break (DSB) repair[17-19]. We therefore hypothesized that the maternal-aging associated clusters arise via a DSB-associated mechanism and investigated the occurrence of cDNMs at regions that are associated with DSBs. As proxies for DSB sites we used (1) sites of *de novo* meiotic gene conversion (MGC), (2) the flanking regions of *de novo* CNV breakpoints in our cohort, and (3) known recombination hotspots[20].

We used MGC sites from Halldorsson et al.[21] and found that these events co-localize with maternal cDNMs significantly more often than expected by chance (p=0.002, permutation testing, **Figure 3a**, **Supplementary Table 14**). This association is not significant for paternal MGCs with paternal cDNMs (p=0.609).

In our primary cohort, we identified 45 high-quality *de novo* CNVs, of which 5 have a total of 17 DNMs within 100kb flanking the breakpoints (**Figure 3b, Supplementary Methods**). Exactly 12 of these 17 DNMs are cDNMs, which constitutes a high enrichment ($p = 2.58 \times 10^{-16}$, Fisher's exact test). For 6 of these DNMs the parent-of-origin was resolved and in all cases the DNMs arose from the maternal allele (p=0.03, Fisher's exact test). In concordance with this, all 5 CNVs are deletions of the maternal allele (**Supplementary Table 15**). An arrangement of several DNMs and a *de novo* deletion on the same allele within the same generation is very unlikely to occur by chance and suggests a single event as a common cause. In our replication cohort, we also discovered 5 *de novo* deletion events. Two of these CNVs have a total of 4 DNMs from the same individual within 100kb of the CNV breakpoints, and two of these are within 20kb of each other (**Supplementary Figure 9**), again showing an enrichment of cDNMs (p=0.002, Fisher's exact test). Interestingly, cSNPs were significantly closer to CNV breakpoints than expected by chance (**Figure 3c**, Mann-Whitney test on 1% of data $p<10^{-9}$), corroborating the co-segregation of CNV events and clustered mutations.

Finally, we used gender specific recombination scores[20] to assess whether cDNMs occur more often at regions of high recombination. We did not find a significant overlap of maternal cDNMs with regions of high maternal recombination (p=0.204 permutation testing, **Figure 3d, Supplementary Table 14**). Nevertheless, genomic regions with maternal cDNMs had higher sex-matched recombination scores than regions with only unclustered maternal DNMs (primary cohort p=0.019, replication cohort p=0.13) and higher than regions of paternal cDNMs (primary cohort p=0.004, replication cohort p=0.29; **Supplementary Figure 10**). In addition, genomic regions with cSNPs have significantly higher recombination rates than genomic regions without cSNPs ($p=3.91 \times 10^{-49}$). Interestingly, our observed association of cDNMs with recombination rates is much smaller than the observed association with MGCs. This is in line with the maternal age effect of MGCs being larger compared to the maternal age effect of the crossover rate[21,22]. Campbell et al. found that, with increasing maternal age, recombination occurs more frequently outside of recombination hotspots[23]. In addition, these events were increasingly deregulated, appearing in closer proximity of each other than expected based on models of crossover interference. The fact that recombination events have shown to be mutagenic[24-26] suggests that this increase in deregulated recombination events may be the underlying cause of cDNM formation. In this study, we found that that chromosomes 8, 9 and 16 are heavily enriched for maternal clusters and

strikingly these chromosomes also have the highest degree of cross-over events escaping interference[23].

Additionally, cDNM mutational spectra, and in particular those of maternal cDNMs, are very similar to the previously identified signature of somatic mutations caused by deficiency in homologous recombination repair of DSBs[27,28] (Signature 3, **Supplementary Figure 11**). The proband's parents are very unlikely to suffer from DNA repair deficiencies such as those underlying cancer mutation profiles, therefore this finding is in agreement with a key role for imperfect DSB repair after unregulated recombination in the formation of maternal mutation clusters. However, we found no statistical association between variants in genes involved in homologous recombination repair or in establishing recombination sites[29-31] (**Supplementary Tables 16 and 17**).

Although the formation of clustered mutations has the potential to be highly deleterious, there seems to be selection in favor of high recombination rates in ageing oocytes[32,33]. It has been argued that these high recombination rates provide protection against aneuploidies[33-34], the risk of which increases with maternal age. Taken together, our results show that deregulated recombination is a likely cause for DNM clusters, whereas replicative errors are not a likely cause. A recent paper that studied genome-wide *de novo* mutations in a cohort of 1,548 Icelanders also found that clustered mutations increase faster with maternal than paternal age[35]. In addition, the authors observed a non-uniform distribution of these events across the genome[35] corresponding with the regions that we reported here.

URLs

goleft indexcov: https://github.com/brentp/goleft/tree/master/indexcov

agg gvcf aggregation tool: https://github.com/Illumina/agg

Acknowledgements

This study was funded by the Inova Health System with support from Fairfax County and the philanthropic support from the Odeen family. We thank the Inova translational medicine institute staff for supporting the study. We also thank the families who participated in the genomic studies that made this research possible. This work was partly financially supported by grants from the Netherlands Organization for Scientific Research (916-14-043 to C.Gilissen and 918-15-667 to J. Veltman), and the European Research Council (ERC Starting grant DENOVO 281964 to J. Veltman).

This study makes use of data generated by the Genome of the Netherlands Project. A full list of the investigators is available from www.nlgenome.nl. Funding for the project was provided by the Netherlands Organization for Scientific Research under award number 184021007, dated July 9, 2009 and made available as a Rainbow Project of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL). The sequencing was carried out in collaboration with the Beijing Institute for Genomics (BGI).

Author contributions

C.G. and J.E.N. designed the study. J.M.G., V.B.S., and W.S.W.W. performed the data analyses. W.S.W.W carried out QC, and *de novo* mutations calling. T.V. performed the Sanger validation. B.D.S., J.F.D, and J.E.N. supervised the data collection, sequencing and writing of the manuscript. D.B. assisted in data analyses and interpretation. J.M.G., V.B.S., W.S.W.W., J.A.V. and C.G. drafted the manuscript. P.B.N. acquired part of the replication data. All authors contributed to the final version of the paper.

Competing Financial Interests

The authors do not declare any competing financial interests.

References

1. Veltman, J.A. & Brunner, H.G. De novo mutations in human genetic disease. *Nature reviews. Genetics* **13**, 565-75 (2012).
2. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471-5 (2012).
3. Wong, W.S.W. *et al.* New observations on maternal age effect on germline de novo mutations. *Nature communications* **7**, 10486 (2016).
4. Goldmann, J.M. *et al.* Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics* (2016).
5. Crow, J.F. The origins, patterns and implications of human spontaneuos mutation. *Nature Reviews Genetics* **1**, 40-47 (2000).
6. Segurel, L., Wyman, M.J. & Przeworski, M. Determinants of mutation rate variation in the human germline. *Annu Rev Genomics Hum Genet* **15**, 47-70 (2014).
7. Michaelson, Jacob J. *et al.* Whole-Genome Sequencing in Autism Identifies Hot Spots for De Novo Germline Mutation. *Cell* **151**, 1431-1442 (2012).
8. Schrider, D.R., Hourmozdi, J.N. & Hahn, M.W. Pervasive multinucleotide mutational events in eukaryotes. *Current biology : CB* **21**, 1051-4 (2011).
9. Yuen, R.K. *et al.* Genome-wide characteristics of de novo mutations in autism. *npj Genomic Medicine* **1**, 16027 (2016).
10. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLOS Genetics* **12**, e1006315 (2016).
11. Terekhanova, N.V., Bazykin, G.A., Neverov, A., Kondrashov, A.S. & Seplyarskiy, V.B. Prevalence of Multinucleotide Replacements in Evolution of Primates and Drosophila. *Molecular Biology and Evolution* **30**, 1315-1325 (2013).
12. Francioli, L.C. *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nature Genetics* **advance on**(2015).
13. Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet* **48**, 126-133 (2016).
14. Harris, K. & Nielsen, R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome research* **24**, 1445-54 (2014).
15. Bodian, D.L. *et al.* Utility of whole-genome sequencing for detection of newborn screening disorders in a population cohort of 1,696 neonates. *Genetics in medicine : official journal of the American College of Medical Genetics* (2015).
16. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015).
17. Titus, S. *et al.* Impairment of BRCA1-related DNA double-strand break repair leads to ovarian aging in mice and humans. *Science translational medicine* **5**, 172ra21 (2013).
18. White, R.R. & Vijg, J. Do DNA Double-Strand Breaks Drive Aging? *Molecular Cell* **63**, 729-738 (2016).
19. Oktay, K., Turan, V., Titus, S., Stobezki, R. & Liu, L. BRCA Mutations, DNA Repair Deficiency, and Ovarian Aging. *Biology of reproduction* **93**, 67 (2015).
20. Kong, A. *et al.* Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467**, 1099-103 (2010).
21. Halldorsson, B.V. *et al.* The rate of meiotic gene conversion varies by sex and age. *Nature Genetics* (2016).

22. Martin, H.C. *et al.* Multicohort analysis of the maternal age effect on recombination. *Nature Communications* **6**, 7846 (2015).
23. Campbell, C.L. *et al.* Escape from crossover interference increases with maternal age. *Nature Communications* **6**, 6260 (2015).
24. Arbeithuber, B., Betancourt, A.J., Ebner, T. & Tiemann-Boege, I. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences* **112**, 2109-2114 (2015).
25. Lercher, M.J. & Hurst, L.D. Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet* **18**, 337-40 (2002).
26. Webster, M.T. & Hurst, L.D. Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* **28**, 101-9 (2012).
27. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-21 (2013).
28. Zámborszky, J. *et al.* Loss of BRCA1 or BRCA2 markedly increases the rate of base substitution mutagenesis and has distinct effects on genomic deletions. *Oncogene* **36**, 746-755 (2017).
29. Moynahan, M.E., Chiu, J.W., Koller, B.H. & Jasin, M. Brca1 controls homology-directed DNA repair. *Molecular cell* **4**, 511-8 (1999).
30. Patel, K.J. *et al.* Involvement of Brca2 in DNA repair. *Molecular cell* **1**, 347-57 (1998).
31. Baudat, F. *et al.* PRDM9 Is a Major Determinant of Meiotic Recombination Hotspots in Humans and Mice. *Science* **327**, 836-840 (2010).
32. Kong, A. *et al.* Recombination rate and reproductive success in humans. *Nature Genetics* **36**, 1203-1206 (2004).
33. Ottolini, C.S. *et al.* Genome-wide maps of recombination and chromosome segregation in human oocytes and embryos show selection for maternal recombination rates. *Nature Genetics* **47**, 727 (2015).
34. Middlebrooks, C.D. *et al.* Evidence for dysregulation of genome-wide recombination in oocytes with nondisjoined chromosomes 21. *Human Molecular Genetics* **23**, 408-417 (2014).
35. Jonsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519-522 (2017).

Figure 1: Differences between maternal and paternal cDNMs (**a**) The fraction of probands with maternal and paternal clustered mutations (y-axis), grouped by parental age quantiles. Error bars indicate the binomial 95% confidence intervals. Labels on the lower axis indicate age ranges of the respective groups and group sizes. See **Supplementary Figure 1** for graphs and regression lines. (**b**) The number of paternal and maternal cDNMs (y-axis) stratified by the distance to the nearest other cDNM (x-axis). (**c**) The size of paternal and maternal age effect of clusters with at least one phased cDNM (y-axis) by inter-mutational distance (x-axis). Whiskers indicate the 95% confidence interval. (**d**) Age of fathers at conception and (**e**) age of the mothers at conception (y-axis) by the number of mutations in the offspring's largest mutation cluster (x-axis). We considered only clusters where at least one cDNM is on the allele from the respective parent (paternal allele for **d** and maternal allele for **e**). Numbers indicate the size of each group. Boxplot compartments: box: interquartile range; line: median; whiskers: extreme values <1.5 × interquartile ranges from box borders).

Figure 2: Patterns of cDNMs across the chromosomes. (**a**) The proportion of phased cDNMs per chromosome. Error bars indicate the binomial 95% confidence intervals. (**b**) Overview of chromosome 16 region enriched for maternal cluster mutations. X-axis and ideogram indicate chromosomal position. The red and blue histograms indicate the number of maternal cDNMs and paternal cDNMs identified in this study, respectively. The pale red and pale blue histograms indicate the number of maternal and paternal unclustered DNMs (ucDNMs). The lowest track indicates normalized cSNP C>G score, which is predictive for maternal DNMs. (**c**) The nucleotide substitution spectrum of maternal and paternal clusters and unclustered DNMs. The star indicates a significant difference assessed by bootstrapping (**Online Methods**). Error bars indicate the binomial 95% confidence intervals. (**d**) The nucleotide substitution spectrum of cDNMs by location. Error bars indicate the binomial 95% confidence intervals. (**e**) The nucleotide substitution spectrum of polymorphism-derived clustered mutation by location. The star indicates a significant difference assessed by bootstrapping. Error bars indicate the binomial 95% confidence intervals. (**f**) Region with increased maternal mutation rate on chromosome 2 (region displayed chr2:1-100,000,000bp; region with maternal cDNMs chr2:40,000,000-60,000,000).

Figure 3: cDNMs and sites likely affected by DSBs. (**a**) Z-scores of expected and observed overlaps of cDNM clusters in our cohort and sex-matched meiotic gene conversion in another cohort[21]. Diamonds: observed values, boxplot compartments: box: interquartile range; line: median; whiskers: extreme values <1.5 × interquartile ranges from box borders. (**b**) DNMs detected close to sites of *de novo* CNVs. Data of DNMs is listed in **Supplementary Table 15**. (**c**) cSNP density close to CNV breakpoints (**Online Methods**). (**d**) Z-scores of expected and observed overlap cDNM clusters and sex-matched recombination hotspots. Symbols and boxplots as in (**a**).

Main Tables

Table 1: Overview of cohorts

Cohort		Total number	Paternal number	Maternal number
Primary cohort	Probands	1,291		
	DNMs	73,755	20,196	5,547
	cDNMs	1,796	323	377
	Clusters	799	110 (+88)	94 (+108)
Replication cohort	Probands	1,557		
	DNMs	74,395	9,466	2,796
	cDNMs	1,643	133	195
	Clusters	745	40 (+49)	67 (+46)

Numbers of probands, DNMs, cDNMs and clusters of the cohorts used in this study. The numbers in brackets indicate clusters where not all cDNMs could be phased for the respective parent.

Online Methods

Cohort

The cohort used in this study is from Inova Translational Medicine Institute's Longitudinal Childhood Genome Study (previously referred to as the First 1,000 Days of Life and Beyond study), which represents a population cohort in good general health[4,15]. The study was conducted by the Inova Translational Medicine Institute and approved by both the Inova and Western Institutional Review Boards (study 20120204). Parents and the newborns were recruited at Inova Fairfax Hospital between 2012 and 2014. A summary of participants' ages is given in **Supplementary Table 1**.

Whole genome sequencing

Sample preparation, processing and whole-genome sequencing (WGS) have been previously described[4,15]. Briefly, DNA was extracted from peripheral blood obtained from each family member. Whole genome sequencing using paired-end 100bp reads (median fragment length is 375) at an average 40X coverage was performed by Illumina Services (San Diego, CA). The sequenced reads were aligned to the hg19 human reference genome by the ISAAC aligner[36] with the Illumina Whole Human Genome Sequencing Service Informatics Pipeline version 2.01 - 2.03.

To systematically analyze the data quality of all sequencing reactions, a principal component analysis on scaled summary statistics was performed (**Supplementary Figure 12, Supplementary Table 18**). The first principal component is highly correlated to average sequencing coverage; a group of outlying points refers to a group of sequencing reactions with average genome coverage above 70x. The second principal component is associated with the date of sequencing and the version of the software used for analysis, respectively. The third principal component is related to the estimated ancestries of the sequenced individuals.

DNM calling and quality control

Callable regions of each sample were determined by CallableLoci in GATK version 3.1. The number of callable bases by batch is shown in **Supplementary Figure 13.** The batch number does not significantly influence the number of DNMs called (**Supplementary Table 19**). Joint calling using HaplotypeCaller, PhaseByTransmission and ReadBackPhasing in GATK version 3.1 were performed on each of the 1,315 trios in the canonical autosomes[37]. The putative *de novo* mutations were generated from taking PASS filter calls with heterozygous in the proband and homozygous reference in both parents in the PhaseByTransmission results in each trio. We have previously analyzed 816 trios[4], of which, 65 trios were also sequenced by the Illumina services with pipeline version 2.0.0-2.0.1, and are not part of this cohort. These 65 trios sequenced by Illumina have gone through the same pipeline to generate a set of putative DNMs. We defined the positive set as those putative DNMs that overlap with previous identified DNMs identified using Complete Genomics (CG) technology (2,670), as well as those that were validated by Sanger sequencing (34), the total number in the true positive set is 2,704. The negative set consists of 50 random putative DNMs in each of the 65 trios that are not in the previously identified set by CG (50*65=3,250), as well as 4 false positive sites identified by Sanger, the total number of negative sites is 3,254. We note that some of the sites in the negative set are true positives but the number is likely to be low. The test

set which consists of the positive and negative sets was split by 90:10 ratio into training set and test set. The R libraries randomForest version 4.6.10 and caret version 6.0.52 were used to train the random forest classifier. The OOB estimate of error rate on training set is 1.77% and the error rate in the test set is 2.18%. The features used in the classifier and their relative importances are shown in **Supplementary Table 20**. The confusion matrix for the test set is shown in **Supplementary Table 21**.

In order to minimize the bias due to mapping errors and coverage differences, we further filtered the predicted DNMs by (1) callable regions in the cohort: A site is in the callable region if at least 90% of the samples has the PASS status by GATK CallableLoci[37], (2) good mappability regions, where mappable is defined according to the CRG 100mer (file wgEncodeCrgMapabilityAlign100mer.bw from UCSC Table Browser) being equal to 1[38], sites also called by the Illumina Isaac Small Variant Caller, and sites with FS (FisherStrand test score) >=20, and sites with exceptionally high or low PL values (**Supplementary Table 22**). An overview of the filtering procedure is given in **Supplementary Table 2**.

In the initial sequencing cohort, there were 12 monozygotic twin pairs, 29 dizygotic twin pairs and a family of three trizygotic siblings. In order to assess the consistency in *de novo* calling, we investigated the concordance percentages of monozygotic and dizygotic families (**Supplementary Table 23** and **Supplementary Table 24**). DNM calls in monozygotic twins are on average 95% concordant, the dizygotic average concordance is 0.1%. This is similar to concordance ratios observed previously[4].

We removed 1 trio with an exceptional high number of DNM calls, 8 trios with a large chromosomal anomaly in either the proband or one of the parents and removed (arbitrarily) one of the monozygotic twins in each set. After performing simple multiple linear regression, 3 samples have a significant Bonferroni p-value for studentized residuals (Bonferroni corrected $p<0.05$) and are removed from the cohort, resulting in 1,291 trios (**Supplementary Table 2**). We investigated the effect of average genome coverage on the filtered data. The results are shown in **Supplementary Figure 14**.

The method for determining the parent-of-origin of DNMs with Illumina WGS trio data was previously described[3,4]. Briefly, GATK PhaseByTransmission was used to assign parent-of-origin to informative heterozygous SNPs in the proband, GATK ReadBackPhasing was used to link DNMs to these informative SNPs. If contradictory markers were linked to the same DNM, it would not be assigned a parent-of-origin. Overall, 227 of the 25,970 filtered DNMs are linked to contradictory markers (0.87%).

Clustered DNMs

We defined cDNMs as DNMs on the same chromosome of the same individual within 20kb of each other. In order to estimate the chance of two DNMs being closer than 20kb on the same chromosome, we simulated 70,000 mutations at random positions within the callable and mappable genome. The randomized positions were given sample IDs as in the set of observed DNMs and the distances were calculated. We found that the false discovery rate of cluster detection is 0.0375 at a threshold of 20kb (**Supplementary Table 9**). Statistics on the number of cDNMs per cluster are given in **Supplementary Table 3**.

For analyses on clusters we extrapolated the parent-of-origin by considering all cDNM to originate from the same allele.

Sanger validation

We performed Sanger validation on 163 clustered DNM sites on the proband and his or her parents, of which 62 are on chromosomes 8, 9 and 16 (**Supplementary Table 25A**). Overall, 91.3% of the DNMs are validated, 92.7% on chromosomes 8,9 and 16 vs. 90.4% on other chromosomes. The number of sites validated in each pipeline version is proportional to the number of trios sequenced in each pipeline (**Supplementary Table 25B**). There is no significant difference in the proportion of sites validated in each pipeline (P = 0.92, Fisher's exact test). No evidence of the mutations was found at any site in the parents. All of the invalidated sites were due to lack of evidence in the proband.

Clustered polymorphism variants

We use polymorphism data from the 1000 Genomes Project Consortium[16]. We only considered non-singleton variants with below 1% derived allele frequency, using the ancestral variant determined by The 1000 Genomes Project Consortium. Clusters were defined as two or more SNPs at distances between 10-1000 nucleotides from each other, such that all the genotypes carrying the derived allele for one of the SNPs also carry the derived allele for any other SNP within the cluster. We show that cSNP spectra are similar to cDNM spectra: enriched by C>G mutations and depleted by CpG>TpG mutations, compared to unclustered DNMs. We restricted ourselves to distances between cSNPs shorter than 1000 nucleotides, because of two reasons. First, the probability of recombination scales with the distance between SNP positions and thus longer clusters are more frequently disrupted. Second, the probability to observe two independent mutations on the same haplotype would be ~20 fold higher in 1-20 kb range than in 0-1 kb range. In contrast we observe 806 cDNMs in 0-1 kb range and 990 cDNMs in 1-20 kb range. Therefore, we expect a higher noise to signal ratio for larger distances. In line with this, the spectra of larger clusters are progressively less similar to cDNMs (**Supplementary Figure 15**). For analyzing the density of cSNPs around CNV breakpoints, we calculated the distances between cSNPs and CNVs on the chromosomes of each individual. We only considered cSNPs flanking CNVs, but not within its body. These distances were compared to the distances between cSNPs and the CNVs on the same chromosome of a random other individual.

Statistical assessment of the maternal age effect

For analyzing the parental age effects on both the number of clusters as well as the number of cDNMs, linear models were fitted using the R statistical environment version 3.3.3 with standard settings. The reported p-values reflect the difference from zero of the respective age effect.

Extrapolations of DNM phasing were done by assigning a cluster's unphased DNMs the same allele as the phased ones. In order to correct for the false detection rate of 3.75% (**Supplementary Table 9**), we sampled 1,000 subsets of 100%-3.75%= 96.25% of cDNMs and calculated the age effects on all of them. We report the median effect size and the median p-value.

For comparing proband groups' risks for having DNM clusters we used risk ratio statistics as implemented in the R package "epitools". For assessing the enrichment of C>G substitutions on chromosomes 8, 9 and 16, we re-sampled the chromosome annotation 1,000 times and compared the difference of the fractions of C>G mutations on the special chromosomes and the remaining autosomes to the observed value.

Statistical assessment of nucleotide substitution profiles

The significance of differences between nucleotide substitution profiles was assessed by bootstrapping: We resampled the grouping variable 1,000 times and compared the resulting random groups to the observed groups. For assessing C>G enrichment we calculated p-values by counting the number of random groups where the difference in C>G fractions between the groups is equal to or larger than in the observed set and dividing by the number of samplings.

Statistical assessment of DSB proxy regions overlap

For calculating distributions on the expected number of overlaps between DNM clusters and DSB proxy regions we used permutation testing as implemented in the R library RegioneR[39]. DNM cluster regions were defined as the positions of cDNMs and the space between them. Recombination hotspots were defined as genomic sites with a recombination-score above 10[20]. Meiotic gene conversions were filtered for non-crossover gene conversions only detected in the chip dataset[21]. In absence of knowledge about the exact boundaries of the conversion streak and confronted with the majority of meiotic gene conversions being observed only in one SNP, we defined the positions of meiotic gene conversions as the distance between the two SNPs adjacent to the SNPs affected by conversion. The cluster regions were randomized 500 times to genomic positions where at least 1000 base pairs were within the callable and mergable subset of the genome. For every randomization round the number of cluster positions overlapping DSB proxy regions was compared to the observed number of overlaps. For the calculation of z-scores of an overlap count the mean number of overlaps was subtracted before division by the standard deviation of the number of overlaps.

De novo CNVs

In the primary cohort, we called *de novo* CNVs using both coverage-based method FREEC[40] and read-pair based method Manta[41]. We also calculated window based normalized coverage with "goleft indexcov". For each proband, we called CNVs using the default options in FREEC with the proband as the case and one of the parents as control. We then required the CNVs subtracted from each parent to have 90% reciprocal overlap, with copy number equals 1 or 3, both parents have the mean normalized coverage between 0.85 and 1.15 in the region, the proband have mean normalized coverage smaller than 0.85 or greater than 1.15 in the region, with length greater or equal to 10kb. We performed joint calling for each trio with Manta using default options. We then filter for SV type being DEL or DUP, proband with GT equals to 0/1 and both parents with GT equal to 0/0, proband's PR and SR for ALT allele >=3 and the proportion of PR and SR for ALT >=0.2, parents' proportion of PR and SR for ALT <=0.05.

In the complete genomics data in the replication cohort, we required the *de novo* CNV to be called by both coverage-based and read-based methods. For the coverage-based method, we first subtracted CNVs in the proband from one of the parents using the

cnvSegmentsDiploidBeta files, and then we intersect the two putative *de novo* CNV files subtracted from each parent, with 90% overlap, and size >9999. For the read-based method, we subtracted highConfidenceSvEventsBeta file from the proband from allSvEventsBeta file from each of the parents, and intersected the two subtracted files requiring 90% overlap. The final list of *de novo* CNVs is generated by intersecting the coverage-based and read-based files from the same proband, requiring 90% overlap. Bedtools 2.22.0 was used to carry out region subtractions and intersections[42].

Mutation signatures

A large set of mutational signatures is known from cancer studies[27], some of which are well annotated with mutational influences. To fit the patterns of our DNMs to these signatures we used an algorithm similar to the one described in[43]: a non-negative least-squares algorithm finds the mixture of known signatures that describes best the observed pattern. In order to get an indication of the robustness of the fitted mixture of signatures, a bootstrapping analysis was done. The mutations of a group were resampled 1,000 times with replacement and the standard deviation as well as the 95% confidence intervals of each fitted signature were calculated.

Single variant association study of parents genotype in *BRCA1, BRCA2* and *PRDM9* with number of phased cDNMs in the proband

The small variants in the autosomes were merged using "agg" with Illumina genome VCF files using default parameters. No sample had a call rate <90%. In this analysis, only those variants in *BRCA1*, *BRCA2*, *PRDM9*, and marker rs2914276, with call rate >90%, no significant deviation from Hardy-Weinberg equilibrium (P>0.001), and with minor allele frequency >0.005 were included. No LD pruning was performed. If a parent has more than 1 offspring in the cohort (twins or siblings), only one of the sibling's number of phased cDNMs is kept as the phenotype for the respective parent. The association analysis was performed with Plink v.1.90b[44] with additive model on paternal genotypes with paternal number of cDNMs, using paternal age at conception, and father's first 3 PCs as covariates; and on maternal genotypes with maternal number of cDNMs, using maternal age at conception, and mother's first 3 PCs as covariates , respectively. The association study included 1,247 fathers and 1,247 mothers. No variant reached significance (P<0.05) after Bonferroni correction.

Data availability

De novo mutation calls used in this manuscript will be available in dbGaP, by the accession code phs001522.v1.p1.

Code availability

Code available upon request.

Methods-only references

36. Raczy, C. *et al.* Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041-2043 (2013).
37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297-303 (2010).
38. Derrien, T. *et al.* Fast Computation and Applications of Genome Mappability. *PLoS ONE* **7**, e30377 (2012).
39. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* **11**, btv562 (2015).
40. Boeva, V. *et al.* Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**, 423-5 (2012).
41. Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics (Oxford, England)* **32**, 1220-2 (2016).
42. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841-2 (2010).
43. Blokzijl, F., Janssen, R., Van Boxtel, R. & Cuppen, E. MutationalPatterns: an integrative R package for studying patterns in base substitution catalogues. *bioRxiv* (2016).
44. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics* **81**, 559-75 (2007).
45. Glusman, G., Caballero, J., Mauldin, D.E., Hood, L. & Roach, J.C. Kaviar: an accessible system for testing SNV novelty. *Bioinformatics (Oxford, England)* **27**, 3216-7 (2011).

a

Paternal allele

Maternal allele

Parent of origin

Father

Mother

Fraction of probands with phased cDNMs

0.0

0.1

0.2

0.3

1
19-29
(n=259)

2
29-32
(n=258)

3
32-35
(n=258)

4
35-38
(n=258)

5
38-59
(n=258)

1
18-28
(n=259)

2
28-30
(n=258)

3
30-33
(n=258)

4
33-36
(n=258)

5
36-48
(n=258)

Age quantile

b

Parent of origin

Father

Mother

No. cDNMs

0

20

40

60

0bp

1−
10bp

11−
50bp

51−
100bp

101−
500bp

501bp−
1kb

1kb−
5kb

5kb−
20kb

Inter−mutational distance

c

Age effect
(phased clusters/year)

0.0000

0.0025

0.0050

0.0075

0.0100

0bp

1−
10bp

11−
50bp

51−
100bp

101−
500bp

501bp−
1kb

1kb−
5kb

5kb−
20kb

Inter−mutational distance

d

Father's age

20

30

40

50

60

1107

165

14

4

1

0

2

3

4

5

Number of mutations in largest cluster

e

Mother's age

20

30

40

1114

117

39

11

4

2

3

1

0

2

3

4

5

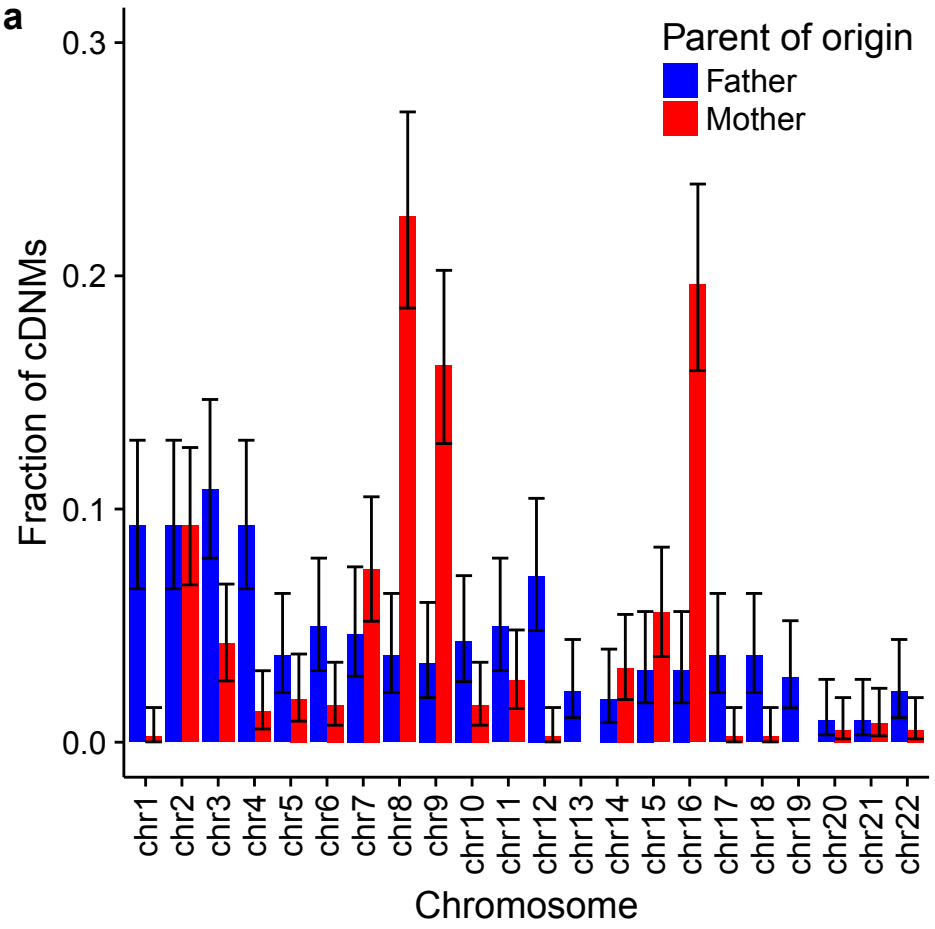
6

7

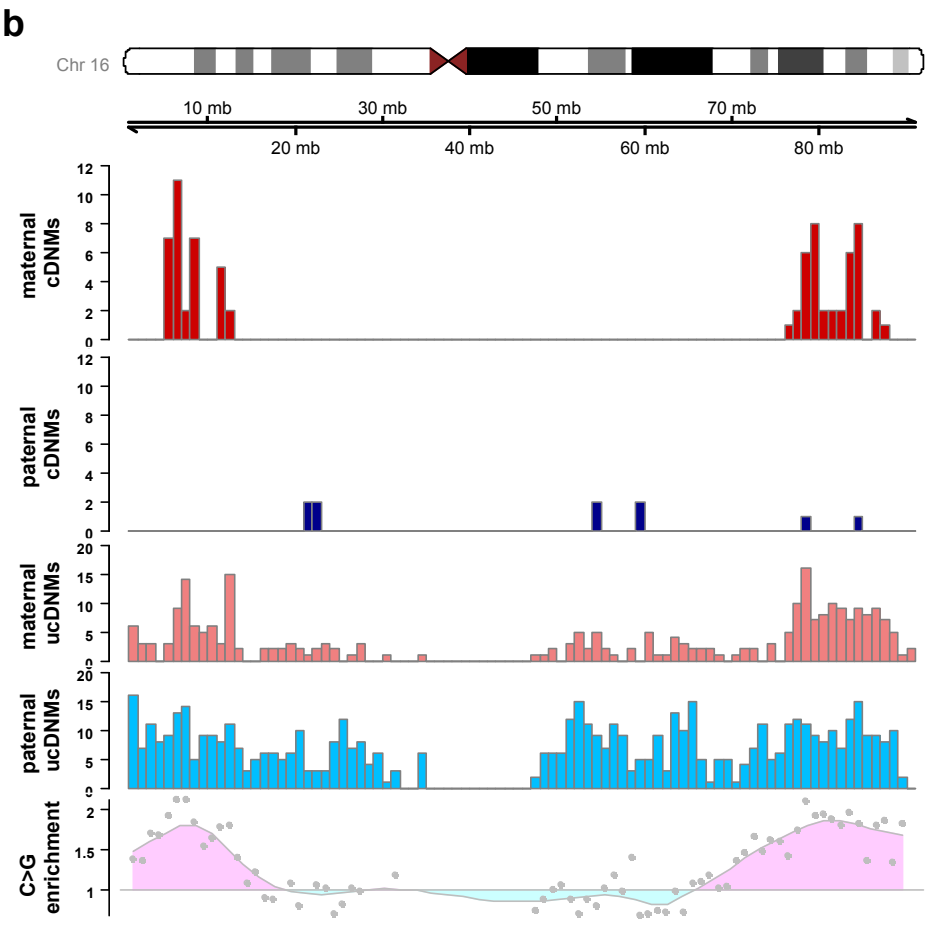
8

Number of mutations in largest cluster

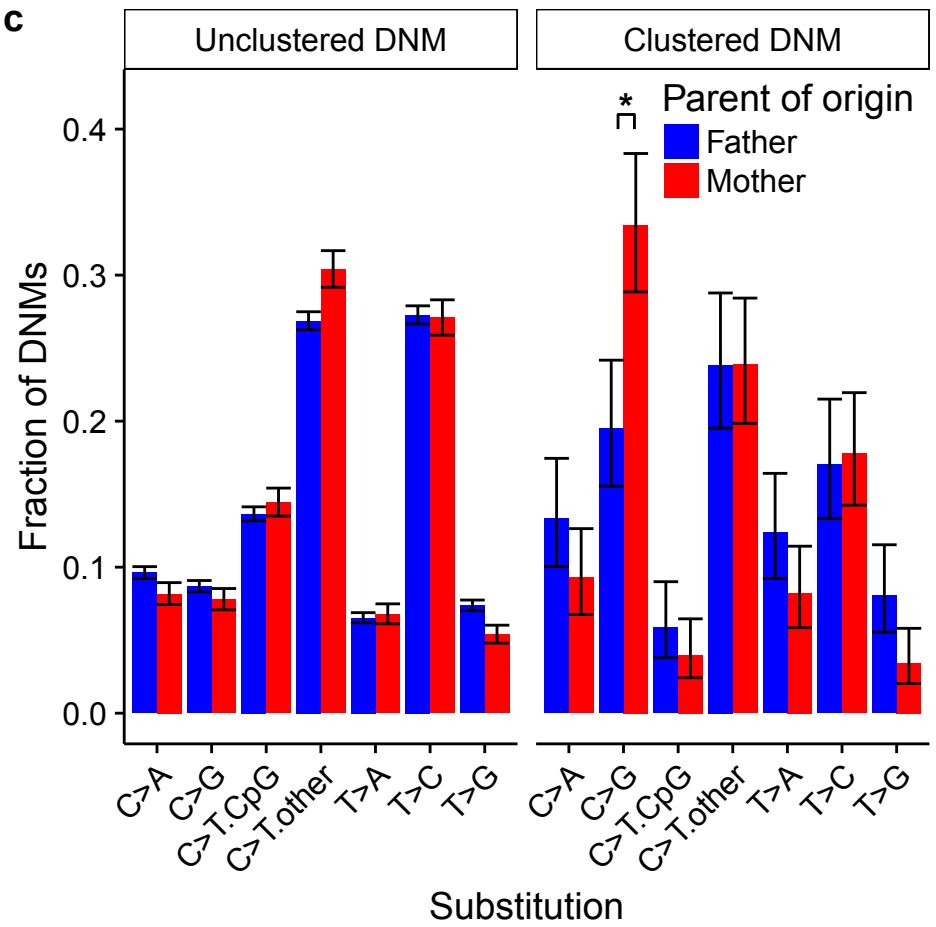
a


Parent of origin
Father
Mother
Fraction of cDNMs
0.0
0.1
0.2
0.3
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
Chromosome

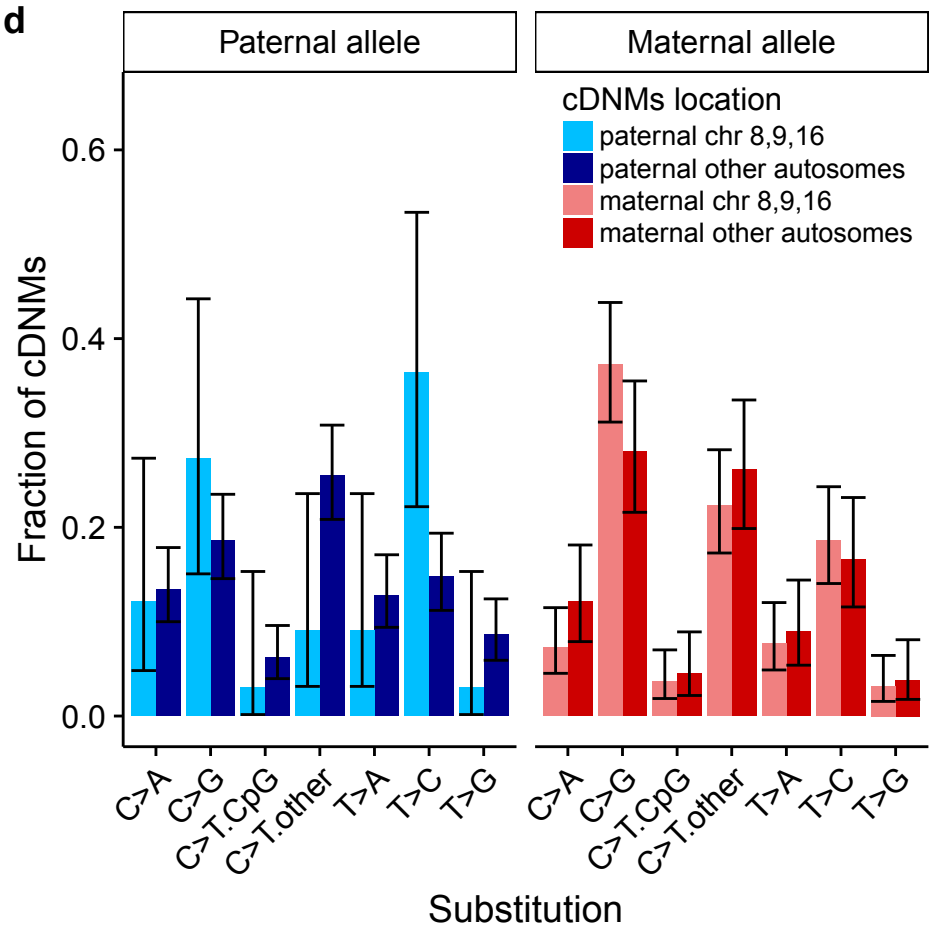
b


Chr 16
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
maternal cDNMs
paternal cDNMs
maternal ucDNMs
paternal ucDNMs
C>G enrichment

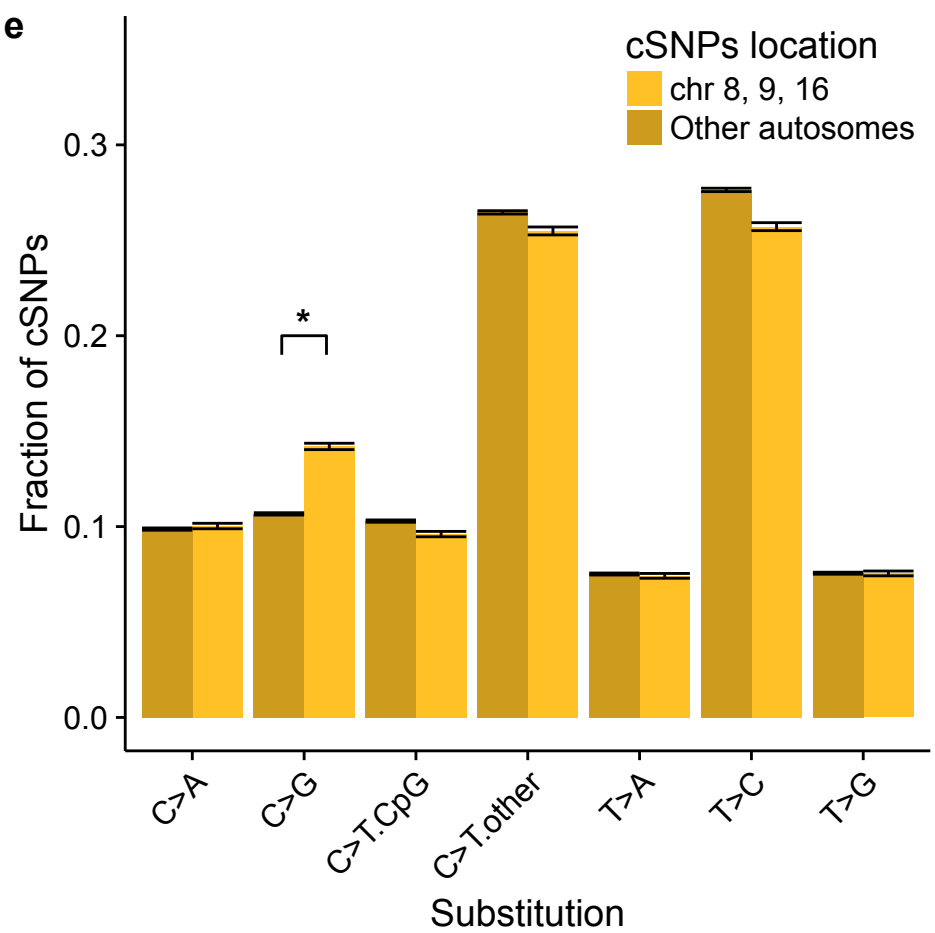
c


Unclustered DNM
Clustered DNM
Parent of origin
Father
Mother
*
Fraction of DNMs
0.0
0.1
0.2
0.3
0.4
C>A
C>G
C>T.CpG
C>T.other
T>A
T>C
T>G
Substitution

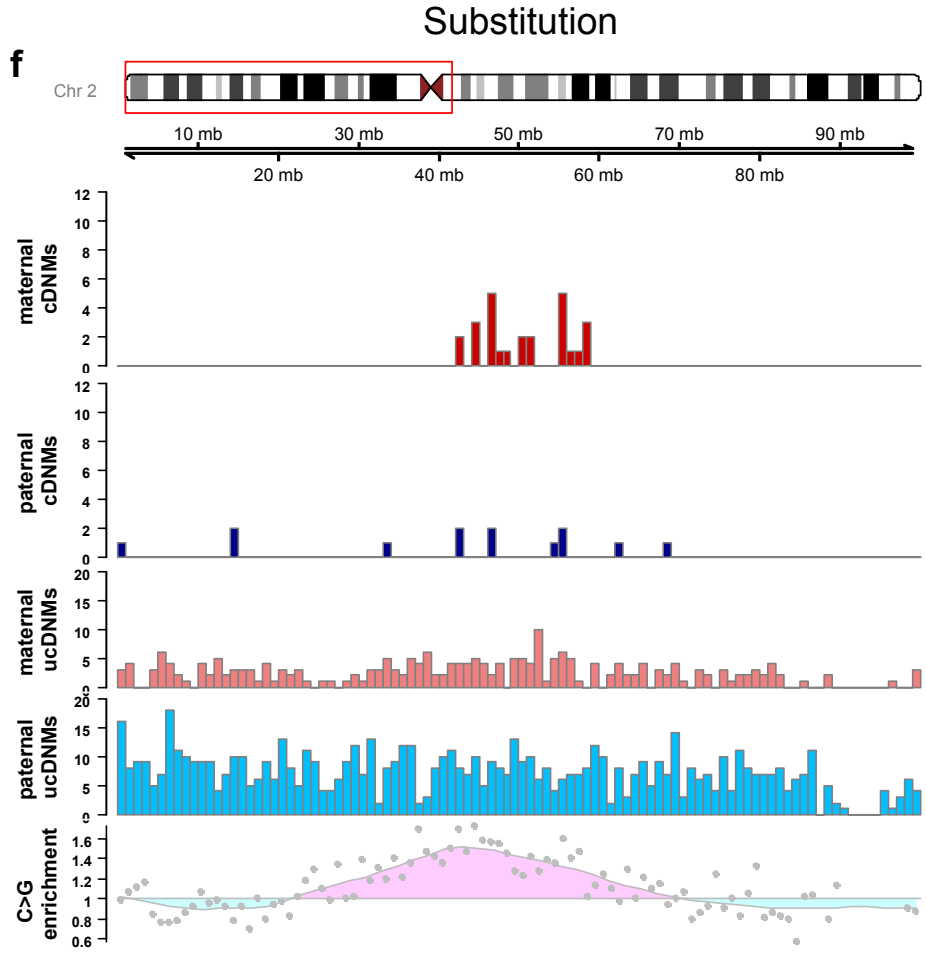
d

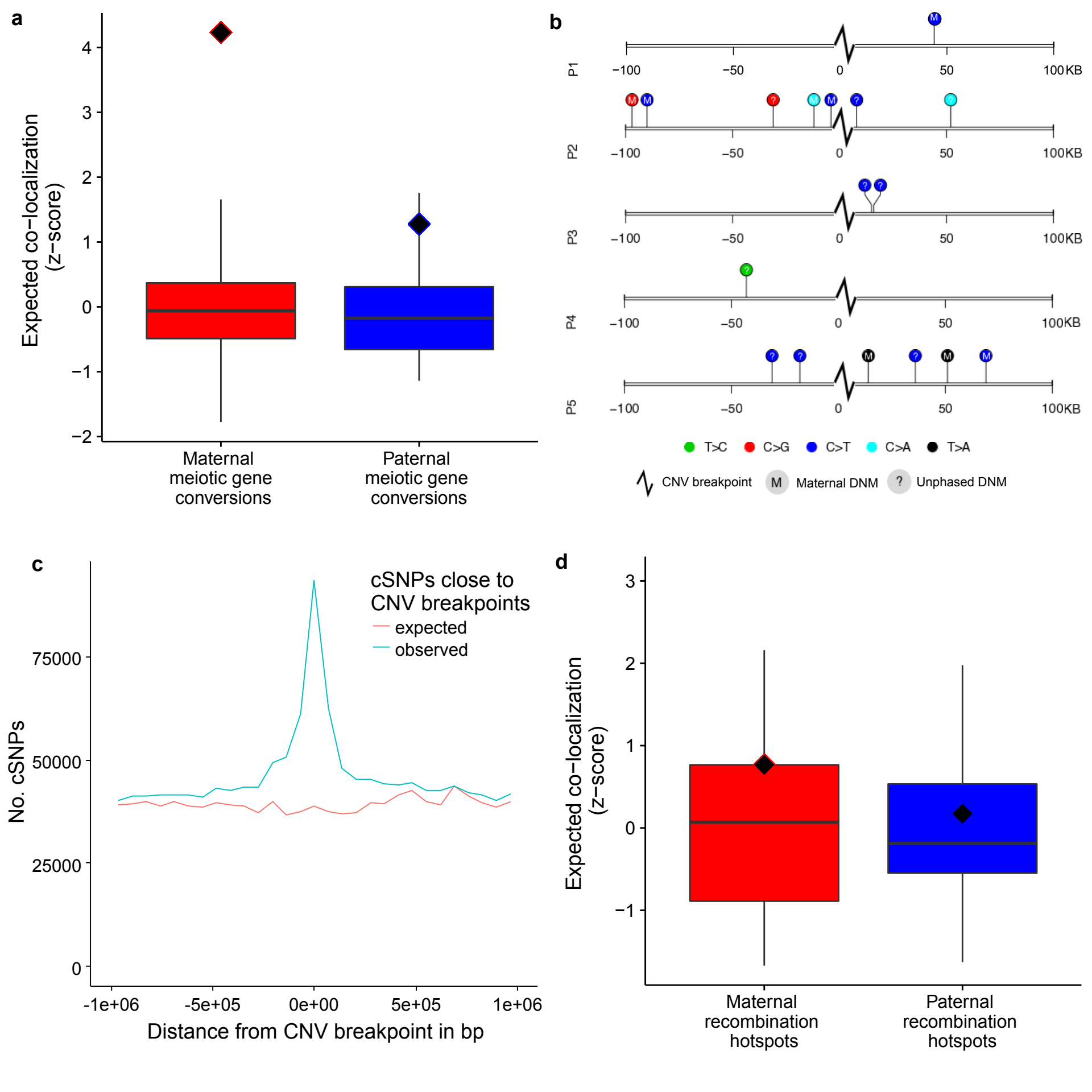

Paternal allele
Maternal allele
cDNMs location
paternal chr 8,9,16
paternal other autosomes
maternal chr 8,9,16
maternal other autosomes
Fraction of cDNMs
0.0
0.2
0.4
0.6
C>A
C>G
C>T.CpG
C>T.other
T>A
T>C
T>G
Substitution

e


cSNPs location
chr 8, 9, 16
Other autosomes
*
Fraction of cSNPs
0.0
0.1
0.2
0.3
C>A
C>G
C>T.CpG
C>T.other
T>A
T>C
T>G
Substitution

f


Chr 2
10 mb
20 mb
30 mb
40 mb
50 mb
60 mb
70 mb
80 mb
90 mb
maternal cDNMs
paternal cDNMs
maternal ucDNMs
paternal ucDNMs
C>G enrichment


a
Expected co−localization (z−score)
4
3
2
1
0
−1
−2
Maternal meiotic gene conversions
Paternal meiotic gene conversions
b
P1
P2
P3
P4
P5
−100
−50
0
50
100KB
T>C
C>G
C>T
C>A
T>A
CNV breakpoint
M
Maternal DNM
?
Unphased DNM
c
cSNPs close to CNV breakpoints
expected
observed
No. cSNPs
75000
50000
25000
0
-1e+06
-5e+05
0e+00
5e+05
1e+06
Distance from CNV breakpoint in bp
d
Expected co−localization (z−score)
3
2
1
0
−1
Maternal recombination hotspots
Paternal recombination hotspots