Association of a germline copy number polymorphism of APOBEC3A and APOBEC3B with burden of putative APOBEC-dependent mutations in breast cancer

Serena Nik-Zainal1,2, David C Wedge1, Ludmil B Alexandrov1, Mia Petljak1, Adam P Butler1, Niccolo Bolli1,3, Helen R Davies1, Stian Knappskog4,5, Sancha Martin1, Elli Papaemmanuil1, Manasa Ramakrishna1, Adam Shlien1,6, Ingrid Simonic7, Yali Xue1, Chris Tyler-Smith1, Peter J Campbell1,3 & Michael R Stratton1

The somatic mutations in a cancer genome are the aggregate outcome of one or more mutational processes operative through the lifetime of the individual with cancer1–3. Each mutational process leaves a characteristic mutational signature determined by the mechanisms of DNA damage and repair that constitute it. A role was recently proposed for the APOBEC family of cytidine deaminases in generating particular genome-wide mutational signatures1,4, and a signature of localized hypermutation called kataegis1,4. A germline copy number polymorphism involving APOBEC3A and APOBEC3B, which effectively deletes APOBEC3B5, has been associated with modestly increased risk of breast cancer6–8. Here we show that breast cancers in carriers of the deletion show more mutations of the putative APOBEC-dependent genome-wide signatures than cancers in non-carriers. The results suggest that the APOBEC3A-APOBEC3B germline deletion allele confers cancer susceptibility through increased activity of APOBEC-dependent mutational processes, although the mechanism by which this increase in activity occurs remains unknown.

In recent analyses of somatic mutational signatures in 21 primary human breast cancers that underwent whole-genome sequencing3, 2 signatures characterized by C>T and/or C>G mutations at TCX trinucleotides were identified (where the underlined base is the mutated base and X can be any base). These signatures were subsequently observed in several other cancer types (Fig. 1a and Online Methods) and are among the most common mutational signatures found in human cancer (Supplementary Fig. 1a,b)1,9. These signatures have been designated signatures 2 and 13 (according to the nomenclature of Alexandrov et al.9). Signature 2 is composed predominantly of C>T transitions with fewer C>G transversions in a TCX sequence context. In contrast, signature 13 is dominated by C>G transversions in a TCX context1,9. A subset of breast cancers and other cancer types have an extremely large number of mutations of these signatures, and we have called these cancers ‘hypermutators’ (refs. 1,9).

The features of the mutations associated with signatures 2 and 13 resemble those of mutations generated by the AID/APOBEC family of cytidine deaminases10,11. Members of this gene family have important physiological roles in antibody diversification (AIDCA) and the restriction of retroviruses and mobile retroelements (for example, APOBEC3A and APOBEC3G) (reviewed in refs. 11–13). However, it has been suggested that the DNA editing capabilities of these enzymes could also underlie undesirable mutagenesis leading to cancer4,11,14,15.

Indeed, in addition to AIDCA, a capacity for the editing of nuclear DNA has been demonstrated for APOBEC3A14,16,17, APOBEC1, APOBEC3C and APOBEC3G18.

A common germline copy number deletion polymorphism involving the APOBEC3 gene cluster on chromosome 22 (Fig. 1b) has been associated with elevated risk of breast cancer. A genome-wide association study (GWAS) of copy number in 16,000 cases of 8 common diseases highlighted this deletion polymorphism in association with breast cancer in a primary screen, although the association was not validated in replication6. Subsequently, a GWAS in the Chinese population demonstrated an association of the deletion polymorphism with breast cancer (odds ratio (OR) = 1.3 for one-copy deletion and 1.8 for two-copy deletion; \( P = 2.0 \times 10^{-24} \))7 that was replicated in a European population (OR = 1.2 for one-copy deletion and 2.3 for two-copy deletion; \( P_{\text{trend}} = 0.005 \))8. The deletion allele has a frequency of ~8% in European populations5,6, 37% in East Asians and 93% in Oceanians5. The ~29,500-bp genomic deletion has delimiting breakpoints in APOBEC3A and APOBEC3B (which are adjacent to each other and in the same orientation on chromosome 22) and results in a chimeric APOBEC3A-APOBEC3B gene. This hybrid gene is predicted to produce a transcript that is predominantly constituted of APOBEC3A sequence but replaces the APOBEC3A 3′ UTR with the APOBEC3 B 3′ UTR (Supplementary Note) and encodes a protein...
that has an identical amino acid sequence to that of APOBEC3A (Fig. 1b). Homozygous carriers of this deletion allele are predicted not to make any APOBEC3B protein. Given the association of the deletion allele with breast cancer, we explored the relationship between this allele (Table 1) and the presence of mutational signatures 2 and 13.

We aggregated a set of 923 breast cancers from multiple sequencing centers in which normal and neoplastic tissues had been sequenced for somatic mutations—123 by whole-genome sequencing and 800 by whole-exome sequencing (Supplementary Table 1a,b). Using next-generation sequence data, we identified 128 cases that were heterozygous and 14 that were homozygous for the APOBEC3A-APOBEC3B deletion allele (Online Methods, Supplementary Fig. 2a–c, Supplementary Table 2a–c and Supplementary Note). Applying the non-negative matrix factorization (NNMF) algorithm, employed to extract mutational signatures1,9,18, to the somatic mutation characteristics with mutations generated by members of the AID/APOBEC family of cytidine deaminases. Both signatures are characterized by C>T transitions and/or C>G transversions in a TCX sequence context. Signature 2 is dominated by C>T transitions. Signature 13 is dominated by C>G transversions. (b) APOBEC3A-APOBEC3B hybrid deletion allele. The APOBEC3A and APOBEC3B genes are located in tandem on chromosome 22. The polymorphism involves a deletion of the APOBEC3B coding sequence (highlighted), fusing the 3' UTR of APOBEC3B to the 3' UTR of APOBEC3A.

Figure 1 The APOBEC3A-APOBEC3B germline deletion polymorphism is associated with increased burden of presumptive APOBEC-related signatures. (a) Signatures 2 and 13 extracted by NNMF11 share sequence-specific mutation characteristics with mutations generated by members of the AID/APOBEC family of cytidine deaminases. Both signatures are characterized by C>T transitions and/or C>G transversions in a TCX sequence context. Signature 2 is dominated by C>T transitions. Signature 13 is dominated by C>G transversions. (b) APOBEC3A-APOBEC3B hybrid deletion allele. The APOBEC3A and APOBEC3B genes are located in tandem on chromosome 22. The polymorphism involves a deletion of the APOBEC3B coding sequence (highlighted), fusing the 3' UTR of APOBEC3B to the 3' UTR of APOBEC3A.

In the cohort of 923 breast cancers, the majority of cases had a mutation rate of signatures 2 and 13 mutations of less than 1 (Table 1, Supplementary Table 3b and the Supplementary Note for discussion of population stratification).

We then examined 1,769 cancers of 11 other cancer types in which signature 2 and 13 mutations have been found (Table 2 and Supplementary Fig. 3b–d)11. Of 40 individuals with acute lymphoblastic leukemia (ALL), 3 were hypermutators (Table 2); all 3 were carriers of the germline deletion allele, with 2 heterozygous and 1 homozygous for the allele ($P = 2.51 \times 10^{-5}$). Enrichment for hypermutators among cases who were heterozygous and homozygous for the deletion allele was also seen in bladder carcinoma, although this enrichment did not reach statistical significance ($P = 0.038$, Bonferroni-corrected $P = 0.452$; Table 2). Thus, the APOBEC3A-APOBEC3B deletion allele may be associated with the signature 2 and 13 mutational burden in cancers other than breast cancer.

In breast and other cancers, several non-carriers of the germline deletion allele had large numbers of signature 2 and 13 mutations (Table 2 and Supplementary Table 2b). Likewise, several carriers of the deletion allele did not have large numbers of signature 2 or 13 mutations in their cancers (Online Methods). It thus seems that the germline deletion allele is neither necessary nor sufficient to generate signature 2 and 13 mutations. This behavior is in keeping with that of a germline susceptibility allele that has a modest effect on a quantitative trait. Indeed, the marked variation in signature 2 and 13 mutation prevalence among different

### Table 1 Relationship between the germline APOBEC3A-APOBEC3B deletion allele and signatures 2 and 13 in breast cancers

| Deletion allele status | Hypermutators | Non-hypermutators | Total | Hypermutators/total cases |
|------------------------|---------------|-------------------|-------|---------------------------|
| Homozygous             | 4             | 10                | 14    | 0.286                     |
| Heterozygous           | 28            | 100               | 128   | 0.219                     |
| Non-carrier            | 74            | 707               | 781   | 0.095                     |

Cochran-Armitage test for trend $P = 6.251 \times 10^{-6}$, $\chi^2$ statistic = 20.4098

In the cohort of 923 breast cancers, the majority of cases had a mutation rate of signatures 2 and 13 of less than 1 mutation per megabase. A subset of cases had mutations comprising mostly (or, in some cases, entirely) signatures 2 and 13 with a very high mutation rate associated with these signatures (hypermutators). A higher proportion of cases were found to be carriers of at least one copy of the germline deletion allele among cases with hypermutator breast cancers. A test for trend demonstrates a correlation between the number of copies of the deletion allele in a breast cancer case and having a hypermutator breast cancer ($P = 6.251 \times 10^{-5}$).
cancers (Supplementary Fig. 3c,d) suggests that multiple factors are likely to influence the burden of signature 2 and 13 mutagenesis, such as inherited variation in the APOBEC genes, APOBEC gene expression, virus or transposon activity, and inflammation, and that these factors may vary in importance in different cancers.

If signatures 2 and 13 are due to APOBEC activity, they should bear the known characteristics of the mutations generated by these enzymes. The substitution classes in signatures 2 and 13 (C>T transitions and C>G transversions, respectively) coupled with the context of the mutations were responsible for the proposed role of the APOBEC enzyme family in generating these signatures. However, APOBEC-induced mutations exhibit other distinctive characteristics, including preferential cytosine-to-uracil deamination on stretches of single-stranded DNA. Consequently, adjacent APOBEC-induced mutations often arise on the same parental allele (in cis with each other) and are on the same DNA strand (where successive mutations may be C>T... C>G... C>T or G>A... G>C... G>A but not C>T... G>A... C>T), a pattern referred to as strand-coordinated mutagenesis (Online Methods and Supplementary Note).

To investigate the presence of strand-coordinated mutagenesis in signatures 2 and 13 (Supplementary Fig. 4a–c), we examined the frequencies of 2 successive mutations arising on the same strand and on different strands in the 123 breast cancers that had undergone whole-genome sequencing. Several cancers had more strand-coordinated pairs of mutations than expected by chance (corrected for the mutation spectrum and mutation burden; Fig. 2a), and this level of strand-coordinated mutagenesis was directly correlated with the proportion of signature 2 and 13 mutations in these cancers (r = 0.74; P = 1.1 × 10−21; Fig. 2b). Furthermore, examination of next-generation sequencing reads in these cancers showed that strand-coordinated mutations usually occurred in cis (P < 0.0001; Supplementary Fig. 4d, Supplementary Table 4 and Supplementary Note), confirming that these mutations are linked to each other on the same parental haplotype. Taken together, these findings are compatible with the model in which signature 2 and 13 mutations often arise on stretches of single-stranded DNA, similar to the mutations induced by APOBEC enzymes.

Table 2 Relationship between the number of copies of the deletion allele and the burden of signature 2 and 13 mutations

| Cancer type | Hom | Het | Non | Total | Hom | Het | Non | Total | Cancer type total | Test for trend | Bonferroni correction |
|-------------|-----|-----|-----|-------|-----|-----|-----|-------|-------------------|---------------|----------------------|
| ALL         | 1   | 3   | 2   | 6     | 0   | 5   | 32  | 37    | 40                | 17.756        | 2.51 × 10⁻⁵         |
| BLCA        | 0   | 3   | 6   | 9     | 0   | 13  | 114 | 127   | 136               | 4.3192        | 0.03769             |
| BRCA        | 4   | 28  | 74  | 106   | 10  | 100 | 707 | 817   | 923               | 20.4098       | 6.25 × 10⁻⁶          |
| CESC        | 0   | 1   | 3   | 4     | 0   | 5   | 29  | 34    | 38                | 0.2852        | 0.5933              |
| HNSC        | 0   | 3   | 33  | 36    | 0   | 33  | 229 | 262   | 298               | 0.5413        | 0.4619              |
| KIRP        | 0   | 0   | 5   | 5     | 0   | 14  | 81  | 95    | 100               | 0.8568        | 0.3546              |
| LUAD        | 0   | 3   | 22  | 25    | 0   | 18  | 260 | 278   | 303               | 1.0856        | 0.2975              |
| LUSC        | 0   | 3   | 12  | 15    | 0   | 17  | 133 | 150   | 165               | 0.9616        | 0.3268              |
| MM          | 0   | 0   | 2   | 2     | 0   | 10  | 51  | 63    | 65                | 0.4152        | 0.5193              |
| STAD        | 0   | 1   | 10  | 11    | 0   | 18  | 104 | 122   | 133               | 0.2643        | 0.6072              |
| THCA        | 0   | 1   | 19  | 20    | 0   | 44  | 220 | 264   | 284               | 1.8977        | 0.1683              |
| UCEC        | 0   | 0   | 12  | 12    | 4   | 35  | 183 | 222   | 234               | 2.319         | 0.1278              |
| Total       | 5   | 45  | 198 | 248   | 16  | 312 | 2143| 2471  | 2719              | 10.9215       | 9.51 × 10⁻⁴          |

A trend was seen for ALL but not for other cancers (test for trend). Hom, homozygous; Het, heterozygous; Non, non-carrier; ALL, acute lymphoblastic leukemia; BLCA, bladder cancer; BRCA, breast cancer; CESC, cervical cancer; HNSC, head and neck cancer; KIRP, kidney papillary cancer; LUAD, lung adenocarcinoma; LUSC, lung squamous cancer; MM, multiple myeloma; STAD, stomach adenocarcinoma; THCA, thyroid cancer; UCEC, uterine cancer.
Figure 2 Additional features of signatures 2 and 13 that are similar to the mutagenic patterns of APOBEC enzymes. (a) Several cancers showed an excess of mutations arising on the same strand or strand-coordinated mutagenesis. Because of space limitations, only a subset of cancers is depicted here. OR values of the observed strand-coordinated mutations over those expected by chance are presented (red boxes) with 95% CIs (gray lines). Each OR value was calculated from the observed number of same-strand/different-strand mutations divided by the number of expected same-strand/different-strand mutations (where expected numbers were corrected for the overall mutation rate of the cancer and the mutation spectrum). A higher OR indicates more same-strand mutations than expected. Cases highlighted in red have hypermutator breast cancer. (b) A direct correlation is seen between the OR of strand-coordinated mutagenesis and the fractional burden of signature 2 and 13 mutations. Cases who are homozygous or heterozygous for the deletion allele are highlighted to show the enrichment of deletion carriers among breast cancers with a high burden of signature 2 and 13 mutations.

in generating these mutations. Furthermore, it provides a plausible biological mechanism by which this breast cancer predisposition allele could confer its effect. The geographic variation in the population frequency of the APOBEC3A-APOBEC3B germline deletion allele suggests that there may be selection in favor of it (Online Methods). As some APOBEC enzymes are involved in innate immunity to infection, it may be that protection against infection is conferred by the deletion allele. This protection may be counterbalanced, to some extent, by predisposition to cancer. If true, this would be remarkable, as both effects would be mediated by the same underlying mechanism—the double-edged sword of the mutagenic activity of the APOBEC proteins.

METHODS

Methods and any associated references are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank M. Hurles and C. Anderson of the Wellcome Trust Sanger Institute for their input. We would like to thank the Wellcome Trust for support (grant 098051). S.N.-Z. is a Wellcome-Beit Prize Fellow and is supported through a Wellcome Trust Intermediate Fellowship (grant WT100183MA). P.I.C. is personally funded through a Wellcome Trust Senior Clinical Research Fellowship (grant WT088340MA). N.B. is a European Hematology Association (EHA) fellow and is supported by a starter grant from the Academy of Medical Sciences. The H.L. Holmes Award from National Research Council Canada and an EMBO (European Molecular Biology Organization) Fellowship support A.S. We would also like to acknowledge funding from Breakthrough Breast Cancer Research (ICGC 08/09) and the BASIS project, funded by the European Community’s Seventh Framework Programme (FP7/2010-2014) under grant agreement 242006. This study was performed within the Research Ethics Approval of 09/h0306/36.

AUTHOR CONTRIBUTIONS

S.N.-Z. and M.R.S. conceived the experiments and wrote the manuscript. S.N.-Z., D.C.W., L.B.A. and P.I.C. carried out analyses and/or statistics with assistance from M.P., A.P.B., N.B., A.S., H.R.D., M.R., E.P., S.K. and I.S. S.M. governed administrative aspects. Y.X. and C.T.-S. advised and performed analysis on selection.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. Cell 149, 979–993 (2012).
2. Nik-Zainal, S. et al. The life history of 21 breast cancers. Cell 149, 994–1007 (2012).
3. Stratton, M.R., Campbell, P.J. & Futreal, P.A. The cancer genome. Nature 458, 719–724 (2009).
4. Nowarski, R. & Kotler, M. APOBEC3 cytidine deaminases in double-strand DNA break repair and cancer promotion. Cancer Res. 73, 3494–3498 (2013).
5. Kidd, J.M., Newman, T.L., Tuzun, E., Kaul, R. & Eichler, E.E. Population stratification of a common APOBEC gene deletion polymorphism. PLoS Genet. 3, e63 (2007).
6. Craddock, N. et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 464, 713–720 (2010).
7. Long, J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J. Natl. Cancer Inst. 105, 573–579 (2013).
8. Xuan, D. et al. APOBEC3 deletion polymorphism is associated with breast cancer risk among women of European ancestry. Carcinogenesis 34, 2240–2243 (2013).
9. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. Nature 500, 415–421 (2013).
10. Harris, R.S., Petersen-Mahrt, S.K. & Neuberger, M.S. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. Mol. Cell 10, 1247–1253 (2002).
11. Conticello, S.G. The AID/APOBEC family of nucleic acid mutators. Genome Biol. 9, 229 (2008).
12. Longerich, S., Basu, U., Alt, F. & Storb, U. AID in somatic hypermutation and class switch recombination. Curr. Opin. Immunol. 18, 164–174 (2006).
13. Koito, A. & Ikeda, T. Intrinsic restriction activity by AID/APOBEC family of enzymes against the mobility of retroelements. Mob. Genet. Elements 1, 197–202 (2011).
14. Suspendé, R. et al. Somatic hypermutation of human mitochondrial and nuclear DNA by APOBEC3 cytidine deaminases, a pathway for DNA catabolism. Proc. Natl. Acad. Sci. USA 108, 4858–4863 (2011).
15. Petit, V., Vartanian, J.P. & Wain-Hobson, S. Powerful mutators lurking in the genome. Phil. Trans. R. Soc. Lond. B 364, 705–715 (2009).
16. Landry, S., Narvaiza, I., Linfesty, D.C. & Weitzman, M.D. APOBEC3A can activate the DNA damage response and cause cell-cycle arrest. EMBO Rep. 12, 444–450 (2011).
17. Suspendé, R., Aynaud, M.M., Vartanian, J.P. & Wain-Hobson, S. Efficient deamination of 5-methylcytidine and 5-substituted cytidine residues in DNA by human APOBEC3A cytidine deaminase. PLoS ONE 8, e63461 (2013).
18. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J. & Stratton, M.R. Deciphering signatures of mutational processes operative in human cancer. Cell Reports 3, 246–259 (2013).
19. Hodge, V.J. & Austin, J. A survey of outlier detection methodologies. Artif. Intell. Rev. 22, 85–126 (2004).
20. Nowarski, R. et al. APOBEC3G enhances lymphoma cell radioresistance by promoting cytidine deaminase–dependent DNA repair. Blood 120, 366–375 (2012).
21. Holtz, C.M., Sadler, H.A. & Mansky, L.M. APOBEC3G cytosine deamination hotspots are defined by both sequence context and single-stranded DNA secondary structure. Nucleic Acids Res. 41, 6139–6148 (2013).
22. Byeon, I.J. et al. NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity. Nat. Commun. 4, 1890 (2013).
23. Petit, V. et al. Murine APOBEC1 is a powerful mutator of retroviral and cellular RNA in vitro and in vivo. J. Mol. Biol. 385, 65–78 (2009).
24. Beale, R.C. et al. Comparison of the differential context-dependence of DNA deamination by APOBEC enzymes: correlation with mutation spectra in vivo. J. Mol. Biol. 337, 585–596 (2004).
25. Taylor, B.J. et al. DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. Elife 2, e00534 (2013).
26. Shinohara, M. et al. APOBEC3B can impair genomic stability by inducing base substitutions in genomic DNA in human cells. Sci. Rep. 2, 806 (2012).
27. Burns, M.B. et al. APOBEC3B is an enzymatic source of mutation in breast cancer. Nature 494, 366–370 (2013).
28. Burns, M.B., Temiz, N.A. & Harris, R.S. Evidence for APOBEC3B mutagenesis in multiple human cancers. Nat. Genet. 45, 977–983 (2013).
29. Roberts, S.A. et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat. Genet. 45, 970–976 (2013).
30. Rousseeuw, P.J., Ruts, I. & Tukey, J.W. The bagplot: a bivariate boxplot. Am. Stat. 53, 382–387 (1999).
ONLINE METHODS

Background information. Cancer samples for next-generation sequencing (whole genome and whole exome) were previously sequenced by members of the International Cancer Genome Consortium (ICGC), The Cancer Genome Atlas (TCGA) and other centers9,18. High-confidence somatic substitutions were obtained from these consortia or other peer-reviewed publications not related to these consortia, filtered further for potential false positive calls using dbSNP, the 1000 Genomes Project database, the National Heart, Lung, and Blood Institute (NHLBI) Grand Opportunity (GO) Exome Sequencing Project and the 69 Complete Genomics panel as well as a bespoke panel of BAM files of unmatched normal tissues containing more than 120 normal genomes and 500 exomes8. These data were then parsed through an algorithm previously developed to extract mutational signatures in human cancers19 called NNMF9,18.

Six main substitution classes (C>A:G>T, C>G:G>C, C>T:A>G, T>A:T>G, T>C:A>G and T>G:A>C) were subdivided according to 5′ and 3′ flanking sequence context. As there were 6 classes of base substitution and 16 possible sequence contexts for each mutated base (A, C, G or T at the 5′ base and A, C, G or T at the 3′ base), there were 96 possible mutated trinucleotides for each cancer. Herein, the convention for describing a mutated trinucleotide was XCX, where X can be any base and the mutated base is underlined1,2,29,18.

A total of 7,042 samples were analyzed from 30 types of cancer, and 21 distinct mutational signatures were extracted. The most common signatures were signatures 1A and 1B, both characterized by C>T mutations in an XCG trinucleotide context (Supplementary Fig. 1a, and signatures 2 and 13, characterized by dominant C>T transitions in a TCX sequence context in signature 2 and C>G transversions in a TCX sequence context in signature 13 (Supplementary Fig. 1a). Signatures 1A and 1B mutations are likely to be caused by deamination at methylated CpGs, whereas signature 2 and 13 mutations are thought to be due to the activity of the APOBEC family of cytidine deaminases. Therefore, for the purpose of this analysis, signatures 2 and 13 were considered together. NNMF is able to estimate the number of cytidine deaminases. Therefore, for the purpose of this analysis, signatures 2 and 13 were considered together. NNMF is able to estimate the number of mutational mechanisms involved in producing the observed mutational spectrum for each cancer type.

In this analysis, a total of 2,719 samples were previously characterized by NNMF and also had BAM files available for inspection (Supplementary Table 1a,b). BAM files were downloaded from CGHub between 9 May 2013 and 26 June 2013. For ease in tracking samples throughout this analysis, we have kept the naming convention attached to the cancer sample for tables and figures, even if the germline deletion polymorphism was sourced from a matched normal sample, as the signatures of somatic mutagenesis would have been identified in the tumor samples in the first place. This notation was also kept for the purpose of continuity between publications. For samples originating from the Sanger Institute, PDXXXX denotes a specific individual, with an ‘a’, ‘c’ or ‘d’ suffix denoting tumor samples and a ‘b’ suffix indicating the matched normal sample.

Detection of the germline APOBEC3A-APOBEC3B deletion polymorphism. To detect the deletion polymorphism from next-generation sequencing data, multiple loci within and flanking the coordinates of the deletion were sampled (Supplementary Fig. 2a and Supplementary Table 2a) from BAM files (the overall workflow and directions for processing data are provided in Supplementary Fig. 2b). Raw short read data had been aligned back to the reference genome (NCBI Build 37), with duplicates and unmapped reads removed. Externally sourced BAM files were obtained from the TCGA data hub.

BAM files for matched normal samples were sought for calling this deletion allele. However, a BAM file for a tumor was used if one for a normal sample was not available (Supplementary Table 2b). Samples that did not have BAM files available for examination were excluded from analysis. In total, detection of the APOBEC3A-APOBEC3B polymorphism was sourced from 561 tumors (99 BLCA, 117 BRCA, 1 CESC, 19 HNSC, 2 KIRP, 303 LUAD, 12 STAD, 2 THCA and 6 UCEC), and 2,158 normal samples.

For samples with whole-genome data, the expected sequencing depth in the absence of the deletion polymorphism—that is, wild-type copy number of 2 or 13—was calculated as the average sequencing depth of the 60 loci in the flanking regions. The expected sequencing depth in the presence of a heterozygous deletion allele, d1, was given by d1 = d2/2, and the expected depth in the presence of a homozygous deletion allele, d0, was set to the expected number of misreads, estimated as d2/20. A maximum-likelihood test was performed to identify the most likely copy number from the set11, with corresponding expected sequencing depths represented as Poisson distributions (Pois(d0), Pois(d1) and Pois(d2)), given the observed sequencing depths within the region of the deletion polymorphism (Supplementary Data Set 1).

For exome-sequenced samples, an expectation maximization algorithm was used, with the copy number of each sample and the ratio of the sequencing depths within and outside the deletion polymorphism region used as latent variables. The copy number of each sample was initialized as 2, and the depth ratio of loci within the deletion polymorphism region relative to loci outside, r, was modeled non-parametrically by bootstrap resampling (n = 1,000). For samples with copy number of 2, 1 and 0, respectively, the expected sequencing depths within the deletion polymorphism region were given by

\[d_{\text{inside},0} = r d_{\text{outside}}/20\]

At each maximization step, the copy number of each sample was assigned as the one that had the distribution showing the most overlap with dinside for that sample, after bootstrap resampling. At each expectation step, r was recalculated using bootstrap resampling of loci within just those samples classified as having copy number = 2 in the previous maximization step. The expectation maximization algorithm was continued until no samples were reclassified from one iteration to the next or for a maximum of 100 iterations (Supplementary Fig. 2b and Supplementary Data Set 2).

Results for the calling of polymorphism status in all samples are provided in Supplementary Table 2b. The reproducibility of the calling method was evaluated by examining the concordance between calling on the tumor and normal BAM files from the same individual (Supplementary Fig. 2c, Supplementary Table 2c and Supplementary Note) as well as the concordance between genome- and exome-sequenced samples from the same individual.

Relationship between the APOBEC3A-APOBEC3B germline deletion allele and somatic mutational signatures in cancer. The data set comprised genome-sequenced (123) as well as exome-sequenced (800) cancers. To perform the analyses, the rate of mutation was calculated for each cancer (rate of signature 2 and 13 mutations per megabase of DNA), correcting for whether the samples had been sequenced over the genome or exome. Because the rates of signature 2 and 13 mutations were not normally distributed (Supplementary Fig. 3a–d), a one-sided Wilcoxon rank-sum test was performed to determine whether carrying one copy of the deletion allele had an overall effect on the mutation rate of the signatures.

We sought to include more cancer samples to increase the power of the analyses. There were no further available breast cancer samples with BAM files ready for download; hence, we sought inclusion of other cancer types that had previously been analyzed9,18. However, it was noted that distributions of the rates of signature 2 and 13 mutations varied considerably among cancer types (Supplementary Fig. 3b–d), and clear outliers were present in all the cancer types, skewing the distributions of mutation rate (Supplementary Fig. 3a,d,e).

Some cancers were observed to have a strikingly high proportion of total mutations associated with signatures 2 and 13 and/or to have higher rates of mutagenesis associated with these signatures (Supplementary Figs. 1c-o and 5c). Using the rate of signature 2 and 13 mutagenesis, outliers were identified as cases with cancers that had a mutation rate exceeding 1.5 times the length of the interquartile range from the 75th percentile for each type of cancer19. We refer to these outliers as hypermutators, although we do not suggest that there is an ongoing biological process attached to this name. Given the considerable variation in the mutation rates for different cancer tissue types (Supplementary Fig. 1a,b), each cancer type was analyzed separately. A summary of the hypermutators versus non-hypermutators is provided in Supplementary Table 3a.
Strand-coordinated mutagenesis. In theory, neighboring mutations could arise on either of two strands of a double helix (Supplementary Fig. 4a), particularly if they had arisen as independent events during different cycles of cell division. If more mutations are observed to occur on the same strand than expected by chance (Supplementary Fig. 4b), this would imply one of two scenarios: either the neighboring mutations arose over different rounds of cell division with preferential targeting of one strand over another or they arose during a single round of cell division and potentially occurred in the same instance.

We sought to formally document that neighboring mutations are occurring on the same strand more often than expected by chance in strand-coordinated mutagenesis. To demonstrate genome-wide strand coordination, analysis was carried out on all whole-genome sequence data for which BAM files were available (Supplementary Table 4). Given a set of mutations, each occurring at a base of type A, C, G or T on the plus strand, we identified all pairs of mutations and classified them as ‘same’ if both mutations were of the same originating base and ‘diff’ if not (first and second mutations of each pair, respectively: A>X and A>G; G>C and G>G; C>T and C>C; and T>T and T>X, with no previous selection for mutations in a TC context). The distance between successive pairs or intermutation distance was also calculated (Supplementary Fig. 4c).

The proportion of the pairs of mutations that were expected to occur on the same base assuming randomly ordered mutations was given by $p_X^2 + p_C^2 + p_G^2 + p_T^2$, where $p_X$ is the fraction of mutations that occurred at nucleotide $X$. To depict the deviation of the observed pairs of mutations found on the same strand from that of the expected pairs of mutations on the same strand, a standard Forest plot was constructed (data for expected and observed same-strand mutations for all 124 samples are provided in Supplementary Table 4, columns b–e).

Because same-strand mutations were ascertained in an unbiased way from any mutation type (not restricted to just cytosine mutations at TC sites), to determine whether strand-coordinated mutations were a particular feature of signatures 2 and 13, we sought a relationship between the degree of strand coordination, given by the OR of strand coordination, and the fractional burden of signature 2 and 13 mutations in each cancer (Supplementary Fig. 4d).

We sought additional characteristics of the mutations in the breast cancers that underwent whole-genome sequencing that supported the suggestion that mutations associated with signatures 2 and 13 had arisen owing to the activity of the APOBEC family of enzymes (Supplementary Fig. 4d, Supplementary Table 4 and Supplementary Note).

Relationship between expression of APOBEC family members and the rates of mutation for signatures 2 and 13. RNA sequencing–derived expression data were obtained from the UCSC Genome Browser for relevant samples. In total, there were 1,691 cases for which comparable data were obtainable. Expression levels for each APOBEC family member were standardized relative to the levels of TBP (encoding TATA-binding protein), and the relationship between APOBEC3B expression levels and germline deletion allele status in these cancers was evaluated (Supplementary Fig. 5a, b and Supplementary Table 5).

Selection for the APOBEC3A-APOBEC3B deletion. The germline APOBEC3A-APOBEC3B deletion polymorphism highlighted in this analysis was reported to display a strikingly differentiated worldwide distribution of allele frequencies. The $F_{ST}$ value (measuring population differentiation) was reexamined using reported deletion allele frequencies and additional SNP genotype data from the Human Genome Diversity Cell Line Panel (CEPH-HGDP) panel published after the CNP study. This value depends on the way populations are grouped and needs to be compared with the values for other variants of similar frequency to measure how unusual it is. We used 2 published grouping schemes, involving division into 5 continental geographic/genetic groups (i) sub-Saharan Africans, (ii) Middle Easterners plus Europeans, (iii) East Asians, (iv) Native Americans and (v) Oceanians) or into 32 population groups and matched SNP frequencies measured as minor allele frequency to ±0.1%. For the five continental groups, $F_{ST}$ was 0.330 (97.4 percent compared with 2,716 frequency-matched SNPs), and for the 32 population groups it was 0.285 (96.6 percentile compared to 2,059 frequency-matched SNPs). The level of population differentiation was thus higher than expected by chance, which can result from positive selection, and we therefore examined other statistics sensitive to positive selection. Cross-population extended haplotype homozygosity (XP-EHH) and integrated haplotype score (iHS) values were obtained from the HGDP Selection Browser. These haplotype-based tests for positive selection use information for sequences 500 kb upstream and downstream of the deletion, and the two flanking regions can thus be examined separately. Neither flanking region showed a significantly high XP-EHH (>2.5) or iHS (|iHS| > 2.0) value in any continental group or individual population. Finally, we looked at allele frequency spectrum-based tests (Tajima’s $D$, Fay and Wu’s $H$ and the composite likelihood ratio test from Nielsen et al.) using 1000 Genomes Project Phase 1 resequenced data in the East Asian populations (CHB, CHS and JPT) (1000G Phase 1) in the regions surrounding the deletion, as described (1000G Pilot). There was no evidence for positive selection in these populations, although, in this case, the power of the tests was limited because the frequency of the deletion in these populations was not high enough.

Overall, this locus showed unusually high differentiation among continents and populations. However, there remained a lack of other evidence for positive selection, and we therefore cannot convincingly conclude that this deletion has been positively selected in human populations.

31. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature 467, 1061–1073 (2010).
32. Li, J.Z. et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 319, 1100–1104 (2008).
33. Perry, G.H. et al. Evolutionary genetics of the human Rh blood group system. Hum. Genet. 131, 1205–1216 (2012).
34. Xue, Y. et al. Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183, 1065–1077 (2009).
35. Pickrell, J.K. et al. Signals of recent positive selection in a worldwide sample of human populations. Genome Res. 19, 826–837 (2009).
36. Nielsen, R. et al. Genome scans for selective sweeps using SNP data. Genome Res. 15, 1566–1575 (2005).
37. 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).