Processed pseudogenes acquired somatically during cancer development

Susanna L. Cooke¹, Adam Shlien¹, John Marshall¹, Christodoulos P. Pipinikas², Inigo Martincorena¹, Jose M.C. Tubio¹, Yilong Li¹, Andrew Menzies¹, Laura Mudie¹, Manasa Ramakrishna¹, Lucy Yates¹, Helen Davies¹, Niccolo Bolli¹, Graham R. Bignell¹, Patrick S. Tarpey¹, Sam Behjati¹, Serena Nik-Zaina², Elli Papaemmanuil¹, Vitor H. Teixeira², Keiran Raine¹, Sarah O’Meara¹, Maryam S. Dodoran¹, Jon W. Teague¹, Adam P. Butler¹, Christine Iacobuzio-Donahue⁴, Thomas Santarius⁵, Richard G. Grundy⁶, David Malkin⁷, Mel Greaves⁸, Nikhil Munshi⁹, Adrienne M. Flanagan²,¹⁰, David BOWTELL¹¹, ICGC Breast Cancer Group¹,¹⁷, Sancha Martin¹, Denis Larsimont¹², Jorge S. Reis-Filho¹³, Alex Boussioutas¹¹,¹⁴, Jack A. Taylor¹⁵, D. Neil Hayes¹⁶, Sam M. Janes², P. Andrew Futreal¹, Michael R. Stratton¹, Ulan McDermott¹,¹⁵, Peter J. Campbell¹,³,⁵

Cancer evolves by mutation, with somatic reactivation of retrotransposons being one such mutational process. Germline retrotransposition can cause processed pseudogenes, but whether this occurs somatically has not been evaluated. Here we screen sequencing data from 660 cancer samples for somatically acquired pseudogenes. We find 42 events in 17 samples, especially non-small cell lung cancer (5/27) and colorectal cancer (2/11). Genomic features mirror those of germline LINE element retrotranspositions, with frequent target-site duplications (67%), consensus TTTTAA sites at insertion points, inverted rearrangements (21%), 5′ truncation (74%) and polyA tails (88%). Transcriptional consequences include expression of pseudogenes from UTRs or introns of target genes. In addition, a somatic pseudogene that integrated into the promoter and first exon of the tumour suppressor gene, MGA, abrogated expression from that allele. Thus, formation of processed pseudogenes represents a new class of mutation occurring during cancer development, with potentially diverse functional consequences depending on genomic context.

DOI: 10.1038/ncomms4644

¹Cancer Genome Project, Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK. ²Lungs for Living Research Centre, Rayne Institute, University College London, London WC1E 6JF, UK. ³University of Cambridge, Cambridge CB2 0XY, UK. ⁴Departments of Pathology and Oncology, Johns Hopkins Medical Institutions, Baltimore, Maryland 21205, USA. ⁵Addenbrooke’s NHS Foundation Trust, Cambridge CB2 0QQ, UK. ⁶Children’s Brain Tumour Research Centre, University of Nottingham, Nottingham NG7 2UH, UK. ⁷Department of Pediatrics, Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada M5G 1X8. ⁸Institute for Cancer Research, Sutton, London SM2 5NG, UK. ⁹Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. ¹⁰Royal National Orthopaedic Hospital, Middlesex HA7 4LP, UK. ¹¹Peter MacCallum Cancer Centre, Melbourne, Victoria 3002, Australia. ¹²Department of Pathology, Jules Bordet Institute, 1000 Brussels, Belgium. ¹³Department of Pathology, Memorial-Sloan-Kettering Cancer Center, New York, New York 10065, USA. ¹⁴Department of Gastroenterology, Royal Melbourne Hospital, University of Melbourne, Parkville, Victoria 3050, Australia. ¹⁵National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina 27713, USA. ¹⁶UNC Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA. Correspondence and requests for materials should be addressed to P.J.C. (email: pc8@sanger.ac.uk).

¹⁷A full list of Participants in ICGC Breast Cancer Group who supplied or verified samples for this project and their affiliations appears at the end of the paper.
Mutation underpins both the evolution of species and the development of cancer. In the germline, repetitive DNA, such as long interspersed elements (LINE) and Alu repeats, form a sizable proportion of the human genome. LINE elements continue to propagate in the genome through their retrotransposition machinery. Essentially, a functional LINE element, when transcribed and translated by the host machinery, encodes two proteins that co-ordinate reverse transcription of the mRNA template and its integration back into the genome at a distant site from the original element. Insertion of transposable elements has considerably reshaped the human genome over evolutionary time. A role for retrotransposition as a mutational force in somatic cells has been increasingly recognized in the last few years, documented to occur during both normal neurogenesis and cancer development.

Germline pseudogenes, representing cDNA copies of mRNA transcripts, are a by-product of LINE-mediated retrotransposition. This copying of DNA sequence through an mRNA intermediate creates several distinctive genomic features of pseudogenes, including the presence of polyA tails, absence of intronic sequence and target-site duplications. In the germline, processed pseudogenes influence evolution through gene duplication, novel exons, gene fusions, sequestration of miRNAs and antisense transcript production.

Formation of processed pseudogenes has not been systematically studied in cancer. From a screen of 660 cancer samples, we find that somatically acquired pseudogenes are present in 2.6% of cancers, especially lung and colorectal cancers. We find a diverse array of transcriptional consequences, including expression from UTR and intronic insertions, as well as inactivation of transcription from exonic insertions.

Results

We developed bioinformatic methods to detect somatically acquired processed pseudogenes in massively parallel sequencing data from both targeted exome and genome-wide studies in cancer. Paired-end sequencing reads were aligned to the genome and transcriptome with a view to identifying reads that split exactly across canonical splice sites, were mapped to exons separated by more than the library insert size, or mapped between a pseudogene and its insertion site (Supplementary Fig. 1). To define a putative pseudogene, we required that at least three exons from a single gene were represented in the tumour DNA, with at least two canonical splice junctions directly observed from either split reads or confirmatory capillary sequencing. To establish that pseudogenes were not germline, we analysed sequencing data from the matched normal DNA for the given patient and screened for identical events in other patients. We performed PCR on tumour and matched normal DNA for all predicted pseudogenes with mapped insertion sites, and excluded variants with a positive germline PCR band from this analysis (Supplementary Fig. 2).

We identified 42 somatically acquired pseudogenes in 14 out of 629 primary cases and 3 out of 31 cell lines sequenced (Table 1, Supplementary Data 1 and Supplementary Table 1). As a typical example, in a lung cancer we identified an insertion of all five exons of the gene FOPNL, including a portion of the 5' UTR, the full coding sequence and the full 3' UTR, into the eleventh intron of SND1 in the opposite orientation (Fig. 1a). All four canonical exon–exon junctions of FOPNL were crossed by sequencing reads in the tumour DNA, and a polyA tail of at least 50 bp was present at the 3' end of the pseudogene. No such evidence was seen in the matched germline DNA from this patient or any other, and PCR confirmed the 5' and 3' insertion points as somatic. The two insertion points were mapped to base-pair resolution and revealed a target-site duplication of 10 bp that included the motif TTTT at either end of the pseudogene.

We observed variations on this theme (Figs 1b and 2a, Supplementary Fig. 3). PolyA tails were usually, but not universally, present, seen in 88% (21/24) pseudogenes with insertion sites fully mapped. Target-site duplications were frequent (67% with insertion sites fully mapped; 16/24), although in five cases target-site deletions occurred. The majority (74%; 31/42) of somatic pseudogenes did not contain the full coding sequence of the gene, as the 5' end was frequently truncated during reverse transcription and insertion. Two somatic pseudogenes were novel isoforms of their template genes not recorded in the Ensembl database. We did not observe any base substitutions in the somatic pseudogenes, implying that the reverse transcriptase has reasonable fidelity and the template transcripts had not undergone RNA editing.

On the basis of these characteristics, somatic pseudogenes, like their germline counterparts, are probably a product of the reverse transcription and transposition capability of endogenous LINE elements acting on mature mRNA transcripts. Most compelling is the presence of target-site duplications of 8–20 bp, with the point of insertion of the polyA tail almost universally occurring within a TTTTAA or very similar motif (Fig. 2b), which is the classic signature of LINE retrotransposition. Where we observed target-site deletions at insertion points the sequence showed less consensus, although there was still predilection for AT-rich sequences (Fig. 2b). We also found a somatic pseudogene inserted at the breakpoint of a somatically acquired genomic rearrangement (Supplementary Fig. 4), a process that can occur in the germline.

Interestingly, 21% (9/42) of somatic pseudogenes showed a single inverted rearrangement within the cDNA occurring away from the canonical splice sites of the gene (Fig. 2a,c, Supplementary Fig. 5). Such internal inverted rearrangements are observed in ~8% of L1 LINE elements in the genome and are thought to occur by a 'twin priming' mechanism. Here, not only does the TTTT overhang prime reverse transcription from the polyA tail, but the opposite strand, nicked 10–20 bp downstream, can fold back onto the mRNA transcript internally and prime another cDNA copy. These two partial copies of the mRNA are then resolved by non-homologous end-joining, leading to an inverted rearrangement. Our data are consistent with this mechanism. The internal rearrangements lead to a shuffling of the exon order and direction in the pseudogene, but without duplication of sequence (Fig. 2c). The insertion points of the 5' end of the inverted pseudogenes showed 1–3 bp of microhomology, consistent with a second priming event. Microhomology of between 1 bp and 4 bp was also usually present at the internal rearrangement breakpoint, indicative of non-homologous end-joining repair.

The 629 primary cases included data from 18 different tumour types (Supplementary Fig. 6, Supplementary Table 2). Somatic pseudogenes were most frequent in non-small cell lung cancer (19%; 5/27 patients) and colorectal cancer (18%; 2/11). The fact that such events occur in colorectal and lung cancers is consistent with the observation of high rates of somatic retrotransposition of LINE elements in these tumour types.

For four patients with somatic pseudogenes, we sequenced more than one tumour sample (Fig. 2d). In one patient, we analysed four clonally related samples from two bronchial lesions, both of which evolved from carcinoma in situ to invasive cancer. We found somatic pseudogenes common to all four lesions and others unique to a single lesion (Fig. 2d). In another case of lung cancer we found four somatic pseudogenes, all of which were present in both the carcinoma in situ and the invasive tumour. Similarly, for two patients with colorectal cancer, we sequenced...
the primary tumour and a liver metastasis. In both cases, the somatic pseudogene was present in both primary and metastasis. Taken together, these data indicate that somatic pseudogenes can form relatively early in cancer development, before the tumour becomes invasive or metastatic, but can also occur in more advanced stages of disease.

Highly expressed transcripts were especially likely to be templates for somatic pseudogenes (Fig. 3a). Overall, 63% of genes acting as the template for somatic pseudogenes were among the top quartile of expressed genes for that tumour type20, and 63% of genes acting as the template for somatic pseudogenes were among the top quartile of expressed genes for that tumour type20. A similar property has been noted for germline pseudogenes21. This may explain why we see several examples in which a gene was recurrently retrotransposed as a somatic pseudogene. For example, we found two somatic pseudogene copies of HDAC1, one in a colorectal cancer and one in a lung cancer (Supplementary Fig. 7a). Similarly, we see somatic pseudogenes involving two aldo-keto reductase genes, AKR1C1 and AKR1C3 (Supplementary Fig. 7b). The aldo-keto reductase proteins activate polycyclic aromatic hydrocarbons in tobacco smoke, a key step in inducing the genotoxicity of these critical carcinogens22,23. Overexpression of these genes has been documented in lung cancer24, which may explain why they recurrently serve as templates for somatic pseudogenes in this disease.

Like any mutational process, the majority of somatic pseudogenes are likely to be passenger mutations, but a few will have functional consequences that may be oncogenic. Somatic pseudogenes could exert functional consequences through many different mechanisms12, including fusion gene formation, increased expression of the pseudogene, disruption of a gene at the insertion site, sequestration of miRNAs from the template gene13 and production of antisense transcripts14. Among the 31 somatic pseudogenes with insertion sites identified, nine were inserted into introns and three into 3' UTRs. Of these, five were in the same orientation as the host gene, three in the opposite orientation and four had internal inverted rearrangements.

To assess transcriptional consequences of somatic pseudogene insertion, we performed RNA-sequencing on two primary cancers and three cell lines, as well as analysed 20 non-small lung cancers sequenced by TCGA25 (Supplementary Table 3). Across these samples, there were 16 somatic pseudogenes, of which 22% were inserted into introns and 16% into 3' UTRs, consistent with the proportions reported for somatic genes26. The remaining 62% of pseudogenes were inserted either intergenically or into 5' UTRs. The frequency of these events was not significantly different from the distribution of insertion sites for somatic genes, suggesting that the mechanism of insertion is similar for both classes of events.
which 10 were inserted in intergenic regions, three in introns, two in 3′ UTRs and one in the first exon of the MGA gene. We found no evidence of expression from somatic pseudogenes inserted in intergenic regions, whereas one of the three intronic insertions was expressed. In this case, a partial KTN1 pseudogene inserted into the last intron of PSD3 in a primary squamous cell lung cancer. We saw a clear peak of expression arising from the PSD3 intron immediately adjacent to the insertion, with aberrantly mapping read pairs aligning to the KTN1 UTR on one end and the PSD3 intron on the other end (Supplementary Fig. 8). Of the two insertions into 3′ UTRs, both were expressed. In one, a KRT6A pseudogene was inserted into the 3′ UTR of MLL, the latter being a well-known fusion gene in leukaemias (Supplementary Fig. 9). The RNA-sequencing data show that the last 1.2 kb of the MLL 3′ UTR is lost from the mature transcript and replaced by the somatic pseudogene, since we found paired-end reads spanning the 5′ insertion point but not the 3′ insertion point (Fig. 3b). The 3′ UTR of MLL has been shown to regulate transcript levels26, a feedback loop that could be disrupted by such a change, although expression of an aberrant transcript does not in itself imply oncogenicity. Similarly, a KIF18A pseudogene inserted into the 3′ UTRs of two overlapping genes on opposite strands, KIAA1967 and BIN3. Reads reporting both KIAA1967-KIF18A and BIN3-KIF18A fusion junctions were found in the RNA-sequencing data (Supplementary Fig. 10).

We also observed that somatic pseudogene insertion could abrogate expression of a target gene at the insertion site. In lung adenocarcinoma cell line NCI-H2009, a PTPN12 pseudogene caused an 8 kb target-site deletion that removed the promoter and first exon of MGA (Fig. 3c). In corresponding RNA-sequencing data, we find only wild-type exon 1 splicing into downstream exons, with no reads linking PTPN12 to MGA. Thus residual
The expression is derived from the intact MGA allele, and loss of the promoter and exon 1 has eliminated expression from the disrupted allele. MGA encodes a MAX-interacting protein, with focal deletions and inactivating mutations in lymphoid malignancies. From a compendium of 7,651 exome sequences, we previously found that MGA is a likely tumour suppressor gene on the basis of a statistically significant excess of nonsense mutations especially in lung adenocarcinoma (Supplementary Fig. 9b).

The diversity, complexity and iniquity of mutational processes operative during the development of cancer have been laid bare by whole-genome sequencing, and here we describe another novel mechanism of somatic mutation. There has been much recent interest in how retrotransposition of repeat elements in somatic cells reshapes the genome during normal brain development and during cancer development. The formation of pseudogenes in somatic cells represents a companion mutational process, with considerable flexibility in potential mechanisms to alter a cell’s transcriptional activity.
Methods

Sequence data. Sequencing data comprised low coverage (2–5 x genome coverage) paired-end sequencing for genomic rearrangements, high-coverage (30–40 x) paired-end shotgun sequencing, and targeted pull-down and sequencing of the coding exome generated at the Wellcome Trust Sanger Institute as described, both published and unpublished, on the Illumina HiSeq platform (Supplementary Table 1). In total, 660 cancer samples (629 primary samples and 31 cell lines) spanning 18 tumour types were analysed (Supplementary Table 2). The identity of cell lines was confirmed by STR testing and were obtained from ATCC. The Cambridgeshire Local Research Ethics Committee approved the studies and all patients gave informed consent.

Pseudogene detection. Data were aligned to both the reference genome (GRCh37) and the reference Ensembl transcriptome using BWA and the alignment coordinates from the transcriptome mapping were converted back to genome space. Owing to the different characteristics of exome versus genome data, the analyses of these alignments were optimized to deal with the different data types (Supplementary Fig. 1). It is important to note that, regardless of the analysis method, exome versus genome data are likely to have differential sensitivities for detecting pseudogenes.

Targeted exome data includes less than 2% (50 Mb) of the human genome and provides high depth over exons but does not contain the intronic and intergenic regions where the majority of genomic structural rearrangements lie. Analysis of genomic alignments through our standard structural variant algorithm was therefore sufficient for pseudogene detection in these data, as it provided high confidence calls for exon–exon junctions with few calls representing other types of genomic rearrangement. Candidate pseudogenes were required to have at least two apparent deletion events in the same gene, that is, involvement of at least three exons, with breakpoints separated by a distance of at least 500 bp but less than 50 kb, a size range that includes the majority of introns of the human genome.

To remove polymorphic germline pseudogenes, groups were excluded if they contained any read pairs from either a matched normal or an unmatched normal panel of nine unrelated individuals. Breakpoints were further excluded if the

Figure 3 | Tissue-specific patterns of somatic pseudogenes. (a) Expression of template genes for somatic pseudogenes (individual points) compared with all genes for the most frequently affected organ sites. The violin plot formulation for all genes shows the median (white point), interquartile range (thick black line) and 1.5 x the interquartile range (thin black line). The coloured shapes denote a kernel density plot of the distribution of gene expression levels for all genes. Due to the non-normal distribution, we used Wilcoxon rank-sum tests to test whether expression levels of template genes for somatic pseudogenes were different to that expected. (b) RNA-sequencing data showing expression of the MLL-KRT6A fusion gene. (c) Deletion of the promoter and first exon of MGA during somatic pseudogene insertion, leading to abrogation of expression from that allele.
intervening intron(s) aligned with a deletion annotated in the Database of Genomic Variants, as this indicates a potential germline pseudogene. Transcriptome alignments were used to identify reads that split across exon–exon boundaries by visual inspection in IGV (Integrative Genomics Viewer).

Due to the large size of the genome, the presence of repetitive elements, which tend to be more prevalent in non-coding sequences, and the presence of multiple genome versions in many genome types, genome data required a custom pipeline to be developed (Perl scripts available on request). After alignment to the transcriptome and conversion of the read mapping coordinates back to genome space, reads that are derived from pseudogenes are characterized by a large insert size, due to either reads mapping to adjacent exons, or a split in the CIGAR string where two halves of a read have aligned across a splice junction in the transcriptome. Read pairs were therefore filtered on these characteristics. Retained pairs were required to have an insert size of between 300 bp and 50 kb, two or fewer splits in the CIGAR string with the first and last match being at least 10 bp, and a mapping quality of at least 10. Read pairs passing these criteria were annotated against gene positions at both the 3' end of both (end 1 and read 2). Read pairs where all four coordinates mapped within the same protein coding gene in Ensembl were retained. The number of supporting read pairs per gene for the tumour, matched normal and an unmatched normal panel of 23 low-depth genomes were calculated and genes with no supporting reads in the normal/normal panel but more than three supporting reads in the tumour were validated by visual inspection in IGV.

**Insertion site mapping.** Somatic pseudogenes were best distinguished from contamination of the genomic DNA library by either RNA-sequence libraries or plasmids containing cDNA expression constructs by mapping the insertion sites of the pseudogene to base-pair resolution. Mappings to the reference genome were used to identify insertion sites using reads mapping between the pseudogene and elsewhere in the genome and/or reads that aligned to one side of the insertion point with soft clipping, and reads with coordinates within the pseudogene at both the 5' and 3' ends of both (end 1 and read 2). Read pairs where all four coordinates mapped within the same protein coding gene in Ensembl were retained. The number of supporting read pairs per gene for the tumour, matched normal and an unmatched normal panel of 23 low-depth genomes were calculated and genes with no supporting reads in the normal/normal panel but more than three supporting reads in the tumour were validated by visual inspection in IGV.

**Statistical analysis of recurrent point mutations.** To assess whether any of the insertions disrupted tumour suppressor genes at the insertion site, we analysed point mutation calls from 7,651 exomes available from previous publications, TCGA, ICGC and in-house for evidence of a higher than expected rate of inactivating mutation, using an adaptation of the method described in Greenman et al.\(^{22}\). This analysis has been described in detail elsewhere\(^{31}\).

**PCR validation and capillary sequencing.** The somatic status of pseudogenes was confirmed by designing primer pairs between the pseudogene and insertion site (where appropriate) and performing PCR on both tumour and matched normal genomic DNA. DNA for both tumour and matched normal samples was available for 12/16 somatic pseudogene insertion sites predicted from whole-genome (>30×) data, all 10 insertion site predictions from low-depth data, and all 3 mapped exome insertions. A total of 92% (11/12) whole-genome predictions validated, with the remaining 8% (1/12) not producing product in either tumour or normal, consistent with PCR/primer failure. All (3/3) exome predictions had confirmed somatic insertion points. Unsurprisingly, low-depth genomes, with no or low coverage sequencing of matching normals, gave a high rate of germline pseudogenes. Overall, 40% (4/10) PCRs showed product in both tumour and normal. In the case of somatic and the one pseudogene that produces product in either tumour or normal, the germline pseudogenes have not been included in this manuscript.

**References**


Acknowledgements
This work was supported by the Health Innovation Challenge Fund, the Wellcome Trust (grant reference 077012/Z/05/Z), and the Kay Kendall Leukaemia Fund. P.J.C. and S.M.J. are personally funded through Wellcome Trust Senior Clinical Research Fellowships. A.S. is funded by the H.L. Holmes Award from the National Research Council Canada and an EMBO fellowship. UAM is personally supported by a Cancer Research UK Clinician Scientist Fellowship. NB was supported by a starter grant from the Academy of Medical Sciences and from Lady Tata Memorial Trust. The breast cancer sequencing was supported by the BASIS project. The BASIS project is a European research project funded by the European Community’s Seventh Framework Programme (FP7/2010–2014) under the grant agreement number 242006. This research was supported in part by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences. We also acknowledge support for the sample banking and processing from UCL/UCLH NIHR Biomedical Centre, the Skeletal Cancer Action Trust (SCAT), especially Miss Dina Halai, and the Rosetrees Trust.

Author contributions

Additional information
Accession codes: Whole-exome sequencing data, whole-genome sequencing data and RNA-seq sequencing data have been deposited in the European Genome-phenome Archive [https://www.ebi.ac.uk/ega/]. DNA-seq accession codes EGAD0000100637, EGAD0000100638 and EGAD0000100639, respectively.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Cooke, S. L. et al. Processed pseudogenes acquired somatically during cancer development. Nat. Commun. 5:3644 doi: 10.1038/ncomms4644 (2014).

This work is licensed under a Creative Commons Attribution 3.0 Unported License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/