Human genome integration of SARS-CoV-2 contradicted by long-read sequencing

Nathan Smits[1,10], Jay Rasmussen[2,10], Gabriela O. Bodea[1,2,10], Alberto A. Amarilla[3,10], Patricia Gerdes[1], Francisco J. Sanchez-Luque[4,5], Prabha Ajjikuttira[2], Naphak Modhiran[3], Benjamin Liang[3], Jamila Faivre[6], Ira W. Deveson[7,8], Alexander A. Khromykh[3,9], Daniel Watterson[3,9], Adam D. Ewing[1], Geoffrey J. Faulkner[1,2]*

[1]Mater Research Institute - University of Queensland, TRI Building, Woolloongabba QLD 4102, Australia.

[2]Queensland Brain Institute, University of Queensland, Brisbane QLD 4072, Australia.

[3]School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane QLD 4072, Australia.

[4]GENYO, Pfizer-University of Granada-Andalusian Government Centre for Genomics and Oncological Research, PTS Granada 18016, Spain.

[5]MRC Human Genetics Unit, Institute of Genetics and Cancer (IGC), University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, United Kingdom.

[6]INSERM, U1193, Paul-Brousse University Hospital, Hepatobiliary Centre, Villejuif 94800, France.

[7]Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Sydney NSW 2010, Australia.

[8]St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney NSW 2052, Australia.

[9]Australian Infectious Diseases Research Centre, Global Virus Network Centre of Excellence, Brisbane QLD 4072, Australia.

[10]These authors contributed equally.

*Corresponding author: faulknergj@gmail.com (G.J.F)

Abstract

A recent study proposed severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) hijacks the LINE-1 (L1) retrotransposition machinery to integrate into the DNA of infected cells. If confirmed, this finding could have significant clinical implications. Here, we applied deep (>50×) long-read Oxford Nanopore Technologies (ONT) sequencing to HEK293T cells infected with SARS-CoV-2, and did not find any evidence of the virus existing as DNA. By examining ONT data from separate HEK293T cultivars, we resolved the complete sequences of 78 L1 insertions arising *in vitro* in the absence of L1 overexpression systems. ONT sequencing applied to hepatitis B virus (HBV) positive liver cancer tissues located a single HBV insertion. These experiments demonstrate reliable resolution of retrotransposon and exogenous virus insertions via ONT sequencing. That we found no evidence of SARS-CoV-2 integration suggests such events *in vivo* are highly unlikely to drive later oncogenesis or explain post-recovery detection of the virus.

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is a single-stranded ~30kbp polyadenylated RNA betacoronavirus[1]. SARS-CoV-2 does not encode a reverse transcriptase (RT) and therefore is not expected to integrate into genomic DNA as part of its life cycle. This assumption is of fundamental importance to the accurate diagnosis and potential long-term clinical consequences of SARS-CoV-2 infection, as demonstrated by other viruses known to incorporate into genomic DNA, such as human immunodeficiency virus 1 (HIV-1) and hepatitis B virus (HBV)[2–4].

LINE-1 (L1) retrotransposons reside in all mammalian genomes[5]. In humans, L1 transcribes a bicistronic mRNA encoding two proteins, ORF1p and ORF2p, essential to L1 mobility[6]. ORF2p possesses endonuclease (EN) and RT activities, and exhibits strong *cis* preference for reverse transcription of L1 mRNA[6–11]. Nonetheless, the L1 protein machinery can *trans* mobilise polyadenylated cellular RNAs, including those produced by non-L1 retrotransposons and protein-coding genes[11–14]. Somatic L1 *cis* mobilisation is observed in embryonic cells, the neuronal lineage, and various cancers[15–20]. By contrast, somatic L1-mediated *trans* mobilisation is very rare[15,16,20] and is likely repressed by various mechanisms[8,16,21–23]. Less than one cellular RNA *trans* insertion is expected for every 2000 *cis* L1 insertions[10]. Both *cis* and *trans* L1-mediated insertions incorporate target site duplications (TSDs) and a 3′ polyA tract, and integrate at the degenerate L1 EN motif 5′-TTTT/AA[6,11–14,24–27]. These sequence hallmarks can together discriminate artifacts from genuine insertions[28].

In a recent study, Zhang *et al.* overexpressed L1 in HEK293T cells, infected these with SARS-CoV-2, and identified DNA fragments of the virus through PCR amplification[29]. These results, alongside other less direct[30,31] analyses, were interpreted as evidence of SARS-CoV-2 genomic integration[29]. Crucially, Zhang *et al.* then detected 63 putative SARS-CoV-2 integrants by Oxford Nanopore Technologies (ONT) long-read sequencing. Of these, only a single integrant on chromosome X was spanned by an ONT read aligned to one locus, and was flanked by potential TSDs (**Extended Data Fig. 1**). However, this SARS-CoV-2 integrant did not incorporate a 3′ polyA tract, as is expected for an L1-mediated insertion, and involved an unusual 28kb internal deletion of the SARS-CoV-2 sequence. The SARS-CoV-2 integrants reported by Zhang *et al.* were 26-fold enriched in exons, despite the L1 EN showing no preference for these regions[32,33]. Zhang *et al.* also used Illumina short-read sequencing to map putative SARS-CoV-2 integration junctions in HEK293T cells without L1 overexpression. A lack of spanning reads and the tendency of Illumina library preparation to produce artefacts[34] leave this analysis open to interpretation.

The application of ONT sequencing to HEK293T cells nonetheless held conceptual merit. ONT reads can span germline and somatic retrotransposition events end-to-end, and resolve the sequence hallmarks of L1-mediated integration[23,35]. Through this approach, we previously found two somatic L1 insertions in the liver tumour sample of an individual positive for hepatitis C virus (HCV, a ~10kbp single-stranded non-polyadenylated RNA virus), including one PCR-validated L1 insertion spanned by a single ONT read[23,36]. HEK293T cells are arguably a favourable context to evaluate L1-mediated SARS-CoV-2 genomic integration. They express L1 ORF1p[37], readily accommodate engineered L1 retrotransposition[16,38,39], and support SARS-CoV-2 viral replication (**Extended Data Fig. 2**). Endogenous L1-mediated insertions can also be detected in cell culture by genomic analysis of separate cultivars derived from a common population[40,41].

We therefore applied ONT sequencing (~54× genome-wide depth, read length N50 ~ 39kbp) to genomic DNA harvested from HEK293T cells infected with SARS-CoV-2 at a multiplicity of infection (MOI) of 1.0, as well as mock infected cells (~28× depth, N50 ~ 47kbp) (**Fig. 1a**, **Extended Data Fig. 2** and **Supplementary Table 1**). As a positive control, we ONT sequenced the tumour and non-tumour liver tissue of a HBV-positive hepatocellular carcinoma patient[36]. To these data, we added those of Zhang *et al.*[29] and, as negative controls, the aforementioned HCV-positive hepatocellular carcinoma and normal liver samples[23] (**Supplementary Table 1**). We then used the Transposons from Long DNA Reads (TLDR)[23] software to call SARS-CoV-2, HBV, HCV and non-reference human-specific L1 (L1HS) insertions spanned by at least one uniquely aligned ONT read. TLDR detected no SARS-CoV-2, HBV or HCV insertions.

In total, TLDR identified 575 non-reference L1 insertions, which were typically flanked by TSDs with a median length of 14bp (**Fig. 1b** and **Supplementary Table 2**). No tumour-specific L1 insertions were found, apart from the two previously detected in the HCV-infected liver tumour[23,36]. Seventy-eight L1 insertions were found only in our SARS-CoV-2 infected HEK293T cells (66) or the mock infected control (12) and produced TSDs with a median length of 14bp (**Fig. 1c**). Of the 78 events, 69 (88.5%) were detected by a single spanning read and 13 carried a 3′ transduction[42,43] (**Supplementary Table 2**). We chose at random 6/69 insertions detected by one spanning read for manual curation and PCR validation. All 6 insertions bore a TSD and a 3′ polyA tract, and integrated at a degenerate L1 EN motif (**Fig. 1d,e** and **Extended Data Fig. 3a-d**). Three were 5′ inverted[44,45] (**Fig. 1e** and **Extended Data Fig. 3b,c**) and one carried a 3′ transduction[42] traced to a mobile[20] full-length non-reference L1HS (**Extended Data Fig. 3b**). Two PCR amplified in the SARS-CoV-2 and

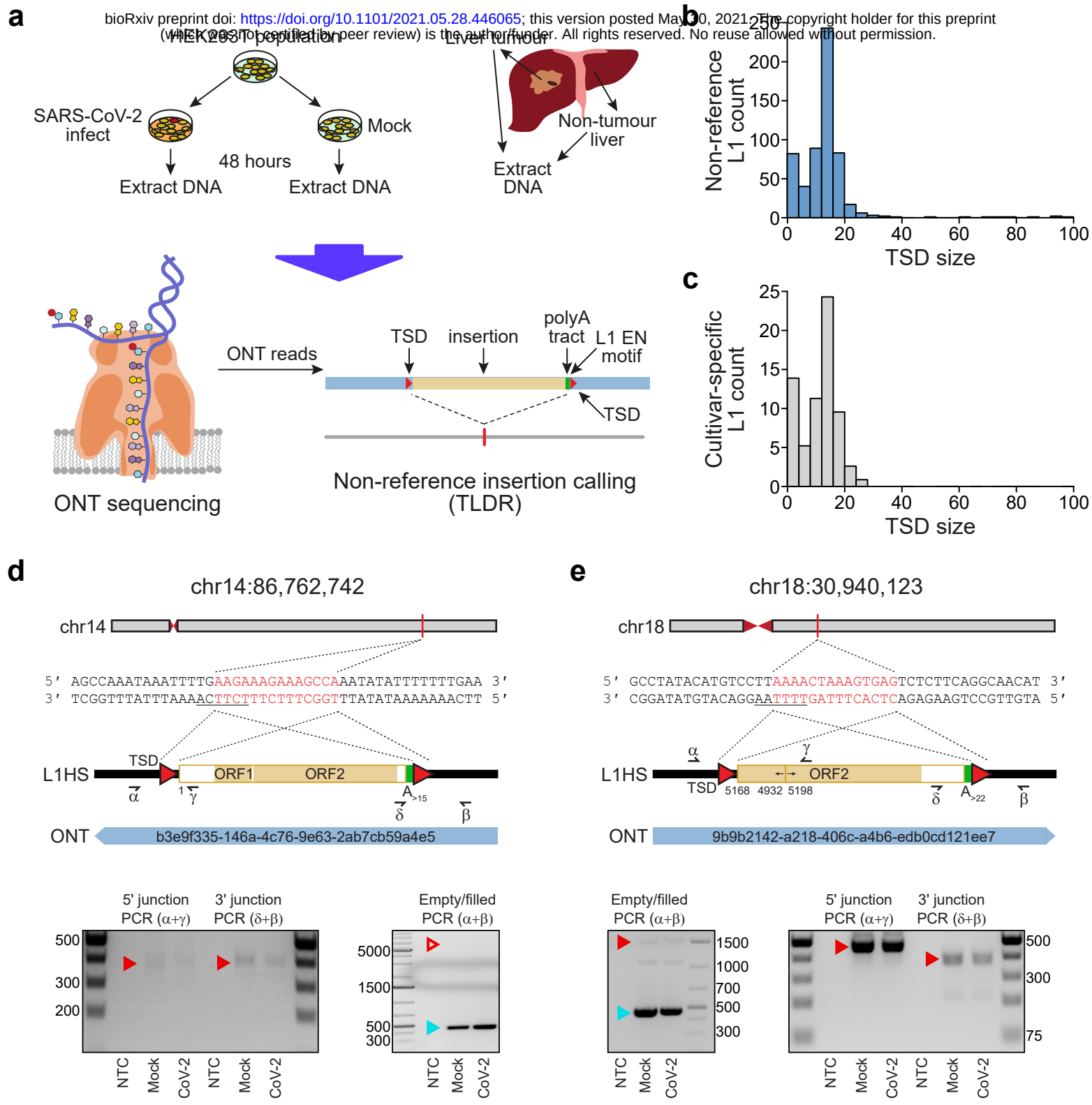


Fig. 1: Detection of endogenous L1-mediated retrotransposition in human cells. a, Experimental design. HEK-293T cells were divided into two populations (cultivars), which were then either SARS-CoV-2 infected or mock infected. DNA was extracted from each cultivar, as well as from hepatocellular carcinoma patient samples, and subjected to ONT sequencing. ONT reads were used to call non-reference L1 and virus insertions with TLDR, which also resolves TSDs and other retrotransposition hallmarks. TSDs: red triangles; polyA tract: green rectangle; ONT read: blue rectangle. Note: some illustrations are adapted from Ewing *et al.*[23]. **b,** TSD size distribution for non-reference L1 insertions, as annotated by TLDR. **c,** As for b, except showing data for L1 insertions found only in either our HEK293T cells infected with SARS-CoV-2 or our mock infected cells. **d,** Detailed characterisation of an L1 insertion detected in SARS-CoV-2 infected HEK293T cells by a single spanning ONT read aligned to chromosome 14. Nucleotides highlighted in red correspond to the integration site TSD. Underlined nucleotides correspond to the L1 EN motif. The cartoon indicates a full-length L1HS insertion flanked by TSDs (red triangles), and a 3′ polyA tract (green). Numerals represent positions relative to the L1HS sequence L1.3[46]. The relevant spanning ONT read, with identifier, is positioned underneath the cartoon. Symbols (α, β, δ, γ) represent the approximate position of primers used for empty/filled site and L1-genome junction PCR validation reactions. Gel images display the results of these PCRs. Ladder band sizes are as indicated, NTC; non-template control. Red triangles indicate the expected size of L1 amplicons (empty triangle: no product observed; filled triangle: product observed). Blue triangles indicate expected empty site sizes. **e,** As for d, except for a 5′ inverted/deleted L1HS located on chromosome 18.

mock infected samples (**Fig. 1d,e**) and four did not amplify in either sample (**Extended Data 3a-d**). The 6 integration sites were on average spanned by 86 reads not containing the L1 insertion (**Extended Data Fig. 3e**), a ratio (1:86) suggesting the L1s were absent from most cells. These and earlier[23,35] experiments show that lone spanning ONT reads can recover *bona fide* retrotransposition events, and highlight endogenous L1 activity in HEK293T cells lacking L1 overexpression systems.

We next tested whether our computational analysis parameters excluded genuine HBV, HCV or SARS-CoV-2 insertions. We directly aligned our ONT reads to the genome of the SARS-CoV-2 isolate (QLD002, GISAID EPI_ISL_407896) used here, as well as to a geographically diverse set of HBV and HCV genomes (**Supplementary Table 1**), and a highly mobile L1HS sequence[46]. In total, 3.6% of our ONT sequence bases aligned to L1HS, whereas no alignments to the SARS-CoV-2 or HCV genomes were observed (**Fig. 2a**). One read from the HBV-infected non-tumour liver sample aligned to 2,770bp of a HBV genotype B isolate, and the remaining 2,901bp aligned to an intergenic region of chromosome 2 (**Fig. 2b** and **Supplementary Table 2**). To validate this HBV insertion, we PCR amplified and capillary sequenced its 3′ junction (**Fig. 2b**). The HBV sequence was linearised and rearranged (**Fig. 2b**) as per prior reports[2–4]. Direct inspection of ONT read alignments thus recovered a HBV integrant, which are found in ~1 per 10^1-10^4 infected hepatocytes[47–49], yet did not reveal reads alignable to the SARS-CoV-2 genome in our ONT datasets.

Reanalysing the ONT data generated by Zhang *et al.*, we found 555 reads (out of ~12 million) that could be aligned to the SARS-CoV-2 genome (**Fig. 2a**), including one matching the aforementioned integrant on chromosome X that lacked a 3′ polyA tract[29] (**Extended Data Fig. 1**). These reads (median length 924bp) were however 65.6% shorter than the overall dataset (2,686kbp) and were comprised of a much higher average proportion of SARS-CoV-2 sequence (52.3%) than the proportion of L1HS sequence found in reads aligned to L1HS (17.1%). An ONT read highlighted by Zhang *et al.* in support of a TSD-bearing SARS-CoV-2 insertion on chromosome 22 was also used to call a SARS-CoV-2 insertion on chromosome 1 (**Extended Data Fig. 1**). Read alignment ambiguity to the genome resulted in TLDR calling neither the putative chromosome X or chromosome 22 SARS-CoV-2 integrants. These analyses confirmed SARS-CoV-2 alignable reads were present in the Zhang *et al.* ONT dataset, yet these reads were unusually short and could include molecular artifacts interpreted by Zhang *et al.* as SARS-CoV-2 integrants.

In sum, we do not observe L1-mediated SARS-CoV-2 genomic integration in HEK293T cells, despite availability of the L1 machinery[16,37–39]. Our approach has several

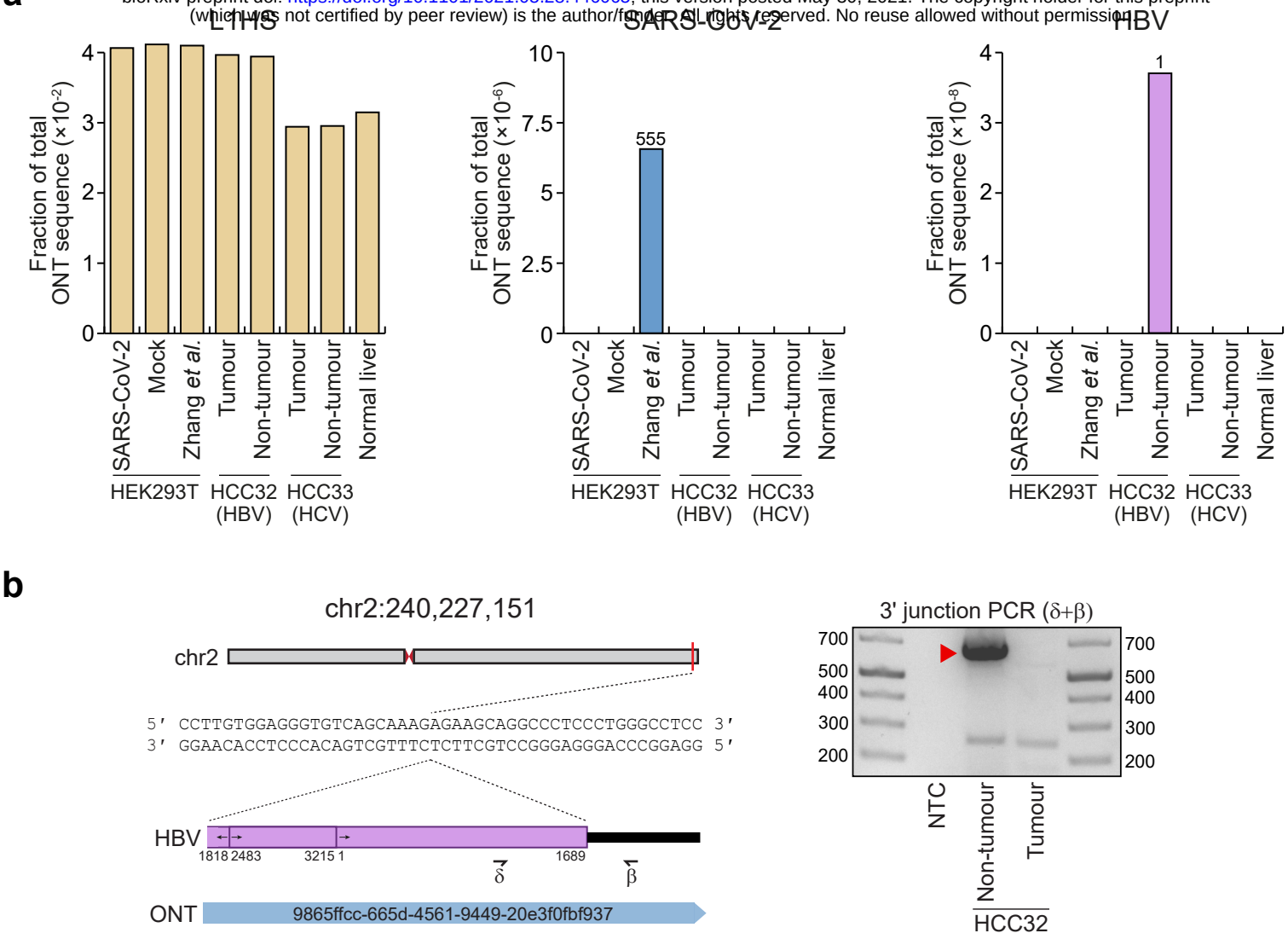


Fig. 2: ONT reads occasionally align to viral genome sequences. a, Fractions of total ONT sequence alignable to L1HS (left), SARS-CoV-2 (middle) and HBV (right) isolate genomes. Read counts for SARS-CoV-2 and HBV are provided above histogram columns. No reads were aligned to the HCV isolate genomes. HEK293T data were generated here (SARS-CoV-2, mock) or by Zhang *et al.*[29]. HCC tumour/non-tumour liver pairs were sequenced here (HCC32; confirmed HBV-positive) or previously[23] (HCC33; HCV-positive). Normal liver ONT sequencing from our prior work[23] was included as an additional control. **b,** A HBV insertion detected in non-tumour liver. In this example, an ONT read from the non-tumour liver of HCC32 spanned the 3′ junction of a HBV integrant located on chromosome 2. Of the HBV isolate genomes considered here (**Supplementary Table 1**), this read aligned best to a representative of genotype B (Genbank accession AB602818). The HBV sequence was rearranged consistent with its linearisation prior to integration[2–4]. Numerals indicate positions relative to AB602818. Symbols (β, δ) represent the approximate position of primers used to PCR validate the HBV insertion. The gel image at right shows the PCR results. Ladder band sizes are as indicated. The red triangle indicates an on-target product.

notable differences and caveats when compared to that of Zhang *et al.*[29]. Each study used different SARS-CoV-2 isolates, and here the multiplicity of infection (MOI 1.0) was double that of Zhang *et al.* (MOI 0.5). The ONT library preparation kit and depth of sequencing applied to HEK293T cells by Zhang *et al.* (SQK-LSK109 kit, ~21× depth, N50 ~ 11kbp) and here (SQK-LSK110 kit, ~54× depth, N50 ~ 39kbp) differed. Zhang *et al.* applied ONT sequencing only to HEK293T cells transfected with an L1 expression plasmid, which human cells would not carry *in vivo*. We do not analyse SARS-CoV-2 patient samples although, arguably, HEK293T cells present an environment far more conducive to L1 activity than those cells accessed *in vivo* by SARS-CoV-2[50,51]. Widespread cell death post-infection also reduces the probability SARS-CoV-2 integrants would persist in the body[52,53]. Finally, the incredible enrichment reported by Zhang *et al.* for putative SARS-CoV-2 insertions in exons, which the L1 EN does not prefer[32,33], contradicts the involvement of L1. We conclude L1 *cis* preference likely disfavours SARS-CoV-2 retrotransposition, making the phenomenon mechanistically plausible but likely very rare, as for other polyadenylated cellular RNAs[6–11].

Acknowledgements

The authors thank the human subjects of this study who donated tissues to the Centre Hépatobiliaire, Paul-Brousse Hospital. We thank S. Richardson and R. Shukla for helpful discussions, K. Chappell and P. Young for project support, and Queensland Health for providing the SARS-CoV-2 virus isolate QLD02. This study was funded by an Australian Government Research Training Program (RTP) Scholarships (N.S.), an NHMRC-ARC Dementia Research Development Fellowship (GNT1108258, G.O.B.), seed funding provided by the Australian Infectious Disease Research Centre to establish SARS-CoV-2 research at the University of Queensland (A.A.K), an Australian Department of Health Medical Research Future Fund (MRFF) Novel Coronavirus Vaccine Development Grant (APP1202445-2020, D.W.), an MRFF Investigator Grant (MRF1175457, A.D.E.), an NHMRC Investigator Grant (GNT1173711, G.J.F.), a CSL Centenary Fellowship (G.J.F.), and the Mater Foundation.

Author contributions

N.S., J.R., G.O.B., A.A.A., P.G., F.J.S-L., P.A., N.M. and B.L. performed experiments and analysed data. A.D.E. and G.J.F. performed bioinformatic analysis. J.F., I.W.D., A.A.K., D.W. and G.J.F. provided resources. G.J.F. designed the project and wrote the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269 (2020).
2. Fujimoto, A. *et al.* Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
3. Nagaya, T. *et al.* The mode of hepatitis B virus DNA integration in chromosomes of human hepatocellular carcinoma. *Genes Dev.* **1**, 773–782 (1987).
4. Jiang, Z. *et al.* The effects of hepatitis B virus integration into the genomes of hepatocellular carcinoma patients. *Genome Res.* **22**, 593–601 (2012).
5. Kazazian, H. H., Jr & Moran, J. V. Mobile DNA in Health and Disease. *N. Engl. J. Med.* **377**, 361–370 (2017).
6. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
7. Kulpa, D. A. & Moran, J. V. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat. Struct. Mol. Biol.* **13**, 655–660 (2006).
8. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol. Cell* **60**, 728–741 (2015).
9. Monot, C. *et al.* The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS Genet.* **9**, e1003499 (2013).
10. Wei, W. *et al.* Human L1 retrotransposition: cispreference versus trans complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001).
11. Garcia-Perez, J. L., Doucet, A. J., Bucheton, A., Moran, J. V. & Gilbert, N. Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res.* **17**, 602–611 (2007).
12. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).
13. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).
14. Hancks, D. C., Ewing, A. D., Chen, J. E., Tokunaga, K. & Kazazian, H. H., Jr. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res.* **19**, 1983–1991

(2009).

15. Evrony, G. D. *et al.* Cell lineage analysis in human brain using endogenous retroelements. *Neuron* **85**, 49–59 (2015).
16. Sanchez-Luque, F. J. *et al.* LINE-1 Evasion of Epigenetic Repression in Humans. *Mol. Cell* **75**, 590–604 (2019).
17. Feusier, J. *et al.* Pedigree-based estimation of human mobile element retrotransposition rates. *Genome Res.* **29**, 1567–1577 (2019).
18. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).
19. Schauer, S. N. *et al.* L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. *Genome Res.* **28**, 639–653 (2018).
20. Rodriguez-Martin, B. *et al.* Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat. Genet.* **52**, 306–319 (2020).
21. Deniz, Ö., Frost, J. M. & Branco, M. R. Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.* **20**, 417–431 (2019).
22. Ahl, V., Keller, H., Schmidt, S. & Weichenrieder, O. Retrotransposition and Crystal Structure of an Alu RNP in the Ribosome-Stalling Conformation. *Mol. Cell* **60**, 715–727 (2015).
23. Ewing, A. D. *et al.* Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. *Mol. Cell* **80**, 915–928 (2020).
24. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 1872–1877 (1997).
25. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**, 595–605 (1993).
26. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol. Cell. Biol.* **25**, 7780–7795 (2005).
27. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).
28. Faulkner, G. J. & Billon, V. L1 retrotransposition in the soma: a field jumping ahead. *Mob. DNA* **9**, 22 (2018).
29. Zhang, L. *et al.* Reverse-transcribed SARS-CoV-2 RNA can integrate into the genome

of cultured human cells and can be expressed in patient-derived tissues. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).

30. Kazachenka, A. & Kassiotis, G. SARS-CoV-2-host chimeric RNA-sequencing reads do not necessarily signify virus integration into the host DNA. *bioRxiv* 2021.03.05.434119 (2021) doi:10.1101/2021.03.05.434119.
31. Yan, B. *et al.* Host-virus chimeric events in SARS-CoV2 infected cells are infrequent and artifactual. *J. Virol.* (2021) doi:10.1128/JVI.00294-21.
32. Sultana, T. *et al.* The Landscape of L1 Retrotransposons in the Human Genome Is Shaped by Pre-insertion Sequence Biases and Post-insertion Selection. *Mol. Cell* **74**, 555–570 (2019).
33. Flasch, D. A. *et al.* Genome-wide de novo L1 Retrotransposition Connects Endonuclease Activity with Replication. *Cell* **177**, 837–851.e28 (2019).
34. Treiber, C. D. & Waddell, S. Resolving the prevalence of somatic transposition in Drosophila. *Elife* **6**, (2017).
35. Siudeja, K. *et al.* Unraveling the features of somatic transposition in the Drosophila intestine. *EMBO J.* **40**, e106388 (2021).
36. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013).
37. Philippe, C. *et al.* Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife* **5**, (2016).
38. Kubo, S. *et al.* L1 retrotransposition in nondividing and primary human somatic cells. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 8036–8041 (2006).
39. Niewiadomska, A. M. *et al.* Differential inhibition of long interspersed element 1 by APOBEC3 does not correlate with high-molecular-mass-complex formation or P-body association. *J. Virol.* **81**, 9577–9583 (2007).
40. Nguyen, T. H. M. *et al.* L1 Retrotransposon Heterogeneity in Ovarian Tumor Cell Evolution. *Cell Rep.* **23**, 3730–3740 (2018).
41. Klawitter, S. *et al.* Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat. Commun.* **7**, 10286 (2016).
42. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H., Jr. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.* **7**, 143–148 (1994).
43. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H., Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).

44. Ostertag, E. M. & Kazazian, H. H., Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).
45. Kazazian, H. H., Jr *et al.* Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
46. Dombroski, B. A., Scott, A. F. & Kazazian, H. H., Jr. Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 6513–6517 (1993).
47. Mason, W. S. *et al.* HBV DNA Integration and Clonal Hepatocyte Expansion in Chronic Hepatitis B Patients Considered Immune Tolerant. *Gastroenterology* **151**, 986–998.e4 (2016).
48. Rydell, G. E. *et al.* Abundance of non-circular intrahepatic hepatitis B virus DNA may reflect frequent integration into human DNA in chronically infected patients. *J. Infect. Dis.* (2020) doi:10.1093/infdis/jiaa572.
49. Tu, T., Budzinska, M. A., Vondran, F. W. R., Shackel, N. A. & Urban, S. Hepatitis B Virus DNA Integration Occurs Early in the Viral Life Cycle in an In Vitro Infection Model via Sodium Taurocholate Cotransporting Polypeptide-Dependent Uptake of Enveloped Virus Particles. *J. Virol.* **92**, (2018).
50. Sungnak, W. *et al.* SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes. *Nat. Med.* **26**, 681–687 (2020).
51. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, Transmission, Diagnosis, and Treatment of Coronavirus Disease 2019 (COVID-19): A Review. *JAMA* **324**, 782–793 (2020).
52. Karki, R. *et al.* Synergism of TNF-α and IFN-γ Triggers Inflammatory Cell Death, Tissue Damage, and Mortality in SARS-CoV-2 Infection and Cytokine Shock Syndromes. *Cell* **184**, 149–168.e17 (2021).
53. Varga, Z. *et al.* Endothelial cell infection and endotheliitis in COVID-19. *Lancet* **395**, 1417–1418 (2020).
54. Amarilla, A. A. *et al.* An optimized high-throughput immuno-plaque assay for SARS-CoV-2. *Front. Microbiol.* **12**, 625136 (2021).
55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

57. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
58. Untergasser, A., Cutcutache, I. & Koressaar, T. Primer3—new capabilities and interfaces. *Nucleic acids* (2012).

Methods

SARS-CoV-2 infection of HEK293T cells

HEK293T cells and African green monkey kidney cells (Vero E6) were maintained in standard Dulbecco's Modified Eagle Medium (DMEM). Culture media were supplemented with sodium pyruvate (11mg/L), penicillin (100U/mL), streptomycin (100µg/mL) (P/S) and 10% foetal calf serum (FCS) (Bovogen, USA). Cells were maintained at 37 ºC with 5% CO_2.

An early Australian SARS-CoV-2 isolate (hCoV-19/Australia/QLD02/2020; GISAID Accession EPI_ISL_407896) was sampled from patient nasopharyngeal aspirates by Queensland Health Forensic and Scientific Services and used to inoculate Vero E6 African green monkey kidney cells (passage 2). A viral stock (passage 3) was then generated on Vero E6 cells and stored at -80ºC. Viral titration was determined by immuno-plaque assay (iPA), as previously described[54]. To verify viral replication in HEK293T cells, a growth kinetic was assessed using a multiplicity of infection (MOI) of 1.0, and showed efficient SARS-CoV-2 replication (**Extended Data Fig. 2**).

HEK293T viral infection was undertaken as follows: 3×10^6 HEK293T cells were seeded onto 6-well plates pre-coated with polylysine one day before infection. Cells were infected at MOI of 1 in 200µL of DMEM (2% FCS and P/S) and incubated for 30min at 37ºC. Plates were rocked every 5min to ensure the monolayer remained covered with inoculum. The inoculum was then removed, and the monolayer washed five times with 1mL of additive-free DMEM. Finally, cells were maintained with 3mL of DMEM (supplemented with 2% foetal bovine serum and P/S) and incubated at 37ºC with 5% CO_2.. Cell supernatant was harvested 0, 1, 2 and 3 days post-infection. The mock infected control differed only in that virus was not added to the inoculum media.

Genomic DNA was extracted from mock and SARS-CoV-2 infected (MOI 1.0) HEK293T cells sampled 2 days post-infection, using a Nanobind CBB Big DNA Kit (Circulomics) following the manufacturer's instructions for high molecular weight (HMW) DNA extraction. DNA was eluted in elution buffer (10mM Tris-Cl, pH 8.5) and concentration measured by Qubit dsDNA High-Sensitivity Assay Kit on a Qubit Fluorometer (Life Technologies).

Hepatocellular carcinoma patient samples

Liver tumour and non-tumour tissue were previously obtained from a HBV-positive patient (HCC32, male, 73yrs) who underwent surgical resection at the Centre Hepatobiliaire, Paul-Brousse Hospital, and made available for research purposes with approval from the French Institute of Medical Research and Health (Reference: 11-047). Further ethics approvals were provided by the Mater Health Services Human Research Ethics Committee (Reference: HREC-15-MHS-52) and the University of Queensland Medical Research Review Committee (Reference: 2014000221). DNA was extracted from the HCC32 tissues in our earlier study[36] with a DNeasy Blood and Tissue Kit (QIAGEN, Germany) and stored at -80ºC. To enrich for HMW DNA, 4.5µg of DNA from the patient HCC32 tumour and non-tumour liver samples was diluted to 75ng/µL in a 1.5mL Eppendorf DNA LoBind tube and processed with a Short Read Eliminator XS Kit (Circulomics) following the manufacturer's instructions.

ONT sequencing

DNA libraries were prepared at the Kinghorn Centre for Clinical Genomics (KCCG) using 3-4µg HMW input DNA, without shearing, and a SQK-LSK110 ligation sequencing kit. 350-500ng of each prepared library was sequenced separately on one PromethION (Oxford Nanopore Technologies) flow cell (FLO-PRO002, R9.4.1 chemistry) (**Supplementary Table 1**). SARS-CoV-2 infected HEK293T DNA was sequenced on two flow cells. Flow cells were washed (nuclease flush) and reloaded at 24hr and 48hr with 350-500ng of additional library to maximise output. Bases were called with guppy 4.0.11 (Oxford Nanopore Technologies). Sequencing data were deposited in the Sequence Read Archive (SRA) under project PRJEB44816.

ONT bioinformatic analyses

To call non-reference insertions with TLDR[23], ONT reads generated here, by Zhang *et al.*[29], and by our previous ONT study of human tissues[23] (**Supplementary Table 1**) were aligned to the human reference genome build hg38 using minimap2[55] version 2.17 (index parameter: -x map-ont; alignment parameters: -ax map-ont -L -t 32) and samtools[56] version 1.12. BAM files were then processed as a group with TLDR[23] version 1.1 (parameters -e virus.fa -p 128 -m 1 --max_te_len 40000 --max_cluster_size 100 --min_te_len 100 --wiggle 100 --keep_pickles -n nonref.collection.hg38.chr.bed.gz). The file virus.fa was composed of: representative HBV and HCV isolate genomes (**Supplementary Table 1**), the SARS-CoV-2 isolate used here (GISAID Accession EPI_ISL_407896) and the mobile L1HS sequence

L1.3[46] (Genbank Accession L19088). The file nonref.collection.hg38.chr.bed.gz is a collection of known non-reference retrotransposon insertions available from github.com/adamewing/tldr/. The TLDR output table was further processed to remove calls not passing all TLDR filters, representing homopolymer insertions, where MedianMapQ < 50 or family = "NA" or remappable = "FALSE" or UnmapCover < 0.75 or LengthIns < 100 or EndTE-StartTE < 100 or strand = "None" or SpanReads < 1 or L1HS insertions where EndTE < 6017. The filtered TLDR output table is provided as **Supplementary Table 2**. L1HS insertions detected in only our mock or SARS-CoV-2 infected HEK293T datasets, but not in both experiments, and not matching a known non-reference L1HS element, were designated as putative cultivar-specific insertions (**Supplementary Table 2**). Many if not most of these insertions were likely to have occurred in cell culture prior to the cultivars being separated.

To identify L1HS and viral sequences, we directly aligned all reads to the virus.fa file with minimap2 (index parameter: -x map-ont; alignment parameters: -ax map-ont -L -t 32). Reads containing alignments of ≥100bp to a sequence present in virus.fa were counted with samtools idxstats. Alignments to HBV, HCV or SARS-CoV-2 were excluded if they overlapped with a genomic alignment of ≥100bp. Read alignments were visualised with samtools view and the Integrative Genomics Viewer[57] version 2.8.6.

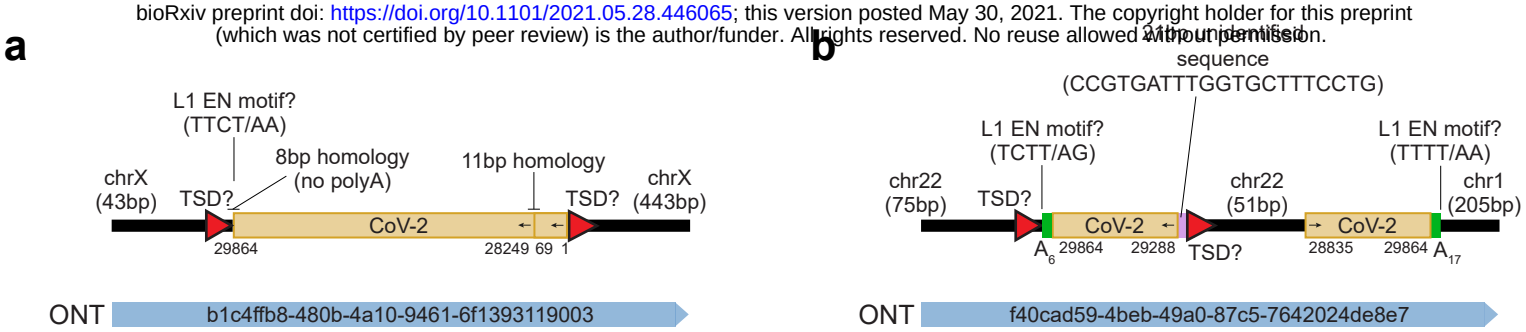
PCR validation

We used Primer3[58] to design PCR primers for 6 L1 insertions found by a single spanning ONT read, using the reference genome and L1HS sequences as inputs (**Supplementary Table 2**). These validation experiments were conducted in three phases. Firstly, we performed an "empty/filled site" PCR using primers positioned on either side of the L1, where the filled site is the L1 allele, and the empty site is the remaining allele(s). Each empty/filled reaction was performed using a DNA Engine Tetrad 2 Thermal Cycler (Bio-Rad) and Expand Long Range Enzyme Mix, with 1X Expand Long Range Buffer with $MgCl_2$, 50pmol of each primer, 0.5mM dNTPs, 5% DMSO, 100ng of template DNA and 1.75U of enzyme, in a 25µL final volume. PCR cycling conditions were as follows: (92˚C, 3min)×1; (92˚C, 30sec; 54-57˚C, 30sec; 68˚C, 7min)×10; (92˚C, 30sec; 52-55˚C, 30sec; 68˚C, 7min + 20sec/cycle)×30; (68˚C, 10min; 4˚C, hold)×1. Amplicons were visualised on a 1% agarose gel stained with SYBR SAFE (Invitrogen). GeneRuler™ 1kb plus (Thermo Scientific) was used as the ladder. Secondly, we combined each empty/filled primer with a primer positioned within the L1 sequence, to amplify the 5′ and 3′ L1-genome junctions.

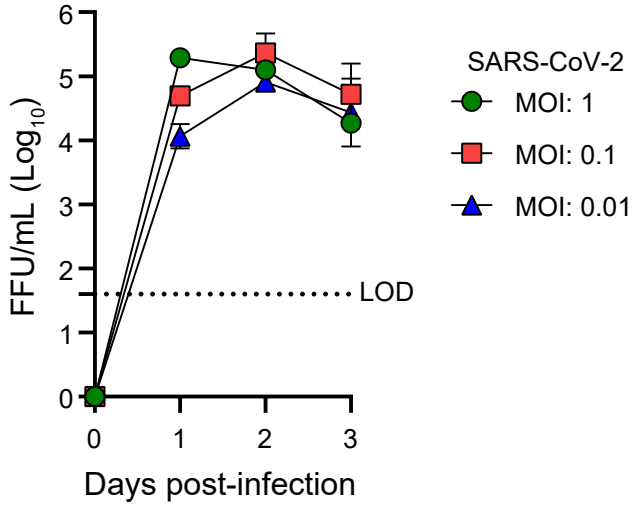
These reactions were conducted on a T100 Thermal Cycler (Bio-Rad), with MyTaq HS DNA polymerase, 1X MyTaq Reaction Buffer, 10pmol of each primer, 10ng of template DNA, and 2.5U of enzyme, in a 25µL final volume. PCR cycling conditions were as follows: (95˚C, 1min)×1; (95˚C, 15sec; 53-55˚C, 15sec; 72˚C, 15sec)×35; (72˚C, 5min; 4˚C, hold)×1. Amplicons were visualised on a 1.5% agarose gel stained with SYBR SAFE (Invitrogen). Thirdly, we repeated the 5′ L1-genome junction-specific PCR using 200ng template DNA. All PCRs were performed with non-template control, as well as DNA extracted from the same HEK293T cells (SARS-CoV-2 and mock) subjected to genomic analysis. Notably, L1 insertions that did not amplify in either cultivar were still likely to be genuine events as they carried all of the relevant sequence hallmarks of L1-mediated retrotransposition.

PCR primers for the HBV insertion 3′ junction (**Fig. 2b** and **Supplementary Table 2**) were designed with Primer3 using the reference genome and closest match HBV sequence (Genbank accession AB602818) as inputs. PCR amplification and capillary sequencing was conducted as per the L1 insertions, except using Expand Long Range polymerase (Roche) with 1X Expand Long Range buffer with $MgCl_2$, 10pmol of each primer, 100ng of template DNA, 500µM of PCR Nucleotide Mix, and 3.5U of enzyme, in a 25µL final volume. PCR cycling conditions were as follows: (92˚C, 2min)×1; (92˚C, 15sec; 65˚C, 15sec; 68˚C, 7:30min)×10; (92˚C, 15sec; 65˚C, 15sec; 68˚C, 7min+ 20sec per cycle)×35 (68˚C, 10min; 4˚C, hold)×1. Amplicons were visualized on a 1.2% agarose gel.

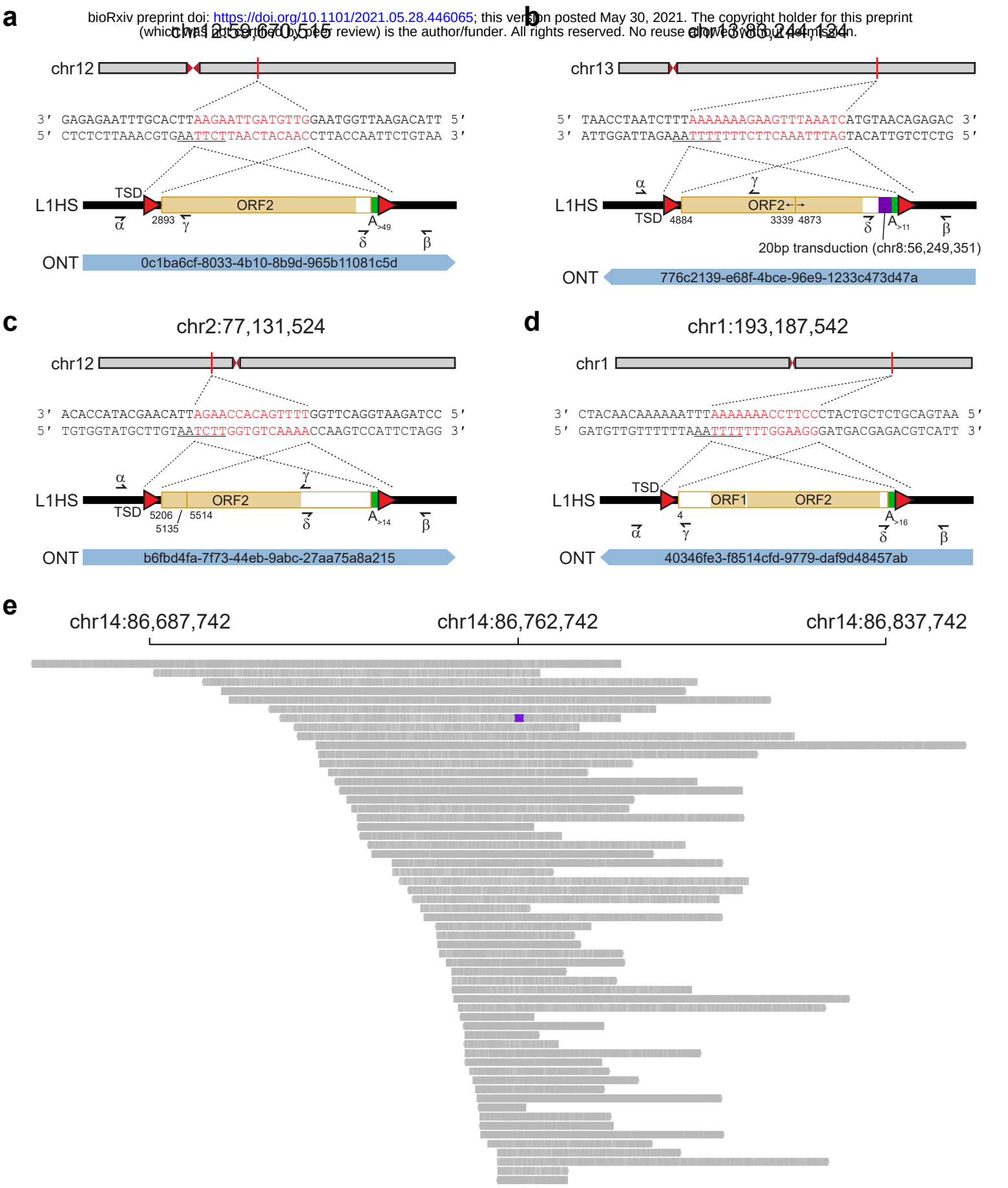
Amplicons in each experiment were visualised using a GelDoc (Bio-Rad) and, if of the correct size, gel extracted using a Qiagen MinElute Gel Extraction Kit and capillary sequenced by the Australian Genomics Research Facility (Brisbane).



Extended Data Fig. 1: Key SARS-CoV-2 insertions reported by Zhang *et al.*. a, A cartoon summarising the features of a putative SARS-CoV-2 integrant on chromosome X. Numerals underneath the SARS-CoV-2 sequence represent positions relative to the QLD02 virus isolate. Potential TSDs are shown as red triangles. No 3′ polyA tract was found. Homologous regions at sequence junctions are marked. One spanning ONT read is positioned underneath the cartoon and its identifier is displayed. **b,** As for a, except showing an ONT read spanning two SARS-CoV-2 insertions, on chromosome 22 and chromosome 1. The alignments to chromosome 22 are flagged as supplementary by the minimap2 aligner. 3′ polyA tracts are represented as green rectangles. Note: the chromosome 22 and chromosome X instances are the key examples reported by Zhang *et al.* in support of SARS-CoV-2 genomic integration. Neither example has a complete set of retrotransposition hallmarks (TSD, 3′ polyA tract, L1 EN motif) *and* the support of a uniquely aligned ONT read.



Extended Data Fig. 2: SARS-CoV-2 is replication competent in HEK293T cells. HEK293T cells were infected with SARS-CoV-2 isolate QLD02 at an MOI of 0.01, 0.1 and 1.0. Inoculum was removed after infection and cells were washed before the addition of growth media. Supernatant was collected at the indicated time points and viral titres were quantified as focus-forming units (FFU) per mL by immuno-plaque assay (iPA)[54] with a limit of detection (LOD) as indicated.



Extended Data Fig. 3: Additional HEK293T cell L1HS insertions. a, A 5′ truncated L1. **b,** A 5′ inverted/deleted L1 carrying a 3′ transduction (purple rectangle) traced to a known non-reference L1 present in HEK293T cells. **c,** A 5′ inverted/deleted L1. **d,** A near full-length L1. Each panel shows the genomic coordinates of an L1 insertion, as well as the sequence at the insertion site. Nucleotides highlighted in red correspond to the integration site TSD. Underlined nucleotides correspond to the L1 EN motif. Cartoons summarise the features of each L1, with numerals representing positions relative to L1.3[46], TSDs shown as red triangles, and 3′ polyA tracts coloured as green rectangles. One spanning ONT read with its identifier is positioned underneath each cartoon. Symbols (α, β, δ, γ) represent the approximate position of primers used for empty/filled and L1-genome junction PCR validation reactions. No L1 amplicons were recovered by these assays. **e.** Integrative Genomics Viewer[57] visualisation of read alignments spanning the L1 integration site displayed in Fig. 1d. The L1 is coloured purple.