Structural biology

Snapshots of genetic copy-and-paste machinery

Gael Cristofari

LINE-1 DNA elements self-duplicate, inserting the copy into new regions of the genome — a key process in chromosome evolution. Structures of the machinery that performs this process in humans are now reported. **See p.186 & p.194**

DNA sequences known as LINE-1 (L1) elements replicate themselves and spread throughout mammalian genomes using a biological 'copy and paste' mechanism. On pages 186 and 194, respectively, Thawani *et al.*[1] and Baldwin *et al.*[2] report X-ray and cryoelectron-microscopy structures of the remarkable complex that enables L1 to propagate itself — providing insight into how and where this DNA element is duplicated.

Enzymes known as reverse transcriptases (RTs) enable retroviruses, such as HIV, to complete their replication cycle. These enzymes convert the viral RNA genome into DNA, which is then integrated into the chromosomes of the infected host. The discovery of RTs in 1970 revolutionized molecular biology by revealing the existence of a reverse flow of genetic information (RNA to DNA), challenging the dogma that only the forward flow — DNA is transcribed into RNA, which is translated into proteins — can occur.

Nearly two decades later, some cases of haemophilia A were attributed to a mutation caused by the insertion of an L1 DNA fragment into a gene that encodes a blood coagulation factor[3]. Given that many copies of L1 are present in human chromosomes, and that this fragment encodes an enzyme distantly related to retrovirus RTs, it was this discovery that suggested that L1s actively replicate in humans and spread themselves throughout the genome using a copy-and-paste mechanism. We now know that L1s not only cause genetic diseases, but are also involved in many cancers, in ageing and probably in several neurodegenerative diseases[3].

L1 sequences constitute one of the many families of DNA elements known as retrotransposons, which have invaded the genomes of all eukaryotes (organisms that include plants, animals and fungi). The L1 family has been remarkably successful throughout evolution, with the activity of these sequences accounting for roughly one-third of human chromosomes; by contrast, only about 1% of human chromosomes consist of DNA that encodes other cellular proteins[3].

The copy-and-paste machinery for L1 is a ribonucleoprotein particle (RNP) — a complex that contains the L1 RNA (known as the template) and two proteins, named ORF1p and ORF2p. ORF1p binds to the template and probably aids the assembly or proper folding of the complex, whereas ORF2p acts as both an RT and an endonuclease (an enzyme that cuts DNA)[3]. ORF2p cuts one of the strands of chromosomal DNA in the cell nucleus, and initiates reverse transcription of the template, starting from the cut[3] (Fig. 1a). The L1 RNP preferentially targets a short DNA sequence that is found in all types of genomic region[4,5]. The overall process is known as target-primed reverse transcription (TPRT), and differs in various respects from retroviral reverse transcription (which generally occurs in the cytoplasm)[6].

Phylogenetic studies indicate that the central catalytic domain of ORF2p is distantly related to retroviral RTs. The structures reported by Baldwin *et al.* and Thawani *et al.* now confirm that this central domain adopts the hand-shaped structure typical of the retroviral enzymes, comprising fingers, palm and thumb subdomains (Fig. 2). However, L1 ORF2p also harbours amino- and carboxy-terminus regions that extend the central domain.

At the N-terminus, the RT domain is connected to the endonuclease domain by a flexible linker that Baldwin *et al.* call the tower, which extends and partially covers the fingers subdomain. At the C-terminus, the thumb is prolonged by a 'wrist' that contacts the template bound to the cut end of the chromosomal DNA. ORF2p ends with a C-terminal domain (CTD; also known as a C-terminal segment) that harbours a structural motif called a zinc finger (often found in proteins that bind to DNA or RNA).

The function of the CTD has long been mysterious, but the new structures show that it makes contact with the template at a position distant from the RT active site, and might help to unwind the RNA, preventing it from becoming tangled and obstructing DNA synthesis. The flexibility of the tower allows the endonuclease to rotate relative to the CTD, thus switching the structure from an 'open' to a 'closed' form, in which the template is clamped[2].

Our understanding of TPRT stems mainly from studies of R2, a retrotransposon found in silkworms (*Bombyx mori*) that is related to L1, but is easier to purify and study[7]. The structural details of R2-mediated TPRT have been reported in the past year[8,9]. The two new structures of the L1 machinery therefore enable a comparison of how the key enzymes of L1 and R2 are used for TPRT. This reveals some notable differences.

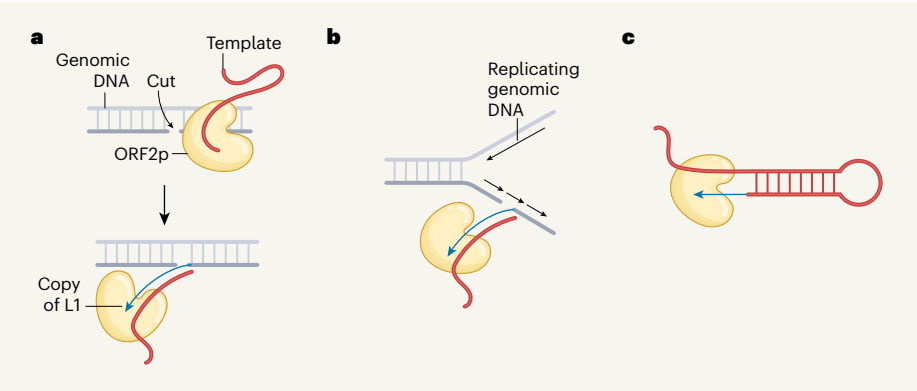

Figure 1 | Enzymatic activities mediated by LINE-1 elements. a, DNA sequences known as LINE-1 (L1) elements, found in mammalian genomes, encode enzymes that enable a copy of the element to be inserted at a new position in the genome. One of these enzymes (ORF2p) forms a complex with the L1 transcript (the template) and binds to a target site in the genomic DNA, cutting one of the DNA strands. It then makes a copy of L1 (blue arrow) starting from the cut, using the sequence encoded in the template (a process known as target-primed reverse transcription). **b**, Thawani *et al.*[1] report that, *in vitro*, the single-strand cleavage is enhanced close to junctions between single- and double-stranded DNA, which are commonly found at sites of DNA replication. **c**, Baldwin *et al.*[2] observe that ORF2p can initiate reverse transcription *in vitro* from 'hairpin' templates that fold back on themselves — possibly explaining how ORF2p synthesizes DNA in the cytoplasm, far from genomic DNA in the nucleus.

First, the enzyme encoded by R2 (ORFp) has an N-terminal DNA-binding domain that guides integration of R2 DNA into one specific genomic site. By contrast, ORF2p inserts L1 DNA into a short DNA motif that is broadly distributed throughout the genome, and lacks the DNA-binding domain found in ORFp[4,5].

Second, ORFp harbours a 'tower-like' domain (NTE-1) that recruits and positions the R2 RNA, ready for reverse transcription, through binding of a structured region found only in that RNA. ORF2p, by comparison, associates with the tail of the RNA template[10]; this tail contains a sequence that is found in many RNA molecules. And third, the endonuclease domains of ORF2p and ORFp are at the N-terminal and C-terminal ends, respectively, of these proteins — a striking difference in 3D organization. Similarly to ORF2p, ORFp has a CTD-like domain with a zinc finger downstream of the RT. However, this zinc finger 'unzips' the target double-stranded DNA, allowing only one of the two strands to access the endonuclease active site, rather than unwinding entangled regions of the template, as it does in ORF2p[1].

Previous studies involving the *in vitro* reconstitution of the L1 RNP with model templates and DNA substrates have also offered mechanistic insights into the activity of ORF2p. The two new studies[1,2] confirm that ORF2p is highly processive[11,12] (capable of copying long RNA templates into DNA in a single step). They also confirm that reverse transcription by ORF2p requires a minimal base-pairing of four to six nucleotides between the DNA substrate and RNA template[11], but can also tolerate some level of mispairing[11,13].

Thawani *et al*. successfully reconstituted TPRT reactions *in vitro* and demonstrated that the initial cleavage of a single strand of the DNA substrate is enhanced when the L1 target-site sequence is close to a junction between single- and double-stranded DNA (Fig. 1b). Such junctions are commonly found at structures called replication forks, which form during replication of the host DNA. This finding corroborates previous proposals that replication forks are preferred sites for retrotransposon integration[4,5], and suggests that cleavage of the second strand is not essential to complete integration.

Baldwin *et al*. found that ORF2p can initiate reverse transcription directly from very short single-stranded DNA or RNA molecules, or from 'hairpin' RNA substrates that fold back on themselves (Fig. 1c). These findings suggest a possible explanation for how L1 DNA is formed in the cytoplasm — something that can lead to the activation of inflammatory pathways during cellular senescence (a condition linked to ageing in which cells cease to proliferate) and in certain inflammatory diseases[14,15]. Consistent with this mechanism, RT activity — but not endonuclease activity — is essential for the accumulation of these cytoplasmic DNA molecules[2], ruling out the possibility that they originate from abortive TPRT products that escaped the nucleus.

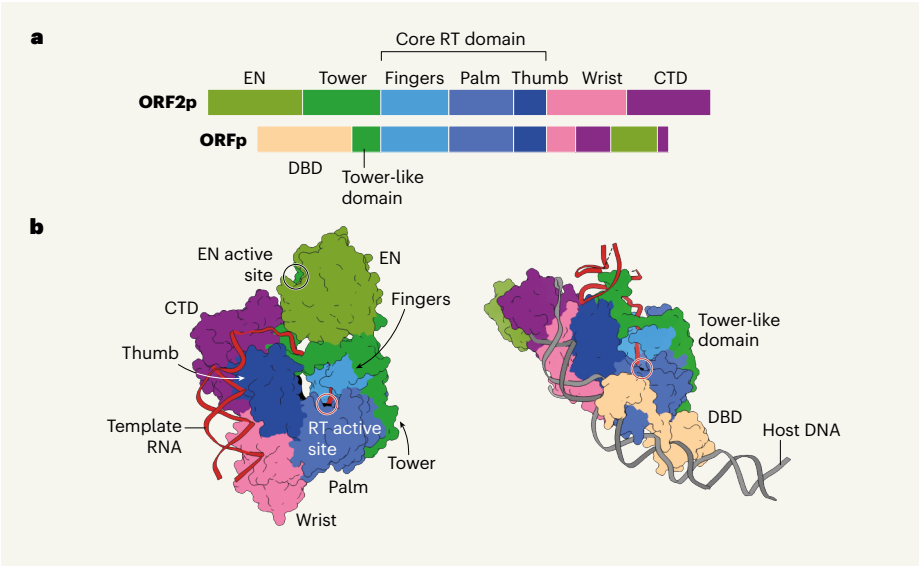

Figure 2 | Comparison of the ORF2p protein with the ORFp protein. a, The ORFp protein, produced by silkworms (*Bombyx mori*), is related to ORF2p. Comparison of the amino-acid sequences of the two proteins shows that they have a similar 'core' (the reverse transcriptase (RT) domain) consisting of subdomains known as the fingers, palm and thumb, next to a wrist domain. ORF2p contains a 'tower' domain[2] that connects the core to the endonuclease domain (EN, used to cleave one strand of target DNA), whereas ORFp has a shorter, tower-like domain. ORF2p lacks the DNA-binding domain (DBD) of ORFp, and its EN domain is found at the amino-terminus of the protein, rather than in the carboxy-terminus domain (CTD). **b**, New structures[1,2] of the human ORF2p protein show that its 3D structure (left) differs substantially from that of ORFp[8,9] (right). Active sites of the EN and RT domains are shown, where visible (the EN active site cannot be seen in this view of ORFp). The 3D structures are drawn from Protein Data Bank accessions 8GH6 (ref. 9) and 8UW3 (ref. 1).

Overall, it seems that ORF2p potentially acts on a variety of DNA substrates beyond those in the conventional TPRT model. It remains to be seen whether replication forks — or other structures containing junctions of duplex DNA with single strands — are the main targets for L1 DNA insertion. *In vitro* experiments that test more DNA substrates, or that incorporate ORF1p and other factors known to intervene in L1 replication[16,17], will help to refine our understanding of this retrotransposon's activity.

"Overall, it seems that the ORF2p protein potentially acts on a variety of DNA substrates."

The discovery of RTs paved the way for methods now used to produce proteins for research and medical applications, including insulin, growth hormone and the hepatitis B vaccine. It also enabled the development of the RT-PCR and RNA-sequencing techniques, which are used to detect RNA-virus infections and to measure gene expression. Expanding the molecular-biology toolbox to include RTs that are highly processive (or exhibit other previously unavailable biochemical properties) might benefit technologies used for genomics — for instance, by enabling the sequencing of full-length RNA molecules[18]. Elucidating the substrates and mechanisms of TPRT might also aid the design of genome-engineering tools[9].

Finally, RT inhibitors were among the first medicines for AIDS, and are still key to therapy regimens that have transformed this devastating disease into a manageable chronic condition. Baldwin *et al*. report that some drugs designed to target HIV also inhibit ORF2p (albeit with moderate binding affinities), and modelled the mode of ORF2p inhibition in light of the structure they had determined. The new ORF2p structures will enable the design of more-specific inhibitors that target both the RT and the endonuclease activities of ORF2p; such inhibitors might be useful for cancer treatment and research into ageing[19,20].

Gael Cristofari is at the University Cote d'Azur, Inserm, CNRS, Institute for Research on Cancer and Aging of Nice (IRCAN), Nice 06107, France.
e-mail: gael.cristofari@univ-cotedazur.fr

1. Thawani, A., Florez Ariza, A. J., Nogales, E. & Collins, K. *Nature* **626**, 186–193 (2024).
2. Baldwin, E. T. *et al*. *Nature* **626**, 194–206 (2024).
3. Kazazian, H. H. & Moran, J. V. *N. Engl. J. Med*. **377**, 361–370 (2017).
4. Sultana, T. *et al*. *Mol. Cell* **74**, 555–570 (2019).
5. Flasch, D. A. *et al*. *Cell* **177**, 837–851 (2019).
6. Sultana, T., Zamborlini, A., Cristofari, G. & Lesage, P.

Nature Rev. Genet. **18**, 292–308 (2017).
7. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. *Cell* **72**, 595–605 (1993).
8. Deng, P. *et al. Cell* **186**, 2865–2879 (2023).
9. Wilkinson, M. E., Frangieh, C. J., Macrae, R. K. & Zhang, F. *Science* **380**, 301–308 (2023).
10. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. *Mol. Cell* **60**, 728–741 (2015).
11. Monot, C. *et al. PLoS Genet.* **9**, e1003499 (2013).
12. Piskareva, O. & Schmatchenko, V. *FEBS Lett.* **580**, 661–668 (2006).
13. Kulpa, D. A. & Moran, J. V. *Nature Struct. Mol. Biol.* **13**, 655–660 (2006).
14. De Cecco, M. *et al. Nature* **566**, 73–78 (2019).
15. Thomas, C. A. *et al. Cell Stem Cell* **21**, 319–331 (2017).
16. Taylor, M. S. *et al. Cell* **155**, 1034–1048 (2013).
17. Benitez-Guijarro, M. *et al. EMBO J.* **37**, e98506 (2018).
18. Zhao, C., Liu, F. & Pyle, A. M. *RNA* **24**, 183–195 (2018).
19. Brochard, T. *et al. Ageing Res. Rev.* **92**, 102132 (2023).
20. Rajurkar, M. *et al. Cancer Discov.* **12**, 1462–1481 (2022).

The author declares competing interests; see go.nature.com/3sfap2x for details.
This article was published online on 15 January 2024.

Public health

Contact-tracing app predicts transmission risk

Justus Benzler

The risk of catching COVID-19 as calculated by a smartphone app scales with the probability of subsequently testing positive for the coronavirus SARS-CoV-2, showing that digital contact tracing is a useful tool for fighting future pandemics. **See p.145**

The COVID-19 pandemic was the first global outbreak in human history to unfold in a world where many countries had more than 70% smartphone coverage[1]. This generated many initiatives for smartphone applications that could complement or replace established measures to control the spread of infection in pandemics. For instance, apps were developed that could deliver test results, provide proof of vaccination[2] or trace the recent contacts of an infected person. On page 145, Ferretti *et al.*[3] report that data from a smartphone app used for contact tracing can provide valuable epidemiological information.

Contact tracing is a well-established process that public-health authorities follow during outbreaks of diseases that are transmitted directly between humans. The aim is to find people who were in contact with infected individuals, so that those with potential exposures can receive recommendations or interventions, such as quarantine, to prevent further disease transmission. Manual contact tracing is particularly resource-intensive and is not easily scalable[4]. Hence, in a pandemic, it quickly reaches its limits. Digital contact tracing is an alternative solution that relies on data gathered by personal mobile devices. However, it can pose a particular threat to privacy, because it involves collecting sensitive information about an individual's health status and relationships[5].

Various approaches to digital contact tracing were debated at the onset of the COVID-19 pandemic. Ultimately, the public-health authorities in many countries chose to base their contact-tracing apps on an integrated feature of smartphone operating systems provided by Google and Apple, known as the Exposure Notification framework[6]. This feature relies on Bluetooth signals that are exchanged between participating smartphones when they are physically close to each other. The signals transmit unique, randomly generated codes that are then temporarily stored locally on the other phone.

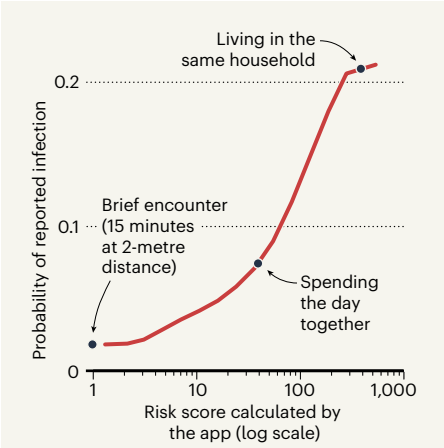

Figure 1 | A smartphone app used for contact tracing during the COVID-19 pandemic can predict the probability of SARS-CoV-2 transmission. Using anonymized data from the NHS COVID-19 app for England and Wales, Ferretti *et al.*[3] show that a risk score for infection – calculated in the app on the basis of the amount of time spent with and proximity to an infected person, and how infectious that person is – scales with the probability of an infection subsequently being reported. A high risk score might result from living in the same household as an infectious person. A brief encounter at a distance of 2 metres for 15 minutes, the threshold for a 'relevant' contact defined in manual contact tracing during the COVID-19 pandemic in most countries, results in a low risk score. In future pandemics, harnessing data from contact-tracing apps might help public-health authorities to understand how infections spread.

If a smartphone user tests positive for SARS-CoV-2, they can opt to upload a set of codes to a server. These codes are renewed daily, and are used to generate the codes that are sent over Bluetooth. The smartphones of other users can then compare the codes on the server to the ones that they had received and stored locally. In the case of a match, the app notifies the second user about past encounters with a potentially infectious person. The threshold for a notification is based on a risk score calculated by the app from the estimated proximity and duration of these exposures, and the infectiousness level of the notifier. This is estimated from the date of the encounter in relation to the notifier's test and the onset of their symptoms.

There is ongoing controversy[7] over the extent to which digital contact tracing – and other non-pharmaceutical interventions – actually contributed to slowing the spread of infections. The aim was to prevent health-care systems being overwhelmed, and to buy time for the development, production and delivery of vaccines. There was little information collected that could address this controversy, mainly because of privacy concerns and the decision to use a decentralized architecture for digital contact-tracing systems, which meant that contact data collected by the apps were not stored in centralized databases. Furthermore, in many countries it was not clear how widely the apps were adopted, because public-health authorities often used downloads as a (poor) proxy for an app's use (see go.nature.com/42q6axc). For the same reasons, the systems also scarcely reached their potential for monitoring key epidemiological indicators for the spread of COVID-19 in the population.

A notable exception is the NHS COVID-19 app rolled out by the National Health Service in England and Wales, which was pioneered by a strong partnership between app developers and academic institutions. Early in the pandemic, the team behind the app created a tool[8] to model the impact of non-pharmaceutical interventions, including digital contact tracing, and published empirical evidence showing that the app helped to prevent COVID-19 cases and COVID-related hospitalizations and deaths[9]. In the current study[3], researchers in the same team analysed data recorded by the app to answer a fundamental question that arose during the COVID-19 pandemic: how is the probability of SARS-CoV-2 transmission from one individual to another related to the proximity and duration of the exposure (Fig. 1)?

Ferretti and colleagues analysed 'packets'