COMMENTARY

HERV-K HML-2 diversity among humans

Jack Lenz[a,1]

Retroviruses comprise over 8% of the human genome (1, 2). Human endogenous retroviruses (HERVs) exist as DNA remnants of infections that occurred in germ lineage cells of our ancestors. Most of this viral DNA is mutated, often including various large disruptions, but some components are intact or otherwise functional. What viral components exist in human genomes, and what gene products do they encode that might interact with the nonviral parts of us? When did they arrive in the genomes of our ancestors, and are they still active today? Does an intact, infectious, retroviral provirus lurk in the genomes of some of us? Wildschutte et al. (3) shed new light on these issues by characterizing the most recently acquired proviruses in human genomes, a subset of the virus HERV-K called HML-2 (for human mouse mammary tumor virus like-2), which are present at various allele frequencies <1 in the human population, i.e., not in everyone.

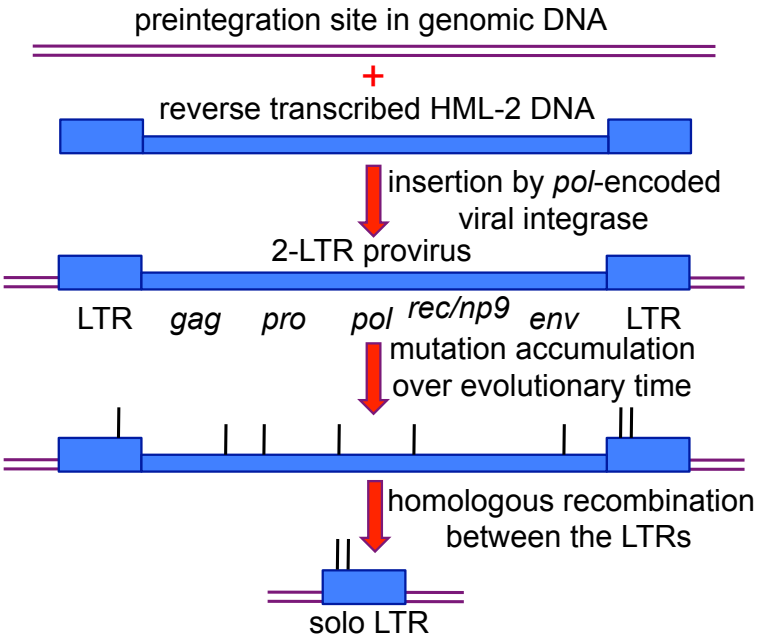

Fig. 1. Successive states of an endogenous retrovirus DNA. The small vertical lines represent mutations of unspecified nature.

Retrovirus replication involves reverse transcription of the RNA genome from viral particles into DNA, which then integrates into host cell DNA. The viral DNA contains two long terminal repeats (LTRs) (Fig. 1). Full-length integrated DNAs, called proviruses, are permanently associated with the infected cell and its descendants unless stochastic mutational events delete them. Endogenous retroviruses are adapted to infect germ lineage cells, and their inserted DNAs become parts of the genome of the infected species, and are subject to selection processes and genetic drift over evolutionary time. In the absence of selective pressure on the host to maintain the viral DNAs or their components in intact condition, these elements inevitably accumulate mutations over evolutionary time that cause functional decay, including the very common event of homologous recombination between the two LTRs that generates solo LTRs (Fig. 1). Each viral DNA insertion also has unique mutations that occurred either during reverse transcription or, more commonly, after the viral DNA became part of the host genome, and these have played a key role in inactivating the infectivity of HML-2 elements (4, 5). The 8% of the human genome that is endogenous retrovirus DNA (also called LTR elements) represents hundreds of thousands of individual insertions of a multitude of different retroviruses over evolutionary time. HML-2 is the most recently active type to infect the germ line of the human lineage and is the subject of the study by Wildschutte et al. Each HML-2 insertion can be defined by its position within the human genome. The human reference genome contains over 120 HML-2 insertions that are not present in chimpanzees, bonobos, or gorillas, indicating that the virus was active until at least fairly recently in human evolution.

Wildschutte et al. computationally analyzed short-length, genomic DNA, sequence reads obtained in the 1000 Genomes Project (that actually encompassed closer to 2,500 genomes) and the Human Genome Diversity Project, and characterized 36 HML-2 insertions in addition to those previously identified

[a]Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461
Author contributions: J.L. wrote the paper.
The author declares no conflict of interest.
See companion article on page E2326.
[1]Email: jack.lenz@einstein.yu.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1603569113

in the human reference genome. Almost all of these were confirmed by PCR amplification and Sanger sequencing of the complete elements. Most were solo LTRs. Five were 2-LTR viruses. They were present at various allele frequencies ranging from 0.75 to as low as one found in only a single individual. Coupled with the previous analysis of reference genome HML-2 insertions from the same investigators (6), Wildschutte et al. provide the authoritative source for the comprehensive identification of the individual HML-2 elements in the human population. Future analyses of sequenced human genomes will likely identify more insertions that exist as low-frequency alleles.

How might these recent insertions matter? Individual endogenous retrovirus DNAs can have significant consequences for their host species. It is probably wisest to think of endogenous retroviruses foremost as genome invaders that, like any parasites, exploit the host for their own propagation and survival. Indeed, it was proposed that the evolution of many key, unique features of eukaryotic gene expression and other processes were, at least initially, due to selection of defensive responses to resist invasive nucleic acids and other parasites (7). Despite many defenses that have evolved, endogenous retroviruses together with the nonviral retrotransposable elements, long interspersed nuclear elements and short interspersed nuclear elements, comprise nearly one-half of the human genome (2), and thus have been indisputably successful. Although integration itself is inescapably mutagenic, successful invaders may have been selected for having limited pathogenic effects on their hosts, such as having weak transcriptional elements and being subject to epigenetic silencing. Once a novel DNA is inserted into a host genome, it can provide functional components that may evolve to have advantageous consequences that may at least reduce the detrimental effects on host fitness (8). Important examples include viral infection resistance factors, trophoblast syncytialization factors, and regulatory elements including transcriptional regulatory networks (9–16). A popular type of study with HERVs is to try to correlate expression with pathogenic states. How humans have coped with the acquired HML-2 elements is an area worth more study. One notion that is strongly reinforced by Wildschutte et al. is that single-nucleotide resolution is essential in such work for distinguishing individual HML-2 elements, some of which are >99% identical in pairwise comparisons.

HML-2 viruses infected the human lineage throughout much of the period of hominid evolution starting in a common ancestor of humans and orangutans over 13 million years ago, and continued to do so after the divergences of the gorilla and bonobo/chimpanzee lineages. By counting the number of differences between the two LTRs and applying an estimate of mutation rate over time, Wildschutte et al. found that the new 2-LTR insertions formed about 0.67–1.8 My ago. Thus, HML-2 infections continued until approximately the time that Neanderthals and Denisovans emerged, archaic hominins and sister taxa that recent studies suggest diverged from the lineage leading to modern humans roughly 650,000 y ago (17, 18). The results from Wildschutte et al. (3) and others (19) show that many, but not all, nonreference genome HML-2 insertions that were originally identified in the archaic hominins (20, 21), are also present in the modern human population today, several at low allele frequencies. One mechanism for this might be incomplete lineage sorting, i.e., a failure of one of the two alleles to win out and become fixed as the sole allele in a population due to genetic drift or selection acting over 650,000 y of evolutionary time. This was previously suggested as a possibility for these insertions (20) and was also invoked to explain the presence of an HML-2 provirus in humans and gorillas but not chimpanzees or bonobos (22). A second possible mechanism to explain the shared insertions is introgression, i.e., more recent interbreeding among the hominin lineages. Wildschutte et al. make a sound case for incomplete lineage sorting being the likely mechanism based on most insertions predating the time of the lineage divergence and their presence predominantly in individuals of African ancestry among the thousands of genomes sampled, populations lacking evidence for introgression. One of the recent interesting findings of the 1000 Genomes Project was that infrequent alleles in modern humans tended to be of recent origin (23). It will be interesting to see how consistent HML-2 insertions are with this paradigm.

The results from Wildschutte et al. and others show that many, but not all, nonreference genome HML-2 insertions that were originally identified in the archaic hominins, are also present in the modern human population today, several at low allele frequencies.

Eight insertions originally detected in the archaic hominins were not detected in any of the thousands of modern human genomes sequenced. These may eventually be found as infrequent alleles once enough human genomes are sequenced, or they may represent insertions that occurred in the Neanderthal and/or Denisovan lineages after divergence of the modern human lineage. It is trickier to tell if HML-2 insertions occurred in the modern human lineage after separation from the archaic hominins, because the latter have not been sequenced to anything remotely approaching the number of individuals needed to determine whether those insertions are present at low frequency in them, as Wildschutte et al. did with the multitude of modern human genomes.

Does an infectious endogenous retrovirus reside in some genomes within the human population? Wildschutte et al. discovered a candidate, a low allele frequency, HML-2 provirus on the X chromosome that has full-length ORFs for all viral proteins and no obviously lethal mutations, i.e., no premature stop codons, frameshifts, or substitutions in conserved functional elements. Experiments to test its infectivity are undoubtedly underway. Until direct evidence emerges, caution and perhaps tentative relief should reign, as even a subtle single amino acid substitution can inactivate an HML-2 provirus (24). Because Wildschutte et al. just discovered this provirus, it is also possible that sequencing of more human genomes will lead to the discovery of more such viral DNAs, albeit at low allele frequency. Also, it must be kept in mind that the components to assemble an infectious HML-2 provirus by just two recombination events (4) exist in the genomes of a substantial fraction of humans. The conclusion that emerges from Wildschutte et al. that HML-2 was not particularly active at reinfecting the genome of the human lineage during the last quarter- to half-million years or so suggests that such events might no longer occur, or that they are strongly selected against if they do. However, somehow HML-2 was active in the human lineage for 13 million years, and it may still possess surprising abilities.

1 Jern P, Coffin JM (2008) Effects of retroviruses on host genome function. *Annu Rev Genet* 42:709–732.
2 Xing J, Witherspoon DJ, Jorde LB (2013) Mobile element biology: New possibilities with high-throughput sequencing. *Trends Genet* 29(5):280–289.
3 Wildschutte JH, et al. (2016) Discovery of unfixed endogenous retrovirus insertions in diverse human populations. *Proc Natl Acad Sci USA* 113:E2326–E2334.
4 Dewannieux M, et al. (2006) Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res* 16(12):1548–1556.
5 Lee YN, Bieniasz PD (2007) Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathog* 3(1):e10.
6 Subramanian RP, Wildschutte JH, Russo C, Coffin JM (2011) Identification, characterization, and comparative genomic distribution of the HERV-K (HML-2) group of human endogenous retroviruses. *Retrovirology* 8:90.
7 Madhani HD (2013) The frustrated gene: Origins of eukaryotic gene expression. *Cell* 155(4):744–749.
8 Mager DL, Stoye JP (2015) Mammalian endogenous retroviruses. *Microbiol Spectr* 3(1):MDNA3-0009-2014.
9 Best S, Le Tissier P, Towers G, Stoye JP (1996) Positional cloning of the mouse retrovirus restriction gene Fv1. *Nature* 382(6594):826–829.
10 Blaise S, de Parseval N, Bénit L, Heidmann T (2003) Genomewide screening for fusogenic human endogenous retrovirus envelopes identifies syncytin 2, a gene conserved on primate evolution. *Proc Natl Acad Sci USA* 100(22):13013–13018.
11 Mi S, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* 403(6771):785–789.
12 Rebollo R, Romanish MT, Mager DL (2012) Transposable elements: An abundant and natural source of regulatory sequences for host genes. *Annu Rev Genet* 46:21–42.
13 Suntsova M, et al. (2015) Molecular functions of human endogenous retroviruses in health and disease. *Cell Mol Life Sci* 72(19):3653–3675.
14 Chuong EB, Elde NC, Feschotte C (2016) Regulatory evolution of innate immunity through co-option of endogenous retroviruses. *Science* 351(6277):1083–1087.
15 Chuong EB, Rumi MA, Soares MJ, Baker JC (2013) Endogenous retroviruses function as species-specific enhancer elements in the placenta. *Nat Genet* 45(3): 325–329.
16 Lynch VJ, et al. (2015) Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep* 10(4):551–561.
17 Meyer M, et al. (2016) Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* 531(7595):504–507.
18 Stringer CB, Barnes I (2015) Deciphering the Denisovans. *Proc Natl Acad Sci USA* 112(51):15542–15543.
19 Marchi E, Kanapin A, Magiorkinis G, Belshaw R (2014) Unfixed endogenous retroviral insertions in the human population. *J Virol* 88(17):9529–9537.
20 Agoni L, Golden A, Guha C, Lenz J (2012) Neandertal and Denisovan retroviruses. *Curr Biol* 22(11):R437–R438.
21 Lee A, et al. (2014) Novel Denisovan and Neanderthal retroviruses. *J Virol* 88(21):12907–12909.
22 Barbulescu M, et al. (2001) A HERV-K provirus in chimpanzees, bonobos and gorillas, but not humans. *Curr Biol* 11(10):779–783.
23 Auton A, et al.; 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.
24 Heslin DJ, et al. (2009) A single amino acid substitution in a segment of the CA protein within Gag that has similarity to human immunodeficiency virus type 1 blocks infectivity of a human endogenous retrovirus K provirus in the human genome. *J Virol* 83(2):1105–1114.