Papers Shown in the Animation:

About Placental ERV:

- https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.3000028
- https://www.pnas.org/content/112/5/E487
- https://pubmed.ncbi.nlm.nih.gov/19474286/

Revival of Extinct ERV in Human DNA:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1665638/

Chimpanzee & human DNA comparisons:

- https://personal.broadinstitute.org/sfs/personal/Science-1982-Yunis-1525-30.pdf
- https://pubmed.ncbi.nlm.nih.gov/16136131/

Studies on how likely it is for an ERV to insert itself in same location of different hosts:

- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933515/
- https://pubmed.ncbi.nlm.nih.gov/7926746/
- https://www.semanticscholar.org/paper/Retroviral-DNA-Integration%3A-ASLV%2C-HIV%2C-and-MLV-Show-Mitchell-Beitzel/f2c061e75d8ee0ddcf34edf93e9c986cbe854aba

HERV-W, exact insertion locations in humans and other apes

- https://www.researchgate.net/publication/322608448_HERV-W_group_evolutionary_history_in_non-human_primates_Characterization_of_ERV-W_orthologs_in_Catarrhini_and_related_ERV_groups_in_Platyrrhini

Estimation of possible retrovirus insertion sites in the human Genome

When we see two species with the same ERV insertions in the same locations of their genomes, what does this mean?

1. Maybe it's a coincidence.
2. Maybe ERVs aren't really virus insertions, but are designed or evolved sequences of DNA that simply happen to look like virus insertions.
3. Maybe it is evidence of common ancestry. The shared insertions were inserted before the birth of the most recent common ancestor shared by the two species.
4. Maybe the viruses in question targeted specific DNA sequences during insertion. This would mean they always (or almost always) insert themselves into the same spots of different host's genomes.

The huge number of identical insertions in humans and chimps means luck is out of the question. Option number 2 has been long ruled out by researchers for many reasons. This leaves us with options 3 and 4.

To see which of the two remaining options might be true, I looked at the literature to see if retroviruses insert randomly (anywhere in the 3 billion nucleotides of the human genome) or if they are selective about where they insert. If they are selective, how specifically selective are they?

Luckily, these questions have been heavily studied because they have a tremendous potential influence on virus treatment and prevention. I found 3 papers to be most helpful:

One which looked at a broad number of retrovirus species in birds.
https://pubmed.ncbi.nlm.nih.gov/7926746/

One which looked in depth at 3 species of retrovirus (a human, a bird, and a rodent retrovirus).
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC509299/?tool=pubmed

One which did a deep dive into our most hated retrovirus, HIV.
https://genome.cshlp.org/content/17/8/1186.long

In summary, retroviruses don't have specific target sequences, but also don't insert completely randomly. The enzyme they use to insert their genes into a host's genome doesn't bind to a specific sequence in the host's DNA, but it does interact with specific host proteins bound to DNA, and seems to interact with specific 3-dimensional structures in a cell's folded genome. In short, only a certain percentage of our 3 billion nucleotides are available for the virus to attack. The question is, how big is that percentage?

The best way to answer this question is to use the HIV paper. This is because it had the largest sample size (by far), and HIV seems to be a good model for retroviruses in general. Like all the others, it's roughly specific about where it inserts, but it's not sequence specific.

The study looked at 165,572 Genomes of Jurkat (human) T cells that had been incubated with HIV. 40,610 had HIV integration sites that were found and deemed usable for the study. 40,569 insertions were unique, 41 were duplicates (identical, independent integrations).

Math not being my strong suit, I needed help using those numbers to estimate the total possible insertion sites in the human genome. This number was not directly provided in the paper. Luckily, I was helped by Dr. John Coffin (genetics), Dr. Brian T. Luke (biomedical computation), and PhD student, Alice Zhang (applied physics).

Together they came up with the following equation which is modified from the birthday paradox. A variation is also used in Mark-Recapture experiments.

$$\frac{1}{n}\binom{40610}{2} \approx 41$$

Dr. Coffin realized that, to err on the side of caution, we should double that 41 number. This is because of the nature of PCR work done in the HIV paper. PCR is expected to hide up to half of the repeat insertions.

$$\frac{1}{n}\binom{40610}{2} \approx 82$$

When you solve for *n*, you get an estimate of about **10,000,000 potential insertion spots** in the human genome.

1 in 10 million (or 0.0000001) is a super rough estimate but where it errs, it errs on the side of being too high of a chance (it errs in favor of the “fixed-species” hypothesis). This is because in the HIV experiment, most cells only had one infection, meaning that only the hottest of hot spots were reported on in the experiment. This means that in reality, there are likely far more insertion spots than just 10 million. 10 million simply represents the easiest spots for a virus to insert itself.

Erring on the side of the Fixed-Species-View is important here, because I want to be as generous as possible to that idea.

Probability of 2 individuals getting 12 of 12 independent insertions in the same spots of their genomes

To get this number, I simply used math for combining the probability of independent events. I take the probability of the event happening once, and multiply it by itself 12 times (raise to the power of 12).

$$\frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot$$
$$\frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000} \cdot \frac{1}{10{,}000{,}000}$$

In scientific notation, the answer is 1 in 1x10^84

To put that in perspective, here's that number written out fully and compared to the number of atoms roughly estimated to exist in the observable universe (the Eddington number):

Arthur Stanley Eddington's estimate of atoms in the observable universe

15,700,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

1,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000,000

Chance of 2 species getting 12 out of 12 independent virus insertions in identical locations

Calculating the odds of humans and chimps sharing the number of HERV-W insertions that we know they share

In this paper, researchers looked for members of the ERV group called HERV-W. They found 211 in humans. 205 of those were found in chimps in identical locations. 3 more were found in chimps but not humans. This gives us 214 ERVs all together, 205 shared, 9 not shared (misses).

To calculate the odds of 205 hits with 9 misses, we can't simply multiply as we did when figuring out the probability of independent events. This would not properly account for the 9 misses. Instead, we have to plug the following into a binomial distribution calculator.

Probability x > 204
n= 214
p= 0.00000010101010101010101010101

Note: in most calculators, you can't simply put in p=0.0000001. If you do it will won't give you back an accurate response (most round to zero). Adding 0101ect at the end was a hack we found that forces the calculator to actually do the full calculation. I know this makes the results slightly wrong, but again, it errs in favor of the Fixed-Species view. We can afford to be generous. That view clearly loses anyway.

Results: 1.7x10^-1419

To convert that to the more comfortable 1 in x, that we've been using earlier, we must divide 1 by our result to solve for x. This gives us 5.88x10^1418 or

1 in
58,800,000

In short, it's impossible that this happened by chance.

Even when we account for the fact that retroviruses don't insert themselves completely randomly into a host's genome, this cannot explain why there are so many matches between humans and chimps.

The matching ERVs found in humans and chimps are yet another line of rock solid evidence for common ancestry.