Past and present: How has infection shaped the human genome?

Humans have long been fascinated by the mechanics of infection. The Ancient Greek physician Hippocrates, father of modern medicine, developed the miasma theory of disease, positing that a nebulous "bad air" could provoke symptoms of illness.[1] From humorism to the eventual advent of germ theory, societies through time have always attempted to understand how infection occurs and what exactly its effect is on the human body. Alongside this, we have sought to understand our origins: where and what did we come from? In the last half-century, science has arrived at a conclusion which would no doubt have astonished both Hippocrates and Darwin - that not only is most disease directly caused by organisms which invade the body, but that these have the ability to alter our genetic material.

Pathogens of all types have had a significant effect on human genetics over time. Currently, the body of published research on viral DNA in the human genome outweighs that of other infectious agents, but the role of disease in selecting for certain alleles should not be understated. Bacterial influence on the genome, too, offers a fertile area for exploration.

Viruses are often described as "hijacking" cells or their DNA, like a gangster performing a hold-up. This image of viruses as wantonly malevolent, while engaging, is of course a simplistic and rather anthropomorphic presentation. In reality, viral infection is an ingenious and complex process, which has shaped the human genome throughout our history. Viruses are able to directly insert DNA into their host, which is known as a form of horizontal gene transfer (HGT) as the genes move "laterally" from one species to another, rather than being passed down "vertically" through a reproductive lineage. HGT therefore appears to conflict with classical Darwinian inheritance, and the idea of non-vertical heredity has become a polarising debate in certain scientific circles.

The effect of viral HGT can only be fully appreciated through an understanding of its mechanics. The most recent estimates suggest that around 8% of the human genome consists of retroviral DNA.[2] Retroviruses are viruses which use single-stranded RNA as their genetic material and which replicate by transcribing their viral RNA into DNA. While other RNA viruses have their RNA directly transcribed by the host cell, retroviruses use a reverse transcriptase enzyme to "reverse" the process of transcription, producing a complementary strand of DNA. The reverse transcriptase then copies the strand, forming a double-stranded DNA molecule. The enzyme integrase, as the name suggests, then integrates this viral DNA into the host genome, forming a provirus and allowing the host's mechanisms of transcription and translation to propagate it in order to form new virions. During this process, therefore, the retrovirus inserts an essentially random portion of its DNA into the host genome. If this provirus is introduced into a cell in the germ line, then the viral DNA is passed down as if it were a host gene, and hence enters the lineage of the species as an endogenous retrovirus (ERV).[3]

Since the first human endogenous retrovirus (HERV) was identified from human brain DNA in 1981,[4] 22 HERV families have been identified: though endogenous retroviruses were originally dismissed as "impossible",[5] HERV research is now well-established. A complete HERV consists of four genes which code for various proteins. The *gag* gene encodes the matrix, capsid and nucleocapsid proteins, which are structural elements of the virion. The *pro* gene encodes protease, the *pol* gene reverse transcriptase and integrase, and *env* encodes the envelope protein which protects the virus. These are flanked by a long terminal repeat (LTR) at each end.[6] The majority of

HERVs as they appear in the human genome are incomplete fragments of this structure (usually solo LTRs) and largely defective due to premature stop codons, mutations or deletions.[7] Although these were interpreted as meaningless "junk DNA" during initial sequencing of the human genome,[7] the fact that they have not been lost throughout millions of years of evolution is the first indication that they are more significant than originally thought.

Most HERVs are ancient - the oldest HERVs were probably infectious between 60 and 70 million years ago, at the beginning of primate development.[8] However, Aashish Jha and his team used population genomics to show that some of the youngest members of the HERV-K family (the youngest and most active of the HERV families) entered the genome during crucial points in the evolution of modern humans: HERV-K113 entered during or just before the divergence of Neandarthals and Denisovans from their common ancestor.[9] Jha went on to lead further research which provided evidence that HERV-K106 actually entered the genome of *Homo sapiens sapiens*, appearing around 150,000 years ago – almost 100,000 years after our species' theorised arrival in Europe.[10] Thus, HERVs have been consistently involved in primate and human evolution.

In fact, some scientists theorise that viruses have been involved, not only in human evolution, but in the evolution of all eukaryotes. This hypothesis assumes that viruses preceded eukaryotes, proposing that the eukaryotic nucleus arose when an Archaean virus infected a cell, and that this relationship eventually resulted in the virus becoming an intracellular parasite. There are two main branches of this theory: Masaharu Takemura of Tokyo University suggests that the nucleus evolved when the archaeon defended its genome against the virus, while biotechnological researcher Philip Bell champions the idea that the virus was subsumed into the archaeon itself, usurping and destroying its genome entirely. According to this theory, therefore, infection *created* the human genome – and in Bell's words "at the heart of every human cell is a virus".[11]

The relationship between HERVs and their human host is sometimes a symbiotic or mutualistic one; there is strong evidence that the influence of HERVs provided an evolutionary advantage to our mammalian ancestors. The most notable example of HERV mutualism is the case of syncytin-1 and 2. The *env* protein of the HERV-W family of retroviruses, which is thought to have entered the genome of Old World monkeys over 23 million years ago, is homologous to syncytin-1, while the HERV-FRD-*env* gene codes for syncytin-2. These are glycoproteins which assist the fusion of cytotrophoblast cells in the human placenta to form the syncytiotrophoblast, the epithelial covering of the placental villi. Without this function, the placenta would be unable to invade the lining of the uterus in order to provide nutrients to the foetus. The syncytiotrophoblast also plays a role in preventing foetus rejection – an immune function which is still not fully understood, so the involvement of HERVs is highly significant. HERVs also have high levels of expression in brain tissue, where they may serve a neuroprotective function. Therefore, the very processes which allow the propagation of our species and its genome could not exist without viral DNA.[6]

Viruses, however, are far from the only microorganisms to have affected the human genome. Even when they do not actually insert DNA into their host, pathogens are able to alter human genetics over time: Dominic Kwiatkowski describes malaria, caused by the protist *Plasmodium falciparum*, as "the strongest known force for evolutionary selection in the recent history of the human genome", which is best demonstrated by its effect on the HbS allele. HbS is a variant of the HBB gene, which encodes ß-globin, a component of haemoglobin. Individuals who are homozygous for HbS suffer from sickle-cell disease, as haemoglobin S tends to polymerise at low oxygen concentrations, which causes the red blood cells carrying it to deform. However, HbS heterozygotes

have an almost ten-fold higher protection against severe malaria. Crucially, despite these often-fatal consequences for homozygotes, the allele is at a frequency of almost one in five in some regions of Africa; these regions correspond broadly to those with the highest levels of malaria endemicity (Fig. 1). Other alleles which confer resistance to malaria also have deleterious effects, but continue to be selected for due to the extraordinarily high selective pressure exerted by the disease.[13]
Additionally, research conducted by a team from the Pasteur Institute used data from over 1,000 European genomes over 10,000 years to show the effect of the strong negative selection pressure of tuberculosis (*Myobacterium tuberculosis*) on the prevalence of the P1104A variant of the TYK2 gene over time. Homozygotes for the allele are at a higher risk of developing TB. Gaspard Kerner and his team revealed that, though it had been at a frequency of almost 10% 3,000 years ago, P1104A dropped drastically around 0BCE (around which time the modern tuberculosis bacterium became widespread).[13] Throughout history, infection has slowly but surely shaped our genetic make-up.

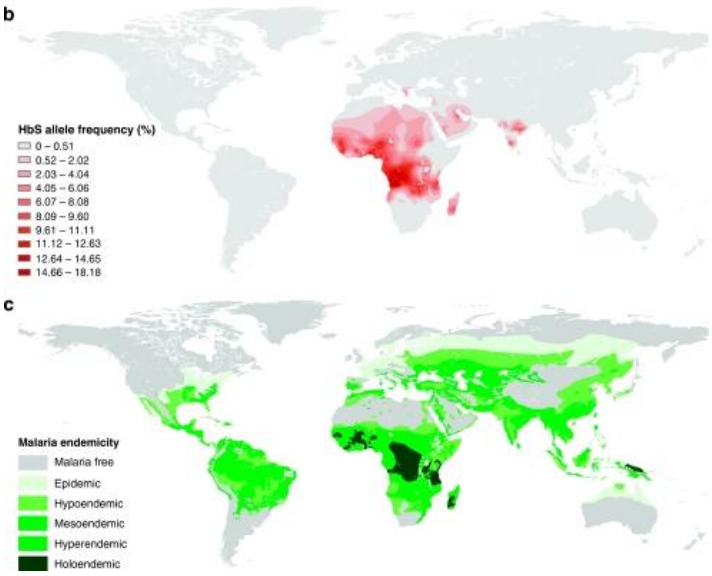

Figure 1: HbS allele distribution compared with malaria endemicity [12]

The widely accepted endosymbiotic theory of eukaryotic organelles, developed by Lynn Margulis in 1967, proposes that mitochondria and plastids evolved from ancient bacteria which infected an archaeon and eventually became dependent upon it.[15] This theory therefore implies that mitochondrial DNA, a vital element of the human genome, is carried by the remnant of an ancient bacterium – once again, infection is integral to our genetic history. The question of actual horizontal gene transfer between bacteria and humans, however, is still controversial.

Although the International Human Genome Sequencing Consortium initially suggested in 2002 that 223 human genes originated from bacterial DNA via horizontal transfer, subsequent research by biostatistician Steven Salzberg and his team countered that this was a hasty impression due to faults in scientific procedure. Salzberg et al. cut the number of genes by over 80% and predicted that it would fall further, probably to zero, with more information. This error was widely publicised and has set the tone for scientific discussions on the topic since.[7] Regardless, research into the extent of HGT between bacteria and human body cells is ongoing; a 2013 study led by Julie Hotopp found evidence that such integrations had indeed taken place, especially in cancerous human cells. This could either indicate that cancerous cells are more susceptible to HGT from bacteria or that it is the bacterial intrusions that somehow trigger the over-proliferation of cancer cells, which would of course be a milestone discovery for the development of new cancer therapy. However, there is still scepticism among many scientists: Hotopp describes how the Consortium's erroneous identification of the 223 genes, and its refutation, had "a chilling effect on the field".[16][17] HERVs have been studied for over 40 years, but it may take more time for the same acceptance to be gained for the idea of bacterium-to-human HGT; in the meantime, this fascinating area of study will continue to develop our understanding of the extent of microbial input into the human genome.

Nobel Laureate Sir Peter Medawar once memorably described a virus as “a piece of bad news wrapped in a protein coat”.[18] Certainly, infection is broadly seen as a “negative” event for the host. However, as I have detailed here, a wealth of research has shown the nuance in its relationship to the human body; infective heredity both defies the principles of Darwinian evolution and helps us to better understand it. Future research into this field will no doubt further develop this idea - that, in researching our past and understanding our present, infection is, occasionally, good news.

Annabelle Sanouillet

References:

1. Kannadan, Ajesh (2018) "History of the Miasma Theory of Disease," ESSAI: Vol. 16, Article 18. Available at: https://dc.cod.edu/essai/vol16/iss1/18
2. Mao, J., Zhang, Q., & Cong, Y. (2021). Human endogenous retroviruses in development and disease. *Computational And Structural Biotechnology Journal*, *19*, 5978-5986. https://doi.org/10.1016/j.csbj.2021.10.037
3. Bannert, N., & Kurth, R. (2006). The Evolutionary Dynamics of Human Endogenous Retroviral Families. *Annual Review Of Genomics And Human Genetics*, *7*(1), 149-173. https://doi.org/10.1146/annurev.genom.7.080505.115700
4. Martin, M. A., Bryan, T., Rasheed, S., & Khan, A. S. (1981). Identification and cloning of endogenous retroviral sequences present in human DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *78*(8), 4892–4896. https://doi.org/10.1073/pnas.78.8.4892
5. Weiss R. A. (2006). The discovery of endogenous retroviruses. *Retrovirology*, *3*, 67. https://doi.org/10.1186/1742-4690-3-67
6. Luganini, A., & Gribaudo, G. (2020). Retroviruses of the Human Virobiota: The Recycling of Viral Genes and the Resulting Advantages for Human Hosts During Evolution. *Frontiers In Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.01140
7. Quammen, D. (2018). *The tangled tree: A radical new history of life*. Harper Collins.
8. Tristem, M. (2000). Identification and Characterization of Novel Human Endogenous Retrovirus Families by Phylogenetic Screening of the Human Genome Mapping Project Database. *Journal Of Virology*, *74*(8), 3715-3730. https://doi.org/10.1128/jvi.74.8.3715-3730.2000
9. Jha, A. R., Pillai, S. K., York, V. A., Sharp, E. R., Storm, E. C., Wachter, D. J., Martin, J. N., Deeks, S. G., Rosenberg, M. G., Nixon, D. F., & Garrison, K. E. (2009). Cross-sectional dating of novel haplotypes of HERV-K 113 and HERV-K 115 indicate these proviruses originated in Africa before Homo sapiens. *Molecular biology and evolution*, *26*(11), 2617–2626. https://doi.org/10.1093/molbev/msp180
10. Jha, A. R., Nixon, D. F., Rosenberg, M. G., Martin, J. N., Deeks, S. G., Hudson, R. R., Garrison, K. E., & Pillai, S. K. (2011). Human endogenous retrovirus K106 (HERV-K106) was infectious after the emergence of anatomically modern humans. *PloS one*, *6*(5), e20234. https://doi.org/10.1371/journal.pone.0020234
11. Wilcox, C. (2020). *Did Viruses Create the Nucleus? The Answer May Be Near.*. Quanta Magazine. Retrieved 10 March 2022, from https://www.quantamagazine.org/did-viruses-create-the-nucleus-the-answer-may-be-near-20201125/.

12. Piel, F. B., Patil, A. P., Howes, R. E., Nyangiri, O. A., Gething, P. W., Williams, T. N., Weatherall, D. J., & Hay, S. I. (2010). Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nature communications*, *1*, 104. https://doi.org/10.1038/ncomms1104
13. Kwiatkowski D. P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. *American journal of human genetics*, *77*(2), 171–192. https://doi.org/10.1086/432519
14. Kerner, G., Laval, G., Patin, E., Boisson-Dupuis, S., Abel, L., Casanova, J., & Quintana-Murci, L. (2021). Human ancient DNA analyses reveal the high burden of tuberculosis in Europeans over the last 2,000 years. *The American Journal Of Human Genetics*, *108*(3), 517-524. https://doi.org/10.1016/j.ajhg.2021.02.009
15. Gray, M. (2017). Lynn Margulis and the endosymbiont hypothesis: 50 years later. *Molecular Biology Of The Cell*, *28*(10), 1285-1287. https://doi.org/10.1091/mbc.e16-07-0509
16. Yong, E. (2013). *Bacterial DNA in Human Genomes*. The Scientist Magazine. Retrieved 10 March 2022, from https://www.the-scientist.com/news-opinion/bacterial-dna-in-human-genomes-39147.
17. Riley, D. R., Sieber, K. B., Robinson, K. M., White, J. R., Ganesan, A., Nourbakhsh, S., & Dunning Hotopp, J. C. (2013). Bacteria-human somatic cell lateral gene transfer is enriched in cancer samples. *PLoS computational biology*, *9*(6), e1003107. https://doi.org/10.1371/journal.pcbi.1003107
18. Medawar, P., & Medawar, J. (1985). *Aristotle to zoos*. Oxford University Press.