

ARTICLE OPEN



Genetic ancestry and population structure of vaccinia virus

Cristian Molteni¹ , Diego Forni¹, Rachele Cagliani¹, Mario Clerici^{2,3} and Manuela Sironi¹ 

Vaccinia virus (VACV) was used for smallpox eradication, but its ultimate origin remains unknown. The genetic relationships among vaccine stocks are also poorly understood. We analyzed 63 vaccine strains with different origin, as well horsepox virus (HPXV). Results indicated the genetic diversity of VACV is intermediate between variola and cowpox viruses, and that mutation contributed more than recombination to VACV evolution. STRUCTURE identified 9 contributing subpopulations and showed that the lowest drift was experienced by the ancestry components of Tian Tan and HPXV/Mütter/Mulford genomes. Subpopulations that experienced very strong drift include those that contributed the ancestry of MVA and IHD-W, in good agreement with the very long passage history of these vaccines. Another highly drifted population contributed the full ancestry of viruses sampled from human/cattle infections in Brazil and, partially, to IOC clones, strongly suggesting that the recurrent infections in Brazil derive from the spillback of IOC to the feral state.

npj Vaccines (2022)7:92; <https://doi.org/10.1038/s41541-022-00519-4>

INTRODUCTION

Smallpox, caused by variola virus (VARV, genus *Orthopoxvirus*, family *Poxviridae*), was the first human infectious disease to be eradicated and, to date, the only one. VARV was also the first infectious agent against which a vaccine was developed. Indeed, vaccination was introduced in 1796 by Edward Jenner, who showed that humans could be protected against smallpox by inoculation of material deriving from animal lesions¹. It was initially assumed that the vaccine developed by Jenner was based on cowpox virus (CPXV), another member of the *Orthopoxvirus* genus. However, it is now clear that the virus used to immunize against smallpox, now referred to as vaccinia virus (VACV), is more closely related to horsepox virus (HPXV)^{1,2}. Thus, the biological origin of the smallpox vaccine is unsure and the situation is further complicated by the fact that neither cows nor horses are likely to represent the original hosts of CPXV and HPXV³. As a consequence, the natural host of VACV is unknown. Notably, VACV infections of animals (primarily cattle in South America and water buffalo in Asia) and humans have been repeatedly reported³. It is still unsure whether such infections derive from viruses that originated from the spillback of vaccine strains or if they represent natural VACV populations circulating in an unknown wild host³.


In the years that followed Jenner's discovery, different inocula used as vaccinating agents were repeatedly passaged in humans and, later, in animals. The most common practices involved virus inoculation in calves, heifers, rabbits, sheep, mice, or chick embryos^{2,4}. The inocula were also transported and distributed throughout the world to be used in local vaccination campaigns. For instance, at the end of the 19th century, the New York City Board of Health (NYCBH) was producing the smallpox vaccine from inocula originally transported from England and Cuba². Derivatives of the NYCBH vaccine were distributed to other laboratories and received different names. Supposedly, the Dryvax vaccine, which is a mixture of different clones and was widely used in the USA, was derived from the NYCBH stock. However, historical records suggest that the Beaugency lymph, used to seed vaccine production in France, was also (or mainly) used to manufacture Dryvax². The Beaugency lymph reached many

countries, including Brazil, where it was used to derive the IOC (Institute Oswaldo Cruz) vaccine². The WR (Western Reserve) and IHD (International Health department) strains are instead thought to derive from the NYCBH stock². In particular, IHD underwent a long passaging procedure and later reached Japan, where the IHD-J and IHD-W strains were derived^{5,6}.

Other vaccines, including Lister, Tian Tan, and Tashkent found diffusion in Europe and Asia. Their origin and passage histories are often unknown or murky. As an example, the Tian Tan vaccine, which was widely used in China, was reportedly isolated in Beijing's Temple of Heaven from a patient with smallpox and subsequently passaged in cows, monkeys, and rabbits. However, the host range of VARV is known to be restricted to humans and the sequencing of Tian Tan clones clearly indicated that the virus used for vaccination was VACV and not VARV^{7,8}. Thus, the origin of the Chinese vaccine is unknown.

Over the years, a number of VACV clones and strains have been sequenced. These include historical samples, such as a stock manufactured in 1902 by the Philadelphia company H.K. Mulford⁹ and vaccination kits preserved at the Mütter Museum of the College of Physicians of Philadelphia, dating to the mid-to-late nineteenth century¹⁰. Phylogenetic analyses showed that these historical vaccine genomes cluster with a HPXV strain, isolated from a horse in an 1976 outbreak in Mongolia^{10,11}. The availability of several sequenced VACV genomes thus allows analysis of their evolutionary histories. This is relevant from an historical perspective, but it is also topical, as zoonotic orthopoxviruses such as CPXV and monkeypox virus (MPXV) are increasingly reported as causes of human disease^{12,13}. In particular, an unprecedented MPXV multi-country outbreak outside the endemic area in Africa has been expanding since the first cases were reported in the UK, in May 2022¹⁴. Because smallpox vaccination with VACV provides cross-protection against multiple orthopoxviruses, the raise in MPXV and CPXV human cases is likely to be partially the result of discontinuation of routine smallpox vaccination¹⁵. Thus, a vaccine based on the modified vaccinia virus Ankara (MVA) is currently being offered to subjects at risk of exposure and to healthcare workers in the UK^{16,17}.

¹IRCCS E. MEDEA, Bioinformatics, Bosisio Parini, Italy. ²University of Milan, Milan, Italy. ³Don C. Gnocchi Foundation ONLUS, IRCCS, Milan, Italy.

email: cristian.molteni@lanostrafamiglia.it

RESULTS

Genetic diversity of VACV

We obtained a list of 64 complete or almost complete VACV genomes from public databases (Supplementary Table 1). These include the Mütter Museum and Mulford historical samples, several Dryvax, Lister, and Tian Tan clones, the single HPXV genome, the chorioallantois vaccinia virus Ankara (CVA, the parental virus of MVA), and other vaccine strains (e.g., WR, IOC, IHD-W). The dataset also comprises isolates from human, cattle, or buffalo infections in Brazil and Asia, as well as some strains derived from rabbit or mouse outbreaks in laboratory settings. Generation of a neighbor-net split network indicated clustering by genome origin (Fig. 1). In line with previous observations, a virus sequenced from mice during a lab outbreak and a human infection in the USA clustered with Lister and Dryvax, respectively^{18,19}.

To compare the genetic diversity of VACV to other orthopoxviruses deriving from natural transmissions, neighbor-net split networks were also generated for 50 modern VARV sequences and 90 CPXV sequences (Supplementary Table 2). VACV was definitely more diverse than VARV, but much less than CPXV (Fig. 1).

Genetic relationships among VACV genomes and population structure

To gain insight into the genetic relatedness of VACV genomes, we applied principal component analysis (PCA). The first two components clustered sequences based on origin or type (e.g., most vaccine strains were grouped by vaccine type). In line with previous phylogenetic analyses, HPXV, Mulford and Mütter Museum sequences clustered together (Fig. 2). The first PC clearly separated most American strains (Dryvax clones, Mulford and Mütter Museum sequences) from all other strains, with IOC having an intermediate placement. The second component separated the Brazilian infections (both cattle and human) from the other sequences, with IOC and HPXV/Mütter/Mulford genomes in intermediate position (Fig. 2). In line with previous phylogenetic analyses, two strains thought to be derived from the NYCBH stock, namely IHD-W and WR, clustered with the European/Asian sequences^{2,6,7}. IOC sequences were the closest to the viral genomes deriving from the Brazilian infections (Fig. 2). The third PC contributed to the further separation of Eurasian sequences, while adding little to the separation of the American strains. The viruses responsible for human/buffalo infections were placed distant from all other genomes. The divergent Lister clone is a Lister-derived recombinant oncolytic virus²⁰ (Fig. 2).

To gain further insight into the structure of VACV populations, we used the program STRUCTURE, which was widely applied to infer ancestry and admixture patterns^{21–23}. STRUCTURE relies on a Bayesian statistical model for clustering genotypes into populations without prior information on their genetic relatedness. The program can identify distinct subpopulations (or clusters, K) that compose the overall population. Subpopulations can then be related to specific features such as origin, genotype classification, or phenotype. We thus considered that this approach might provide insight into the ancestry and evolutionary history of VACV genomes.

Because STRUCTURE is ideally suited for weakly linked markers, we first analyzed the level of linkage disequilibrium with LIAN v3.7, which tests the null hypothesis of linkage equilibrium across loci²⁴. Statistically significant LD was detected (Monte Carlo simulations, 1000 repetitions, $p < 10^{-3}$), but the standardized index of association (I_A^S) resulted equal to 0.038. This value indicates weak LD, most likely as a cause of recombination and other processes, and warrants the application of STRUCTURE models. Specifically, we used the linkage model with correlated allele frequencies, which assumes that discrete genome “chunks” were inherited from K ancestral populations²².

To estimate the optimal number of subpopulations in the VACV dataset, STRUCTURE was run for values of K from 1 to 14. The ΔK method yielded a major peak at $K = 9$ (Supplementary Fig. 1). Analysis of ancestry components was thus performed for 9 subpopulations and by plotting genomes according to their origin or type (Fig. 3a). Results showed a very good clustering by origin, as well as concordance with the neighbor-net split network and PCA analyses. Thus, the nine subpopulations roughly identified the HPXV/Mütter/Mulford strains, Dryvax sequences (together with the USA infection), Lister strains (together with the lab mouse infection), Tian Tan, IHD-W, MVA, and Tashkent clones. Two additional populations identified the Brazilian and Asian infections (Fig. 3a). All these sequences formed distinct clusters in the network and PCA analyses (Figs. 1 and 2). Evidence of extensive admixture was observed for other VACV genomes (i.e., IOC, WR, Copenhagen, CVA, and the lab rabbit strain) (Fig. 3a) that formed distinct branches in the neighbor-net split network and were separated in the PCA (Figs. 1 and 2).

We next used the linkage model in STRUCTURE to estimate the level of drift of each subpopulation from a hypothetical common ancestral population. Specifically, we estimated the F parameter, which represents a measure of genetic differentiation between populations based on allele frequencies. Results indicated that the lowest drift was experienced by the two subpopulations that account for the largest ancestry component of Tian Tan genomes and for the full ancestry of HPXV/Mütter/Mulford sequences (hereafter referred to as subpopulations HPXV and Tian Tan) (Fig. 3b). In line with the PCA analysis, subpopulation HPXV accounted for variable proportions of the ancestry of the IOC and Dryvax clones, of Tian Tan strains, and of the WR sequence (Fig. 3a). The subpopulation highly represented in Tian Tan strains also contributed to the WR genome, as well as to most Lister and IOC clones, to the Copenhagen strain, and to the ancestry of CVA (Fig. 3a). Some level of uncertainty in the estimation of F was observed for one of the populations showing low drift (Fig. 3b). This component accounted for the major ancestry of Lister clones, and it was mainly restricted to Lister samples, with the exception of a contribution to WR (Fig. 3a). Two populations with slightly higher drift represented the major ancestry of Dryvax clones and the full ancestry of buffalo and human infections sampled in Asia (Fig. 3b). Whereas the Dryvax component was evident in several other strains, the human/buffalo infection component was virtually restricted to the Asian strains (Fig. 3a).

Finally, four subpopulations were found to have experienced very strong drift (Fig. 3b). One of these almost fully contributed to the Tashkent ancestry and minimally to other genomes. Two other highly drifted subpopulations accounted for the full ancestry of MVA and IHD-W sequences (Fig. 3a). This is in very good agreement with the notion that MVA and IHD-W are among the most passaged VACV strains. Interestingly, the fourth high-drift population contributed the full ancestry of viruses responsible for the Brazilian human and cattle infections. Other than these sequences, this component was present at appreciable frequency only in the IOC clones, in line with the PCA analysis (Figs. 2 and 3a). These data strongly support the view that the recurrent infections in Brazil derive from the spillback of IOC to domestic and possible wild hosts.

To gain further insight into the ancestry contribution to VACV genomes, we exploited the site-by-site inference in STRUCTURE, which allows population-of-origin assignment for individual variants. The analysis was performed for the four components (HPXV, Tian Tan, Dryvax, and Lister) that are shared across several VACV genomes (Fig. 4 and Supplementary Fig. 2). The HPXV component is the most widely shared among the VACV genomes we analyzed. Clearly, this component contributes to the almost full ancestry of HPXV, Mütter and Mulford sequences (Fig. 4). In all other instances, irrespective of their origin, contributing HPXV variants are not clustered but rather scattered throughout the

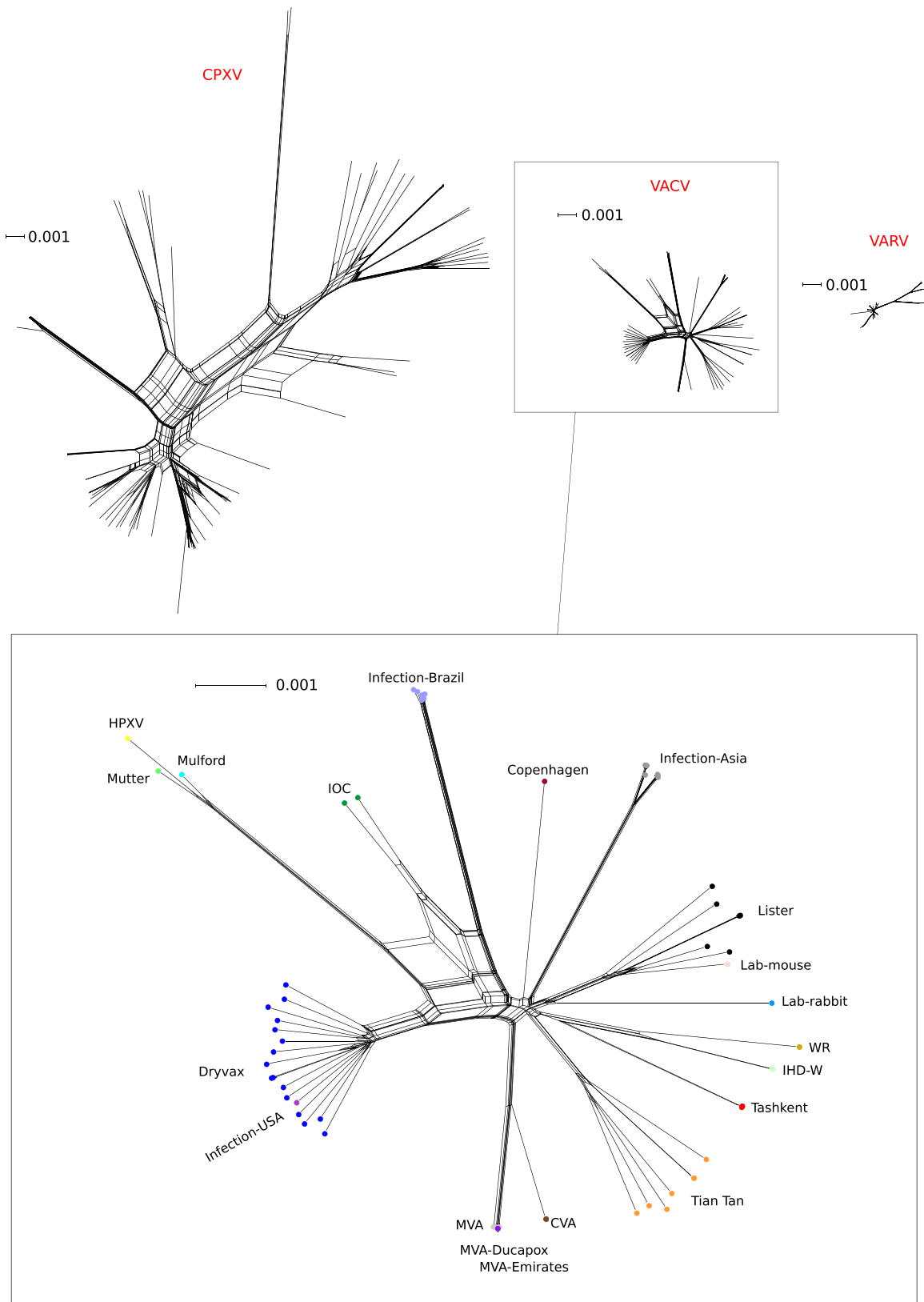


Fig. 1 Genetic variability within different orthopoxvirus species. Neighbor-net split network of 90 CPXV, 50 VARV, and 64 VACV genome sequences; the three networks are plotted at the same scale (expressed as substitutions/site). An enlargement of VACV genomes is also provided, with samples shown as dots and colors representing group membership.

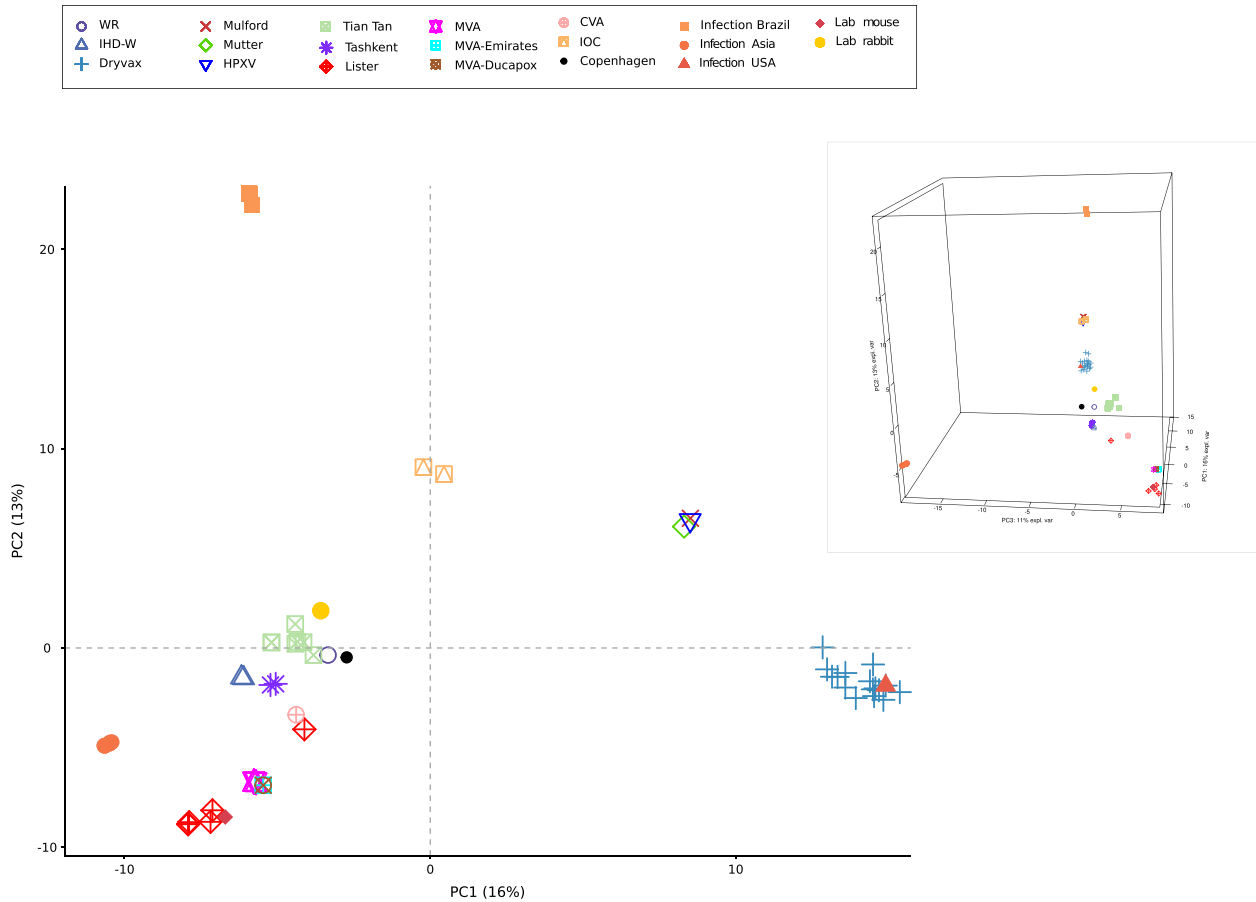


Fig. 2 **PCA of vaccinia virus genomes.** A principal component analysis of VACV samples is shown. Each VACV group/type is colored and displayed with a different symbol, as described in the legend. The two first PCs are plotted in the large figure. In the insert, the two axes (PC1 and PC2) are rotated to visualize the contribution of PC3.

genome. This observation is also true for the Tian Tan, Lister, and Dryvax components (Fig. 4 and Supplementary Fig. 2). Thus, the admixture observed in VACV genomes is not the result of the exchange of large genomic fragments. Conversely, these patterns of intermixed small blocks suggest a scenario of recombination of short genomic regions or evolution by mutation from a common population(s), or both.

Recombination and mutation in VACV evolution

To analyze the relative contribution of recombination and mutation in the evolution of VACV strains, recombination was inferred genome-wide for the 64 genomes using ClonalFrameML²⁵. In total, 988 recombination events were detected along the branches of the inferred phylogeny. The average length of the recombinant segments was 36.9 bp; the average divergence between donor and recipient was 0.175. The genome-wide rate of recombination to mutations (R/θ) was estimated to be equal to 0.07, meaning that mutations happened about fourteen times more often than recombination events. Overall, the relative effect of recombination to mutation (r/m) was 0.46. So, nucleotide changes in the VACV phylogeny were twice more likely to arise from mutation than by recombination.

DISCUSSION

The events that unfolded since Jenner's seminal intuition until the eradication of smallpox have no equal in the annals of infectious diseases and of medicine at large. The evolutionary history of

VACV and its convolutions are intimately linked to those events. VACV reached every corner of the world in the form of vaccine inocula, but its ultimate origin and natural host(s) remain unknown. Thus, the genetic makeup of VACV is the result of artificial virus growth, passaging, selection, and migration. Most likely, extant VACV strains and clones derive from a limited number of seed viruses^{2,6}, a situation that clearly introduced founder effects and limited genetic variability. The genetic diversity of VACV is intermediate between VARV and CPXV, two viruses with very different host range and epidemiology. Like CPXV, VACV can infect many different hosts, suggesting that, had it evolved naturally, it might display a diversity comparable to that of CPXV. However, the possible extinction of natural VACV lineages implies that important differences with CPXV might exist. Indeed, it is unsure whether any natural VACV sequence has ever been sampled, as even the HPXV strain might represent a vaccine escapee to the feral state. Nonetheless, based on the pattern of gene inactivation, it has been suggested that HPXV or a HPXV-like virus was the progenitor of several early inocula that were turned into vaccines. Our data strongly support this hypothesis, as the ancestry of HPXV, as well as those of the Mütter Museum and Mulford vaccines, are fully accounted for by a population showing low drift. The HPXV ancestry component is also one of the most widespread among the VACV genomes we analyzed. In particular, it contributed a significant ancestry portion in IOC vaccines, which are thought to be derivatives of the Beaugency lymph². This strengthens the tie between this latter and the HPXV/Mulford/Mütter cluster^{2,10}. These data are also in agreement with the view that Dryvax, also showing a HPXV

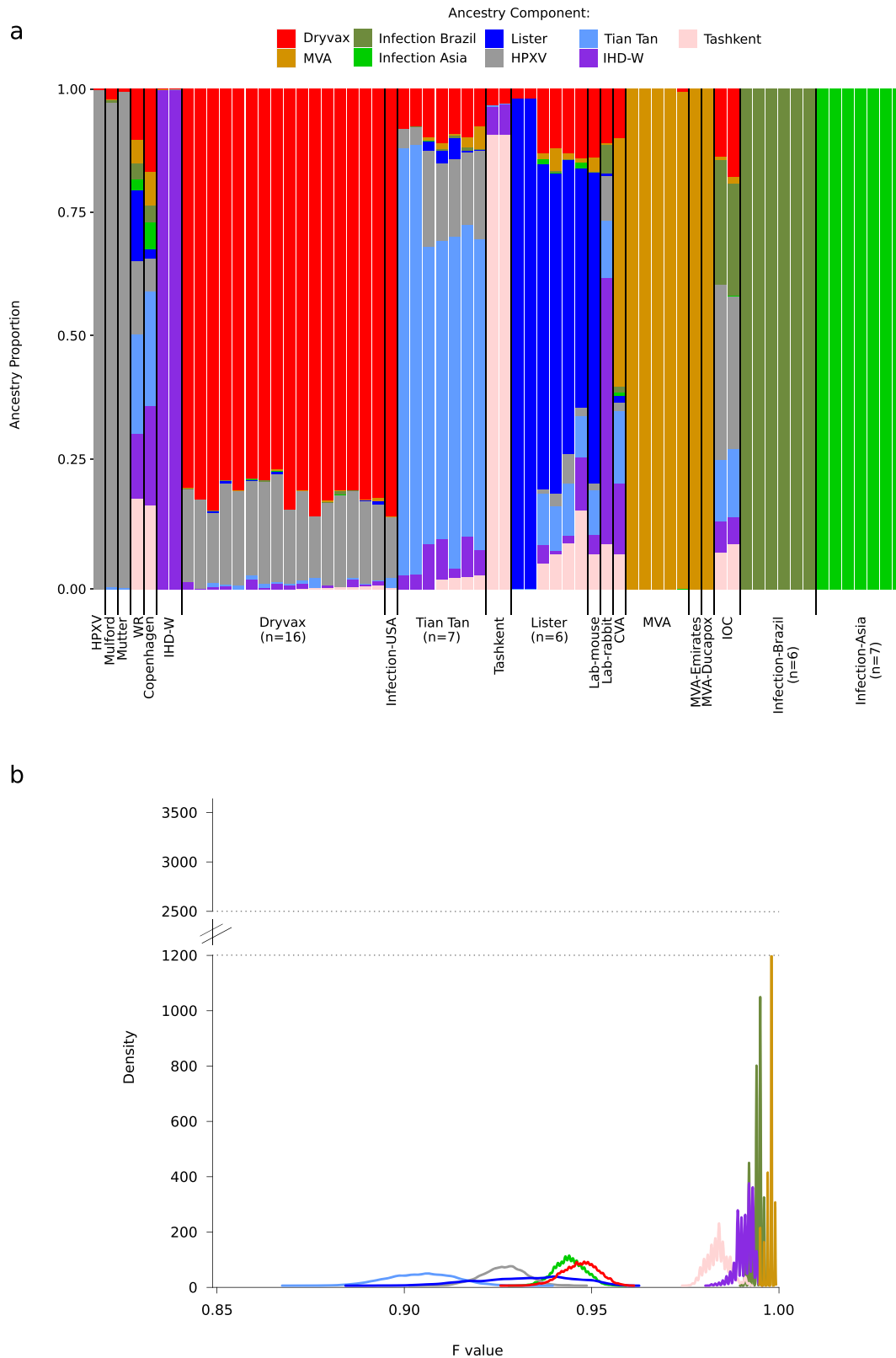


Fig. 3 Population structure of vaccinia virus. a Bar plot representing the proportion of ancestral population components for $K = 9$. Each vertical line represents a VACV genome and it is colored by the proportion of sites that have been assigned to the nine populations by STRUCTURE. Ancestry components are named based on the genomes where they are more prevalent. **b** Distributions of F values for the nine populations. Colors are as in panel a. Y axis is cut to better display all density distributions.

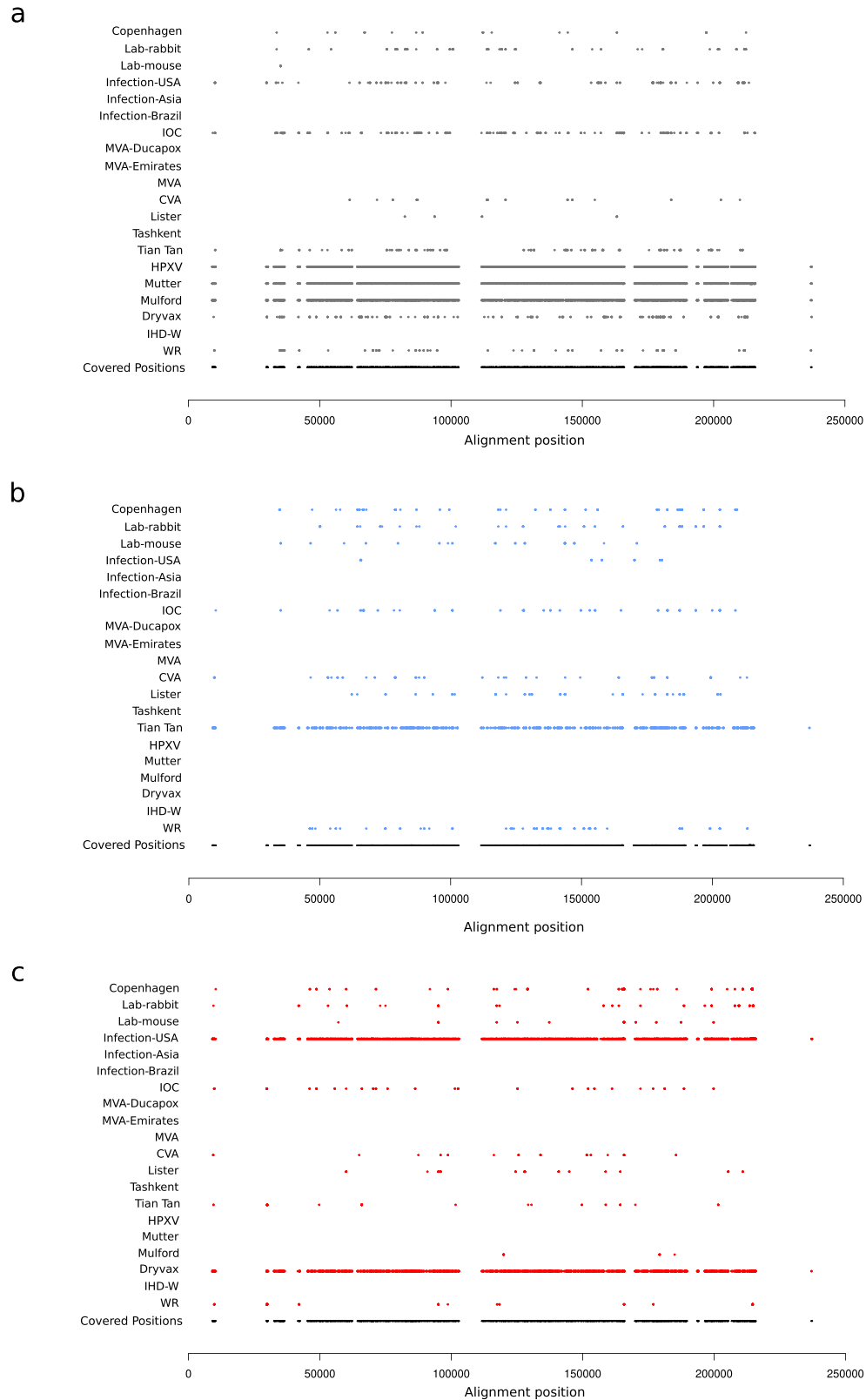


Fig. 4 Ancestral component probabilities at nucleotide resolution. Site by site probability of ancestry components for a subset of VACV genomes. Each line represents a randomly selected virus for each of the clones/populations of VACV; each dot represents a PI site with a probability >0.75 to derive from the **a** HPXV, **b** Tian Tan, and **c** Dryvax ancestry component. Colors are the same as in Fig. 3. The bottom line represents the position of all PI sites (as black dots) analyzed in this study. Positions refer to the genome alignment. Regions that are not covered by PIs are highly dynamic and sequence information was present for less than 90% of sequences.

ancestry component, was mainly derived from the Beaugency lymph². Nonetheless, the HPXV population is not the one showing the lowest *F* values. The Tian Tan component has lower drift, although the *F* distribution partially overlaps with that of HPXV, as well as with the Lister component. Reportedly, the Tian Tan strain was first isolated in 1923, but its origin is unknown. In line with previous phylogenetic analyses^{7,8}, we found the Tian Tan component to be also present in the Copenhagen, Lister and WR sequences, as well as in the two IOC strains. Overall, it is tempting to speculate that HPXV-like and Tian Tan-like viruses, whether they were part of the same or of distinct populations, have served as the ancestors of most vaccine lineages that subsequently evolved independently.

Our results showing strong drift of the MVA and IHD-W populations is in good agreement with the passage history of these vaccine strains. Indeed, MVA was derived from CVA by more than 570 passages in chicken embryo fibroblasts. As a result, MVA acquired a restricted tropism for avian cells²⁶. The ancestor of IHD-W underwent several rounds of intracerebral infection in mice and chorioallantoic membranes before reaching Japan^{5,6}. High drift was also observed for the two Tashkent clones. Tashkent, whose use in humans was discontinued due to severe adverse effects, is generally considered a relatively old and pristine vaccine⁶. However, its origin is unknown, and our analyses suggest that it experienced a level of drift comparable to MVA and IHD-W, possibly indicating a long passage history or a strong bottleneck resulting from other effects. Notably, the fourth population that experienced strong drift accounts for the ancestry of viruses responsible for cattle and human infections in Brazil (Br-VACV). This component is also present in the IOC clones, but not in other VACV sequences. Thus, these data are in line with the PCA results and with previous indications that Br-VACV sequences are more closely related to IOC than to HPXV^{1,27}. These observations strongly suggest that Br-VACV derived from the escape of an IOC-like vaccine strain and that the spillback caused a bottleneck in the viral population. It should however be mentioned that VACV infections in Brazil are caused by viruses belonging to two distinct lineages (referred to as Br-VACV group 1 and Br-VACV group 2)²⁸. We only analyzed Br-VACV group 1 (which includes Cantagalo virus and Serro 2), as no complete viral genome is available for group 2. Thus, the history of Br-VACV is necessarily more complicated than the one we can reconstruct, and the sequencing of Br-VACV group 2 genomes will be necessary to reach to the full scenario. Even more enigmatic is the situation of the Asian infections. The ancestry component accounting for the VACV genomes sampled in Asia is in the low range of *F* values and is virtually absent from any other genome analyzed herein. These data therefore do not support the view that these infectious strains derive from the Lister vaccine and, in PCA, they were similarly distant from many Eurasian vaccines. Thus, if these infections represent another spillback event, the source vaccine strain may have remained unsampled. Alternatively, these infections, which were first described in 1934, might be caused by natural viruses circulating in an unknown reservoir(s)³.

Previous studies that analyzed recombination patterns in Dryvax genomes^{29,30} suggested that recombination contributed to the diversification of VACV clones and described a patchy pattern consistent with the exchange of short genomic fragments. Our analysis with ClonalFrameML across the entire VACV phylogeny is consistent with this possibility and we found that the average length of recombining fragments is less than 40 bp. Indeed, the same analysis revealed that mutation played a more relevant role than recombination in the evolution of VACV. This might seem counter-intuitive, as the production of VACV in artificially infected animals and cells probably often resulted in high multiplicities of infection, which might promote recombination. However, microscopy studies have shown that the mode of

orthopoxvirus replication within infected cells creates physical constraints that reduce opportunities for forming recombinants^{30,31}. Also, it should be mentioned that recombination events that occurred between distantly related genomes are easier to detect than events involving close relatives. In particular, ClonalFrameML is based on a model of extra-population recombination, although it can also detect a large proportion of intra-population recombination³². Thus, the program might have missed some events that occurred among closely related VACV genomes (e.g., those that may have occurred among the clones of a vaccine stock).

Finally, we should add that most of our analyses were based on PI sites for which at least 90% of sequences had non-missing information. Whereas this had the purpose of excluding poorly aligned regions, we certainly missed information from the more dynamic portions of VACV genomes, which evolve by gene gains and losses.

In summary, our data provide new insight into the evolutionary history of smallpox vaccines. We suggest that early vaccine strains were related to HPXV and Tian Tan, and that their distribution worldwide originated the two major VACV clades. Thus, HPXV, possibly closely related to the Beaugency lymph, served as the seed for Dryvax and IOC, whereas a Tian Tan-like population contributed to vaccine clones in the Eurasian clade. We consider that a better understanding of orthopoxvirus evolutionary trajectories may also help shed light into the ongoing MPXV outbreak, as this orthopoxvirus most likely underwent a bottleneck (the transmission that brought it out of Africa) and is probably evolving at accelerated rate³³.

METHODS

Sequences, alignments, networks

The complete sequences of VACV genomes were obtained from the NCBI database (<https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/>). The number of strains retrieved by NCBI consisted of 71 viruses. Five of these were omitted because they were reassembled genomes^{34,35}. Also, two sequences (MG012795 and MG012796) that were previously shown to determine exceptionally long terminal branches in the VACV phylogenetic tree were excluded from the analyses³⁶. The final dataset of 64 sequences (Supplementary Table 1) was aligned using MAFFT (v.7.475) with default parameters³⁷. SplitsTree4 (v4.16.2)³⁸, with HKY85 distances, all polymorphic sites, and without gap sites, was used to generate a neighbor-net split network. Using the same parameters, networks were also generated for 90 CPXV and 50 VARV sequences obtained from NCBI (Supplementary Table 2).

Linkage disequilibrium and population structure

From the VACV alignment, biallelic parsimony-informative (PI) sites were extracted. In particular, we selected biallelic sites, each with a minimum frequency of two, for those genomic positions where at least 90% of sequences had non-missing information. Gaps and all nonstandard nucleotide bases were considered as missing values. This generated a list of 3794 variants. Some genes in highly dynamic genome regions were not covered by any PI (Supplementary Table 3). To evaluate the level of linkage disequilibrium (LD) in the dataset, the LIAN software was used (v.3.7)²⁴. This software tests for independent assortment by computing the number of loci at which each pair of haplotypes differs. Significance was assessed by Monte Carlo simulations (1,000 iterations). The standardized index of association (I_A^s) generated by LIAN was associated to the LD interpretation, with zero meaning linkage equilibrium. For the VACV dataset, we obtained a I_A^s of 0.0379, therefore allowing the application of population structure analyses as implemented in the STRUCTURE (v.2.3.4) suite²¹. To run STRUCTURE, we first estimated the allele frequency spectrum parameter (λ) by running the program with $K=1$, as suggested²². The λ parameter was estimated to be equal to 0.8762. Using this value, the linkage model with correlated allele frequencies was performed²², estimating K from 1 to 14. In particular, for each K , ten runs were run with a MCMC total chain length of 500,000 iterations and 50,000 iterations as burn-in. The map distances were set equal to PI site physical

distances. The optimal K was evaluated using the HARVESTER tool³⁹, according to Evanno's method⁴⁰. The CLUMPAK⁴¹ software was used to combine replicate runs from the same K and to generate the Q value matrix. The amount of drift that each subpopulation experienced was quantified by the F parameter calculated for the optimal k value²². A linkage model analysis run with the optimal K was also performed with the SITEBYSITE option selected, allowing to assign ancestry contribution to every biallelic position for each viral strain.

PCA

Principal component analysis (PCA) was performed using the same PI matrix used for STRUCTURE. PCA can find patterns in a sample without prior knowledge⁴². The purpose of PCA analysis is to reduce the complexity in high-dimensional data while retaining trends and patterns, by transforming the data into fewer dimensions, which act as summaries of features⁴². PCA was carried out with the mixOmics R package⁴³.

Recombination

Recombination events were analyzed using ClonalFrameML v.1.12²⁵. ClonalFrameML requires a sequence alignment and a phylogenetic tree as input files. We used the whole genome alignment as input sequences, and we generated a maximum-likelihood (ML) tree with phyML v.20120412⁴⁴, using a HKY85 substitution model. PhyML also estimates the transition/transversion ratio; this value, along with the mean branch length, were also provided to ClonalFrameML. A 100 pseudo-bootstrap replicates were performed and R/theta (relative rate of recombination to mutation), 1/delta (inverse mean DNA import length), and nu (mean divergence of imported DNA) were estimated. The r/m parameter (the rate at which nucleotides become substituted as a result of recombination or mutation) was calculated as follow: $r/m = \text{delta} * R/\text{theta} * \text{nu}$.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The list of NCBI IDs of the viral sequences analyzed is provided in Supplementary Table 1–2.

Received: 27 May 2022; Accepted: 28 July 2022;

Published online: 11 August 2022

REFERENCES

- Esparza, J., Nitsche, A. & Damaso, C. R. Beyond the myths: novel findings for old paradigms in the history of the smallpox vaccine. *PLoS Pathog.* **14**, e1007082 (2018).
- Damaso, C. R. Revisiting Jenner's mysteries, the role of the Beaugency lymph in the evolutionary path of ancient smallpox vaccines. *Lancet Infect. Dis.* **18**, e55–e63 (2018).
- Silva, N. I. O., de Oliveira, J. S., Kroon, E. G., Trindade, G. S. & Drumond, B. P. Here, there, and everywhere: the wide host range and geographic distribution of zoonotic orthopoxviruses. *Viruses* **13**, 43 (2020).
- Esparza, J., Lederman, S., Nitsche, A. & Damaso, C. R. Early smallpox vaccine manufacturing in the United States: Introduction of the "animal vaccine" in 1870, establishment of "vaccine farms", and the beginnings of the vaccine industry. *Vaccine* **38**, 4773–4779 (2020).
- Ichihashi, Y. & Dales, S. Biogenesis of poxviruses: interrelationship between hemagglutinin production and polykaryocytosis. *Virology* **46**, 533–543 (1971).
- Qin, L., Favis, N., Famulski, J. & Evans, D. H. Evolution of and evolutionary relationships between extant vaccinia virus strains. *J. Virol.* **89**, 1809–1824 (2015).
- Qin, L., Liang, M. & Evans, D. H. Genomic analysis of vaccinia virus strain TianTan provides new insights into the evolution and evolutionary relationships between Orthopoxviruses. *Virology* **442**, 59–66 (2013).
- Zhang, Q. et al. Genomic sequence and virulence of clonal isolates of vaccinia virus Tiantan, the Chinese smallpox vaccine strain. *PLoS One* **8**, e60557 (2013).
- Schrick, L. et al. An early American smallpox vaccine based on Horsepox. *N. Engl. J. Med.* **377**, 1491–1492 (2017).
- Duggan, A. T. et al. The origins and genomic diversity of American Civil War Era smallpox vaccine strains. *Genome Biol.* **21**, 175-020-02079-z (2020).
- Tulman, E. R. et al. Genome of horsepox virus. *J. Virol.* **80**, 9244–9258 (2006).
- Simpson, K. et al. Human monkeypox - After 40 years, an unintended consequence of smallpox eradication. *Vaccine* **38**, 5077–5081 (2020).
- Silva, N. I. O., de Oliveira, J. S., Kroon, E. G., Trindade, G. S. & Drumond, B. P. Here, There, and Everywhere: The Wide Host Range and Geographic Distribution of Zoonotic Orthopoxviruses. *Viruses* **13**, 43 (2020).
- Kraemer, M. U. G. et al. Tracking the 2022 monkeypox outbreak with epidemiological data in real-time. *Lancet Infect. Dis.* **22**, 941–942 (2022).
- Zumla, A. et al. Monkeypox outbreaks outside endemic regions: scientific and social priorities. *Lancet Infect. Dis.* **22**, 929–931 (2022).
- Mahase, E. Monkeypox: Gay and bisexual men with high exposure risk will be offered vaccine in England. *BMJ* **377**, o1542 (2022).
- Mahase, E. Monkeypox: Healthcare workers will be offered smallpox vaccine as UK buys 20 000 doses. *BMJ* **377**, o1379 (2022).
- Mavian, C., López-Bueno, A. & Alcamí, A. Genome Sequence of WAU86/88-1, a New Variant of Vaccinia Virus Lister Strain from Poland. *Genome Announc.* **2**, e01086–13 (2014).
- Martin, R. M. et al. Contact transmission of vaccinia to an infant diagnosed by viral culture and metagenomic sequencing. *Open Forum Infect. Dis.* **7**, ofaa111 (2020).
- Zhang, Q. et al. The highly attenuated oncolytic recombinant vaccinia virus GLV-1h68: comparative genomic features and the contribution of F14.5L inactivation. *Mol. Genet. Genomics* **282**, 417–435 (2009).
- Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
- Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
- Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
- Haubold, B. & Hudson, R. R. LIAN 3.0: detecting linkage disequilibrium in multilocus data. *Linkage Analysis. Bioinformatics* **16**, 847–848 (2000).
- Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
- Meisinger-Henschel, C. et al. Genomic sequence of chorioallantois vaccinia virus Ankara, the ancestor of modified vaccinia virus Ankara. *J. Gen. Virol.* **88**, 3249–3259 (2007).
- Medaglia, M. L. et al. Genomic Analysis, Phenotype, and Virulence of the Historical Brazilian Smallpox Vaccine Strain IOC: Implications for the Origins and Evolutionary Relationships of Vaccinia Virus. *J. Virol.* **89**, 11909–11925 (2015).
- Oliveira, J. S. et al. Vaccinia Virus Natural Infections in Brazil: The Good, the Bad, and the Ugly. *Viruses* **9**, 340 (2017).
- Qin, L., Upton, C., Hazes, B. & Evans, D. H. Genomic analysis of the vaccinia virus strain variants found in Dryvax vaccine. *J. Virol.* **85**, 13049–13060 (2011).
- Paszczkowski, P., Noyce, R. S. & Evans, D. H. Live-Cell Imaging of Vaccinia Virus Recombination. *PLoS Pathog.* **12**, e1005824 (2016).
- Lin, Y. C. & Evans, D. H. Vaccinia virus particles mix inefficiently, and in a way that would restrict viral recombination, in coinfecting cells. *J. Virol.* **84**, 2432–2443 (2010).
- Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
- Isidoro, J. et al. Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus. *Nat. Med.* (2022).
- Brinkmann, A., Souza, A. R. V., Esparza, J., Nitsche, A. & Damaso, C. R. Re-assembly of nineteenth-century smallpox vaccine genomes reveals the contemporaneous use of horsepox and horsepox-related viruses in the USA. *Genome Biol.* **21**, 286-020-02202-0 (2020).
- Duggan, A. T., Holmes, E. C. & Poinar, H. N. Response to Brinkmann et al. "Re-assembly of 19th century smallpox vaccine genomes reveals the contemporaneous use of horsepox and horsepox-related viruses in the United States". *Genome Biol.* **21**, 287-020-02203-z (2020).
- Wei, C., Chen, Y. M., Chen, Y. & Qian, W. The missing expression level-evolutionary rate anticorrelation in viruses does not support protein function as a main constraint on sequence evolution. *Genome Biol. Evol.* **13**, evab049 (2021).
- Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
- Huson, D. H. & Bryant, D. Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
- Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Cons. Genet Res* **4**, 359–361 (2012).
- Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

41. Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A. & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol. Ecol. Resour.* **15**, 1179–1191 (2015).
42. Lever, J., Krzywinski, M. & Altman, N. Points of significance: principal component analysis. *Nat. methods* **14**, 641–643 (2017).
43. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K. A. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
44. Guindon, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).

ACKNOWLEDGEMENTS

This work was supported by the Italian Ministry of Health ("Ricerca Corrente 2021-2022" to MS).

AUTHOR CONTRIBUTIONS

M.S., C.M., and D.F. designed the study. C.M., R.C. and D.F. performed the analyses. M.S., C.M., D.F., R.C. and M.C. interpreted the data. M.S., C.M., D.F. and M.C. wrote the manuscript. All authors contributed to and approved the final manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41541-022-00519-4>.

Correspondence and requests for materials should be addressed to Cristian Molteni.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022