SCIENTIFIC REPORTS

Correction: Author Correction

OPEN

Received: 31 May 2017
Accepted: 31 January 2018
Published online: 22 February 2018

Whole genome diversity of inherited chromosomally integrated HHV-6 derived from healthy individuals of diverse geographic origin

Marco Telford[1], Arcadi Navarro[1,2,3,4] & Gabriel Santpere[1,5]

Human herpesviruses 6-A and -B (HHV-6A, HHV-6B) are ubiquitous in human populations worldwide. These viruses have been associated with several diseases such as multiple sclerosis, Hodgkin's lymphoma or encephalitis. Despite of the need to understand the genetic diversity and geographic stratification of these viruses, the availability of complete viral sequences from different populations is still limited. Here, we present nine new inherited chromosomally integrated HHV-6 sequences from diverse geographical origin which were generated through target DNA enrichment on lymphoblastoid cell lines derived from healthy individuals. Integration with available HHV-6 sequences allowed the assessment of HHV-6A and -6B phylogeny, patterns of recombination and signatures of natural selection. Analysis of the intra-species variability showed differences between A and B diversity levels and revealed that the HHV-6B reference (Z29) is an uncommon sequence, suggesting the need for an alternative reference sequence. Signs of geographical variation are present and more defined in HHV-6A, while they appear partly masked by recombination in HHV-6B. Finally, we conducted a scan for signatures of selection in protein coding genes that yielded at least 6 genes (4 and 2 respectively for the A and B species) showing significant evidence for accelerated evolution, and 1 gene showing evidence of positive selection in HHV-6A.

Human herpesvirus 6 (HHV-6) is a globally dispersed dsDNA virus that infects more than 90% of the adult human population[1–3]. HHV-6 presents two species (HHV-6A and HHV-6B) with high nucleotide identity[4,5] but with different epidemiology, biology and immunologic profiles[6].

HHV-6 achieves latency by integration at the subtelomeric end of the telomeres of very few host cells[7], making it arduous to detect in the absence of primary infection or reactivation[8,9]. While the preferential replication site for HHV-6 are CD4$^+$T lymphocytes *in vitro*[10] and *in vivo*[11], they are characterized by a broad tropism for human cells[12–15]. The genomes of HHV-6A and -6B are similarly organized, with most of the genes encoded in a long non-repetitive region, enclosed between two identical long repeats (ca 8 Kb) called DR_L and DR_R. The two DRs are flanked by additional repeats (T1 and T2) composed of perfect telomere-like repeats (TTAGGG)n in variable copy number. The T1 region is longer than T2, and presents as well degenerate telomere-repeats[16,17]. These telomere-like repeat allow for homologous recombination with the subtelomeric end of the telomere region[7], and has been found in different chromosomes, but usually in a single copy per host[7–9,18–26]. While both T1 and T2 take part in the integration process, only T2 has been shown to be necessary[16].

The integration in the telomere region could have negative effects on the host cell as it could interfere with its protective role against chromosome shortening or incorrect identification of the chromosome end as a

[1]Institute of Evolutionary Biology (UPF-CSIC), Departament de Ciències Experimentals i la Salut, Universitat Pompeu Fabra, PRBB, Barcelona, Catalonia, Spain. [2]National Institute for Bioinformatics (INB), PRBB, Barcelona, Catalonia, Spain. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), PRBB, Barcelona, Catalonia, Spain. [4]Center for Genomic Regulation (CRG), PRBB, Barcelona, Catalonia, Spain. [5]Department of Neuroscience, Yale School of Medicine, New Haven, CT, 06510, USA. Correspondence and requests for materials should be addressed to A.N. (email: arcadi.navarro@upf.edu) or G.S. (email: gabrielsantperebaro@gmail.com)

double-strand break[27,28]. Huang *et al.* showed that while telomeres *in vitro* (lymphoblastoid cell lines (LCL)) presenting HHV-6 integration were shorter than average, *in vivo* (sperm DNA) showed no different sign of erosion[29]. Nevertheless, the impact of HHV-6's integration on telomere functionality and on the host cell is yet to be fully understood and other aspects unrelated to chromosome shortening might be affected[30]. Integrated HHV-6 has been shown to be in instances able to fully reactivate *in vitro*[7,31] and *in vivo*[32–34].

The integration occasionally occurs in germinal cells allowing it to enter the germ line. The virus can then be transmitted vertically in subsequent generations and be present in every cell of the offspring[33] as a congenital condition. This condition is classically referred to as inherited chromosomally integrated HHV-6 (iciHHV-6) or endogenous HHV-6, and it affects 0.2–1% of the world adult population[20,33,35,36]. The most common, non-congenital, form of the virus is referred to as exogenous HHV-6. The congenital form of the virus can consist of an intact genome, which can potentially lead to the expression of the whole viral gene set within every cell of the carrier[7,37]. The implication of the presence of iciHHV-6 and its relation with HHV-6-related diseases are currently being explored[24,26,29,38].

While only the HHV-6B species is a recognized etiological factor for *exanthema subitum*[39], both species are linked to viral fever, febrile seizure[40], graft rejection[41,42], and are common causes of *status epilecticus*[43]. Both species present a wide range of putative disease associations, with different degree of characterization and support. Among these we find multiple sclerosis (MS)[44–46], Hodgkin's lymphoma (HL)[23,26,47,48], and encephalitis/meningitis (EM)[49–51].

The study of the pathological association of HHV-6 is limited by its high seroprevalence and detection rates that hamper the determination of causal links between virus and diseases. Also, there are numerous confounding variables that can influence the interpretation of results or mask an association, primarily their variability, their putative geographical stratification and the differences between exogenous and endogenous HHV-6. To estimate these factors, a much more complete set of sequences is needed. Both prevalence of HHV-6 and HHV-6-associated diseases show worldwide geographical variation[52–57]. For instance, MS has higher prevalence in high-latitude and Caucasian populations[58], while HL and non-HL have higher incidence in Europe and North America compared to Africa, Asia and South America[59].

Genetic differences among HHV-6 strains could contribute to these marked geographical patterns in epidemiology of the above-mentioned diseases. However, and due to the scarcity of full HHV-6 sequences, little is known about the genetic stratification among HHV-6 viruses in latency within individuals from different populations.

We performed a systematic search for the presence of HHV-6 within individuals from different populations sequenced in the 1000 Genome Project[60] (1KGP), a large-scale database of whole human genomes from healthy individuals (http://www.internationalgenome.org/). We identified 11 individuals presenting HHV-6 in a possible congenital state from which we could obtain total DNA of derived LCLs. We then performed target enrichment of the entire HHV-6 genome, allowing us to reconstruct 9 almost complete genomic sequences. Through droplet digital PCR we quantified the virus copy number, supporting the congenital state of the viruses, producing a new data set of iciHHV-6. We carried out a comprehensive comparative analysis across HHV-6 genomes to reveal the geographical structure of genetic diversity, the patterns of recombination along the virus' genome and putative targets of positive selection in protein coding genes.

Results

A small proportion of 1KGP individuals show HHV-6 presence. The detection of the whole HHV-6 genome using reads from the 1KGP low-coverage dataset (mean depth = 7,4x[61]) would be a solid sign of the presence of latent integrated HHV-6. Out of 2,535 individuals scanned from the 1KGP, 11 showed higher number of reads mapping to HHV-6, which amounts to 0.44% of the whole data set.

We searched for signs of the presence of a single or multiple species within one sample by two approaches: a) counting uniquely mapping reads and density of mismatches against both species. b) constructing phylogenetic trees to provide bootstrap support for the separation between species using the whole genome or using only the U83 ORF. U83 is the most divergent ORF between HHV-6A and HHV-6B[4,5,62]. We observed a clear pattern of mismatches discriminating both species and a clear separation between HHV-6A and –B with maximum bootstrap support (Fig. 1). We concluded that in all HHV-6 positive individuals only a single HHV-6 species (Table 1) was detected.

Target enrichment produces high coverage HHV-6 genomes. We obtained total DNA from the LCL from the 9 of the 11 HHV-6-positive individuals which were available from Coriell Institute for Medical Research. We performed target enrichment of the viral genome using baits designed to capture the two species of HHV-6. Captured DNA was multiplexed and sequenced on a single Illumina MiSeq System flow-cell. Reads from each sample were mapped against the two HHV-6's species references. Median coverages obtained ranged from 77x to 229x. Lastly, we checked for species-specific Single Nucleotide Variants (SNVs) within the U65 gene for both HHV-6A and HHV-6B[63], confirming in all cases our assignment predictions, and the absence of the two species in a single individual.

To avoid ambiguities in variant calling, we masked long repeats and low-complexity regions (>500 bp), which resulted in the masking of the two long terminal Direct Repeats DR_L and DR_R (left and right), and the internal ones R1, R2 and R3 in both species, plus R0 for the B species (Repeats coordinates shown in Supplementary Table S1). That left a total of 138,144 bp and 140,628 bp for HHV-6A and HHV-6B respectively (~87% of the whole genome for both species) for further analysis.

Sequences were then scanned for large structural variation, but no signs of duplications, inversions or large insertions/deletions were found (Supplementary Figure 2).

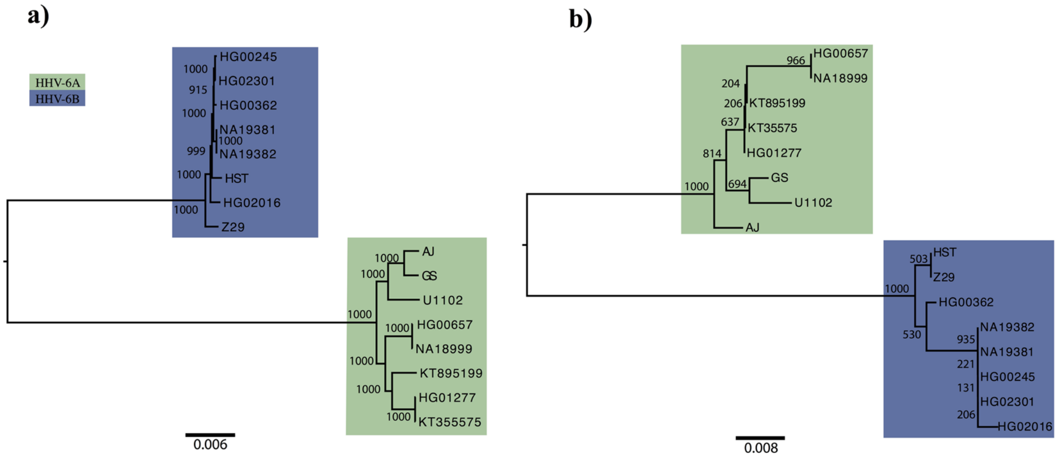

Figure 1. Species assignation through Neighbor-joining tree. The trees were built using the whole HHV-6 data set (**a**), and using only the U83 ORF (**b**). The different virus species are colour-coded, showing the clear separation between HHV-6A and HHV-6B strains. U83-based tree presents lower bootstraps due to the low number of variants present in the limited U83 region.

		1000 GP low-coverage data			Target enrichment data	
Individual	**Sample origin**	**Covered Genome (%)**	**Infecting Variant**	**Infecting/ Alternative**	**HHV-6A**	**HHV-6B**
iciHG00245	**UK**	**56**	B	1.68	12	**125**
iciHG00362	**Finland**	**72**	B	2.04	15	**150**
iciHG01058	**Puerto Rico**	**42**	B	2.04	Nd*	Nd*
iciHG01162	**Puerto Rico**	**62**	B	1.98	Nd*	Nd*
iciHG02016	**Viet Nam**	**58**	B	2.06	18	**159**
iciHG02301	**Peru**	**80**	B	1.92	13	**141**
iciNA19381	**Kenya**	**62**	B	2.04	11	**89**
iciNA19382	**Kenya**	**69**	B	1.97	9	**94**
iciHG00657	**China**	**50**	A	1.86	**229**	33
iciHG01277	**Colombia**	**68**	A	1.81	**140**	16
iciNA18999	**Japan**	**61**	A	1.70	**77**	9

Table 1. Median viral coverage before and after target enrichment. Results for HHV-6 mapping performed on the 1000 Genome Project selected individuals (on the left), and on the sequences generated through target enrichment. On the left the length of the complete virus genome covered after quality filters is shown as percentage of the whole genome. Proportion of mapping reads between the species for which there is a higher number of mapping reads (dominant) and the alternative one (alternative) is also shown. On the right the median values for the whole genome are shown for both species. First-degree family members are marked in italic. *No data due to unavailability of biological sample for target enrichment.

We further masked base pairs with low depth of coverage or extreme allele balance for each sample (see "Materials and method section"). In order to do comparative analyses, we calculated the "comparable region", also known as common "callable" region, defined as all the bases of the virome that were covered after the quality filters by all individuals in the data set. The resulting comparable region showed lowly fragmented genomes (135 fragments for A and 94 for B) that account for 83.0% of the HHV-6B complete genome (95.2% of the repeat-masked reference), and for 74.0% of the HHV-6A complete genome (85.4% of the repeat-masked reference).

We produced *de novo* assemblies of our sequences and compared the variant calling reference-based and the one produced from aligning assemblies, obtaining an almost perfect overlap. We manually curated the few ambiguous variants and deposited the final sequences in Genbank.

Inherited chromosomally integrated HHV-6 identification.

The LCLs from the sequenced samples were tested for the presence of iciHHV-6. Following the method proposed by Sedlack *et al.*[25], we performed droplet digital PCR (ddPCR) in order to quantify the absolute number of HHV-6 copies and human genome copy. If the virus was present in an inherited form, we would expect to have a copy of its genome per each cell. The results showed around 1 HHV-6 copy per cell (range: 0.92–1 copy/cell), supporting the inherited condition of these viruses. While very unlikely, the clonal expansion of a lymphocyte bearing integration of non-inherited form of HHV-6 cannot be excluded, as it would produce similar results in a ddPCR analysis.

HHV-6A and HHV-6B species show different diversity patterns. We integrated all available HHV-6 complete sequences published up to the date of performing this study to our dataset and compared intra-species diversity values. A second data set, referred to as iciHHV-6, was built selecting only the proven congenital sequences. All diversity analyses were repeated in this subset obtaining similar results unless we indicate otherwise.

Overall variability was calculated and corrected for the different number of sequences in the data sets using the Watterson estimator[64] Θ_w applied to the SNV density of each data set. The diversity within the A and B species resulted notably higher in the former. The ratio of the number of polymorphism was three-fold higher in HHV-6A compared to HHV-6B. The ratio was similarly higher in HHV6 in all genomic features analyzed (HHV-6A/HHV-6B ratios: exons = 3.12, introns = 2.73, UTR = 2.97, Unclassified = 2.89; iciHHV-6A/iciHHV-6B: exons = 3.60, introns = 4.51, UTR = 3.01, Unclassified = 2.93; all values expressed as ratio of Θ_w calculated on the number of SNV/bp). The absolute SNV density reflected this ratio, with 0.020 SNVs/bp for HHV-6A, compared to the 0.007 SNVs/bp of HHV-6B, and with 0.010 SNVs/bp for the iciHHV-6A, compared to the 0.003 SNVs/bp of the iciHHV-6B.

Within a single species, exon showed the highest level of constraints, where we measured the lowest SNV density, while introns showed the higher values (Supplementary Table S2).

Overall transition/transversion ratio (Ti/Tv) was high for both species (>2.3; Supplementary Table S3). The Ti/Tv ratio measured for both species were higher than for other *Herpesviridae* subfamilies, but similar to members of the *Bethaherpesvirus* subfamily such as Human herpesvirus 5 (HHV-5; a.k.a. Human cytomegalovirus (HCMV)), and Human herpesvirus 7 (HHV-7)[65].

Non-sense SNVs and the need of choosing a new reference sequence. SNVs that create or disrupt a stop codon can dramatically change the translation product and function of a gene. Only two SNVs of our datasets produced non-sense mutations. A stoploss in *U12* was present in all available HHV-6B sequences, except for the reference Z29, at the end of the gene's coding region. In order to support this statement, the U12 stoploss mutation was validated by Sanger sequencing in all the newly produced HHV-6B sequences. In HHV-6A, a different stoploss in *U47* was detected in 4 of the 8 sequences (4 of the 5 iciHHV-6A sequences). These were the only mutations found affecting translation, hinting together with the lack of large structural variations to the presence of essentially intact genes and potentially functioning viruses, in agreement with Zhang *et al.*[66].

It is remarkable that the non-sense variant in HHV-6B is only absent in the Z29 sequence. Additional evidences indicated that the Z29 is quite divergent from the rest of sequences; SNVs that are exclusive of the Z29 strain add up to 37% of the total number of SNVs in the HHV-6B comparable data set (337/905 SNVs), compared to the 15% (370/2442) of exclusive SNVs of the U1102 strain (HHV-6A reference) in the HHV-6A data set. The presence of stratification in HHV-6B populations could explain, to some extent, the divergence shown by Z29 but, importantly, the observation holds when considering only the strains with the same geographical origin.

HHV-6 population structure: stratification, phylogeny and recombination. We performed Principal Component Analysis (PCA) separately in HHV-6A and -6B. In HHV-6A the Asian sequences cluster closely together when plotting the first two components (Fig. 2a), of which variation explained sums up to 66.3% of the total. While AJ and U1102, both of African origin (Gambia and Uganda respectively), cluster together, AJ falls also very close to the North American strain GS. This similarity is consistent with what had been reported by Tweedy *et al.* in the study where the AJ sequence was first published[67]. We observed a clear separation between Asian and African sequences explained by the first principal component, a pattern found also in the sister Herpesviridae Human herpesvirus 4 (HHV-4)[68,69]. While the two European sequences (KT355575 from UK and KT895199 from Germany or Czech Republic) appeared relatively close, the former clustered with the South American iciHG01277 sequence, a pattern that was evident also in HHV-6B, but that could be explained by the reduced cohort of this study. HHV-6B shows little sign of geographic stratification, and the first principal component, explaining 39.1% of the variance, mainly separates the Z29 reference sequence from the rest of the data set, while the second principal component, explaining 22.9% of the variance, separates the Vietnamese sequence iciHG02016. The other sequences cluster together, with iciNA19381 and iciNA19382 appearing completely overlapped (Fig. 2c), an expected result due to the first-grade relationship of the individuals from where these two sequences derived[70].

We built NJ-phylogenetic trees for both HHV-6 species resulting in a highly robust tree for HHV-6A, where the two African strains clustered together, even though the North American GS strain remained the most similar sequence to AJ. The Asian cluster remains well supported and well separated from the other sequences (Fig. 2b). The separation between Asian and African strains is well defined in this analysis. The tree built on HHV-6B sequences, in contrast, shows little signs of population structure by geography. Only two of the same-origin sequences clustered together. These are the African iciNA19381 and iciNA19382, which comes from individuals with first-degree parent-son relationship[70]. Since the endogenous form is transmitted vertically, the two sequences are virtually identical, and are thus uninformative for this analysis. As seen in the PCA, the two European sequences do cluster with each other, but together to the South American one, a pattern that appeared less explicitly in HHV-6A. Finally, this tree confirmed the separation of the reference Z29 from the rest of HHV-6B sequences (Fig. 2d).

The lack of a clear geographical pattern suggests the action of recombination. Analyses on recombination were conducted separately for the two species, for which every sequence has been analysed one at a time as the putative recombination product of the rest of sequences (see "Materials and methods" section). The analysis on the HHV-6A confirmed the genome-wide similarity of the Asian strains compared to the others, and the high whole genome identity between the GS and AJ strains. Interestingly, when the African U1102 sequence is analysed, patterns of similarities are evident with both the other African strain, AJ, and with the North American strain,

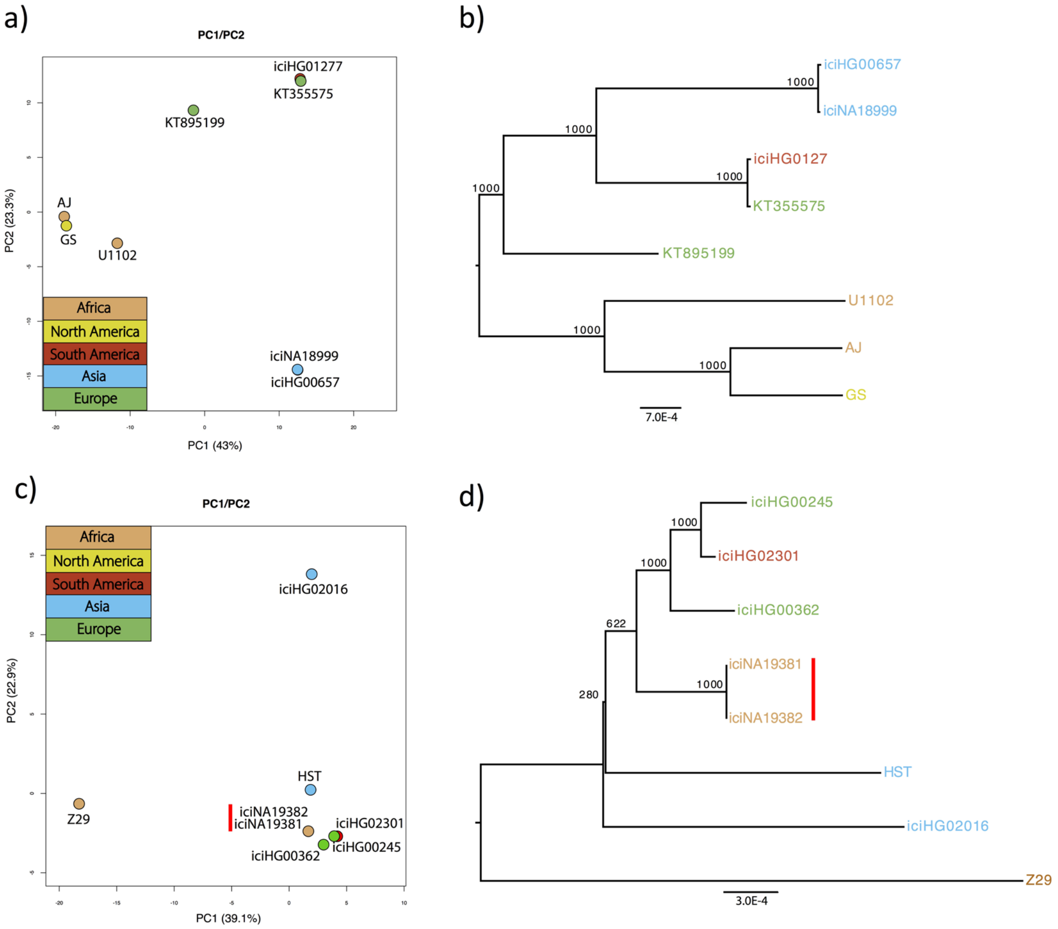

Figure 2. Geographical stratification in HHV-6. (**a**) PCA and (**b**) Neighbour joining tree, built using all the SNVs in HHV-6A strains. (**c**) PCA and (**d**) Neighbour joining tree, built using all the SNVs in HHV-6B strains. Individuals are color-coded based on the geographical origin. The vertical red line indicates the first-degree family members duo.

GS. The latter presents similarity mainly with the two African sequences, sign of a possible African origin for this strain (Fig. 3a–c). The same analysis performed on the B species data set showed that Z29, the African sequence separated from all the other sequences in previous analysis, is very similar to the other African sequences compared to the rest of the data set (Fig. 3d). It should be noted that iciNA19381 and iciNA19382, the two congenital sequences derived from parent and son, have identical SNVs, so using them together in a block similarity analysis that scale itself on the minimum block distance would mask their putative similarity with the other sequences. To avoid this, we used only one of the two sequences, iciNA19381, as representative of the family. The region between 90–130 kbp is where most of the significant recombination breakpoints lie, which translates into stronger influence of this region in analysis such as PCA and phylogeny and, possibly, in concealing of clusters such as the African one. To confirm this, we constructed another tree excluding the 90–130 kb region. Remarkably, this resulted in the African sequences clustering together, as shown in Supplementary Figure S3.

The recombination analysis plots for each sequence of the data set are shown in Supplementary Figures S1 and S2.

Selection footprints in coding genes: stronger selection constraints in HHV-6B compared to HHV-6A.

We measured rates of molecular evolution in protein coding genes in the iciHHV-6 data set for both species. Global dN/dS values (ω) were estimated for each gene, providing average ω values of 0.39 and 0.11 for iciHHV-6A and iciHHV-6B, respectively. While this value falls within the boundaries of the average human value for iciHHV-6B (average: 0.13, standard deviation: 0.11[71]), iciHHV-6A shows a remarkably higher value. The difference observed between iciHHV-6A and iciHHV-6B was significant (Fig. 4; Student's T-test p-value, 0.00008).

Signals of positive or purifying selection were explored for single genes independently in both species. As described before, diversity within iciHHV-6B sequences was lower compared to iciHHV-6A. The number of SNVs in each gene was often low, leading to less robust estimations. We used different models in codeml to test for significantly high or low ω estimates, comparing the models using likelihood ratio tests (LRTs). We obtained the likelihood of a free-estimated ω with that obtained from an identical model but where ω was fixed at the average value reported above. This would identify an ω value that is at the tail of distribution. A second LRT was performed between the likelihood of the model leaving ω free for estimation, and the same model were ω was

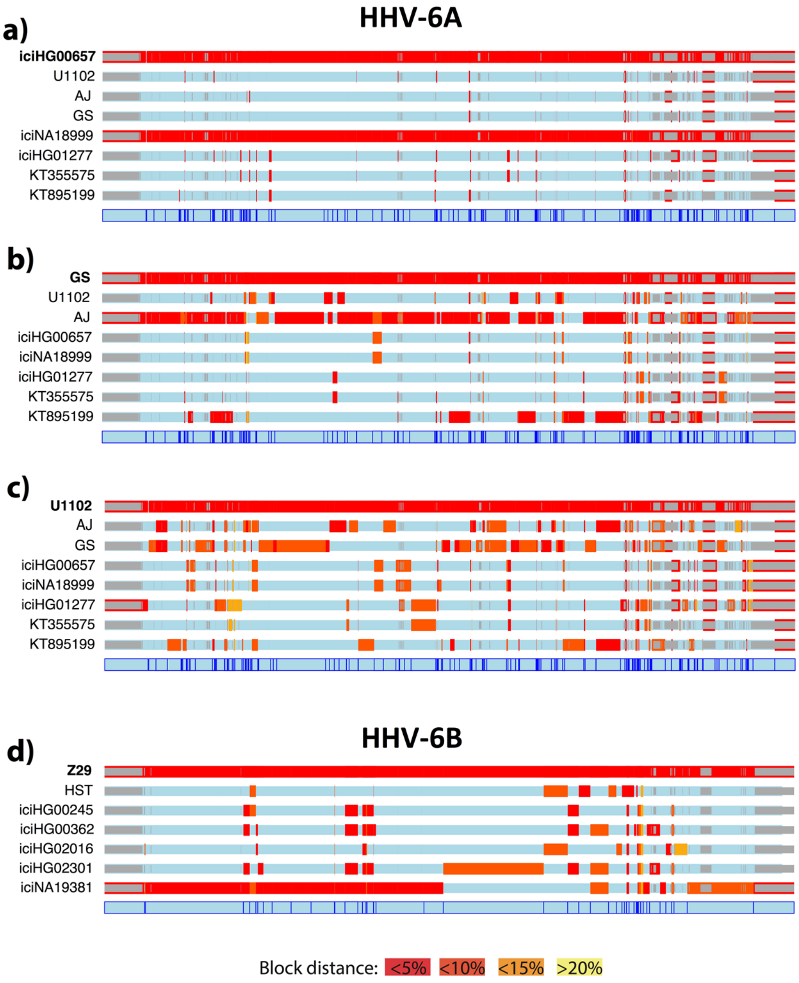

Figure 3. Recombination footprints. Recco analysis of the most cryptic HHV-6A (**a**–**c**), and HHV-6B (**d**) strains. The first row in each panel shows the analysed sequence, considered a recombination product of the rest of the species data set. The colour scale shows the proportional genetic distance, as shown in the appendix at the bottom of the figure. The masked regions of the genome are marked in grey.

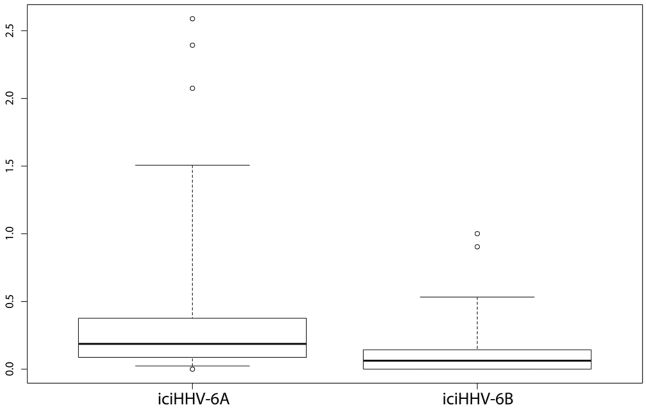

Figure 4. Selection footprints. The boxplots shows the gene ω values in iciHHV-6A and iciHHV-6B.

fixed to 1, a value considered a threshold to identify positive selection (Table 2). Two genes stood out for high dN/dS ratio in iciHHV-6B: *U90* and *U100*, with an ω value of 0.90, and 1.00 respectively. iciHHV-6A showed more diversity than its sister species, increasing the power to detect signatures of selection. There was no clear pattern of sequence constraints when observing the distribution of the ω values along the genome, with the exception of the IE-A region ω values tended to be higher. The highest ω values supported by LTR analysis was obtained for

Gene	Omega (ω)	LRT(est.ω VS avg ω)	LRT(est.ω VS ω = 1)
iciHHV-6A			
U11	0.9704	0.025957	0.986413
U24	1.3613	0.045580	0.986414
U47	1.0043	0.005118	0.986415
U90	1.0139	0.008566	0.986416
U95	2.3926	0.000001	0.003436
U4	0.0666	0.035232	3.40E-06
U24A	0.0001	7.06E-05	5.51E-08
U33	0.0232	0.007389	4.52E-06
U39	0.0799	0.027790	1.22E-06
U48	0.0860	0.003596	1.09E-09
U51	0.0597	0.020087	1.78E-06
U56	0.0001	0.045580	0.000595
U63	0.0001	0.004029	1.47E-05
icIHHV-6B			
U90	0.903	0.006543	0.998871
U100	1.000	0.027501	0.998871

Table 2. Genes showing strong selection footprint. Genes with estimated ω significantly different from the average value of the species is shown. The LRT results between a model that estimates ω for each branch of the tree and the same model with ω fixed at the average value of the species, and fixed at 1, are reported.

U95 (ω = 2.39), the only gene showing clear signs of positive selection. Signs of accelerated evolution were found in genes *U11*, *U24*, *U47* and *U90*, with ω estimates of 0.97, 1.00, 1.36, and 1.01 respectively (see Table 2). This is consistent with Dominguez *et al.* findings[4], where *U90* and *U95* were also pointed out as the genes with highest Ka/Ks. Significantly low ω values were found in various genes: *U24A*, *U56*, *U63*, *U33*, *U51*, *U4*, *U39*, *U48*, here ordered *per* lower value of ω.

Discussion

In the present study we performed a genome-wide comparative analysis on HHV-6A and HHV-6B sequences. We report here nine new iciHHV-6 sequences derived from healthy 1KGP individuals. Using targeted genomic enrichment, we could generate deep-coverage iciHHV-6 genomes. These new sequences combined with previously published HHV-6 sequences, make it possible to perform phylogenetic, recombination and selection analysis. This data set has been used to assess variability within, and compare it between, HHV-6A and HHV-6B, and iciHHV-6A and iciHHV-6B. We have also explored geographical stratification and proposed a set of genes as candidates to be under trends of either accelerated evolution or positive, or purifying selection.

Mutation rates become an important factor when analysing virus variability, and its estimation would help to pinpoint the divergence time between species, and to explore the possibility of a single or multiples germ-line integration events. In our data set we presented the sequences of iciHHV-6B belonging to two individuals with a first-degree relationship (iciNA19381, iciNA19382), which could be used to infer the mutations fixed per generation. The two-sequence showed one single different SNV after variant calling, where iciNA19382 presents the mutation, while iciNA19381 does not. A Sanger sequencing of an amplicon covering the position of the SNV performed on both individuals showed that the mutation was indeed present in both. The family duo thus resulted identical when the SNVs were compared, precluding the estimation of a mutation rate. The lack of SNVs between the two sequences is nevertheless expected when considering the estimation of the mutation rate described for other herpesviruses[72,73], and the relatively short length of the HHV-6 genome. More of such duos would make it possible to estimate a mutation rate per generation for iciHHV-6.

SNV densities were statistically significantly different between the A and B species, in both the whole HHV-6 data set and the iciHHV-6 subgroup, where the variation was homogeneously distributed along the genome. The difference in divergence is consistent with what has been reported by Tweedy *et al.* for iciHHV-6[74].

Analysis made on SNV and their frequency distribution rose doubts on the convenience of keep using Z-29 sequence as a reference genome for HHV-6B; this sequence showed a disproportionate number of private mutations compared to the rest of sequences from the same or different continents. The HHV-6A reference strain instead (U1102), shows comparable number of private mutations. Non-sense mutations support this hypothesis, showing in our dataset that while in HHV-6A the stoploss in *U47A* is equally probable than the reference allele, the elongated form of *U12* in HHV-6B was missing only in Z29, indicating that the event could rather represent a stop-gain occurred in the reference strain than a stop-loss occurred in all other sequences. The change in *U12* length is notable, with the two coding regions of this gene fusing into one. *U12* encodes a G protein-coupled receptor involved ultimately in cellular response. No functional consequence of the change in length of this protein can be hypothesised due to the limited information available. An additional hint of the dissimilarity of the reference sequence Z29 can be found in the study by Stanton *et al.*[75], in which two HHV-6B groups are described based on the variation within the *IE1* gene in 14 sequences. The two groups clustered around Z-29 or HST. Notably, the HST-like group was much larger than Z-29, including 10 of the 14 sequences of the data set. As a whole, and taking into consideration the whole set of published HHV-6B genomes, these evidences support that Z-29 is not a common sequence, and that HST, or any of our newly produced sequences, would be a better candidate to serve as reference sequence for HHV-6B.

Previous studies have interrogated putative HHV-6 genetic stratification[36,55,76], but none of these studies ever attempted the analysis at a whole-genome scale. We found evidence of a solid Asian cluster in HHV-6A. The reference U1102 and the AJ strain, from Uganda and Gambia respectively, show strong relationship whereas the GS strain from North America is highly similar to AJ, as previously described[67]. Nevertheless, the African and Asian strains appeared well separated from each other, as described in the sister *Herpesviridae* Human herpesvirus 4[68,69]. The European sequences, from Germany or Czech Republic, and from UK, show low divergence, but the latter cluster deeply with the South American strain. This pattern could be driven by recombination, or be a consequence of the low resolution of our analysis with a small data set.

On the contrary, no clear geographical pattern was observed in HHV-6B, with the two European strains clustering together, but with the UK sample showing more similarity to the South American one. Nevertheless, the other European sample originates from Finland, a population that is known to segregate from the other population of the same continent[77,78]. The African family iciNA19381-iciNA19382 also clusters near the Europeans and South American sequences. Recombination events among strains from different geographic origin might be masking global signals of stratification. Recombination analysis indicated that the Z-29 sequence from Africa was the most similar to the other African sequence in a sizeable part of the evaluated genome, supporting the presence of an African component. Sample size is key in this kind of analyses and our low resolution precludes us from exploring in full detail the presence and structure of geographical stratification in these viruses.

Evidence of accelerated evolution was found in a higher number of genes in iciHHV-6A compared to iciHHV-6B. We listed the genes with strong footprints of a history of accelerated or negative selection. The two genes showing significant evidence of accelerated evolution in iciHHV-6B were *U90* and *U100*, for both of which the results of the LRTs showed significant differences between the estimated ω and the average for the species, but almost no sign of difference when the comparison was against the model using a fixed ω of 1 (Table 2). Both genes had already been described by Dominiguez *et al.*[4] as presenting high dN/dS (>1) when comparing HHV-6A and HHV-6B reference sequences. The protein encoded for by *U90* is an Immediate-Early transactivator, putatively regulating RNA replication, transcription and modification during the first phase of HHV-6 active infection[4]. *U100* instead encodes for the gp82-gp105[79,80], a glycoprotein related to CD46 substrate binding and fusion, and is thus involved in the different tropism of HHV-6 species. In iciHHV-6A, where selection footprints were more distinct due to the higher variability, evidence for signals of positive selection were obtained for *U95*, while *U24*, *U11*, *U47* and *U90* showed signs of accelerated evolution. Two of the five genes have already been detected as having high dN/dS values when comparing the HHV-6A and -6B references[4] (U90, U95). In the same study, also *U54* and *U91* showed high dN/dS values, genes for which our estimated ω was high (>1.5), although in our set of sequences these high values were not statistically significant. *U24* is an important gene related to the host immune response, inhibiting T-cells activation[81], and blocking early endosomal recycling[82], thus minimizing immune recognition. U11 is instead a necessary antigenic protein for virion reconstitution[83]. Lastly, U95 is also a putatively important gene, having been shown to indirectly modulate cell death signals in HHV-6B[84].

One of the most constrained genes in iciHHV-6A was expectedly *U48*, the gene encoding for the glycoprotein H of the gL-gH-gQ complex, and that is known to be conserved across the *Herpesviridae* family.

Materials and Methods

Individuals showing HHV-6 integration and variant identification.

We downloaded Illumina-sequenced, paired-end reads from the individuals of the 1KGP Phase 3 non-mapping to the human genome reference (b37). We then mapped these reads against HHV-6A and B reference strains (NCBI accession numbers: NC_001664.2 and NC_000898.1 respectively) using Burrow-Wheelers Aligner[85] (BWA) and setting the phred trimming quality threshold to 15. Only uniquely mapping reads were selected for HHV-6 genome reconstruction and comparative analysis. BAMs were masked for low-complexity and main repeats regions using NCBI annotations, RepeatMasker[86], and Tandem Repeat Finder[87]. Duplicated reads were removed using the SAMtools package rmdup function[85]. SNVs and Insertion/Deletions callings were performed on the obtained alignments using VarScan 2[88] set with default parameters. The resulting SNVs files have been used as known polymorphic sites for quality score recalibration of the alignments themselves through Genome Analysis ToolKit[89] (GATK). Recalibrated files were re-mapped with the same settings, and depth of coverage was calculated using GATK's DepthOfCoverage function. Individuals were considered as presenting HHV-6 latency when the median coverage along the whole masked virus genome was above or equal to 2.

HHV-6A and -6B typing was achieved by checking for species-specific variants in the U65 open reading frame. Starting from position 101,548 of the reference sequence U1102 coordinates, the HHV-6A species was identified by 5′-**T** TGT G**T**G TT**G** TTT T**A**-3′. The HHV-6B species was identified by 5′-**G** TGT G**C**G TT**A** TTT T**C**-3′ starting from position 102,848 of the reference sequence Z29 coordinates. In order to avoid incorrect species assignment, we mapped the reads belonging to each individual against both HHV-6A and HHV-6B species, and calculated the relative coverage ratios along the whole genome. We did not found any case with ambiguous variant assignment.

Whole-genome target enrichment and repository upload.

Available DNA samples belonging to 1KGP individuals presenting HHV-6 integration were retrieved from established blood-derived LCL at Coriell Institute for Medical Research (Camden, NJ, USA).

Illumina sequencing libraries were produced following Illumina TruSeq® DNA Library Prep LT protocol starting from 1 µg of the retrieved LCL DNA per sample. Library preparation was followed by target enrichment using custom RNA baits covering more than 99% of the two HHV-6 species genomes (SeqCap EZ Library

SR, Nimblegen). The Genomic Unit of the Centre for Genomic Regulation (CRG, Barcelona, Spain) performed library preparation and target enrichment experiments. The resulting libraries were sequenced on an Illumina MiSeq System at Pompeu Fabra University (UPF, Barcelona, Spain) Genomic Core Facility using the 300-cycle MiSeq reagents kit V2 (Illumina), and base calling was performed using the MiSeq Reporter Illumina pipeline.

The sequences obtained by target enrichment were scanned for repeats longer than 500 bp, which were masked from the sequence and excluded from all analyses. The catalogue of repeats was obtained by merging to the NCBI reference sequences annotated ones the results of Repeat Masker, and Tandem Repeat Finder.

Reads were processed and mapped following the same protocols as above (see "Individuals showing HHV-6 integration and species identification" section). Using the information provided by the paired-end mapping we interrogated large structural variation (such as insertions or deletions, inversions and duplications). Inversions: Identifying clustered mate reads with same orientation, with depth of coverage falling within 3 standard deviations from the mean of the sample. Insertions/Deletions: Insert size distributions were calculated for each individual. Insert size peaks separated from the average insert size by at least 3 standard deviation were considered as putative large deletions or insertions. Putative insertions were scanned for the presence of high-density SNV clusters. Large insert sizes due to circularity of the viral genomes were manually excluded. Duplications: Identifying regions with depth of coverage doubling the median value in windows of 1000 bp (shift of 100 bp).

Genomic GC content was calculated in the same 1000 bp sliding windows mentioned above, using the GC function of the R package Seqinr[90].

Inherited chromosomally integrated HHV-6 identification.

The samples were tested for inherited HHV-6 integration presence using digital droplet PCR (ddPCR). The experiments followed the method proposed by Sedlack *et al.*[25], starting from 150 ng of sample DNA per reaction.

De novo assemblies.

Contigs for each individual were produced starting from the raw sequencing reads using Discovar *de novo* assembler[91]. The contigs were filtered for a minimum length of 1000 bp, and their reference-based coordinates were determined aligning them to the species reference genome using Geneious V5.6.7[92]. Variants not supported by the reference-based variant calling were manually corrected. A final reference-based scaffolding was achieved using Ns for gaps, and annotations were transferred from the specie reference sequence to the assemblies using RATT[93].

Variability analysis.

Variant calling was performed using VarScan 2, setting a minimum base coverage of 20, with the criterion that the loss of the callable genome was less than 15%. Variants were additionally filtered for extreme allele balance (frequency $< 5\%$). Transition/transversion ratios were calculated using vcftools[94].

Diversity values between species were obtained by counting the number of SNV divided by masked feature length, and correcting for the number of sequences with the Watterson estimator.

Variability across the genome was calculated by counting SNVs in 10 Kbp bins of the repeat-masked genomes.

Genomic intervals of the two viruses that were common across samples in terms of 'callability' were calculated and all comparative analysis (*i.e.* PCA, phylogenetic trees generation, recombination, variability patterns and selection footprints) were performed using this common set of positions.

Sanger sequencing SNV validations.

The first-degree relationship individuals present in our data set, iciNA19381 and iciNA19382, showed a single SNV of difference after the variant calling. Being the locus one with decreased base quality compared to the neighbouring regions, we checked the status of the SNV sequencing through Sanger technology the position in both individuals.

A 500 bp amplicon was amplified in 25 µl reactions using the primers FamMutFWD = 5′- GGACATCTCTTT GTTGTGTGCC-3′ and FamMutREV = 5′-GGCTGGTATTAGAACAATTAGGACA-3′ (20 mM Tris HCl, 50 mM KCl, 2 mM MgCl2, 200 µM dNTPs, 1.5 U BioTherm Taq DNA Polymerase (GeneCraft), 900 nM each of primer, 50ng sample). The PCR reaction was purified using QIAquick Purification Kit, and run on a 2% agarose gel to control the amplicon length. Each amplicon was then sequenced separately two times, one for each primer. This was performed following the Big Dye Terminator v3.1 kit (Thermo Fisher) guidelines, starting from 10 ng of the purified PCR reaction.

Following the same protocol, the stoploss found in all newly generated sequences in the U12 ORF was validated. The primers were: U12stlssFWD = 5′-GTAAGCAGACCGAAAGTAAAAC-3′ and U12stlssREV = 5′-TAGAGAAACAATGTACCTGTGG-3′.

Population structure: stratification, phylogeny and recombination.

All HHV-6 complete sequence published at the beginning or during the performance of this project were added to our newly assembled sequences for a comprehensive comparative analysis, for a total of 8 sequences for each HHV-6 species. The reduced list of HHV-6 sequences included U1102[95–97] (GenBank Accession number: NC_001664.2), GS[98] (GenBank Accession number: KC465951.1), AJ[67] (GenBank Accession number: KP257584.1), KT895199[74] (GenBank Accession number: KT895199.1) and KT355575[74] (GenBank Accession number: KT355575.1) for the A species, and Z29[4] (GenBank Accession number: NC_000898.1) and HST[5] (GenBank Accession number: AB021506.1) for the B species.

PCA was performed using the Prcomp function in Stats R package, and plotted using basic R plot functions[99]. Partial genome sequences were aligned with MAFFT[100] and phylogenetic analysis were performed on transformed sequences obtained by substituting each identified SNV to the correspondent base in the reference genomes. Because members of the Herpesviridae family are well known to undergo homologous recombination among viromes[101–103], the phylogenetic trees were built using neighbour joining algorithms, with no assumptions regarding recombination. Using the MEGA7 software[104], the alignments were read, the genetic distances and, the neighbour-joining trees were computed, and the bootstrap values calculated. Trees were rooted at midpoint, and bootstrap analysis were set at 1000 replicates per tree. The trees were visualized using FigTree. Recombination breakpoint calculations were performed with Recco[105]; genetic identities between predicted recombinant blocks and plots were performed as in Santpere *et al.*[68].

Selection on coding genes. The phylogenetic trees generated for the enriched genomes in the previous phase were used to calculate dN/dS values for each gene of the two viral strains using the codeml program implemented in PAML package[106], together with the sequences available in NCBI for the published strains. codeml was set to run applying M0 model (Nsites = 0) with ω estimated for every branch of the tree, and the same model fixing ω at 1 and at the average of the estimated ω values for every gene. Genes that presented no SNV were excluded from the analysis of the diversity between species in ω distributions, leading to the exclusion of 9 genes in iciHHV-6A, and 18 in iciHHV-6B.

We obtained maximum likelihoods values for every gene using the aforementioned different models and LRTs was used to compare them. P-values were calculated from the cumulative chi-square distribution with a number of degree of freedom equal to the difference in free parameters in the model (1 in our case). In order to correct for multiple comparisons and limit the false discovery rate, we applied Benjamini & Hochberg method[107] to our p-values using the P.adjust function of the Stats package bundled with R. The comparison between ω distributions in iciHHV-6A and iciHHV-6B was plotted using the Adjbox function of the Robustbase R package[108].

SNV functional annotation, including coding stop-gain/loss, was performed using Annovar[109].

Data availability. Target enrichment Illumina raw reads are available at the Short Read Archive under the BioProject ID: PRJNA412600. Assembled sequences are available at GenBank and identified by the accession number MG894368-76.

Ethical approval and informed consent. The present study involves human genotyping data made publically available by the 1KG project with no need of ethics approval. It also involved LCL from Coriell Institute, to obtain the samples from Coriell we produced the required Statement of Research and Assurance Form for Biomaterials approved by the Institutional Official of the Pompeu Fabra University.

References

1. Ihira, M. *et al*. Serological examination of human herpesvirus 6 and 7 in patients with coronary artery disease. *J. Med. Virol.* **67**, 534–537 (2002).
2. Okuno, T. *et al*. Seroepidemiology of human herpesvirus 6 infection in normal children and adults. *J. Clin. Microbiol.* **27**, 651–3 (1989).
3. Saxinger, C. *et al*. Antibody reactivity with HBLV (HHV-6) in U.S. populations. *J. Virol. Methods* **21**, 199–208 (1988).
4. Dominguez, G. *et al*. Human herpesvirus 6B genome sequence: coding content and comparison with human herpesvirus 6A. *J. Virol.* **73**, 8040–52 (1999).
5. Isegawa, Y. *et al*. Comparison of the complete DNA sequences of human herpesvirus 6 variants A and B. *J. Virol.* **73**, 8053–63 (1999).
6. Ablashi, D. *et al*. Classification of HHV-6A and HHV-6B as distinct viruses. *Arch. Virol.* **159**, 863–870 (2014).
7. Arbuckle, J. H. *et al*. The latent human herpesvirus-6A genome specifically integrates in telomeres of human chromosomes *in vivo* and *in vitro*. *Proc. Natl. Acad. Sci.* **107**, 5563–5568 (2010).
8. Daibata, M., Taguchi, T., Taguchi, H. & Miyoshi, I. Integration of human herpesvirus 6 in a Burkitt's lymphoma cell line. *Br. J. Haematol.* **102**, 1307–13 (1998).
9. Luppi, M. *et al*. Three cases of human herpesvirus-6 latent infection: integration of viral genome in peripheral blood mononuclear cell DNA. *J. Med. Virol.* **40**, 44–52 (1993).
10. Lusso, P. *et al*. *In vitro* cellular tropism of human B-lymphotropic virus (human herpesvirus-6). *J. Exp. Med.* **167**, 1659–70 (1988).
11. Takahashi, K. *et al*. Predominant CD4 T-lymphocyte tropism of human herpesvirus 6-related virus. *J. Virol.* **63**, 3161–3 (1989).
12. Cermelli, C. *et al*. Growth of human herpesvirus 6 in HEPG2 cells. *Virus Res.* **45**, 75–85 (1996).
13. Chen, M. *et al*. Human herpesvirus 6 infects cervical epithelial cells and transactivates human papillomavirus gene expression. *J. Virol.* **68**, 1173–8 (1994).
14. He, J., McCarthy, M., Zhou, Y., Chandran, B. & Wood, C. Infection of primary human fetal astrocytes by human herpesvirus 6. *J. Virol.* **70**, 1296–300 (1996).
15. Luka, J., Okano, M. & Thiele, G. Isolation of human herpesvirus-6 from clinical specimens using human fibroblast cultures. *J. Clin. Lab. Anal.* **4**, 483–6 (1990).
16. Wallaschek, N. *et al*. The Telomeric Repeats of Human Herpesvirus 6A (HHV-6A) Are Required for Efficient Virus Integration. *PLoS Pathog.* **12**, e1005666 (2016).
17. Gompels, U. A. & Macaulay, H. A. Characterization of human telomeric repeat sequences from human herpesvirus 6 and relationship to replication. *J. Gen. Virol.* **76**, 451–458 (1995).
18. Torelli, G. *et al*. Targeted integration of human herpesvirus 6 in the p arm of chromosome 17 of human peripheral blood mononuclear cells *in vivo*. *J. Med. Virol.* **46**, 178–88 (1995).
19. Daibata, M., Taguchi, T., Nemoto, Y., Taguchi, H. & Miyoshi, I. Inheritance of chromosomally integrated human herpesvirus 6 DNA. *Blood* **94**, 1545–9 (1999).
20. Tanaka-Taya, K. *et al*. Human herpesvirus 6 (HHV-6) is transmitted from parent to child in an integrated form and characterization of cases with chromosomally integrated HHV-6 DNA. *J. Med. Virol.* **73**, 465–73 (2004).
21. Clark, D. *et al*. Transmission of integrated human herpesvirus 6 through stem cell transplantation: implications for laboratory diagnosis. *J. Infect. Dis.* **193**, 912–6 (2006).
22. Nacheva, E. P. *et al*. Human Herpesvirus 6 Integrates Within Telomeric Regions as Evidenced by Five Different Chromosomal Sites. *J. Med. Virol.* **80**, 1952–1958 (2008).
23. Strenger, V., Aberle, S. W., Nacheva, E. P. & Urban, C. Chromosomal integration of the HHV-6 genome in a patient with nodular sclerosis Hodgkin lymphoma. *Br. J. Haematol.* **161**, 594–595 (2013).
24. Endo, A. *et al*. Molecular and virological evidence of viral activation from chromosomally integrated human herpesvirus 6A in a patient with X-linked severe combined immunodeficiency. *Clin. Infect. Dis.* **59**, 545–548 (2014).
25. Sedlak, R. H. *et al*. Identification of Chromosomally Integrated Human Herpesvirus 6 by Droplet Digital. *Clin. Chem.* **60**, 765–772 (2014).
26. Bell, A. J. *et al*. Germ-line transmitted, chromosomally integrated HHV-6 and classical hodgkin lymphoma. *PLoS One* **9**, 9–15 (2014).

27. Arnoult, N. & Karlseder, J. Complex interactions between the DNA-damage response and mammalian telomeres. *Nat. Struct. Mol. Biol.* **22**, 859–866 (2015).
28. Lazzerini-denchi, E. & Sfeir, A. Stop pulling my strings — what telomeres taught us about the DNA damage response. *Nat. Rev. Mol. cell Biol.* **17**, 364–378 (2016).
29. Huang, Y. *et al*. Human telomeres that carry an integrated copy of human herpesvirus 6 are often short and unstable, facilitating release of the viral genome from the chromosome. *Nucleic Acids Res.* **42**, 315–27 (2014).
30. Kim, W. *et al*. Regulation of the human telomerase gene TERT by telomere position effect over long distances (TPE-OLD): Implications for aging and cancer. *PLoS Biol.* **14**, e2000016 (2016).
31. Arbuckle, J. H. *et al*. Mapping the Telomere Integrated Genome of Human Herpesvirus 6A and 6B. *Virology* **442**, 3–11 (2013).
32. Hall, C. *et al*. Congenital infections with human herpesvirus 6 (hhv6) and human herpesvirus 7 (hhv7). *J. Pediatr.* **145**, 472–477 (2004).
33. Hall, C. B. *et al*. Chromosomal Integration of Human Herpesvirus 6 Is the Major Mode of Congenital Human Herpesvirus 6 Infection. *Pediatrics* **122**, 513–520 (2008).
34. Gravel, A., Hall, C. B. & Flamand, L. Sequence Analysis of Transplacentally Acquired Human Herpesvirus 6 DNA Is Consistent With Transmission of a Chromosomally Integrated Reactivated Virus. *J. Infect. Dis.* **207**, 1585–1589 (2013).
35. Griffiths, P. D. *et al*. Human herpesviruses 6 and 7 as potential pathogens after liver transplant: Prospective comparison with the effect of cytomegalovirus. *J. Med. Virol.* **59**, 496–501 (1999).
36. Leong, H. M. *et al*. The prevalence of chromosomally integrated human herpesvirus 6 genomes in the blood of UK blood donors. *J. Med. Virol.* **79**, 45–51 (2007).
37. Arbuckle, J. H. P. G. M. The molecular biology of human herpesvirus-6 latency and telomere integration. *Microbes Infect.* **13**, 731–741 (2011).
38. Zhang, E. *et al*. HHV-8-unrelated primary effusion- like lymphoma associated with clonal loss of inherited herpesvirus-6A from the telomere of chromosome 19q. 2–10, https://doi.org/10.1038/srep22730 (2016).
39. Yamanishi K., Okuno T., Shiraki K., Takahashi M., Kondo T., Asano Y., K. T. Identification Of Human Herpesvirus-6 As a Causal Agent for Exanthem Subitum. 1065–1067 (1988).
40. Hall, C. B. *et al*. Human Herpesvirus-6 Infection in Children – A Prospective Study of Complications and Reactivation. *N. Engl. J. Med.* **331**, 432–438 (1994).
41. Appleton, A. L. *et al*. Human herpes virus-6 infection in marrow graft recipients: role in pathogenesis of graft-versus-host disease. Newcastle upon Tyne Bone Marrow Transport Group. *Bone Marrow Transplant.* **16**, 777–82 (1995).
42. Dulery, R. *et al*. Early Human Herpesvirus Type 6 Reactivation after Allogeneic Stem Cell Transplantation: A Large-Scale Clinical Study. *Biol. Blood Marrow Transplant.* **18**, 1080–1089 (2012).
43. Ward, K. N., Andrews, N. J., Verity, C. M., Miller, E. & Ross, E. M. Human herpesviruses-6 and -7 each cause significant neurological morbidity in Britain and Ireland. *Arch. Dis. Child.* **90**, 619–23 (2005).
44. Berti, R. *et al*. Increased detection of serum HHV-6 DNA sequences during multiple sclerosis (MS) exacerbations and correlation with parameters of MS disease progression. *J. Neurovirol.* **8**, 250–256 (2002).
45. Sola, P. *et al*. Human herpesvirus 6 and multiple sclerosis: survey of anti-HHV-6 antibodies by immunofluorescence analysis and of viral sequences by polymerase chain reaction. *J. Neurol. Neurosurg. Psychiatry* **56**, 917–9 (1993).
46. Soldan, S. S. *et al*. Association of human herpes virus 6 (HHV-6) with multiple sclerosis: increased IgM response to HHV-6 early antigen and detection of serum HHV-6DNA. *Nat. Med.* **3**, 1394–7 (1997).
47. Maeda, A. *et al*. The evidence of human herpesvirus 6 infection in the lymphnodes of Hodgkin's disease. *Virchows Arch. A. Pathol. Anat.* **423**, 71–75 (1993).
48. Siddon, A., Lozovatsky, L., Mohamed, A. & Hudnall, S. D. Human herpesvirus 6 positive Reed-Sternberg cells in nodular sclerosis Hodgkin lymphoma. *Br. J. Haematol.* **158**, 635–643 (2012).
49. Shahani, L. HHV-6 encephalitis presenting as status epilepticus in an immunocompetent patient. *BMJ Case Rep*. **2014** (2014).
50. Yao, K. *et al*. Detection of human herpesvirus-6 in cerebrospinal fluid of patients with encephalitis. *Ann. Neurol.* **65**, 257–267 (2009).
51. Yoshikawa, T. *et al*. Exanthem Subitum-Associated Encephalitis: Nationwide Survey in Japan. *Pediatr. Neurol.* **41**, 353–358 (2009).
52. Bhattarakosol, P., Pancharoen, C., Mekmullica, J. & Bhattarakosol, P. Seroprevalence of anti-human herpes virus-6 IgG antibody in children of Bangkok, Thailand. *Southeast Asian J. Trop. Med. Public Health* **32**, 143–7 (2001).
53. Linhares, M. I., Eizuru, Y., Tateno, S. & Minamishima, Y. Seroprevalence of human herpesvirus 6 infection in Brazilian and Japanese populations in the north-east of Brazil. *Microbiol. Immunol.* **35**, 1023–7 (1991).
54. Nielsen, L. & Vestergaard, B. F. Competitive ELISA for detection of HHV-6 antibody: seroprevalence in a danish population. *J. Virol. Methods* **56**, 221–30 (1996).
55. Politou, M. *et al*. Seroprevalence of HHV-6 and HHV-8 among blood donors in Greece. *Virol. J.* **11**, 153 (2014).
56. Tolfvenstam, T. *et al*. Seroprevalence of viral childhood infections in Eritrea. *J. Clin. Virol.* **16**, 49–54 (2000).
57. Wu, Z., Mu, G. & Wang, L. Seroprevalence of human herpesvirus-6 in healthy population in two provinces of north China. *Chinese Med. Sci. J. =Chung-kuo i hsueh k'o hsueh tsa chih* **12**, 111–4 (1997).
58. Ascherio, A. & Munger, K. L. Epstein–Barr Virus Infection andMultiple Sclerosis: A Review. *J. Neuroimmune Pharmacol.* **5**, 271–277 (2010).
59. Huh, J. Epidemiologic overview of malignant lymphoma. *Korean J. Hematol.* **47**, 92–104 (2012).
60. Altshuler, D., Lander, E. & Ambrogio, L. A map of human genome variation from population scale sequencing. *Nature* **476**, 1061–1073 (2010).
61. 1000 Genomes Project Consortium, T. 1000 G. P. *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
62. Tweedy, J. *et al*. Analyses of germline, chromosomally integrated human herpesvirus 6A and B genomes indicate emergent infection and new inflammatory mediators. *J. Gen. Virol.* **96**, 370–89 (2015).
63. Boutolleau, D. *et al*. Identification of human herpesvirus 6 variants A and B by primer-specific real-time PCR may help to revisit their respective role in pathology. *J. Clin. Virol.* **35**, 257–263 (2006).
64. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–76 (1975).
65. Sijmons, S. *et al*. High-Throughput Analysis of Human Cytomegalovirus Genome Diversity Highlights the Widespread Occurrence of Gene-Disrupting Mutations and Pervasive Recombination. *J. Virol.* **89**, 7673–7695 (2015).
66. Zhang, E. *et al*. Inherited chromosomally integrated human herpesvirus 6 genomes are ancient, intact and potentially able to reactivate from telomeres. *J. Virol*. **44** (2017).
67. Tweedy, J. *et al*. Complete Genome Sequence of the Human Herpesvirus 6A Strain AJ from Africa Resembles Strain GS from North America. **3**, 1–2 (2015).
68. Santpere, G. *et al*. Genome-wide analysis of wild-type Epstein-Barr virus genomes derived from healthy individuals of the 1,000 Genomes Project. *Genome Biol. Evol.* **6**, 846–60 (2014).
69. Palser, A. L. *et al*. Genome Diversity of Epstein-Barr Virus from Multiple Tumor Types and Normal Infection. *J. Virol.* **89**, 5222–5237 (2015).
70. Pemberton, T. J., Wang, C., Li, J. Z. & Rosenberg, N. A. Inference of unexpected genetic relatedness among individuals in HapMap phase III. *Am. J. Hum. Genet.* **87**, 457–464 (2010).

71. Gayà-Vidal, M. & Albà, M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15**, 599 (2014).
72. Drake, J. W. & Hwang, C. B. C. On the mutation rate of herpes simplex virus type 1. *Genetics* **170**, 969–70 (2005).
73. Sanjuán, R. & Domingo-Calap, P. Mechanisms of viral mutation. *Cell. Mol. Life Sci.* **73**, 4433–4448 (2016).
74. Tweedy, J. *et al.* Complete genome sequence of germline chromosomally integrated human herpesvirus 6A and analyses integration sites define a new human endogenous virus with potential to reactivate as an emerging infection. *Viruses* **8** (2016).
75. Stanton, R., Wilkinson, G. W. G. & Fox, J. D. Analysis of human herpesvirus-6 IE1 sequence variation in clinical samples. *J. Med. Virol.* **71**, 578–84 (2003).
76. Potenza, L. *et al.* Prevalence of human herpesvirus-6 chromosomal integration (CIHHV-6) in Italian solid organ and allogeneic stem cell transplant patients. *Am. J. Transplant* **9**, 1690–7 (2009).
77. Bauchet, M. *et al.* Measuring European Population Stratification with Microarray Genotype. *Data.* **80**, 948–956 (2007).
78. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nat. Publ. Gr.* **536**, 285–291 (2016).
79. Pfeiffer, B. & Thomson, B. Identification and Characterization of a cDNA Derived from Multiple Splicing That Encodes Envelope Glycoprotein gp105 of Human Herpesvirus 6. **69**, 3490–3500 (1995).
80. Pfeiffer, B. *et al.* Identification and Mapping of the Gene Encoding the Glycoprotein Complex gp82-gplO5 of Human Herpesvirus 6 and Mapping of the Neutralizing Epitope Recognized by Monoclonal Antibodies. **67**, 4611–4620 (1993).
81. Sullivan, B. M. & Coscoy, L. Downregulation of the T-Cell Receptor Complex and Impairment of T-Cell Activation by Human Herpesvirus 6 U24 Protein. **82**, 602–608 (2008).
82. Sullivan, B. M. & Coscoy, L. The U24 Protein from Human Herpesvirus 6 and 7 Affects Endocytic Recycling. **84**, 1265–1275 (2010).
83. Mahmoud, N. F. *et al.* Human herpesvirus 6 U11 protein is critical for virus infection. *Virology* **489**, 151–157 (2016).
84. Yeo, W. M., Isegawa, Y., Chow, V. T. K. & Irol, J. V. The U95 Protein of Human Herpesvirus 6B Interacts with Human GRIM-19: Silencing of U95 Expression Reduces Viral Load and Abrogates Loss of Mitochondrial Membrane Potential. *J. Virol.* **82**, 1011–1020 (2008).
85. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
86. Smit, A. F. A., Hubley, R. & Green, P. Repeatmasker Open-4.0. http://www.repeatmasker.org. (2013–2015).
87. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–80 (1999).
88. Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
89. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–303 (2010).
90. Charif, D. & Lobry, J. R. In 207–232, https://doi.org/10.1007/978-3-540-35306-5_10 (Springer Berlin Heidelberg, 2007).
91. Weisenfeld, N. I. *et al.* Comprehensive variation discovery in single human genomes. *Nat. Genet.* **46**, 1350–1355 (2014).
92. Kearse, M. *et al.* Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
93. Otto, T., Dillon, G., Degrave, W. & Berriman, M. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res.* **39**, e57 (2011).
94. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
95. Downing, R. G. *et al.* Isolation of human lymphotropic herpesviruses from Uganda. *Lancet (London, England)* **2**, 390 (1987).
96. Gompels, U. A. *et al.* The DNA sequence of human herpesvirus-6: structure, coding content, and genome evolution. *Virology* **209**, 29–51 (1995).
97. Tedder, R. S. *et al.* A novel lymphotropic herpesvirus. *Lancet (London, England)* **2**, 390–2 (1987).
98. Gravel, A., Ablashi, D. & Flamand, L. Complete Genome Sequence of Early Passaged Human Herpesvirus 6A (GS Strain) Isolated from North America. *Genome Announc.* **1** (2013).
99. R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: the R Foundation for Statistical Computing, at http://www.r-project.org/ (2011).
100. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
101. Haberland, M., Meyer-Konig, U. & Hufert, F. T. Variation within the glycoprotein B gene of human cytomegalovirus is due to homologous recombination. *J. Gen. Virol.* **80**, 1495–1500 (1999).
102. Thiry, E. *et al.* Recombination in alphaherpesviruses. *Rev. Med. Virol.* **15**, 89–103 (2005).
103. Walling, D. M., Perkins, A. G., Webster-Cyriaque, J., Resnick, L. & Raab-Traub, N. The Epstein-Barr virus EBNA-2 gene in oral hairy leukoplakia: strain variation, genetic recombination, and transcriptional expression. *J. Virol.* **68**, 7918–26 (1994).
104. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7. 0 for Bigger Datasets Brief communication. **33**, 1870–1874 (2016).
105. Maydt, J. & Lengauer, T. Recco: recombination analysis using cost optimization. *Bioinformatics* **22**, 1064–1071 (2006).
106. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
107. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the royal statistical society series B* 289–300 at, https://www.jstor.org/stable/2346101?seq=1#page_scan_tab_contents (1995).
108. Rousseeuw, P. *et al.* robustbase: Basic Robust Statistics. At http://cran.r-project.org/package=robustbase/ (2015).
109. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164–e164 (2010).

Acknowledgements

This work was supported by Ministerio de Ciencia e Innovación, Spain (BFU2012-38236 and BFU2015-68649-P to A.N.), by Direcció General de Recerca, Generalitat de Catalunya (2014SGR1311 and 2014SGR866), by the Spanish National Institute of Bioinformatics (PT13/0001/0026) and the REEM (RD12/0032/0011, RD16/0015/0017) of the Instituto de Salud Carlos III and by FEDER (Fondo Europeo de Desarrollo Regional)/FSE (Fondo Social Europeo).

Author Contributions

M.T. designed and performed the experiments and analysis. G.S. conceived and supervised the project, designed the analysis and together with A.N. reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-21645-x.

Competing Interests: The authors declare no competing interests.

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