# Dynamics, evolutionary and epidemiological patterns of RNA viruses

**Juan Ángel Patiño Galindo**
**Doctorat en Biotecnologia**


VNIVERSITAT DE VALÈNCIA

**Directors: Fernando González Candelas, María Alma Bracho Lapiedra, Francesc Xavier López Labrador**

**Març 2017**


VNIVERSITAT DE VALÈNCIA

**Doctorat en Biotecnologia**

# Dynamics, evolutionary and epidemiological patterns of RNA viruses

**Juan Ángel Patiño Galindo, tesi Doctoral.**

**Març 2017**

**Directors: Fernando González Candelas, María Alma Bracho Lapiedra, Francesc Xavier López Labrador**

**Los Doctores Fernando González Candelas, María Alma Bracho Lapiedra y Francesc Xavier López Labrador, certifican que la presente memoria titulada "Dynamics, evolutionary and epidemiological patterns of RNA viruses" ha sido realizada bajo su dirección por D. Juan Ángel Patiño Galindo, con el fin de optar al grado de doctor en Biotecnología por la Universidad de València.**

**Y para que así conste, firman el presente certificado.**

**Fdo. Dr. Fernando González Candelas**

**Fdo. Dra. María Alma Bracho Lapiedra**

**Fdo. Dr. Francesc Xavier López Labrador**


VNIVERSITAT DE VALÈNCIA

**Doctorat en Biotecnologia**

# Dynamics, evolutionary and epidemiological patterns of RNA viruses

**Memòria presentada per Juan Ángel Patiño Galindo, per optar al grau de Doctor per la Universitat de València**

**Directors: Fernando González Candelas, María Alma Bracho Lapiedra, Francesc Xavier López Labrador**

**Aknowledgements**

In first place, I would like to thank my advisors, Fernando, Xavier and Alma for their help in my PhD. To Fernando and Xavier for their patience, commitment and personal support. To Alma for all the things she taught me when I was an undergrad, and also for her great work obtaining HIV sequences, as well as recording patients' information. Of course, I also thank Manoli for this.

I am especially grateful to Oliver and his team (Nuno, Jayna, Sarah, Kenneth) from the University of Oxford for their warm welcome to their lab, and the knowledge they transmitted to me, both during and after my stay. I could not have had better hosts!

I am also thankful to my friends from the University of València, especially to Lluís, Irene, Laura Parra, Ángeles and Marta Pinto for their help and company. Also, to the IGEMmers (Jose, Dani Tamarit, Manel and Sari).

Of course, no need to say that this thesis would have never been written without the unconditional help and love of my family (including my cat). Thank you so much for everything!

Finally, I would like to acknowledge all the teachers I had, from school to the university, because thanks to their invaluable work I am what I am today.

*To Ludwig.*

*“Justice is indivisible. You can't decide who gets civil rights and who doesn't”*

Angela Davis

# Index

**Summary**

Viral infections, specifically those caused by RNA viruses such as Human Immunodeficiency virus (HIV), Hepatitis C virus (HCV) or Influenza, are among the most important public health concerns to humans due to their high prevalence and associated mortality. Prevention and treatment campaigns against these viruses usually had limited efficacy, partly because their biological features allow them to reach very high levels of diversity, both at the within- and between-host levels.

Research focused on understanding the processes and mechanisms involved in the evolution of RNA viruses, and on the clinical and/or epidemiological consequences of their diversification, is important for improving the management of their epidemics. The aim of this PhD thesis is to study different aspects of the mid- and long-term evolution of RNA viruses, with special interest in molecular epidemiology. For this, different datasets (viral alignments, obtained either from public databases or by sequencing patient's derived samples from the studied populations) were obtained and analyzed by means of evolutionary and statistical approaches.

Firstly, phylogenetic, coalescent and statistical analyses were performed to depict the HIV epidemic in two different Spanish regions: Euskadi (Basque Country) and Comunitat Valenciana (Valencian Community). A significant number of patients from both regions, especially those from the Comunitat Valenciana, were included in local transmission clusters. Men who have unprotected sex with men (MSM) were significantly more prone to form transmission clusters than other risk groups. The high vulnerability of MSM to HIV infection was also evidenced by the detection of an extraordinarily large transmission cluster, affecting more than 100 patients solely in

the city of Valencia. Interestingly, the recent expansion of the highly pathogenic HIV CRF19_cpx among local Valencian MSM was also reported, for the first time outside Cuba.

Secondly, by means of Bayesian coalescent analyses, the genomic evolutionary rates of different HIV-1 subtypes (A1, B, C, D, G) and CRFs (CRF01_AE, CRF02_AG) were estimated and compared. The results obtained revealed that HIV-1 A1, C and CRF01_AE evolve significantly faster than subtypes B, D, G and CRF02_AG.

Thirdly, datasets containing sequences from the 6 major genotypes causing the HCV pandemic were analyzed, inferring the prevalence, evolutionary history and genetic barrier of naturally-occurring resistance mutations (RAVs) to direct acting antivirals (DAAs) in these genotypes. The obtained results demonstrate that RAVs are common in all HCV genotypes, and that there is an overall low genetic barrier for the selection of RAVs. Interestingly, some of these resistance mutations present a high potential to be transmitted between patients at risk.

In the fourth place, the distribution of positively selected sites along the genomes of HCV subtypes 1a and 1b was analyzed. The results show that positive selection is acting in all HCV genes, and that positively selected sites are associated with the presence of CD8 epitopes, while conserved sites are associated with RNA secondary structure and CD4 epitopes.

Finally, the effect of using RNA substitution models on the phylogenetic inference of viroids and RNA viruses was assessed. Such models were found to fit best for all the species analyzed. Compared to viral phylogenies inferred only from DNA

models, using RNA models usually leads to significantly longer tree length estimates, while has no significant effect on tree topology inference.

The results obtained from this work will not only have direct applications to HIV control campaigns in Spain and HCV treatment refinement, but also provide new insights into different aspects of the evolution of RNA viruses.

**Resum**

Les infeccions víriques, específicament aquelles causades per virus d'ARN com el Virus de la Inmunodeficiència humana (VIH), el Virus de la Hepatitis C (VHC) i els virus de la grip, es troben entre els problemes de salut pública més importants per als humans, a causa de la seua alta prevalença i mortalitat associada. Les campanyes de prevenció i tractament contra aquests virus solament han tingut una eficàcia limitada, en part perquè les seues característiques biològiques els permeten aconseguir nivells molt alts de diversitat, tant a nivell intra- com al inter- hoste.

La recerca centrada a entendre els processos i mecanismes involucrats en l'evolució dels virus d'ARN, i en les conseqüències clíniques i/o epidemiològiques de la seua diversificació, és important per a millorar la gestió de les epidèmies que causen. L'objectiu d'aquesta tesi doctoral és estudiar diversos aspectes relacionats amb l'evolució de virus d'ARN a mitjà i llarg termini, amb especial interès en epidemiologia molecular. Per a açò, diferents conjunts de dades (alineaments, obtinguts de bases de dades públiques o per seqüenciació de mostres obtingudes a partir de pacients provinents de les poblacions estudiades) van ser obtingudes i analitzades per mitjà d'aproximacions evolutives i estadístiques.

En primer lloc, es van realitzar diferents anàlisi filogenètics, de coalescència i estadístics per a descriure la epidèmia de VIH en dues regions espanyoles: Euskadi i la Comunitat Valenciana. Un nombre rellevant de pacients de les dues regions, especialment d'aquells de la Comunitat Valenciana, estaven associats amb grups de transmissió local. Els homes que practiquen sexe amb altres homes sense protecció (HSH) tenien major predisposició a formar grups de transmissió que altres grups de

risc. L'alta vulnerabilitat dels HSH a la infecció per VIH va ser evidenciada amb la detecció d'un grup de transmissió extraordinàriament gran, que afecta a més de 100 individus només a la ciutat de València. També es va reportar, per primera vegada fora de Cuba, l'expansió recent d'una variant altament patògena de VIH (CRF19) entre HSH valencians.

En segon lloc, les taxes d'evolució genòmica de diferents subtipus de VIH-1 (A1, B, C, D, G) i CRFs (CRF01_AE, CRF02_AG) van ser estimades, per mitjà d'anàlisi de coalescència bayesiana, i comparades. Els resultats obtinguts van revelar que VIH-1 A1, C i CRF01_AE evolucionen significativament més ràpid que els subtipus B, D, G i CRF02_AG.

En tercer lloc, es va analitzar conjunts de dades que contenien seqüències dels 6 genotips principals que causen la pandèmia d'hepatitis C, per a estimar la prevalença, història evolutiva i la barrera genètica de diferents mutacions associades amb resistència a antivirals d'acció directa en aquests genotips. Els resultats van demostrar que les mutacions de resistència són comunes en tots els genotips del VHC i que, en general, hi ha una barrera genètica baixa per a la seua selecció. Algunes d'aquestes mutacions, tenen un alt potencial per a ser transmeses entre pacients de risc.

En quart lloc, es va analitzar la distribució de codons seleccionats positivament al llarg dels genomes del VHC subtips 1a i 1b. Els resultats mostren que la selecció positiva actua en tots el gens de VHC, y que la presència de epítops CD8 està associada amb selecció positiva, mentre que els epítops CD4 i l'estructura secundaria d'ARN ho estan amb conservació.

Finalment, es va avaluar l'efecte de l'ús de models de substitució d'ARN a la inferència filogenètica dels virus d'ARN i viroids. L'ús d'aquests models s'ajustava millor que la seua exclusió en totes les espècies analitzades, i en la majoria dels casos resultava en arbres amb branques significativament més llargues encara que no tenia un efecte significatiu en la topologia.

Els resultats obtinguts en aquest treball tenen una aplicació directa, relacionada amb el desenvolupament de campanyes de control de VIH a Espanya i amb el refinament de tractaments efectius contra HCV. També aporten coneixements nous sobre diferents aspectes relacionats amb l'evolució dels virus d'ARN.

## Glossary

**AIDS:** Acquired immune deficiency syndrome.

**CRF:** Circulating recombinant form.

**CV:** Comunitat Valenciana.

**DAA:** Direct acting antiviral.

**dN:** Number of non-synonymous nucleotide substitutions per non-synonymous site.

**DNA:** Deoxyribonucleic acid.

**DRC:** Democratic Republic of Congo.

**ds:** Double-stranded.

**dS:** Number of synonymous nucleotide substitutions per synonymous site.

**HAART:** Highly Active Antiviral Therapy.

**HCV:** Hepatitis C virus.

**HIV:** Human immunodeficiency virus.

**HKAT:** Hudson-Kreitman-Aguadé test.

**HPD:** High Posterior Density.

**HT:** Heterosexual.

**IDU:** Intravenous drug user.

**MCMC:** Markov chain Monte Carlo.

**ML:** Maximum Likelihood.

**MSM:** Men who have sex with men.

**NGS:** Next-generation sequencing.

**PP:** Posterior probability.

**PR/RT:** Protease/retrotranscriptase.

**P/R:** Pegylated interferon + ribavirin.

**RAV:** Resistance associated variant.

**RNA:** Ribonucleic acid.

**SIV:** Simian immunodeficiency virus.

**ss:** Single-stranded.

**s/s/y:** Substitutions per site per year.

**tMRCA:** Time to the most recent common ancestor.

# *1.* Introduction and objectives

### 1.1- RNA viruses: evolution and molecular epidemiology fundamentals

Viruses are the most abundant and diverse biological entities on Earth (Abecasis et al. 2013; Edwards and Rohwer 2005; Holmes 2011). Unlike cellular life forms (cellular organisms), viral genomes can be based either on DNA or RNA, which can form single-stranded (ss, either sense or antisense) or double-stranded (ds) molecules. They also display a broad variety of genomic strategies, replicating and transcribing either RNA or DNA, with some viral species being able to perform the reverse transcription of RNA to DNA (Koonin & Dolja 2012). Hence, although all these genome properties are used to classify viruses into seven groups (Baltimore 1971) the broadest way to classify viruses is differentiating between viruses with a DNA genome (DNA viruses) and those with a RNA genome (RNA viruses).

RNA viruses account for most of the known viruses. They are characterized by extremely high mutation rates, the highest among all living beings. Only ssDNA viruses such as Phi X174 present similar rates (~$1x10^{-5}$ to $1x10^{-4}$ substitutions/nucleotide/generation) (Drake & Holland 1999; Holmes 2003; Raney et al. 2004; Holmes & Drummond 2007; Duffy et al. 2008; Sanjuán 2012). These high error rates, enhanced by the lack of proofreading activity of their replicases, have been proposed to be important factors limiting their genome size, which are generally smaller than that of DNA viruses.

RNA viruses are also characterized by short generation times, which usually generate a large population size in the infected individuals. All these characteristics, along with recombination and reassortment, are responsible for the high genetic variability and evolutionary rates of between hosts and intrahost viral populations

(Moya et al. 2000; Trifonov & Rabadan 2010). As a counterpart, viral variability is limited by two different phenomena: i) the selective constraints that the host's immune response imposes to the virus, and ii) bottlenecks at transmission between hosts, which will determine the viral diversity transmitted to other hosts (Grenfell 2004).

The fast evolutionary rates of RNA viruses, which have been estimated to be between $10^{-4}$ and $10^{-3}$ substitutions per site per year(s/s/y) (Jenkins et al. 2002), make them good experimental models for studying evolution: they have been used as model systems to address basic questions in evolutionary biology, such as the Red Queen's hypothesis (Clarke et al. 1994) or the phenomenon of convergent evolution (Moya et al. 2004). Their quick pace of change has also important public health implications: whereas DNA viruses establish persistent infections more often (in which the virus is not cleared but remains in the infected individuals), RNA viruses are more commonly associated with emerging diseases (those caused by viruses that jump species barriers, causing infections in different host species) (Holmes 2004). Their capability to emerge in new hosts could be explained by different factors, such as their high capability to be transmitted, and the fact that fast-evolving organisms are more successful at exploiting new niches (Cleaveland et al. 2001). The fast evolutionary rates of RNA viruses also play an important role in other public health concern: RNA viruses can develop antigenic variation and/or resistance to antiviral agents quickly as a result of natural selection acting on their continuously generated genetic variation (Holmes 2003; Moya et al. 2004). In this way, RNA viruses have a high potential to evade antiviral treatments and vaccines, exemplified by HIV, HCV or Influenza viruses.

It is possible to combine viral molecular data with clinical and epidemiological information from the host by means of a discipline called molecular epidemiology, a term coined by Edwin D. Kilbourne in 1973 (Kilbourne 1973). By means of evolutionary analyses, this branch of epidemiology aims to answer different questions, such as the typing of viral populations, the reconstruction of the history of an epidemic, the study of the dynamics of variants associated to resistance against antivirals as well as the detection of groups of individuals most vulnerable to a certain infection (Moya et al. 2004). Molecular epidemiology analyses often involve the detection of transmission clusters (groups of individuals infected by viral variants that derive from a common, recent ancestor, which evidences that they are epidemiologically related), which can be used to infer the minimum number of introductions of the virus in a given population. Their detection is usually performed by means of phylogenetic analyses, as clades of high support in the phylogenetic tree obtained with the sequences derived from the infecting viruses (Hué et al. 2005). Alternatively, transmission clusters can also be detected as groups of individuals of low genetic distance (Wertheim et al. 2014).

Often, transmission clusters are studied within population groups sharing a specific practice or behavior which enhances the risk of viral transmission. In the context of this thesis I will call them “risk groups” and, considering the transmission routes of HIV and HCV, “MSM” will be used to refer to men who have unprotected sex with other men, “IDUs” to users of intravenous drugs who share needles, “HTs” to people who have unprotected heterosexual sex, and so on.

## 1.2- Evolutionary analyses for molecular epidemiology

Phylogenetic reconstruction is a basic procedure in the molecular epidemiology of RNA viruses. There are different phylogenetic methods: i) based on distance matrices: neighbor-joining and UPGMA phylogenetic reconstruction; and ii) based on an optimality criterion: parsimony, maximum-likelihood (ML) or Bayesian phylogenetic inference. The later are usually known as statistical methods, because they seek to maximize a statistical parameter (the likelihood of the data or the posterior probability) by explicitly modelling the molecular evolutionary processes involved in the phylogenetic reconstruction using stochastic processes.

Statistical phylogenetic reconstructions are currently the most widely used methods in molecular epidemiology, because they allow the rigorous testing of phylogenetic hypotheses, such as evolutionary models, the quality of the trees or the confidence values assigned to subtrees (Whelan et al. 2001). Furthermore, they tend to resolve more accurately the evolutionary relationships of highly divergent sequences (Huelsenbeck & Hillis 1993). In the case of ML inference (Felsenstein 1981), the aim is to find the evolutionary tree and model which yields the highest probability of observing the empirical data (sequence alignment) under the chosen model of evolution. Assuming independence of evolution at each site in the alignment, the probability of a given alignment arising on a given tree is computed site by site, and all the probabilities obtained are multiplied to yield the sought likelihood of the alignment. As the number of analyzed taxa increases, the number of possible tree topologies grows exponentially. In order to make the process more computationally affordable, tree search of the potential ML trees (which usually starts with an initial

tree obtained by a distance-based method which is then modified and improved) is performed heuristically using different tree-pruning algorithms; this means that the chosen tree is the "best known" ML tree, but not necessarily the global ML tree (which may be not found). Thus, computation time is one of the main limitations in ML phylogenetics (Guindon & Gascuel 2003).

Bayesian phylogenetics (Rannala & Yang 1996) is based on Bayes' theorem and probabilities of tree states are calculated as posterior probabilities (PP), which are probabilities conditioned on the DNA alignment and the prior information provided by the user, and can be used as a measure of the reliability of the model (including the priors) given the data. PPs are calculated considering both the likelihoods of the tree states and their prior probability ("prior"), a probability which derives from prior information on the evolutionary parameters to be estimated. Whereas the phylogeny with the highest PP could be chosen as the best estimate of the evolutionary relationships, results are usually shown as a set of most probable trees (High posterior density distribution -HPD-), thus retaining uncertainty.

In Bayesian phylogenetics, as well as in ML, computing the PPs of all possible trees that may arise from a given dataset would be computationally unbearable, even at relatively small datasets. This problem is addressed, generally, by using the Markov chain Monte Carlo (MCMC) algorithm which generates a posterior distribution of candidate phylogenetic trees, in which those trees with higher PPs are sampled more frequently (Yang & Rannala 1997). With the implementation of MCMC, Bayesian phylogenetics became faster than ML, a factor that had a key influence on the current

popularity of this phylogenetic approach among molecular epidemiologists of RNA viruses.

A critical concern in Bayesian phylogenetics, absent in ML, is the sensitivity of PP to the specified prior distributions. When there is enough information in the alignment, the PP is mainly determined by the likelihood (Rannala & Yang 1996), but specifying inappropriate priors could bias our results. Thus, it is important to take into account that informative prior distributions should only be set if there is prior information on the evolutionary parameters to be estimated. If not, an uninformative prior (e.g., uniform distribution) should be used. It is important to point out that the impact of the priors in phylogenetic analyses should always be assessed.

Importantly, in phylogenetics the inference of evolutionary distances between taxa only from their observed nucleotide differences usually leads to underestimate the actual number of substitutions that have occurred. These errors are corrected by using models of nucleotide (or amino acid) evolution, which are assumptions about the process of nucleotide substitution, such as the existence of differences in base frequencies, in the substitution rates between nucleotides (Posada & Crandall 2001) or codon positions (Shapiro et al. 2006). Furthermore, models which account for the influence of RNA secondary structures on sequence evolution have been developed (Savill et al. 2001; Allen & Whelan 2014). In this thesis, the impact of RNA models on viral evolution has been assessed: whereas RNA viruses present conserved, RNA secondary structures of considerable length along their genomes, these models have never been applied in their phylogenetic analyses.

### 1.3- The coalescent: applications to molecular epidemiology of viruses

If mutations are fixed in a viral population on the same time scale as the ecological and epidemiological processes that are responsible for its genetic diversity, the divergence times from the common ancestor of the sampled individuals in a population can be associated with its demographic history (Holmes 2008) . This can be done by incorporating the coalescent theory of population genetics (Kingman 1982; Donnelly & Tavaré 1995) into ML or Bayesian phylogenetics. Assuming that a population accomplishes the Fisher-Wright assumptions (generations are discrete and non-overlapping, there is no effect of natural selection, recombination is absent, population remains constant along generations), the coalescence process starts with a sample of *n* lineages from a population and goes backwards in a genealogical process, linking these lineages when they share a common ancestor (this is called a coalescence event). The number of lineages decreases until there is only one left, the common ancestor to all the individuals analysed (Figure 1A). Coalescent times are directly proportional to the number lineages and inversely proportional to population size (Figure 1B). In this way, applying the coalescent model to viral phylogenetics allows estimating the times when new viral lineages were established in a host population.

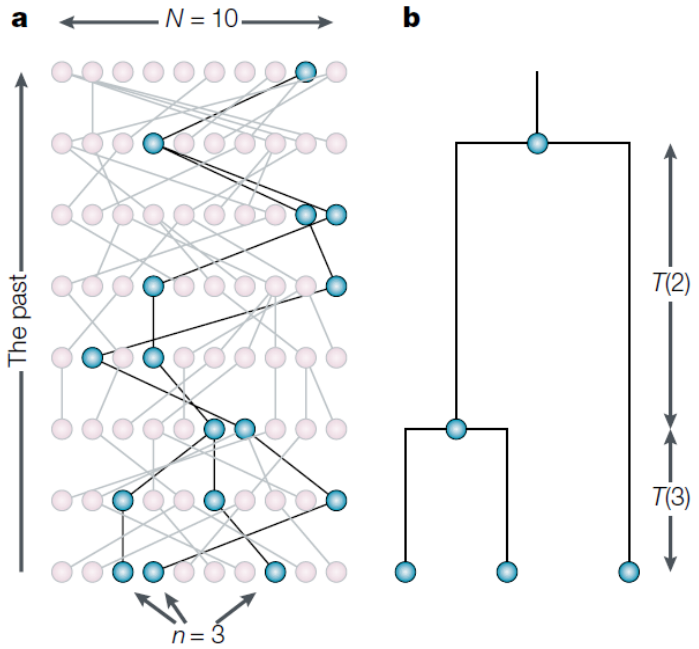



**Figure 1.A** graphical interpretation of the coalescent process. A) Black lines trace the ancestries of three contemporaneous lineages back to a single ancestor. B) Coalescent tree of these for these three lineages: coalescent times will depend on the sample size (retrieved from Rosenberg & Nordborg 2002).

The dynamics of demographic change affects coalescent times. Given that population sizes usually change over time, alternative growth models to the constant population size have been designed. Hence, coalescent phylogenetic analyses can also be used to estimate how the number of infected individuals changes along the epidemic (Kuhner et al. 1995; Hué et al. 2005). The most widely used demographic models are:

- Exponential growth model: It assumes an exponential growth of the population.
- Logistic growth model: Exponential growth would only be possible if there is no limitation by the availability of resources. It takes into

account that population size growth is limited by the competence between the individuals from the population.

- The Skyline Plot framework (Pybus et al. 2000): Unlike previous demographic models, this is a non-parametric model. Assuming that population sizes can change only at coalescent events but remain constant between coalescent events, it allows estimating the different evolutionary parameters without the need of a priori restrictions on demographic trends. Refined variations of the Skyline Plot, such as the Bayesian skyride, which assumes that population sizes change gradually (Minin et al. 2008) or the Extended Bayesian Skyline, which allows multilocus analyses (Heled & Drummond 2008) have been designed.

Phylogenetic methods which use the coalescent also allow estimating the rate at which a population evolves, taking into account the number of substitutions that have occurred between the dated coalescent events. For this goal, the initial assumption to consider is the presence of a molecular clock: the number of differences between taxa should be proportional to their divergence times. Ideally, the evolutionary rate remains constant (fixed) along the different lineages in the genealogy (Zuckerkandl & Pauling 1962). The molecular clock is then calibrated with information added by the researcher (e.g., sampling date of each sequence or tMRCA of an internal node). However, in reality, many organisms, including RNA viruses, evolve at rates which significantly deviate from the fixed molecular clock assumptions. They undergo, along their evolution, accelerations or decelerations in their pace of evolution caused by changes in selective pressures and/or population sizes (Jenkins et al. 2002). For this

reason, alternative molecular clock models taking into account such changes in the evolutionary rate have been generated:

- Auto-correlated relaxed molecular clock: A model that allows the rate to vary in an auto-correlated manner across the tree, in which the rate in each branch is drawn from a distribution of rates, whose mean is a function of the rate of the parental branch. Hence, the rate along a given branch is more similar to its parent branch than a branch chosen at random (Lepage et al. 2006; Thorne et al. 1998).
- Uncorrelated relaxed molecular clock (Drummond et al. 2006): It does not assume correlation between rates in neighbouring branches. Rates at each lineage are drawn independently from the same distribution (e.g., Gamma, exponential or lognormal). Under this model, each lineage will have a distinct rate. In the case of rapidly evolving organisms, such as RNA viruses, it appears to fit better than the autocorrelated model (Drummond et al. 2006).
- Local clock (Drummond & Suchard 2010): It allows different lineages in the tree to have different evolutionary rates, but in some adjacent lineages the evolutionary rate can be similar (Yoder & Yang 2000). Drummond & Suchard (2010) created a "random local clock" model, which proposes and compares a series of alternative local molecular clocks, which can arise on any branch of the phylogeny and then extend along adjacent lineages. This model has been proven to be the most powerful model in finding rate shifts along phylogenies (Fourment & Holmes 2014).

Undoubtedly, the most popular program for estimating dates of divergence events and evolutionary rates is BEAST (Drummond & Rambaut 2007). It infers dated phylogenies using a Bayesian MCMC coalescent method, providing the aforementioned demographic and molecular clock models. Alternatively, other programs based on ML such as r8s (Sanderson 2002) or Physher (Fourment & Holmes 2014) can be used.

### 1.4- Analyzing the evolutionary constraints of RNA viruses

The high mutation rate of RNA viruses produces new viral variants constantly (Elena & Sanjuán 2005). This compromises viral fitness, because most mutations are lethal, and such effect is intensified with the occurrence of bottlenecks at transmission events, which cause a fast accumulation of deleterious mutations. Despite this handicap, the survival of viral populations can be explained by a mutation-selection equilibrium, in which the action of positive and negative selection leads to a changing population composed by a large amount of different, yet related, viral variants (also called viral quasispecies) (Domingo et al. 1978; Manrubia et al. 2005). In this way, although the high mutation rates of RNA viruses could increase their adaptive capacity, the effect of negative selection has been demonstrated to be very strong, even stronger than for DNA viruses (Hughes & Hughes 2007).

Human RNA viruses are subjected to selective pressures at different levels: RNA and protein secondary structures have been previously reported to be associated with conservation, which would facilitate persistent infection by masking the viral genome from its degradation be RNAse L and innate antiviral defenses (Washenberger et al. 2007; Li & Lemon 2013; Snoeck et al. 2011; Sanjuán & Bordería 2011; Mauger et al.

2015). On the other hand, epitopes are usually associated with positive selection, thus favoring escape mutants from immune system cells (Snoeck et al. 2011; Fares et al. 2001). Interestingly, several studies have observed very conserved epitopes in different viruses which, at least in the case of HIV, have been associated with a benefit from immune activation (Sanjuán et al. 2013; Sarobe et al. 2001; Snoeck et al. 2011).

There are different types of tests that (considering neutral evolution as the null hypothesis) can be used for detecting genes or positions within a given gene evolving under selection, either at within-species (population) level or between-species.

Tests based in allele frequencies consider that, in a given population, positive selection changes the patterns of genetic variation with respect to the expected variation under a neutral model, and such an effect would skew the allele frequency distribution, reduce genetic variability and increase the level of linkage disequilibrium (Biswas & Akey 2006). Some of the most frequently used methods are Tajima's D (Tajima 1989), Fu and Li's D and F (Fu & Li 1993) and linkage disequilibrium decay (Sabeti et al. 2002) calculations. One of the main caveats of these methods is that the changes associated to selection could also be caused by demographic fluctuations, and it is generally recommended to analyze different loci along the genome to distinguish between the two alternatives.

Another test used for detecting selection at population level is the Hudson-Kreitman-Aguadé test (HKAT; Hudson et al. 1987). It is a goodness-of-fit which compares the levels polymorphism that exist within a given species with the divergence from an outgroup. This test analyzes at least two loci in order to determine significant variations from neutrality.

dN/dS tests can be used for the detection of selection either at the population level (comparing levels of polymorphism) or between-species (comparing levels of divergence). These tests assume that in protein-coding sequences an excess of non-synonymous nucleotide substitutions per non-synonymous site (dN) with respect to the number of synonymous nucleotide substitutions per synonymous site (dS) is a signature of positive selection. The rationale is that synonymous substitutions are assumed to be always neutral (they do not alter the amino acid), but non-synonymous substitutions can be subjected to evolutionary constraints. If non-synonymous substitutions are neutral, dN and dS should be similar. Consequently, an excess of dN as compared to dS suggests that natural selection promotes amino acid changes (positive selection). On the other hand, and excess of dS respect dS would suggest that natural selection discards amino acid changes (purifying selection).

dN/dS tests are commonly used either comparing dN and dS at individual branches of a phylogenetic tree ("branch methods"; Yang and Nielsen 1998) or searching for codons with dN/dS > 1 ("codon-based methods";Yang 1998; Kosakovsky Pond and Frost 2005). Other approaches search for codons with dN/dS >1 at certain lineages ("branch-site methods"; Zhang et al. 2005; Pond & Frost 2005). It is also worth mentioning the McDonald-Kreitman test (McDonald & Kreitman 1991), which is similar to the HKAT but differentiating between synonymous and nonsynonymous sites. This test checks for the presence of positive selection within a species comparing its amount of variation with the divergence between species at synonymous (neutral) and non-synonymous sites from a given locus, assuming that positive selection leads to an increase in non-synonymous divergence (dN/dS within species < dN/dS between species). One of the potential problems associated to dN/dS tests is that a number of

codons may present dN > dS just by chance, thus producing false positives (Hughes & Friedman 2005).

### 1.5- HIV: Phylodynamics and epidemiology

HIV, the causing agent of acquired immune deficiency syndrome (AIDS), is an enveloped virus, a retrovirus (family *Retroviridae*) of the genus *Lentivirus.* Its single-stranded positive-sense RNA genome comprises >9,700 nucleotides that encode 8 different genes: *gag*, *pol*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env* and *nef* (Figure 2).

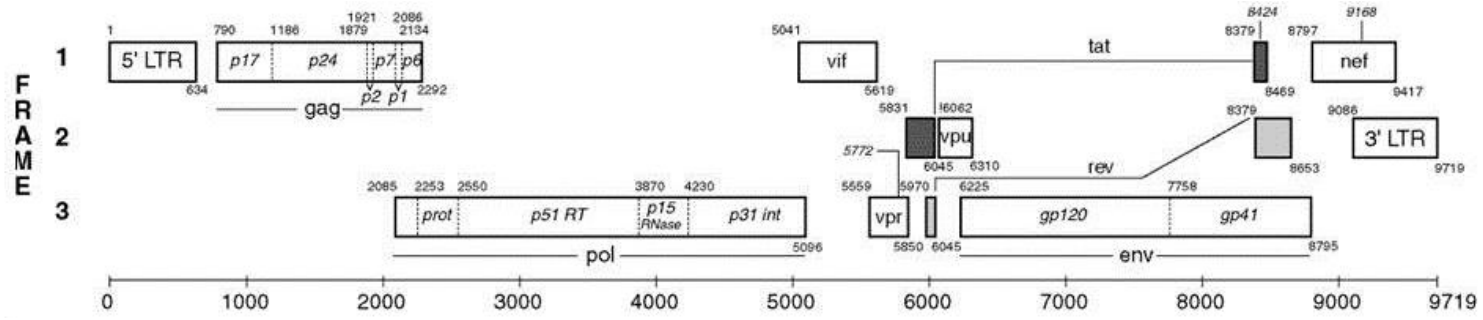



**Figure 2.** Structure of the HIV-1 genome (reference sequence HXB2, Genbank accession number K03455). The start and end of each open reading frame (shown as rectangles) are indicated in the upper left and lower right corners, respectively (Modified from Kuiken et al. 1999).

The HIV genome is characterized by a very high genetic diversity. There are two types of HIV: HIV-1 and HIV-2, the former being more pathogenic and the main cause of the AIDS pandemic. HIV-1 comprises four phylogenetically distinct groups: M, N, O, and P. Groups N and O are confined almost exclusively to West-Central Africa (Hahn et al. 2000). Only two strains from group P have been reported so far, both in Cameroon (Plantier et al. 2009; Vallari et al. 2011). The HIV pandemic is mainly driven by HIV-1 group M. Within this group, there are nine subtypes (denoted as A, B, C, D, F, G, H, J, and K) and at least 61 circulating recombinant forms (Kuiken et al. 2012) (Figure 3).

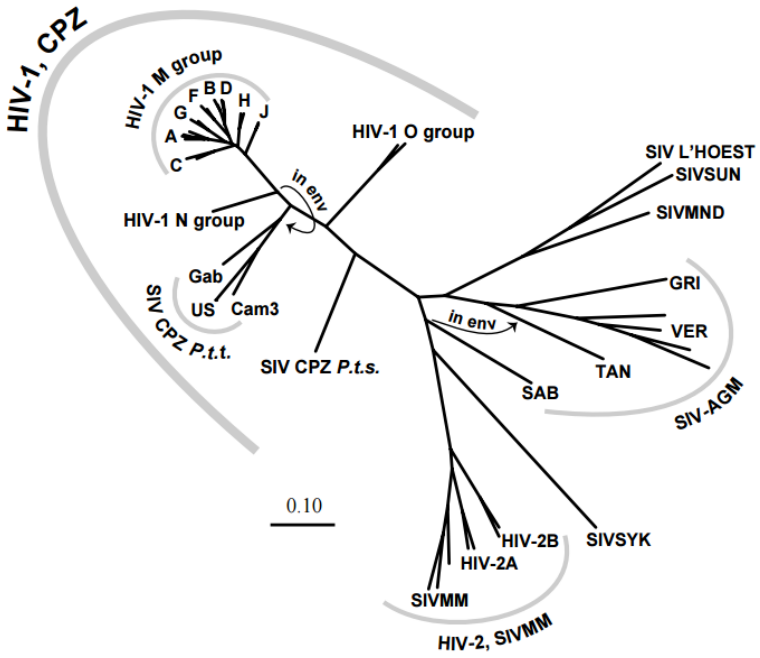



**Figure 3.** Evolutionary relationships of different lentiviruses infecting mammals, including HIV: HIV-1 (groups O, N and different pure subtypes from group M) and HIV-2. The phylogenetic tree reveals that HIV-1 N and M are more closely related to SIVcpz than they are to HIV-1 O, and HIV-2 is more closely related to SIVmm (SIV infecting sooty mangabey) than it is to HIV-1 (Retrieved from Kuiken et al. 1999).

Evolutionary analyses have revealed that HIV-1 lineages do not have a monophyletic origin: HIV-1 M and N are more closely related to SIVcpz (Simian immunodeficiency Virus of chimpanzees) than they are to HIV-1 O or P (Gifford et al.

2008). Furthermore, different recombination events occurred in the SIV lineages that gave rise to HIV-1 (Courgnaud et al. 2002; Bailes et al. 2003; Paraskevis et al. 2003; Heeney et al. 2006). Thus, AIDS is an emerging infectious disease: HIV-1 had a zoonotic origin, occurring from cross-species transmission events between chimpanzees and humans (Figure 3) (Hahn et al. 2000).

Faria et al. (2014) inferred the common ancestor of HIV-1 M to have occurred in the 1920s in Kinshasa (Democratic Republic of Congo, DRC), which subsequently diversified and expanded resulting in different variants (subtypes and recombinant forms) across Africa and the rest of the world. Phylogenetic analyses of HIV-1 subtype B, the most widespread HIV-1 subtype, revealed that it jumped from Africa to Haiti in the 60s, and then to other American countries and the rest of the world from the late 60s (Gilbert et al. 2007). Nevertheless, the most prevalent HIV subtype worldwide is C, accounting for around 50% of all the infections. Its expansion occurred mainly from Kinshasa to other DRC regions in the 1930s and then expanded around Africa and from there to Asia (Faria et al. 2014; Wilkinson et al. 2015) (Figure 4).

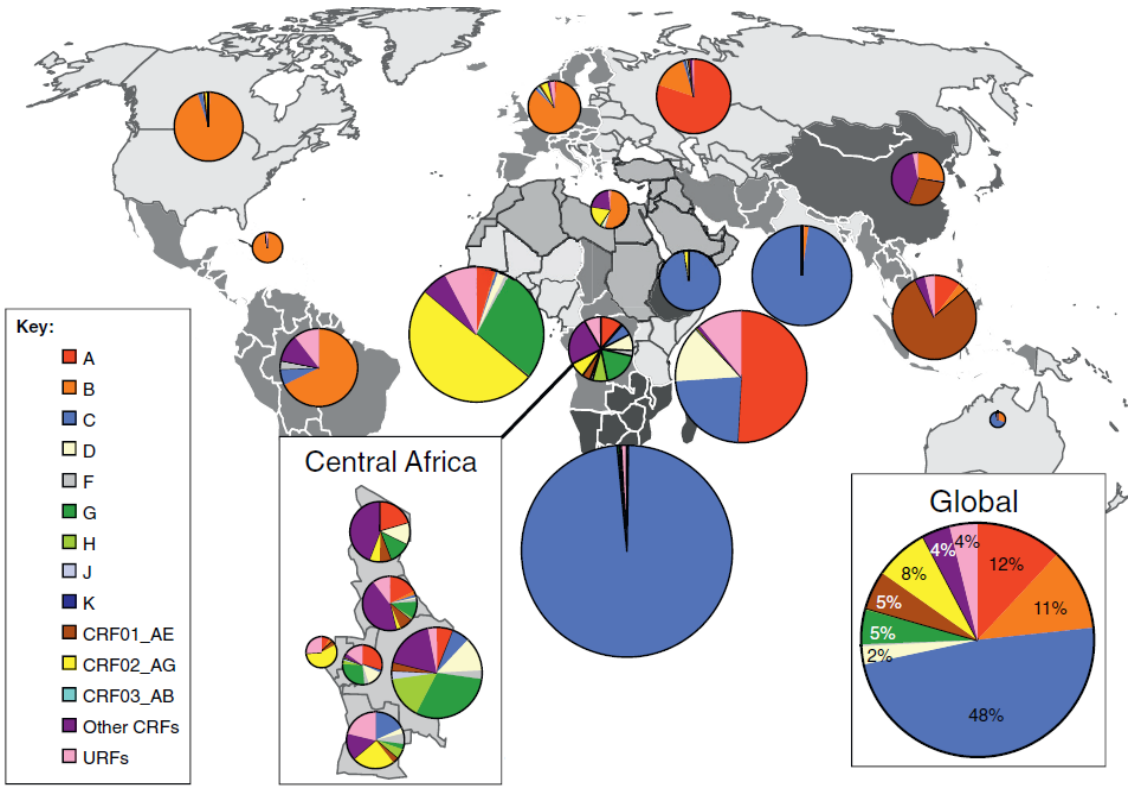



**Figure 4.** The global distribution of the HIV-1 subtypes and CRFs (Retrieved from Hemelaar 2012).

Thus, HIV-1 spread for > 50 years before the first AIDS case was recognized in 1981 and its subsequent discovery as the causative agent of this disease in 1983. In 1997, the AIDS pandemic had already caused more than 11.7 million of deaths in the world, and 30.6 million people were living with HIV, most of them in sub-Saharan Africa (UNAIDS 1998). In 1996, Highly Active Antiviral Therapy (HAART) was introduced as a therapeutic option in economically developed countries. This treatment consists of a combination of at least three drugs that inhibit HIV replication. Consequently, since then the global incidence of new HIV cases and AIDS deaths has slowed down, especially in those countries with high HAART coverage. In 2003 the number of people living with HIV was 38 million (UNAIDS 2004) and by 2012 it was estimated to be 35.3 million (UNAIDS 2013).

In Spain, as in other western countries, the first AIDS case was reported in 1981, and since then its incidence grew year by year until reaching a maximum in 1996, when HAART was introduced. In that year, the incidence started to decrease: by the end of 1997, the estimated number of people living with HIV in Spain reached 120,000 (UNAIDS 1998) while in 2012 the estimated number of people living with HIV was around 150,000 (UNAIDS 2013).

Despite that in western countries, such as Spain, IDU was considered the main transmission route from the beginning of the pandemic to the first years of the 21st century, a change has occurred lately. In contrast to IDUs and HTs, the number of new HIV diagnosis among MSM in the European Union and European Economic Area has increased, and this has become the main transmission route (ECDC 2013).

Although the HIV epidemic in Europe, particularly among MSM, is driven by the B subtype of the virus (Abecasis et al. 2013), with transmission clusters being reported frequently (Hué et al. 2005; Lewis et al. 2008; Kouyos et al. 2010; Leigh Brown et al. 2011; Bezemer et al. 2015), there is evidence for an increased introduction of non-B subtypes (Paraskevis et al. 2006), and several non-B HIV transmission clusters affecting MSM have been reported recently (Thomson et al. 2012; Antoniadou et al. 2014; Bracho et al. 2014). The vulnerability of MSM to HIV infection is also reflected in their shorter time between infections compared to that of HTs (Hughes et al. 2009).

Whereas the introduction of HAART had a clear impact on the control of the AIDS pandemic, and the combination of different drugs tries to prevent the raise of resistance mutations, drug efficacy is hampered by their emergence in treated patients (Costagliola et al. 2007). Genotypic tests of resistance to antiretroviral drugs, based on

sequencing the protease and reverse transcriptase (PR/RT) regions, are carried out routinely in many countries, including Spain, both for the design of individualized antiretroviral treatments and for the assessment of the frequency of certain resistance mutations in the population (Costagliola et al. 2007). The widespread use of these tests has led to large, publicly accessible datasets of HIV-1 sequences that, by means of molecular epidemiology studies, allow depicting the epidemic in a given population as well as characterizing its phylodynamics (Hué et al. 2004; Bello et al. 2010; Kouyos et al. 2010).

In the Spanish regions of Euskadi (Basque Country) and Comunitat Valenciana (Valencian Community), these genotypic tests of resistance have been carried out, for all HIV diagnosed patients, for over 10 years. These tests have generated HIV datasets comprising hundreds, even thousands, of sequences from different patients. In this thesis these datasets were used aiming at depicting the local epidemics occurring in these regions, by inferring the local subtype distribution and by detecting the existence of transmission clusters (which represent outbreaks) by means of phylogenetic and coalescent analyses. Some of the results from these analyses are very difficult to obtain by means of traditional epidemiological analysis: in the particular case of the Comunitat Valenciana the detection of transmission clusters is especially helpful, since no outbreaks had been detected by the Public Health Institutions until 2016 (unpublished results).

Another question addressed in this thesis is the evolutionary rate of the different HIV-1 subtypes and CRFs. Generally, diversification dates and evolutionary rates estimates for the different HIV-1 subtypes have been inferred only from single genes (Abecasis et al. 2009; Hemelaar 2012; Wertheim et al. 2012), thus ignoring

possible differences in mutation rate and/or selective constraints existing in other genomic regions. The genomic tMRCAs and evolutionary rates of some of the most prevalent HIV-1 subtypes and CRFs were estimated, expecting to obtain more accurate estimates thanks to the higher phylogenetic signal that is usually associated with longer nucleotide sequences. For this project, different HIV datasets that were representative of each HIV-1 subtype/CRF were retrieved from public databases, and then analyzed them by means of Bayesian coalescent phylogenetics.

### 1.6- HCV: Phylodynamics and epidemiology

HCV is an enveloped, positive-sense, single-stranded RNA virus which belongs to the family *Flaviviridae,* genus *Hepacivirus.* The HCV genome is about 9,600 nucleotides long and encodes a 3,011 amino acid polyprotein which is cleaved into three structural (Core, E1, E2) and six nonstructural proteins (P7, NS2, NS3, NS4A, NS5A, NS5B) (Figure 5). It also encodes an alternate reading frame protein (F/ARFP) of unclear function, which is synthesized by ribosomal frameshift (Xu et al. 2001).

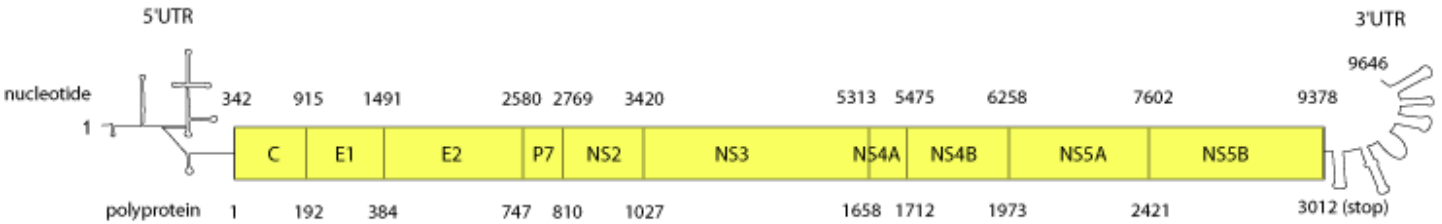



**Figure 5.** Structure of the HCV genome (reference sequence H77, accession NC_004102). The genomic RNA contains an open reading frame (shown in yellow), flanked by the 5' and 3' untranslated regions 5' (5'UTR and 3'UTR). The polyprotein generated after translation is cleaved into the 10 different HCV proteins (shown as rectangles). Numbers located in the upper left and lower left corners of each rectangle represent, respectively, the position of the first nucleotide (with respect to the HCV genome) and amino acid (with respect to the polyprotein) of each HCV protein. The Protein F/ARFP starts 4 nucleotides after the start of CORE, and continues for 201 codons (603 nucleotides) (modified from the Los Alamos HCV database, hcv.lanl.gov; Kuiken et al. 2005).

HCV is classified in 7 genotypes (denoted as 1-7) and 67 subtypes (Smith et al. 2014; Figure 6). Nucleotide divergence between complete genomes of different genotypes is >30%, and between subtypes of the same genotype it ranges between 20% and 25% (Simmonds et al. 1993; Simmonds 2004).

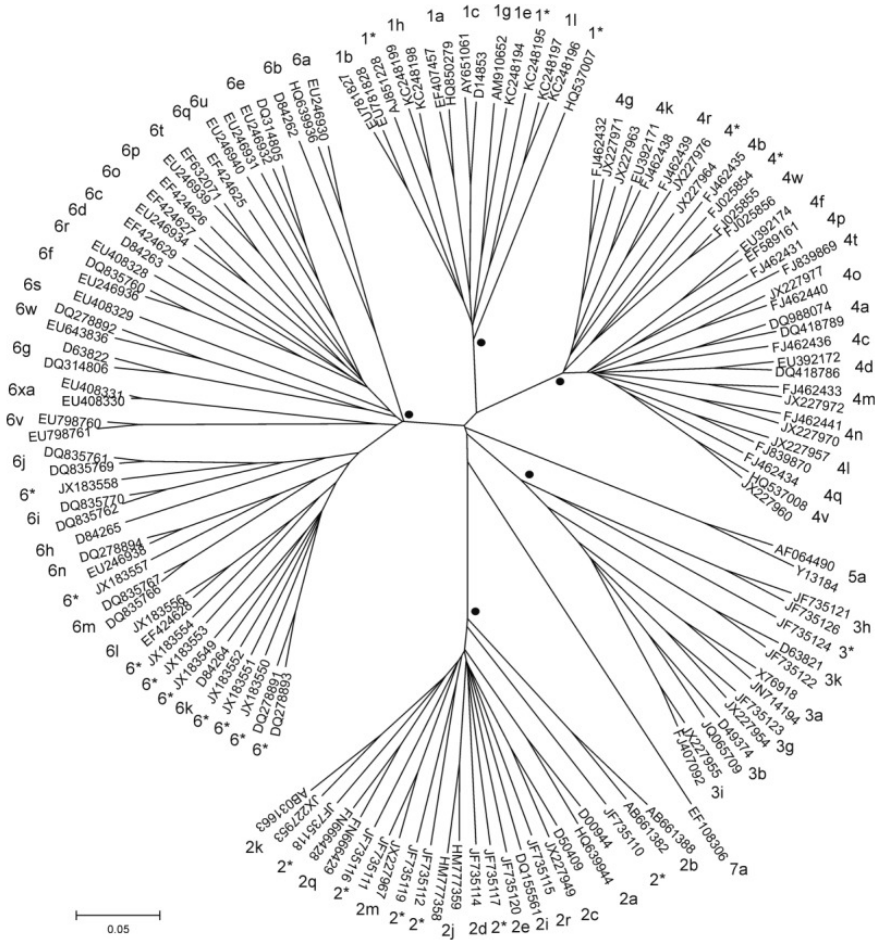



**Figure 6.** Phylogenetic tree of the different HCV genotypes and subtypes (Retrieved from Smith et al. 2014).

HCV genotype 1 (including subtypes 1a and 1b) is the most prevalent worldwide (46.2%), followed by genotype 3 (30.1%). Genotypes 2, 4 and 6 account for around 23% of all cases, and genotype 5 comprises <1%. Only one infection with genotype 7 has been

reported so far, and it corresponds to a case of a Central African migrant living in Canada (Messina et al. 2015).

There are important epidemiological differences between HCV genotypes/subtypes, along with differences in prevalence. HCV genotypes are differentially distributed geographically: genotypes 1, 2 in West Africa, 4 in Central Africa and the Middle East, and genotypes 3, 5 and 6 in South Asia, South Africa and East Asia, respectively (Smith et al. 1997; Murphy et al. 1996). Also, there are differences in the transmission routes associated to different variants: Subtypes 1a, and 3a are linked to IDUs in Western countries, while subtypes 1b, 2a and 2b are more frequently linked to blood transfusions and other nosocomial infections (Simmonds 2013) and genotype 4 with unsafe injections in Egypt (Frank et al. 2000).

Differences between HCV genotypes also exist at the clinical level: sustained virological response to the traditional pegylated interferon + ribavirin treatment (P/R) differs significantly between genotypes, ranging between 80% (genotype 2) to 46% (genotype 1) (Mangia et al. 2005; Pang et al. 2009). P/R inefficacy and toxicity urged to search for new HCV treatments, which has resulted in the development of direct-acting antiviral drugs (DAAs). These drugs target three viral proteins: the NS3/4A protease , NS5A and the NS5B polymerase) and have been demonstrated to overcome P/R limitations (AASLD/IDSA HCV Guidance Panel 2015; European Association for Study of Liver 2015). However, naturally occurring DAA resistance-associated variants (RAVs) have already been reported in DAA-naïve patients, with some RAVs showing differential prevalence between

genotypes and subtypes (López-Labrador et al. 2008; Di Maio et al. 2014). Thus, DAA efficacy is potentially hampered by the virus variability.

Unlike HIV-1, no closely related animal virus homologue to HCV has been identified. To date, the most closely-related known virus to HCV infects horses (Lyons et al. 2012). However, the high divergence between both species suggests that a zoonosis to humans from horses is unlikely to have occurred. The other hepaciviruses discovered, which infect dogs (Kapoor et al. 2011), rodents (Kapoor et al. 2013) and bats (Quan et al. 2013) are even more distantly related to HCV. While the scientific community is searching for new viruses, more closely related to HCV than the equine hepacivirus, to uncover the origin of HCV, the hypothesis that this virus has been always infecting humans is also plausible. Such hypothesis would be supported by evidences of long-term virus/host co-adaptations, such as the specific expression of microRNA 122 in human liver, which enhances virus replication (Jopling et al. 2005; Simmonds 2013). Furthermore, evolutionary analyses estimates suggest that HCV has been infecting humans for at least 500-2,000 years (Smith et al. 1997). For instance, the origin of genotype 6 has been estimated to have occurred from 600 years ago to 3,000 years ago (Pybus et al. 2009). Although these estimates are far from the tMRCA of humans (100,000- 200,000 years), inferring the true date of old viruses is hampered by the loss of signal due to substitution saturation and convergent evolution, which leads to underestimate the tRMCA (Aiewsakun & Katzourakis 2016). An alternative hypothesis, supported by the high genetic distance between genotypes and their geographic delimitations, is that the different HCV genotypes arose from independent zoonotic events (Pybus & Thézé 2016).

Interestingly, despite being older to humans than HIV, HCV was not discovered until 1989. The main epidemic HCV subtypes (1a, 1b and 3a) started their global spread in the early 1,900s, and then grew exponentially from the second half of the 20$^{th}$ century (Magiorkinis et al. 2009; Pybus et al. 2005). Sharing medical or surgical equipment, blood transfusions and injecting drug use facilitated its pandemic spread. 25 years after its discovery, HCV is considered a major public health problem: the estimated number of chronically infected people surpasses 170 million (>900,000 in Spain), with the consequent risk of developing liver diseases such as cirrhosis and liver cancer, which can eventually cause death (Hajarizadeh et al. 2013; Mohd Hanafiah et al. 2013). Although most cases occur through parenteral transmission, there is an increasing trend of sexual transmission among HIV-positive MSM (most commonly of HCV genotypes 4 and 1), despite the known low efficacy of sexual transmission of HCV (Bradshaw et al. 2013).

The high diversity that characterizes HCV raises several questions about this virus which have been addressed in this thesis: i) Several epidemiological and clinical differences have been identified between the most prevalent HCV subtypes (HCV-1a and 1b). Do these subtypes also present differences in genetic variability and in the selective constraints they are subjected to at the genome level? ii) Almost all DAAs have been designed to act against HCV-1, but there is little information on the effectiveness of these antivirals against the other 6 HCV genotypes. Do the different HCV genotypes present the same susceptibility towards DAAs? Public databases allow accessing to thousands of HCV sequences, which were retrieved in order to obtain representative datasets of each genotype/subtype.

## 1.7- Objectives

This PhD thesis is aimed at studying different aspects of the long-term evolution of RNA viruses by means of evolutionary and statistical approaches, with special interest in the molecular epidemiology of HIV-1 and HCV. The specific objectives are the following:

- To depict, by means of evolutionary analyses, the HIV epidemic in different Spanish regions, with especial interest in the Comunitat Valenciana.
- To analyze the diversity of HIV, and detect the existence of local transmission clusters. The results obtained will help to detect the emergence of new HIV variants, not detected in the Comunitat Valenciana before, as well as the most vulnerable groups of people towards this virus.
- To compare the tMRCAs and evolutionary rates of the main HIV-1 variants, by using nearly-complete genomes instead of individual genes.
- To analyze the prevalence and evolutionary history of HCV resistant mutations to DAAs, focusing on the potential differences between genotypes and subtypes.
- To analyze and compare the effect of positive selection on the genomic evolution of HCV subtypes 1a and 1b, analyzing how the presence of epitopes and RNA and protein secondary structures influence the distribution of positively selected and conserved sites along their genomes.

- To investigate whether RNA models outperform the use of DNA models in different sets of genomic alignments from RNA viruses and viroids, and assess the effect of such models on phylogenetic inference.

# *2.* Publications

## 2.1- Chapter 1: Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain

*Infect Genet Evol. (2016) 40:91-97.*

Infection, Genetics and Evolution 40 (2016) 91–97

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

# Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain

J.A. Patiño-Galindo [a,b], Michael M. Thomson [c], Lucía Pérez-Álvarez [c], Elena Delgado [c], María Teresa Cuevas [c], Aurora Fernández-García [b,c], Rafael Nájera [c], José A. Iribarren [d], Gustavo Cilla [d,h], Leyre López-Soria [e], María J. Lezaun [f], Ramón Cisterna [g], F. González-Candelas [a,b,*], on behalf of Group of HIV-1 Antiretroviral Resistance Studies in the Basque Country

[a] Joint Research Unit "Infection and Public Health" FISABIO-Universitat de Valencia, Spain
[b] CIBER in Epidemiology and Public Health, Madrid, Spain
[c] Centro Nacional de Microbiología, Instituto de Salud Carlos III, Majadahonda, Madrid, Spain
[d] Hospital Universitario Donostia, San Sebastián, Spain
[e] Hospital Universitario Cruces, Bilbao, Spain
[f] Hospital Universitario Araba, Vitoria, Spain
[g] Hospital Universitario Basurto, Bilbao, Spain
[h] CIBER for Respiratory Diseases, Madrid, Spain

ARTICLE INFO

*Article history:*
Received 9 December 2015
Received in revised form 15 February 2016
Accepted 22 February 2016
Available online 26 February 2016

*Keywords:*
Transmission cluster
IDUs
MSM
Antiretroviral resistance

ABSTRACT

This work was aimed to study the HIV-1 subtype B epidemics in the Basque Country, Spain. 1727 HIV-1 subtype B sequences comprising protease and reverse transcriptase (PR/RT) coding regions, sampled between 2001 and 2008, were analyzed. 156 transmission clusters were detected by means of phylogenetic analyses. Most of them comprised less than 4 individuals and, in total, they included 441 patients. Six clusters comprised 10 or more patients and were further analyzed in order to study their origin and diversification. Four clusters included men who had unprotected homosexual sex (MSM), one group was formed by intravenous drug users (IDUs), and another included both IDUs and people infected through unprotected heterosexual sex (HTs). Most of these clusters originated from the mid-1980s to the mid-1990s. Only one cluster, formed by MSM, originated after 2000. The time between infections was significantly lower in MSM groups than in those containing IDUs (P-value < 0.0001). Nucleoside RT and non-nucleoside RT inhibitor (NRTI and NNRTI)-resistance mutations to antiretroviral treatment were found in these six clusters except the most recent MSM group, but only the IDU clusters presented protease inhibitor (PI)-resistance mutations. The most prevalent mutations for each inhibitor class were PI L90M, NRTI T215D/Y/F, and NNRTI K103N, which were also among the most prevalent resistant variants in the whole dataset. In conclusion, while most infections occur as isolated introductions into the population, the number of infections found to be epidemiologically related within the Basque Country is significant. Public health control measures should be reinforced to prevent the further expansion of transmission clusters and resistant mutations occurring within them.

© 2016 Elsevier B.V. All rights reserved.

## 1. Introduction

Since the detection of the first cases of acquired immunodeficiency syndrome (AIDS) in the early 1980s, the pandemic caused by its main causal agent, the human immunodeficiency virus type 1 (HIV-1), has become one of the most important global health problems due to its mortality and morbidity. The latest UNAIDS/WHO report (2013) estimates a total of 35.3 (32.2–38.8) million people infected around the world. In 2012, there were 2.3 (1.9–2.7) million new HIV infections globally, which is a 33% decrease with respect to 2001. In Western European countries, such as Spain, subtype B is the most prevalent among the 9 subtypes of HIV-1. The epidemic of this variant started to spread rapidly among specific risk groups, such as men who have sex with men (MSM) and intravenous drug users (IDUs). Although the rate of infections in these groups decreased during the 1990s thanks to the development of adequate prevention campaigns (UNAIDS/WHO, 2013; Zehender et al., 2010), later years have been characterized by a continuous increment of sexually-related infections, mainly among MSM, while parenteral infections have decreased (ECDC/WHO, 2010).

HIV-1 is a retrovirus of the genus *Lentivirus*. Retroviruses present high evolutionary rates that usually lead to a high genetic diversity. These features are caused by three main factors: polymerization errors of the reverse transcriptase (RT) (Roberts et al., 1988), genetic recombination and an explosive proliferation that leads to enormous effective

* Corresponding author at: Joint Research Unit "Infection and Public Health" FISABIO-Universitat de Valencia, Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Parc Científic de la Universitat de València, c/ Catedrático José Beltrán, 2, 46980 Paterna, Valencia, Spain.
*E-mail address:* fernando.gonzalez@uv.es (F. González-Candelas).

http://dx.doi.org/10.1016/j.meegid.2016.02.028
1567-1348/© 2016 Elsevier B.V. All rights reserved.

population sizes, which promote the action of natural selection, favoring those mutations that increase the biological fitness of the virus and eliminating the disadvantageous alleles (Moya et al., 2004). These factors have important clinical consequences, such as the rise and spread of mutations related to resistance to antiretroviral drugs, but they also allow the reconstruction of the epidemic history of the virus by using phylogenetic tools (Holmes, 2004).

In the field of molecular epidemiology it is considered that epidemiologically related sequences should group together in a phylogenetic tree, forming transmission clusters, because they all share a common, recent ancestor (Hue et al., 2004, 2005). Coalescent methods for estimating phylogenetic trees (Kingman, 1982; Donnelly and Tavaré, 1995) are used to associate the divergence times from the common ancestor of the sampled individuals in a population with their demographic history. Thus, the results obtained offer information about dates of the introduction of viral variants in populations, the growth rate of infections during the epidemic and the most vulnerable groups of people to the virus (Moya et al., 2004).

The efficacy of highly active antiretroviral therapy (HAART) introduced in the 1990s is hampered by the emergence of resistance mutations in HIV-1 (Costagliola et al., 2007). Genotypic tests of resistance to antiretroviral drugs, after sequencing of the protease and reverse transcriptase (PR/RT) regions, are carried out routinely in many countries, including Spain, both for the design of individualized antiretroviral treatments and for the assessment of the frequency of certain resistances in the population (Costagliola et al., 2007). The widespread use of these tests has led to large, publicly accessible data sets of HIV-1 sequences that, combining analyses of their evolutionary history with epidemiological data, allow depicting the epidemic as well as characterizing its phylodynamics (Hue et al., 2004; Bello et al., 2010; Kouyos et al., 2010).

Genotypic tests of resistance for samples from the Basque country have been performed since 2001. The epidemic in the Basque Country has previously been analyzed using molecular data by Cuevas et al. (2009), who reported the existence of five major HIV-1 subtype B transmission clusters, with sizes ranging between 7 and 18 patients. Here, we have used a larger number of samples and we have also included a representative set of reference sequences to perform a more detailed phylogenetic analysis.

The objective of the present study was to characterize the epidemic of HIV-1 subtype B in this region and analyze the transmission dynamics of some relevant cases. For this, we have used a dataset of 1727 sequences comprising HIV-1 PR/RT genomic regions obtained from 2001 to 2008 in the Basque Country as part of the genotyping program for the search of drug resistance mutations. The largest HIV-1 subtype B transmission clusters detected were subjected to dated phylogenetic analysis (Drummond and Rambaut, 2007). The prevalence of mutations associated with antiretroviral drug resistance was estimated. The results obtained may help in the design of proper HIV prevention campaigns and treatments in this region.

## 2. Methods

### 2.1. Dataset

A total of 2497 HIV sequences 1200 nt long and spanning the full PR and partial RT coding regions were obtained from patients attending the main health centers in the Basque Country, Spain (cities of Bilbao, San Sebastián and Vitoria) from 2001 to 2008. In cases of multiple sequences from a single patient, only the earliest one was included. Thus, each viral sequence represented a different patient. Additionally, 8504 worldwide sequences were retrieved from Los Alamos HIV dataset (http://www.hiv.lanl.gov) and were used as reference sequences to ensure the validity of the transmission chains detected in the Basque sample. All the sequences were aligned with MUSCLE v3.5 (Edgar, 2004).

### 2.2. Phylogenetic reconstruction

In order to identify transmission clusters, defined as viral lineages derived from the same variant in the Basque population, two phylogenetic trees for the dataset of sequences which included both the Basque and the reference sequences (11,001 sequences in total) were obtained using FastTree 2.1 software (Price et al., 2010) using the GTR + Γ (4 categories) substitution model: (i) a tree obtained from a full codon alignment, in which 40 codons associated with major resistance in PR (30, 32, 46, 47, 48, 50, 54, 58, 74, 76, 82, 83, 84, 88, 90) and RT (41, 62, 65, 67, 69, 70, 74, 75, 77, 100, 101, 103, 106, 108, 115, 116, 151, 181, 184, 188, 190, 210, 215, 219 y 225) (Johnson et al., 2013) were removed, yielding a total length of 1080 nt, and (ii) a tree obtained from only third-codon positions of the original alignment (length of sequences = 400 nt).

We considered as potential transmission clusters those clades formed by at least 2 sequences of Basque origin present in both trees with SH-like local support ≥0.90 (Christin et al., 2012). Furthermore, clusters with support between 0.90 and 0.95, and/or including at least 10 patients, were further validated after their joint analysis with three random datasets of 1000 subtype B reference sequences (full codon alignments without resistance mutations). Basque sequences included in these clusters were incorporated to the three datasets and analyzed by maximum-likelihood with PhyML 3.0 (Guindon et al., 2010). Clusters with Basque Country-only sequences were considered only if they had Chi2-based approximate Likelihood-ratio test (aLRT) support >0.999. Clusters were classified depending on the major transmission route (>50%) for the corresponding patients.

Only HIV-1 subtype B clusters were considered for further analysis. All the sequences were subtyped with the REGA HIV-1 Subtyping Tool — Version 2.0 (http://dbpartners.stanford.edu/RegaSubtyping/; De Oliveira et al., 2005).

### 2.3. Dated phylogenies

The molecular clock signal of each transmission group equal to or larger than 10 individuals was assessed by performing linear regression analyses between the parameters "root-to-tip divergence" and "sampling date" with the software Path-O-Gen v1.4 (Drummond et al., 2003), using the phylogenetic trees from each transmission cluster, obtained as subtrees from the full-codon FastTree tree, as input.

These transmission groups were further analyzed using the full codon alignments of 1080 nt. Dated phylogenies were obtained using a Bayesian MCMC coalescent method, as implemented in BEAST v1.8.1 (http://beast.bio.ed.ac.uk/; Drummond and Rambaut, 2007). The SRD06 model, which partitions by codon position ($HKY_{112} + \Gamma_{112}$), was used for all the BEAST analyses, as it fits better in most viral protein-coding regions (Shapiro et al., 2006). A log-normal prior (median = 0.002 substitutions per site and year, s/s/y, 95% HPD upper limit = 0.0039 s/s/y) was placed on the ucld.mean parameter (Hue et al., 2005; Zehender et al., 2010). Under a relaxed molecular clock model, the most appropriate demographic model [either constant demographic size, exponential growth, logistic growth or Bayesian Skyline Plot (BSP)] was determined as the one with the lowest Akaike Information Criterion (AIC) value (Baele et al., 2012).

For each transmission cluster we performed at least two independent runs of Bayesian MCMC, with chain lengths ranging between 5 and 10 million states, sampling every 10,000 generations. Subsequently, these runs were combined after discarding a 10% burn-in. All the parameters were estimated from an effective sampling size >200 using the software Tracer 1.5 (http://tree.bio.ed.ac.uk/software/tracer/). Trees generated from the two BEAST runs were combined and summarized after discarding a 10% burn-in using TreeAnnotator (http://beast.bio.ed.ac.uk/).

### 2.4. Detection of intra-subtype recombination

Intra-subtype recombination might introduce spurious long branches in the phylogenies of the transmission clusters considered (Hughes et al., 2009). We grouped all sequences from these clusters and checked for the presence of recombination events by performing five different recombination analyses implemented in RDP3 software: RDP, Geneconv, Bootscan, Maxchi and Chimera (Martin et al., 2010; Martin and Rybicki, 2000; Padidam et al., 1999; Martin et al., 2005; Smith, 1992; Posada and Crandall, 2001). The criterion used to consider the existence of recombination was to obtain significant evidence of recombination in at least two different analyses.

### 2.5. Estimates of time between infections in transmission clusters

The internal branch lengths of the transmission clusters allow us to estimate the time between infections (Lewis et al., 2008). 95% HDPs for the median time between infections at each transmission cluster were estimated from the tree files produced with BEAST. We obtained the median and the upper and lower 95%HPD limits for the internal branch lengths of each tree (Lewis et al., 2008) using an in-house Perl script combined with an R-script (R Development Core Team, 2011). The distributions of the median internal branch lengths were compared among transmission groups by ANOVA tests. Tukey's tests were performed as post-hoc analyses.

### 2.6. Estimate of prevalence of drug resistance mutations

Mutations associated with resistance to PR and RT inhibitors (Johnson et al., 2013), both in the total dataset and in each of the transmission groups analyzed, were detected using the Stanford University HIV Drug Resistance Database [http://sierra2.stanford.edu/sierra/servlet/JSierra, (Liu and Shafer, 2006)] and their prevalence was estimated. Only major mutations were taken into account for the protease gene.

## 3. Results

### 3.1. Detection of transmission clusters

Of the 2497 HIV-1 sequences analyzed, 2311 belonged to subtype B and had been obtained from a total of 1727 different patients. Of these, 833 corresponded to IDUs, 340 to people infected through unprotected heterosexual sex (HT), 181 to MSM, and 49 to people vertically infected (VERT). 175 sequences belonged to people infected through sexual contact, not specifying whether it was heterosexual or homosexual. No risk factor was known for 149 people.

A total of 156 HIV-1 subtype B transmission clusters were consistent with the two phylogenetic reconstructions obtained with FastTree (from full-codon and third-positions alignments). In total, 441 (25.5%) sequences in the Basque country dataset were included in a transmission cluster. Most of these clusters (93.6%) contained 4 individuals at most (Fig. 1A). Transmission clusters of IDUs were the most abundant (Fig. 1B), followed by groups formed by people infected through unprotected heterosexual sex (HT), MSMs, and clusters containing both IDUs and HTs (IDU/HT). IDU clusters encompassed the largest number of patients (n = 126), followed by MSM (n = 124), HT (n = 75) and IDU/HT (n = 81) (Fig. 1C). MSM were significantly more likely to group in a transmission cluster than patients from other risk groups (Fisher's exact test: P-value = 2.2E-4; odds-ratio = 1.76, 95% confidence interval = 1.30–2.37). Only 6 of the detected clusters comprised at least 10 individuals that, altogether, represented 4.8% of the Basque sample (83 individuals). Four of these clusters were formed by MSM (clusters C, D, E and F), one was classified as an IDU cluster (cluster B) and another was classified as IDU/HT (cluster A). All of them were validated with the maximum likelihood analyses (all had aLRT >0.99 in the reconstructions with PhyML), and all but cluster B had been reported previously by Cuevas et al. (2009): A (cluster M2 in Cuevas et al.), C (M5), D (M1.3), E (M1.1, M1.2), F (M3). MSM were significantly more likely to group in a large transmission cluster than UDIs and HTs (Fisher's exact test: P-value = 2.00E-15; odds-ratio = 8.95, 95% confidence interval = 5.12–15.84).

### 3.2. Bayesian coalescent analyses

No evidence of intra-subtype recombination events was found in any of the six transmission clusters considered. Fig. 2 shows the dated phylogenies of the transmission clusters reconstructed with BEAST using dated-tips and considering the demographic model that yielded the lowest AIC value in each group. tMRCA estimates differed among groups with the earliest date corresponding to group B (median tMRCA = 1983.8), and the latest to group D (2000.5) (Table 1).

The ANOVA test comparing the median lengths of internal branch concluded that there were significant differences between transmission clusters (F = 37,382.3, df = 5 and 47,463.27, P-value <2.2E-16), being significantly shorter in MSM than in IDU or IDU/HT transmission clusters (all Tukey's test comparisons: P = 0.00; Table 1). The same results (not shown) were obtained with estimates obtained using all the sequences as contemporaneous.

### 3.3. Resistance mutations in transmission clusters

The prevalence of mutations associated with antiretroviral drug resistance in each transmission cluster and in the complete dataset are shown in Table 2. While mutations associated with resistance to protease inhibitors (PIs) were found only in one transmission group (B, most prevalent mutation L90M: 0.30), all transmission groups except D presented mutations associated with resistance to nucleoside and non-nucleoside reverse transcriptase inhibitors (NRTIs and NNRTIs, respectively). Groups A and B presented the largest number of NNRTI and NRTI resistance mutations, respectively. Among NRTI mutations, T215D/Y/F had the highest prevalence (1.0 in cluster F, 0.50 in cluster B), followed by M184I/V (0.50 in cluster B, 0.27 in cluster A) and M41L (0.40, also in group B). The prevalence of NNRTI-resistance mutations was lower than those causing resistance to NRTI, with K103N being the most prevalent one (0.36 and 0.20 in groups A and B, respectively), followed by G190A/S (0.20 and 0.19 in groups B and E, respectively). In the complete subtype B Basque dataset, mutations associated with resistance to PIs were also present with low prevalence except L90M and M46I/L (0.13 and 0.11, respectively). The most frequent NRTI-resistance mutations were M184I/V (0.36), L215Y/F (0.26), M41L (0.23), D67N (0.17), L210W (0.15), K70E/R (0.13), and K219D/Q/E/R (0.12). The most frequent NNRT-resistance mutation was K103N/S (0.22), the only one with prevalence >0.10.

## 4. Discussion

We have analyzed 1727 HIV-1 subtype B sequences from different patients obtained from health centers in the Basque Country, Spain, between 2001 and 2008 to assess the HIV-1 epidemics in this population. The large size of the dataset and the time-span in which these sequences were obtained provide enough confidence to consider the results obtained in this work as representative of the epidemic scenario of HIV-1 in this region.

The results obtained from this work suggest that the HIV-1 subtype B epidemic in the Basque Country is characterized by a majority of infections occurring as isolated introductions of the virus, although a representative 25.5% of the patients were included in transmission clusters, which ranged in size between 2 and 18 individuals. This proportion is much lower than the 47% sequences grouping in transmission clusters found by Cuevas et al. (2009) among newly diagnosed individuals. The smaller size of their dataset (261 vs 1727 patients), methodological

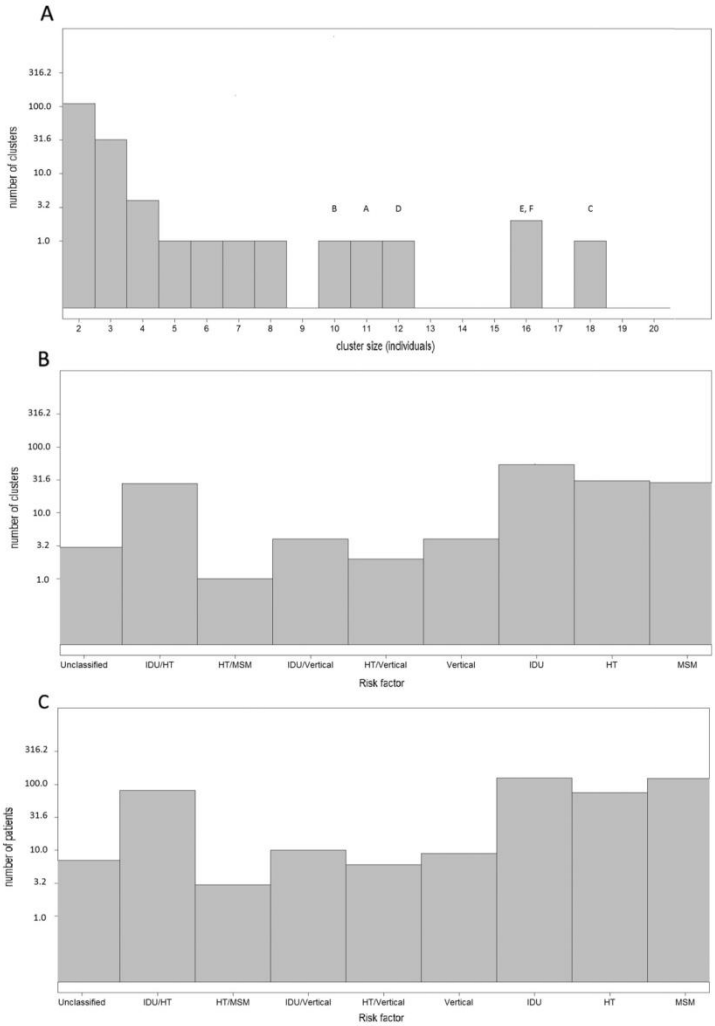



**Fig. 1.** Distribution of sizes, in log scale, of the 156 transmission groups found in the Basque Country (2001–2008) with the phylogenetic analyses. Block letters on each bar indicate the 6 main transmission clusters that were analyzed using BEAST (panel 1 A), number of transmission clusters depending on the risk group in which their members are included (panel 1B) and total number of patients for each risk group included in transmission clusters (n = 441) (panel 1C).

differences in phylogenetic analyses, and the low number of subtype B reference sequences (n = 4) used in that study can explain the markedly different proportions of individuals included in transmission clusters between both analyses. An additional factor that might explain the differences observed is the inclusion of non-newly infected patients in our analyses, which may cluster with less frequency due to higher number of nucleotide substitutions (longer external branches).

We found 6 large clusters, with sizes ranging between 10 and 18 individuals, that represented almost 5% of the total Basque dataset. Five of these clusters had been reported previously by Cuevas et al. (2009). Cluster B was not detected previously, because none of its sequences was analyzed by these authors for the reasons explained above. Previous studies in different European populations found transmission groups that were mainly formed by MSM (Hue et al., 2005; Kouyos et al., 2010; Lewis et al., 2008). In fact, the four largest transmission clusters found in our analysis (C, D, E, F) were also formed by MSM. Hence, although the MSM population was less frequent than other risk groups in the Basque country sampling, they were the major group associated to transmission clusters. IDUs frequently clustered either as the only risk factor or including also transmissions through unprotected heterosexual sex (IDU/HT), HTs and MSM, thus portraying a more diverse scenario in which IDUs were present in most of the smaller transmission groups. Such clustering of IDUs and HTs has seldom been reported (Kouyos et al., 2010; Holmes et al., 1995).

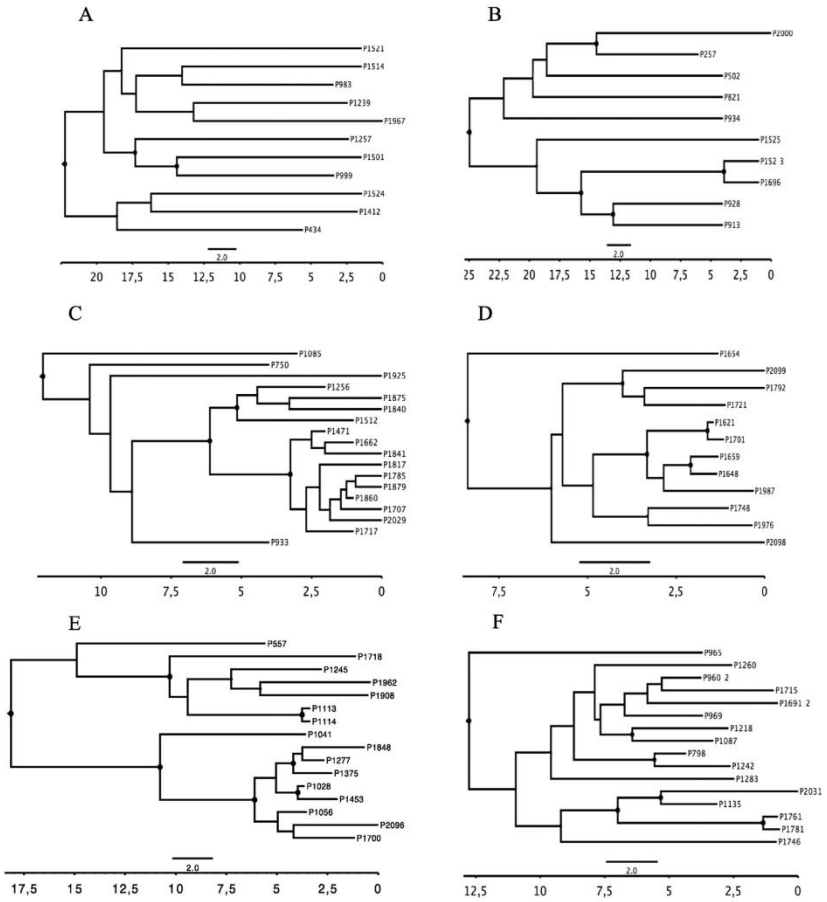



**Fig. 2.** Dated phylogenies of the six transmission clusters (A to F) analyzed with BEAST, as obtained with tip dating. Branch lengths represent years. Black dots represent nodes with posterior probability ≥0.90.

Previous studies in other European regions estimated that HIV-1 subtype B clusters initiated between the late 1960s and the early 1980s (Hue et al., 2005) or between the early 1990s and the beginning of 21st century (Lewis et al., 2008; Zehender et al., 2010). Dated phylogenies showed the MRCAs from most clades to have diversified from the mid-1980s to mid-1990s. The most recent clusters were D and F, both formed by MSM. Their dates of origin (tMRCAs) are coincident with the increase of infections among MSM after the

**Table 1**
Transmission routes, size (number of patients), range of sampling dates and root-to-tip divergence vs sampling date correlation coefficient for each clade and estimates of tree heights, internal branch lengths and substitution rates as obtained with BEAST under the best fitting demographic model, using a relaxed molecular clock model, with tip dating.

| Cluster | Transmission | Taxa | Range* | Range (dates) | R (root-to-tip divergence vs sampling date correlation) | Best fitting demographic model | Median tree height (95% HPD)* | Median internal branch lengths (95% HPD)* | Median substitution rate x10e-3 (95% HPD)† |
|---|---|---|---|---|---|---|---|---|---|
| A | IDU/HT | 11 | 5.57 | Nov 02–Jun. 08 | 0.03 | Exponential | 1986.3 (1980.6–1996.9) | 2.07 (0.88–4.71) | 1.19 (0.54–2.09) |
| B | IDU | 10 | 4.30 | May 04–Oct. 08 | 0.46 | Logistic | 1983.8 (1964.8–1996.1) | 3.08 (1.36–7.18) | 1.54 (0.67–2.73) |
| C | Homosexual | 18 | 3.98 | Nov 04–Nov. 08 | 0.61 | Exponential | 1996.6 (1989.9–2001.6) | 0.61 (0.30–1.31) | 2.20 (1.17–3.38) |
| D | Homosexual | 12 | 1.44 | Jun 07–Nov. 08 | (−0.16) | Constant | 2000.5 (1992.8–2004.9) | 1.00 (0.42–2.44) | 1.93 (0.81–3.36) |
| E | Homosexual | 16 | 5.57 | Apr 03–Nov. 08 | 0.25 | Logistic | 1990.8 (1979.3–1998.9) | 1.39 (0.61–3.09) | 1.66 (0.80–2.73) |
| F | Homosexual | 16 | 4.34 | Apr 04–Sept. 08 | 0.43 | Exponential | 1996.1 (1988.1–2001.2) | 1.22 (0.56–2.65) | 1.57 (0.74–2.58) |

* Time measured in years.
† Substitution per site and year.

**Table 2**
Prevalence (proportion) of PI-NRTI- and NNRTI-resistance mutations in the HIV-1 subtype B Basque dataset (n = 1727 sequences) and the six largest transmission clusters (A to F).

PI

| Cluster | D30N | V32I | M46I/L | I47V/A | G48V | I54 M/L | L76V | V82 A/F/T/S/L | I84V | N88S | L90M |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Full dataset | 0.05 | 0.02 | 0.11 | 0.01 | 0.01 | 0.01 | 0.01 | 0.09 | 0.03 | 0.01 | 0.13 |
| A | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 0 | 0.10 | 0 | 0.10 | 0.10 | 0.10 | 0 | 0 | 0 | 0 | 0.30 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

NRTI

| | M41L | A62V | K65R | D67N | T69A/D/N/T | K70 E/R | L74I/V | V75I | F77L | Y115F | F116Y | Q151M | M184I/V | L210W | T215Y/F | K219 D/Q/E/R |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full dataset | 0.23 | 0.03 | 0.02 | 0.17 | 0.05 | 0.13 | 0.07 | 0.01 | 0.01 | 0.02 | 0.01 | 0.02 | 0.36 | 0.15 | 0.26 | 0.12 |
| A | 0 | 0.09 | 0.18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.27 | 0 | 0 | 0 |
| B | 0.4 | 0 | 0 | 0.30 | 0.10 | 0.10 | 0.10 | 0 | 0 | 0 | 0 | 0 | 0.50 | 0.30 | 0.50 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.06 | 0 | 0 | 0 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0.19 | 0 | 0 | 0.19 | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0.13 | 0 |
| F | 0 | 0 | 0.06 | 0 | 0 | 0.06 | 0.06 | 0 | 0 | 0.06 | 0 | 0 | 0 | 0 | 1.00 | 0.06 |

NNRTI

| | V90I | A98G | L100I | K101 E/H/P | K103N/S | V106 A/I/M | V108I | R138 A/G/K/Q/R | V179D/E/F/T/L | Y181C/I/V | Y188 C/L/H | G190A/S | H221Y | P225H |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Full dataset | 0.05 | 0.02 | 0.04 | 0.06 | 0.22 | 0.03 | 0.06 | 0.04 | 0.02 | 0.08 | 0.02 | 0.08 | 0.03 | 0.03 |
| A | 0.09 | 0 | 0.09 | 0 | 0.36 | 0 | 0.09 | 0 | 0.09 | 0.09 | 0.09 | 0 | 0.09 | 0.09 |
| B | 0 | 0 | 0 | 0.1 | 0.2 | 0 | 0 | 0 | 0.10 | 0.10 | 0 | 0.20 | 0 | 0.10 |
| C | 0.06 | 0 | 0 | 0 | 0.06 | 0 | 0.06 | 0 | 0.06 | 0 | 0 | 0 | 0 | 0.06 |
| D | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0.13 | 0 | 0 | 0 | 0 | 0 | 0.13 | 0 | 0.19 | 0 | 0 |
| F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.06 | 0 | 0.06 | 0 | 0 |

commercialization of antiretroviral treatments and subsequent relaxation of prevention measures in this transmission group, especially a decrease of consistent condom usage (ECDC, 2013).

The estimated time between transmissions was significantly lower in MSM groups than in those including IDUs. Hughes et al. (2009) also found that HTs infected with HIV-1 subtypes A and C presented longer times between infections than MSM infected with HIV-1 subtype B. These results may be explained by the known high transmission risk of unprotected anal sex (Baggaley et al., 2010). In MSM who practice unprotected sex, the risk of HIV-1 infection is also increased due to role reversal during sexual intercourse: many individuals practice both insertive and receptive anal sex. This would increase HIV-1 spread by overcoming the low infection rates from receptive to insertive sexual partners (Beyrer et al., 2012).

NRTI resistance mutations were the most prevalent in the Basque dataset, with seven mutations present in more than 10% of the sampled sequences. PI and NNRTI resistance mutations were less frequent, with only two and one cases with a prevalence >0.10, respectively. For the six large transmission clusters, the prevalence of resistance mutations differed both among type of antiviral drug and among the analyzed transmission clusters. While only cluster B presented PI resistance mutations, all the clusters but one presented NRTI and NNRTI resistance mutations. It is important to mention the case of cluster F, in which all patients carry the low-level NRTI resistance mutation T215D. This possible example of drug resistance transmission in the Basque population has been reported previously (Cuevas et al., 2009; Vega et al., 2015). Hence, these results indicate that the dynamics of resistance mutations to antiretroviral drugs may differ among transmission clusters. However, this study lacks sufficient data to perform a detailed analysis of these patterns, and more extensive analyses are necessary to elucidate the factors originating these differences.

In conclusion, our results suggest an epidemic scenario of HIV-1 subtype B in the Basque Country in which most infections appear to correspond to independent introductions in the population, although there exist at least 6 major long-standing and diverse transmission groups. Most of these groups are characterized by a large proportion of MSM, in a disproportionally large frequency with respect to the presence of this risk group in the global sample. Furthermore, a shorter time between infections among MSM relative to other risk groups demonstrates the vulnerability of this collective to HIV-1 infections. Our results reinforce the need to implement prevention campaigns in the MSM population. This study also highlights the relevance and interest of applying Bayesian methods for phylogenetic and coalescent inference in epidemiology.

## Acknowledgments

We thank Dr. Daniel Zulaica, coordinator of the Plan for AIDS Prevention and Control and Osakidetza-Basque Health Service for their support in the development of the study.

The Spanish Group of HIV-1 Antiretroviral Resistance Studies in the Basque Country: Álava—Hospital Txagorritxu: Agud J. M. Aldamiz M, Ayensa C, Barroso J, Lezaun M. J, Michaus L, Pérez-Ortolá R, and Portu J. J. H. de Santiago: Andia A and Labora A. Guipúzcoa—Complejo Hospitalario Donostia: Arrizabalaga J, Cilla G, Echevarría J, Iribarren J. A, Rodriguez-Arrondo F, Serrano-Bengoechea E, and Von Wichmann M. A; H. de Zumarraga: Bustillo M.A. Vizcaya — H. de Galdákano: López de Goicoechea M. J and Mayo J; H. de Cruces: Aguirrebengoa K, López Soria L, Goikoetxea J, Montejo M, and Bereciartura E; H. de Basurto: Baraia J. B, Cisterna R, Ezpeleta C, Ferrero O, Ibarra S, Imaz M, Muñoz- Sánchez J, Santamaría J. M, Sota M, and Zubero Z; H. San Eloy: Lizarraga M and Silvariño R.

This work was funded by Convenio de Colaboración entre Osakidetza-Servicio Vasco de Salud y el Instituto de Salud Carlos III (MVI 1158/03) and projects BFU2011-24112 and BFU2014-58656-R from MINECO (Spanish Government).

## References


Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M.A., Alekseyenko, A.V., 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29, 2157–2167.

Baggaley, R.F., White, R.G., Boily, M.C., 2010. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. Int. J. Epidemiol. 39, 1048–1063.

Bello, G., Aulicino, P., Ruchansky, D., Guimaraes, M., Lopez-Galindez, C., Casado, C., Chiparelli, H., Rocco, C., Mangano, A., Sen, L., Morgado, M., 2010. Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. Retrovirology 7, 22.

Beyrer, C., Baral, S.D., van Griensven, F., Goodreau, S.M., Chariyalertsak, S., Wirtz, A.L., Brookmeyer, R., 2012. Global epidemiology of HIV infection in men who have sex with men. Lancet 380, 367–377.

Christin, P.A., Besnard, G., Edwards, E.J., Salamin, N., 2012. Effect of genetic convergence on phylogenetic inference. Mol. Phylogenet. Evol. 62, 921–927.

Costagliola, D., Descamps, D., Assoumou, L., Morand-Joubert, L., Marcelin, A.G., Brodard, V., Delaugerre, C., Mackiewicz, V., Ruffault, A., Izopet, J., 2007. Prevalence of HIV-1 drug resistance in treated patients: a French nationwide study. J. Acquir. Immune Defic. Syndr. 46, 12–18.

Cuevas, M.T., Munoz-Nieto, M., Thomson, M.M., Delgado, E., Iribarren, J.A., Cilla, G., Fernandez-Garcia, A., Santamaria, J.M., Lezaun, M.J., Jimenez, L., Lopez-Soria, L.M., Sota, M., Contreras, G., Najera, R., Perez-Alvarez, L., 2009. HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. J. Acquir. Immune Defic. Syndr. 51, 99–103.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M.J., van de Vijver, D.A., Boucher, C.A., Camacho, R., Vandamme, A.M., 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics 21, 3797–3800.

Donnelly, P., Tavaré, S., 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29, 401–421.

Drummond, A., Oliver, G., Rambaut, A., 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. 54, 331–358.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

ECDC, 2013. Men who have sex with men. Monitoring implementation of the Dublin Declaration on Partnership to Fight HIV/AIDS in Europe and Central Asia: 2012 progress report (www.ecdc.europa.eu).

ECDC/WHO, 2010. HIV/AIDS Surveillance in Europe 2009. European Centre for Disease Prevention and Control, Stockholm.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.

Holmes, E.C., 2004. The phylogeography of human viruses. Mol. Ecol. 13, 745–756.

Holmes, E.C., Zhang, L.Q., Robertson, P., Cleland, A., Harvey, E., Simmonds, P., Leigh Brown, A.J., 1995. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. J. Infect. Dis. 171, 45–53.

Hue, S., Clewley, J.P., Cane, P.A., Pillay, D., 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS 18, 719–728.

Hue, S., Pillay, D., Clewley, J.P., Pybus, O.G., 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. U. S. A. 102, 4425–4429.

Hughes, G.J., Fearnhill, E., Dunn, D., Lycett, S.J., Rambaut, A., Leigh Brown, A.J., on behalf of the UK HIV Drug Resistance Collaboration, 2009. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 5, e1000590.

Johnson, V.A., Calvez, V., Gunthard, H.F., Paredes, R., Pillay, D., Shafer, R.W., Wensing, A.M., Richman, D.D., 2013. Update of the drug resistance mutations in HIV-1: March 2013. Top. Antivir. Med. 21, 6–14.

Kingman, J.F.C., 1982. The coalescent. Stoch. Process. Appl. 13, 235–248.

Kouyos, R.D., von Wyl, V., Yerly, S., Böni, J., Taffé, P., Shah, C., Bürgisser, P., Klimkait, T., Weber, R., Hirschel, B., Cavassini, M., Furrer, H., Battegay, M., Vernazza, P.L., Bernasconi, E., Rickenbach, M., Ledergerber, B., Bonhoeffer, S., Günthard, H.F., 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J. Infect. Dis. 201, 1488–1497.

Lewis, F., Hughes, G.J., Rambaut, A., Pozniak, A., Leigh Brown, A.J., 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 5, e50.

Liu, T.F., Shafer, R.W., 2006. Web resources for HIV type 1 genotypic-resistance test interpretation. Clin. Infect. Dis. 42, 1608–1618.

Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.

Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.

Martin, D.P., Posada, D., Crandall, K.A., Williamson, C., 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retrovir. 21, 98–102.

Moya, A., Holmes, E.C., González-Candelas, F., 2004. The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2, 279–288.

Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218–225.

Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U. S. A. 98, 13757–13762.

Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One 5, e9490.

R Development Core Team, 2011. R: A language and environment for statistical computing. (http://www.R-project.org/. Vienna, Austria).

Roberts, J.D., Bebenek, K., Kunkel, T.A., 1988. The accuracy of reverse transcriptase from HIV-1. Science 242, 1171–1173.

Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23, 7–9.

Smith, J.M., 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34, 126–129.

UNAIDS/WHO, 2013. Global Report: UNAIDS Report on the Global AIDS Epidemic (Geneva).

Vega, Y., Delgado, E., Fernández-García, A., Cuevas, M.T., Thomson, M.M., Montero, V., Sánchez, M., Sánchez, A.M., Pérez-Álvarez, L., Spanish Group for the Study of New HIV, 2015. Epidemiological surveillance of HIV-1 transmitted drug resistance in Spain in 2004-2012: relevance of transmission clusters in the propagation of resistance mutations. PLoS One 10, e0125699.

Zehender, G., Ebranati, E., Lai, A., Santoro, M.M., Alteri, C., Giuliani, M., Palamara, G., Perno, C.F., Galli, M., Lo, P.A., Ciccozzi, M., 2010. Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. J. Acquir. Immune Defic. Syndr. 55, 156–160.

## 2.2- Chapter 2: The molecular epidemiology of HIV-1 in the Comunitat Valenciana (Spain): analysis of transmission clusters

***This work is in the final stage of preparation.***

# The molecular epidemiology of HIV-1 in the Comunitat Valenciana (Spain): analysis of transmission clusters

Juan Ángel Patiño-Galindo [1,2], Manoli Torres-Puente [1,2], María Alma Bracho [1,2,7], Ignacio Alastrué [3], Amparo Juan [3], David Navarro [2,4], María José Galindo [4], Dolores Ocete [5], Enrique Ortega [5], Concepción Gimeno [2,5], Josefina Belda [6], Victoria Domínguez [7], Rosario Moreno [7], Fernando González-Candelas [1,2,8,], on behalf of the CRIVIH

[1] FISABIO-CSISP, Valencia.

[2] Universidad de Valencia.

[3] Unidad Prevención del SIDA y otras ITS, Valencia.

[4] Hospital Clínico Universitario, Valencia.

[5] Hospital General Universitario, Valencia.

[6] Unidad Prevención del SIDA y otras ITS, Alicante.

[7] Hospital General Universitario, Castelló.

[8] CIBER Epidemiología y Salud Pública

**Abstract**

A total of 1806 HIV-1 sequences comprising protease and reverse transcriptase (PR/RT) coding regions, sampled between 2004 and 2014 in the Comunitat Valenciana (CV, Spain), were subtyped and subjected to phylogenetic analyses in order to detect transmission clusters. Also, univariate and multinomial comparisons were performed to detect epidemiological differences between HIV-1 subtypes and risk groups.

The HIV epidemic in the CV during the studied period is dominated by subtype B infections among local men who have sex with men (MSM). It also affects disproportionally to immigrants. We identified 270 transmission clusters of sizes ranging from 2 to 111 patients, representing more than 57% of the dataset. 12 clusters included more than 10 patients; 11 of subtype B (9 affecting MSM) and one (n=21) of CRF14, affecting predominately intravenous drug users (IDUs). Dated phylogenies revealed that these large clusters originated from the mid-80's to the early 21st century. After 2000, the only clusters originated were of MSM.

Univariate comparisons indicate that subtype B is more likely to form transmission clusters than non-B variants. MSM were also more likely to cluster than other risk groups. Multinomial analysis revealed an association between non-B variants and different groups of immigrants and people of foreign origin. Furthermore, CRF14 was associated to patients over 50 years of age and those forming large clusters.

In conclusion, the HIV epidemic in the CV is characterized by a majority of local HIV-1 B transmissions occurring among MSM. Non-B variants are not yet established in the local population, and mostly affect immigrants.

**Introduction**

Of the four phylogenetic groups which comprise HIV-1 (M, N, O, P), group M is the causal agent of the AIDS pandemic (Hahn et al. 2000; Plantier et al. 2009). The latest UNAIDS/WHO report (2016) estimated in almost 37 million the number of persons infected with HIV globally, with approximately 2.1 million new infections in 2015 (UNAIDS 2016).

Although the rate of new HIV diagnosis has stabilized from the early 2000s in the European Union and European Economic Area (EU/EEA), transmissions among men who have sex with men (MSM) have experienced a sustained increase, thus following a different trend to other risk groups(ECDC 2013). This is evident in Spain: during the late 90s, most HIV new diagnoses were associated to intravenous drug use (IDU). However, in 2013, 51% of the infections occurred among MSM (UNAIDS 2002; DGSP 2014). This increasing incidence among MSM is remarkably high in the age range 20-35 years (DGSP 2014). HIV infections in Spain also affect disproportionally the foreign population: in 2012, 35% of the new diagnosis corresponded to immigrants or persons of foreign origin (DGSP 2014).

Within HIV-1 group M, there exist nine subtypes (denoted as A, B, C, D, F, G, H, J and K) and at least 61 circulating recombinant forms (CRFs) (Kuiken et al. 2012). There are differences among HIV-1 variants in several biological features. For instance, some subtypes and CRFs are associated to a faster progression to AIDS than others (Kaleebu et al. 2002; Kouri et al. 2015). Genetic, and also antigenic, differences among HIV-1 subtypes and CRFs are also a challenge for the development of an effective HIV-1 vaccine (Nickle et al. 2007).

Worldwide, the most prevalent HIV-1 subtype is C, accounting for around 50% of all cases. However, the HIV epidemic in Europe, particularly among MSM, is mainly driven by subtype B (Abecasis et al. 2013), with frequent reported transmission clusters affecting them (Kouyos et al. 2010; Lewis et al. 2008; Leigh Brown et al. 2011; M. T. Cuevas et al. 2009; Zehender et al. 2010; Hué et al. 2005; Bezemer et al. 2015). However, there is evidence for an increased introduction of non-B subtypes (Paraskevis et al. 2006). For instance, in a sample of 206 patients from Spain, Abecasis el al. (2013) found that CRF02_AG was the second most prevalent HIV-1 variant after subtype B (prevalence < 2%).

The Comunitat Valenciana (CV) is the fourth largest region in Spain (~5 million inhabitants), representing >10% of the total population. Genotypic tests of resistance to antiviral drugs are performed routinely for the design of individualized antiretroviral treatments. Between 2004 and 2014, more than 1800 HIV-1 *pol* sequences have been obtained from different patients at eight public hospitals and HIV testing and counselling centers from the CV. Phylogenetic analyses of this large dataset, combined with

epidemiological data, are a powerful tool for depicting the HIV-1 epidemic in the CV and were used in this study to infer the distribution of HIV-1 subtypes, to analyze the introductions (and further local expansion) of this virus in the CV, to identify the emergence of new viral variants to this region, and also to analyze which population groups are currently more vulnerable to HIV infection. The results obtained from this work may be useful in establishing and reinforcing preventive measures in specific target groups.

## Materials and methods

### *Dataset*

A total of 1806 PR/RT sequences were obtained from newly HIV diagnosed people at six different hospitals and two HIV counseling and testing centers (CIPS) from the three provinces in the CV between 2004 and 2014. The sequences comprised the complete PR and the first 1005 nucleotides (335 amino acids) of the RT (1302 nt in total), and were obtained through viral RNA extraction followed by RT-PCR and direct sequencing using amplification and procedures described previously (Holguin et al., 2005). All sequences were subtyped using the Rega HIV-1 Subtyping Tool Version 3.0 (Pineda-Peña et al. 2013), the COMET HIV-1 Subtyping tool (http://comet.retrovirology.lu/) and an initial phylogenetic tree obtained with FastTree 2.1 (Price et al. 2010) which also included 133 reference sequences downloaded from the Los Alamos HIV Database

(http://www.hiv.lanl.gov), representing the diversity of HIV-1 group M. Nucleotide alignments were obtained with MAFFT version 7 (Katoh & Standley 2013).

*Detection of local transmission clusters*

Independent alignments were generated for each HIV-1 subtype and CRF detected, which included the CV sequences and those sequences from the same variant retrieved from the Los Alamos HIV database, spanning the analyzed PR/RT region and used as references. Phylogenetic analyses were initially performed with FastTree 2.1 (GTR+ Γ, 4CAT) in order to detect potential transmission clusters, which were defined as clades with SH-like support ≥ 0.70, and containing ≥ 90% sequences from the CV dataset.

The potential clusters identified in the first step were then validated as follows. All sequences from the CV in a potential cluster were used as query for a BLASTN search at the NCBI server (https://blast.ncbi.nlm.nih.gov/Blast.cgi). The 10 sequences with the highest similarity to each CV sequence were downloaded and included into the alignments. Then, ML phylogenies were obtained with PhyML 3.0 (Guindon et al. 2010) using the GTR+ Γ(4 CAT) model, and only those potential clusters which contained more than 90% of sequences from the CV and grouped in the ML tree with aLRT support ≥ 0.98 were considered as confirmed. Clusters were then classified depending on the major risk of transmission (≥50%) for the corresponding patients.

*Dated phylogenies*

The molecular clock signal for the transmission cluster containing at least 10 patients from the CV was assessed by performing linear regression analyses between the parameters “root-to-tip divergence” and “sampling date” with the software Path-O-Gen v1.4 (now renamed as TempEST; Rambaut et al. 2016). As input, we used the phylogenetic trees from each transmission cluster obtained as subtrees from the ML tree.

Dated phylogenies for each of these clusters were obtained using a Bayesian MCMC coalescent method as implemented in BEAST v1.8.1 (Drummond & Rambaut 2007), using the $GTR_{112}+\Gamma_{112}$ (4 CAT) model in all the analyses. In those clusters with low root-to-tip divergence vs sampling date correlation ($R<0.4$), a log-normal prior (median = 0.0025 per site and year, s/s/y, 95% HPD upper limit = 0.0035 s/s/y) was placed on the ucld.mean parameter (Hué et al. 2005). Under an uncorrelated relaxed molecular clock model, the most appropriate demographic model [either the exponential growth, logistic growth or Bayesian Skyline Plot (BSP)] was determined as the one with the lowest Akaike Information Criterion (AICM) value (Baele et al. 2012).

For each transmission cluster, at least two independent runs of Bayesian MCMC with chain lengths of at least 10 million states were performed. These runs were sampled every 1000 generations and then combined after discarding a 10% burn-in. All the evolutionary parameters were estimated from an effective sampling size > 200 determined with Tracer version 1.4. The retained trees were summarized using TreeAnnotator (http://beast.bio.ed.ac.uk/).

*Statistical analyses*

A multinomial logistic regression analysis was performed in order to identify the main predictors of HIV-1 subtype/CRF group distribution, considering relevant epidemiologic variables: country of origin, sex, risk group, collection date, age and clustering status. Due to the lack of epidemiological data for many sequences from the dataset, only 907 of the 1806 sequences were included in this analysis. Seven groups of different HIV-1 subtypes and CRFs were used: A1 (n = 14), B (n = 753), F1 (n = 13), G (n = 11), 02_AG (n = 42) and 14_BG (n = 15). All HIV-1 variants with fewer than 10 patients sampled were pooled as "Other variants" (n = 59). Prior to the multinomial analysis, univariate analyses (Fisher's Exact Tests) were performed for the aforementioned variables, in order to exclude from the multinomial analysis those with non-significant p-values. Only the variable "collection date" (p-value > 0.70) was excluded. In the multinomial analysis, the most representative category of each variable was used as "baseline category" (Subtype B, Spaniard, male, MSM, age between 21 and 29 years and not clustering; Supplementary Table S.1). All the statistical analyses were performed using R (R Core Team 2014). The mlogit R package (Croissant 2015) was used for the multinomial analysis.

**Results**

1514 of the 1806 sequences analyzed (83.8%) belonged to subtype B. Among the 292 non-B sequences, the most prevalent HIV-1 variant was CRF02_AG (n = 66, overall

prevalence = 3.6%), followed by subtypes A1, F1 (both n = 34, 1.9%), CRF14_BG (n = 28, 1.6%) and subtype G (n =20, 1.1%). Other variants (n =110) were present with a prevalence lower than 1.0%. Considering those patients for whom epidemiological information was available, 83.05% (931/1121) were male vs 16.95% (190/1121) female; 66.77% (633/948) were native from Spain vs 33.22% (315/948) immigrants or of foreign origin; 66.88% (638/954) were MSM, 21.28% (204/954) heterosexual (HT) and 11.53% (110/954) IDUs. One patient was infected vertically and other one was hemophiliac. The mean age was 34.84 years (range 0 to 76) (Table 1). MSM were more likely to be Spaniards than non-MSM (Fisher's Exact test, FET: p-value=$6x10^{-6}$, odds-ratio, OR = 2 (1.47-2.70).

**Table 1.** Distribution of HIV cases in the dataset (n=1806) classified by viral subtype, gender, nationality, risk group, age and clustering status.

| | B (n=1514) | A1 (n=34) | F1 (n=34) | G (n=20) | CRF02_AG (n=66) | CRF14_BG (n=28) | Others (n=110)* | Total (n=1806) |
|---|---|---|---|---|---|---|---|---|
| **Gender** | | | | | | | | |
| Male | 819 | 10 | 13 | 5 | 24 | 11 | 49 | 931 |
| Female | 115 | 6 | 6 | 8 | 25 | 6 | 24 | 190 |
| UNK | 580 | 18 | 15 | 7 | 17 | 11 | 37 | 685 |
| **Nationality** | | | | | | | | |
| Spain | 575 | 6 | 5 | 3 | 11 | 8 | 25 | 633 |
| Other | 202 | 10 | 11 | 10 | 33 | 8 | 41 | 315 |
| UNK | 737 | 18 | 18 | 7 | 22 | 12 | 44 | 858 |
| **Risk group** | | | | | | | | |
| HT | 124 | 8 | 6 | 9 | 29 | 2 | 26 | 204 |
| MSM | 587 | 3 | 6 | 0 | 11 | 1 | 30 | 638 |
| IDU | 85 | 3 | 0 | 2 | 3 | 12 | 5 | 110 |
| Other | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| UNK | 716 | 20 | 22 | 9 | 23 | 13 | 49 | 852 |
| **Clustering** | | | | | | | | |
| No cluster | 622 | 22 | 24 | 13 | 32 | 5 | 49 | 767 |
| Small cluster (2-3) | 371 | 12 | 5 | 7 | 25 | 2 | 39 | 461 |
| Medium cluster (4-9) | 240 | 0 | 5 | 0 | 9 | 0 | 22 | 276 |
| Large cluster (>=10) | 281 | 0 | 0 | 0 | 0 | 21 | 0 | 302 |
| **Age: mean (min - max)** | 35.24 (14-76) | 32.8 (19-56) | 33.5 (18-62) | 28.9 (21-41) | 30.6 (19-49) | 38.11 (22-65) | 33.2 (0-62) | 34.84 (0-76) |

*The "Others" subset includes 10 subtype C, 2 subtype D, 13 CRF19_cpx, 12 CRF12_BF, 9 CRF06_cpx, 7 CRF47_BF and 57 other (mostly unassigned) recombinant sequences.

Phylogenetic analyses revealed the existence of 270 transmission clusters, with sizes ranging from 2 to 111 patients (Figure 1.A). In total, 1039 patients from the dataset were included in a transmission cluster (57.5%), 302 of them (16.7%) were included in large clusters of 10 or more patients (Table 1; Figure 1.A). Among the 892 patients clustering in transmission groups of subtype B, 407 were MSM, 56 HTs, 38 IDUs and 391 of unknown transmission route. On the other hand, of the 147 patients clustering in non-B clusters 29 were HTs, 29 MSM, 15 IDUs and 74 of unknown transmission route (Figure 1.B). The most prevalent transmission clusters of subtype B were those classified as MSM, followed by IDUs, but for non-B transmission clusters of HTs were the most frequent ones, followed by those of MSM (Figure 1.C). Subtype B sequences were more likely to be part of a transmission cluster than non-B sequences (FET: p-value = 0.008, OR = 1.41, 95% CI = 1.09-1.83). Also, MSM were more likely to be part of a transmission cluster than groups other risk groups (FET: p-value = $1.68 \times 10^{-4}$, OR = 1.56, 95% CI = 1.23-1.99).

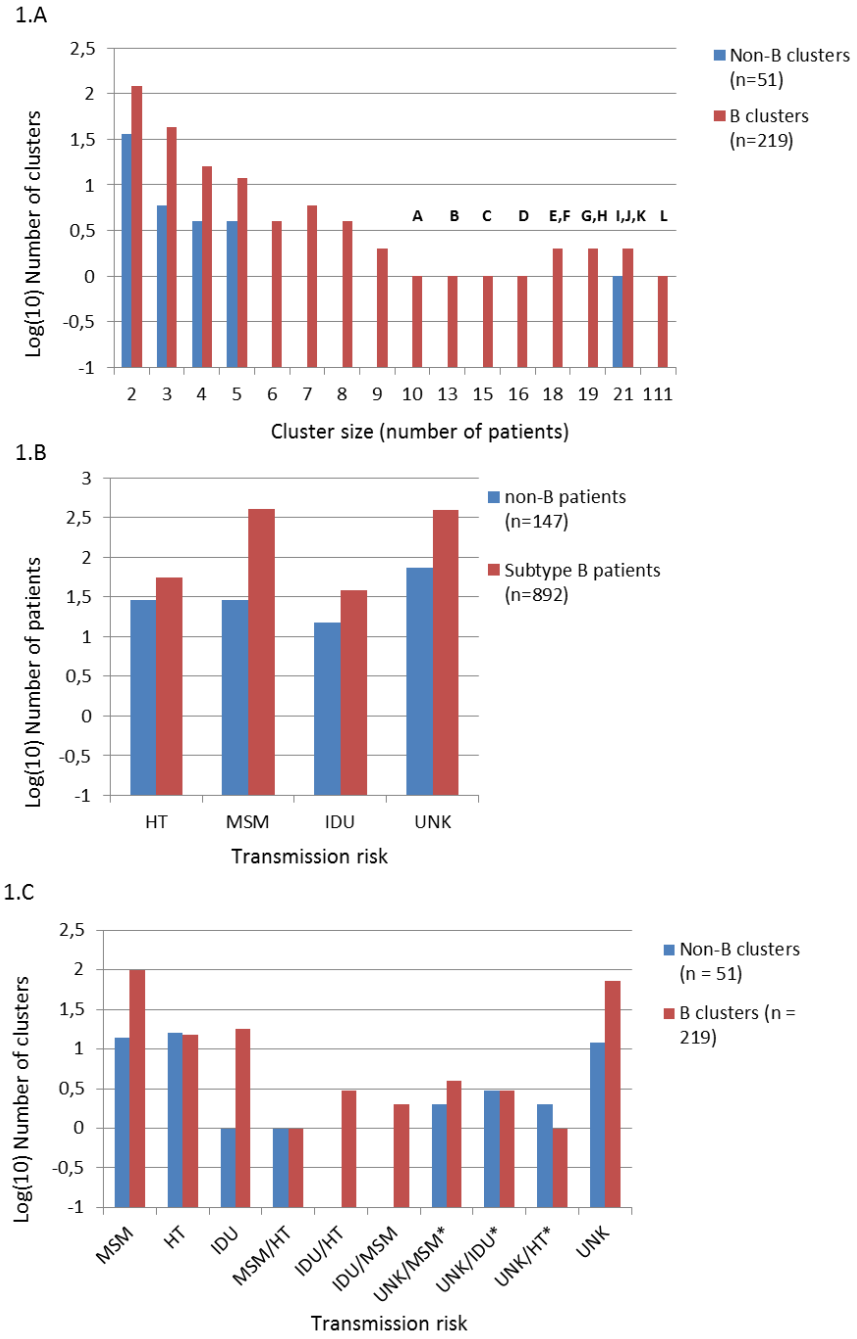
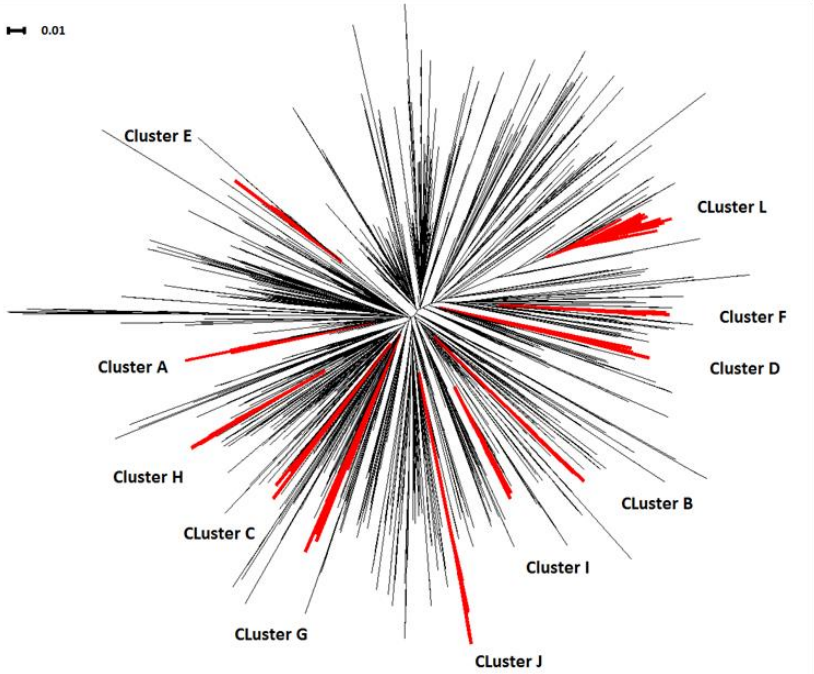



**Figure 1.** (A) Distribution of sizes (log10 scale) of the 270 transmission clusters found in the CV (2004-2014) through phylogenetic analysis. Block letters indicate the 12 clusters that were analyzed with BEAST. (B) Total number of patients for each risk group included in transmission clusters (n=1039). (C) Number of transmission clusters depending on the risk group in which they were classified. *: >1/4 patients shared a known risk group, but they were not enough to classify the cluster.

12 transmission clusters included at least 10 patients from the CV (Figure 2). 11 of them were from subtype B and included a total of 281 patients (Figure 2.A). One cluster of 21 patients corresponded to CRF14_BG (Cluster K, Figure 2.B). Patients infected with subtype B were more likely to form large transmission clusters than those infected with non-B variants (FET: p-value = $3.4 \times 10^{-7}$, OR = 2.94, 95% CI = 1.84-4.92).

2.A. Subtype B tree




2.B. Subtype G + 14_BG tree

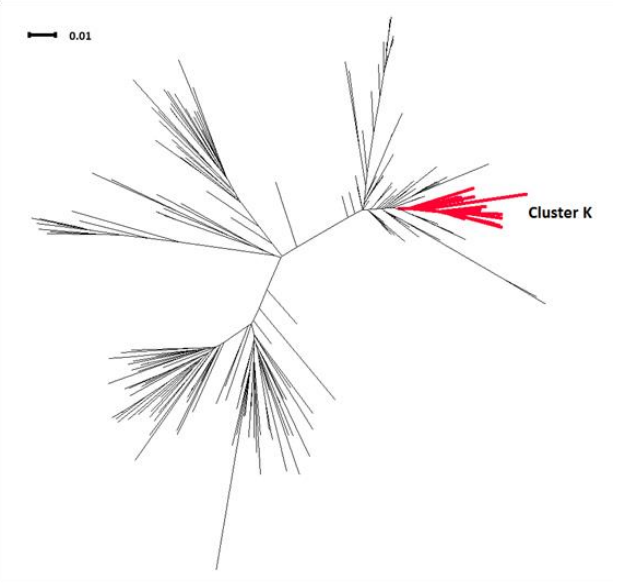



**Figure 2.** Maximum likelihood trees with the 12 largest transmission clusters highlighted in red. (A) Subtype B tree (clusters A-J, L). (B) Subtype G and CRF14_BG tree (cluster K).

9 of the largest clusters were classified as MSM; the other three included mostly IDUs, although in a proportion lower than 50% (Table 2). Dated phylogenies of these clusters revealed they originated between the years 1984 (cluster G, unclassified transmission route) and 2005 (cluster B, MSM). All the clusters originated since 2000 included MSM as the main transmission risk (Table 2; Supplementary Figure 1). Bagplots (bivariate representations of boxplots; Rousseeuw et al. 1999) representing tree height (the time elapsed between tMRCA and the sampling date of the last sequence) and evolutionary rate estimates from the posterior distribution of each large cluster are shown in Figure 3. A significant, negative correlation between median tree height and substitution rate estimates of these large clusters was obtained ($R = -0.70$, p-value = 0.013).

**Table 2.** Size (n, number of patients), risk group, range of sampling dates and root-to-tip vs sampling date correlation coefficient for each transmission cluster (R), and estimates of their tMRCAs (median) and substitution rates as obtained with BEAST under the best-fitting demographic model (EXPO: exponential growth; BSP: Bayesian Skyline Plot; LOG: logarithmic growth).

| Cluster | n | risk group | range # | R | Model | tMRCA (95% HPD)# | Substitution rate (95% HPD)* |
|---|---|---|---|---|---|---|---|
| A | 10 | MSM | 2004 - 2011 | (-0.08) | EXPO | 1984.7 (1971.2 - 1994.5) | 0.0014 (0.0009 - 0.0021) |
| B | 13 | MSM | 2007 - 2013 | 0.61 | BSP | 2005.9 (2003.8 - 2006.9) | 0.0043 (0.0016 -0.0087) |
| C | 15 | UNK/IDU | 2004 - 2014 | (-0.18) | BSP | 1988.0 (1969.0 - 1999.2) | 0.0015 (0.0007 - 0.0025) |
| D | 16 | MSM | 2011- 2013 | (-0.05) | LOG | 1994.9 (1982.4 -2006.3) | 0.0022 (0.0012 - 0.0034) |
| E | 18 | UNK/MSM | 2010 - 2014 | 0.25 | EXPO | 2002.2 (1995.9 -2006.8) | 0.0018 (0.0011 - 0.0027) |
| F | 18 | MSM | 2008 - 2013 | 0.40 | LOG | 1999.3 (1989.6 - 2005.0) | 0.0013 (0.0005 -0.0023) |
| G | 19 | UNK | 2004 - 2013 | (-0.05) | EXPO | 1984.3 (1966.9 -1997.0) | 0.0011 (0.0006 - 0.0019) |
| H | 19 | MSM | 2008 -2014 | 0.58 | BSP | 2002.1 (1995.9 -2006.2) | 0.0019 (0.0010. 0.0029) |
| I | 21 | MSM | 2004 - 2012 | 0.58 | LOG | 1994.6 (1982.2 -2001.3) | 0.0012 (0.0005 -0.0020) |
| J | 21 | MSM | 2004 - 2013 | 0.76 | LOG | 1999.5 (1993.2 - 2003.4) | 0.0023 (0.0010 - 0.0045) |
| K | 21 | UNK/IDU | 2004 -2014 | 0.62 | BSP | 1990.6 (1979.6 -1998.1) | 0.0010 (0.0005 - 0.0016) |
| L | 111 | MSM | 2006 - 2014 | 0.61 | EXPO | 2001.8 (1998.6 -2004.8) | 0.0028 (0.0022 - 0.0035) |

# Time measured in years.

*Substitutions per site and year.

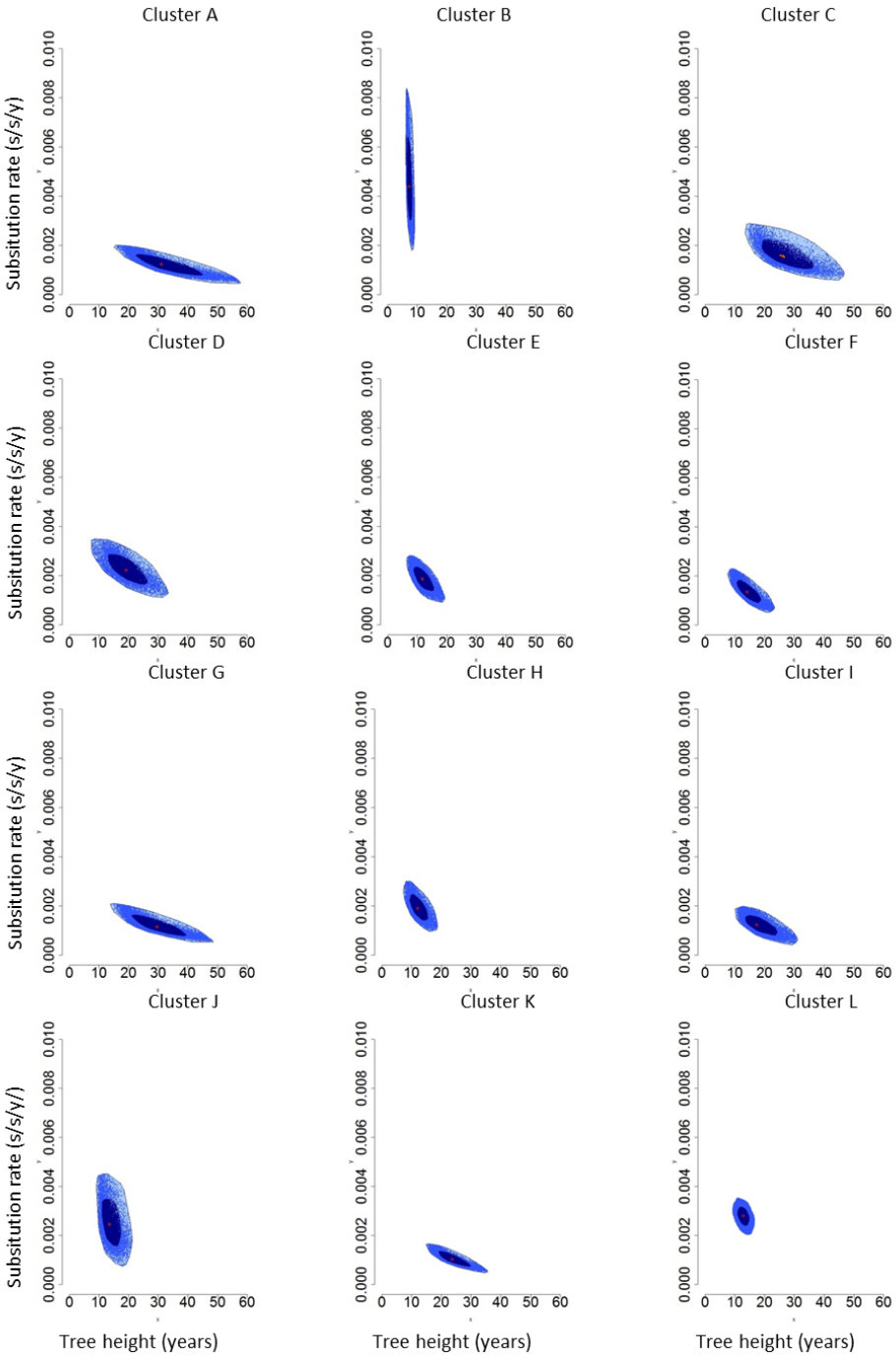



**Figure 3.** Bagplots representing tree height and evolutionary rate estimates, obtained from the posterior distribution of the 12 largest transmission clusters detected (A to L).

Despite most patients infected with subtype B being Spanish natives (575 Spaniards vs 202 foreigners), a majority of those infected with non-B variants were foreigners (58 vs 113; FET: p-value < $2.2x10^{-16}$, OR = 5.50, 95% CI = 3.81-8.01). This difference was also observed when considering only patients belonging to a transmission cluster (subtype B: 370 vs 121; non-B: 31 vs 47; FET: p-value = $1.1x10^{-9}$; OR = 4.62, 95% CI= 2.74 – 7.89). Other univariate analyses for the variables considered for the multinomial analysis revealed differences between subtypes: subtype B sequences were disproportionally more present in males compared to females (FET: p-value < $2.2x10^{-16}$, OR = 4.76, 95% CI = 3.30-6.87), in MSM with respect to HTs and IDUs (p-value < $2.2x10^{-16}$, OR = 7.4, 95% CI = 4.88-11.32 and p-value =$2x10^{-5}$, OR = 3.37, 95% CI = 1.9 – 5.9, respectively), and in IDUs compared to HTs (p-value = 0.0038, OR=2.19, 95% CI = 1.26-3.88). There were also intersubtype differences regarding age distributions (Kruskal-Wallis test: P = 0.0013, $Chi^2$ = 21.8, df = 6), with CRF02_AG patients being significantly younger than B patients (Games-Howell post-hoc test result for this comparison: p = 0.0049, t = 3.88, df = 56). Given that no significant differences were found between subtypes in the distribution of collection dates (P > 0.70), this variable was excluded from the multinomial analysis.

Multinomial analyses were performed using a subset of 907 patients for whom there was information on all the different variables analyzed (sex, risk group, country of origin, clustering status and age) considering as baseline the most frequent category for each variable (Subtype B, Spaniard, male, MSM, age between 21 and 29 years and not clustering), and a significant model was obtained (McFadden $R^2$ = 0.32, $Chi^2$ = 414.55, p-

value < $2.2x10^{-16}$). We checked that this subset was representative of the global set (n=1806) by means of FETs, and the p-values obtained for all the analyzed variants were > 0.05. The multivariate model shows that the chance of being infected with all non-B groups increased in foreign patients, coming from Eastern Europe (A1, F1, 14_BG and the pooled, rare, variants), Africa and the Middle East (F1, G, 02_AG and the rare variants) and Latin America (rare variants), with all p-values < 0.01, all ORs > 3.0. The likelihood of being infected with CRF14_BG also increased in patients older than 50 years (p-value = 0.028, OR = 36.2, 95% CI = 1.45 – 904.0), in IDUs (p-value = $9.43x10^{-5}$, OR = 257, 95% CI = 15.9 – 4160) and in those patients forming large transmission (p-value = $1.6x10^{-4}$, OR = 36.9, 95% CI = 5.67 – 240) (Table 3).

**Table 3.** Results of the multinomial analysis (only significant associations between each HIV variant and the categories compared at each variable are shown)

Coefficients

| | Estimate | Std error | t-value | Pr(>\|t\|) | | OR | 2.50% | 97.50% |
|---|---|---|---|---|---|---|---|---|
| 14_BG:Age(>50) | 3.59 | 1.64 | 2.188 | 0.028696 | * | 36.2 | 1.45 | 904 |
| 14_BG:RISK(HT) | 2.39 | 1.43 | 1.673 | 0.094242 | $ | 11.0 | 0.664 | 181 |
| 02_AG:RISK(HT) | 1.07 | 0.579 | 1.850 | 0.064354 | $ | 2.92 | 0.938 | 9.08 |
| 14_BG:RISK(UDI) | 5.55 | 1.42 | 3.905 | 9.43E-05 | *** | 257.0 | 15.9 | 4.16E+03 |
| 14_BG:Eastern Europe | 3.29 | 1.09 | 3.023 | 0.002502 | ** | 26.9 | 3.18 | 227 |
| F:Eastern Europe | 3.51 | 1.08 | 3.251 | 0.001151 | ** | 33.6 | 4.04 | 280 |
| OTHERS: Eastern Europe | 2.01 | 0.584 | 3.500 | 0.000561 | *** | 7.49 | 2.39 | 23.5 |
| A1:Eastern Europe | 2.78 | 0.795 | 3.494 | 0.000476 | *** | 16.1 | 3.39 | 76.5 |
| F1:Africa and Middle East | 3.51 | 1.05 | 3.355 | 0.000793 | *** | 33.3 | 4.30 | 259 |
| G:Africa and Middle East | 3.78 | 1.17 | 3.236 | 0.001215 | ** | 43.7 | 4.43 | 431 |
| OTHERS:Africa and Middle East | 2.77 | 0.554 | 5.013 | 5.37E-07 | *** | 16.0 | 5.42 | 47.5 |
| A1:Africa and Middle East | 1.92 | 0.993 | 1.934 | 0.053091 | $ | 6.82 | 0.975 | 47.8 |
| 02_AG:Africa and Middle East | 3.76 | 0.580 | 6.481 | 9.11E-11 | *** | 43.0 | 13.8 | 134 |
| 14_BG:Latin America | 2.71 | 1.38 | 1.956 | 0.050527 | $ | 15.0 | 0.994 | 225 |
| OTHERS:Latin America | 1.20 | 0.352 | 3.401 | 0.000672 | ** | 3.31 | 1.66 | 6.60 |
| 14_BG:ClusterLARGE | 3.61 | 0.955 | 3.776 | 0.000159 | *** | 36.9 | 5.67 | 240 |

***: P<0.001; **: 0001 < P < 0.01; *: 0.01 < P < 0.05; $: 0.05 < P < 0.1.

**Discussion**

We have studied the HIV epidemic in the Comunitat Valenciana by analyzing with molecular and evolutionary tools 1806 sequences obtained between 2004 and 2014 from different patients. Our results indicate that the HIV epidemic in the CV is dominated by HIV-1 subtype B infections among local MSM. However, non-B infections represented an important number of cases, with a prevalence higher than 15%, being CRF02_AG the most prevalent non-B variant (prevalence = 3.6%), in agreement with previous estimates for Spain obtained by Abecasis et al. (2013) and Yebra et al. (2012).

Immigrants were disproportionally represented in the dataset (almost 1/3 of the patients were of non-Spanish origin), reflecting their higher vulnerability to HIV infection. Multinomial analyses evidenced the significant association between all non-B groups analyzed and different foreign populations: Eastern Europe (subtypes A1, F1, CRf14_BG and the rare variants), Africa and the Middle East (subtypes F1, G, CRF02_AG and rare variants) and Latin America (rare variants). These associations were in agreement with the geographical distributions of these variants (Hemelaar 2012; Buonaguro et al. 2007; Bello et al. 2012). These results, along with the fact that most non-B patients, either clustering or not, were of non-Spanish origin (CRF14_BG was the only exception) suggest that along the analyzed time-span non-B HIV variants were not well established among Spanish locals in this region. Furthermore, non-B patients clustered in a significantly lower proportion than patients infected with subtype B, especially when considering clusters with at least 10 patients, thus displaying significantly lower local transmission efficiency. The significant association between non-B HIV variants and immigrants reported in this work and in

previous molecular epidemiology analyses of HIV-1 in Madrid, Spain (González-Alba et al. 2011; Yebra et al. 2013), suggest that high migration and tourism rates existing in some Spanish regions may explain the high genetic diversity of their HIV-1 epidemics.

Overall, the detection of transmission clusters demonstrates the importance of the domestic spread of HIV-1 in the CV. Most patients from the whole dataset (>57%) were included in local transmission clusters, of sizes ranging from 2 to 111 individuals. Local transmission was especially important among MSM, who were more likely to belong to a transmission cluster, as well as for Spanish natives, than other risk groups. Considering the 12 largest clusters (size ≥ 10 patients), this transmission risk was the most prevalent in 9 of them. Excluding cluster A (MSM, tRMCA= 1984.7), MSM clusters were of more recent origin as estimated using Bayesian coalescent analyses, especially clusters B (n=13), E (n=18), H (n=19) and L (n=111), which originated after year 2000. Although previous analyses in the Spanish regions of Madrid (Yebra et al. 2013) and the Basque Country (Patiño-Galindo et al. 2016) detected lower proportions of clustering patients (18 and 27% of their analyzed sequences, respectively), suggesting that the importance of local transmissions may not be the same in different Spanish regions, they also found evidences for an increased vulnerability to HIV of the Spanish MSM community in recent years.

Although 11 of the 12 largest transmission clusters were of subtype B, one of the largest clusters found (n=21, median tMRCA= 1990.6) corresponded to the CRF14_BG, and included a high number of IDUs from Spain. CRF14 has been associated to a predominance of CXCR4 tropism, which usually leads to faster AIDS onset (Bártolo et al. 2011; Pérez-Álvarez et al. 2014). The multinomial analysis of our dataset showed that this highly

pathogenic variant is significantly associated to IDUs, immigrants from Eastern Europe, older than 50 years and/or forming large transmission clusters. Previous phylogeographic analyses have suggested that CRF14_BG originated in the Iberian Peninsula (Bártolo et al. 2011), and its prevalence is increasing in some Eastern European countries, such as Romania, boosted by migration events between these countries and Spain (Niculescu et al. 2015).

We also detected two other transmission clusters of smaller size (n=5 and n=4) affecting Spanish-native MSM of another highly pathogenic variant (CRF19_cpx), which we previously reported as the first evidence of expansion of this variant outside Cuba (Patiño Galindo et al. 2015). Despite the prevalence of these CRFs remaining low in the CV, the effective expansion of these highly pathogenic HIV variants, evidenced by the detection of transmission clusters, is of especial interest because it may hamper the control of the local HIV epidemic, especially among vulnerable populations such as IDUs or MSM.

One potential difficulty in the analysis of this dataset was the number of sequences for which no epidemiological information (sex, transmission risk, or country of origin) was available. However, the number of cases with available information for all the variables was > 900. Furthermore, the distributions of all the variables included in the multinomial analysis were not significantly different to those from the whole dataset, including cases with incomplete information. In consequence, the results reported are representative of the sampled population and represent one of the most comprehensive analysis of the HIV pandemics in Spain (cf. González-Alba et al. 2011; Yebra et al. 2012; Yebra et al. 2013; Yebra et al. 2014).

It is noteworthy to mention the significant, negative correlation found between tree height and evolutionary rate estimates obtained when comparing the 12 largest clusters. Several publications have addressed this time-dependency on the evolutionary rate (TDRP), that is, virus lineages estimated to have a recent origin yield higher estimated evolutionary rates than older viruses (Ho et al. 2015; Aiewsakun & Katzourakis 2015; Aiewsakun & Katzourakis 2016). Although temporal changes in selective pressure and/or viral biology could be a reason for TDRP, in our study the most plausible explanation for this phenomenon is the overestimation of evolutionary rates in the most recent clusters, caused by the presence of deleterious mutations over which purifying selection has not had time to act. This phenomenon might be potentiated by the effect of the bottlenecks that occur at viral transmission, which usually cause a fast accumulation of deleterious mutations. Other possible causes, such as errors in calibration, were excluded after repeating the correlation tests between tree height and evolutionary rate without those clusters with low molecular clock signal, because similar results were obtained. Finally, the presence of skewed rate distributions was excluded by checking them graphically. Consequently, these results suggest that TDRP is an important factor to consider in molecular epidemiology, even when datasets are obtained from the same, short timescale (in this case 10 years, from 2004 to 2014).

It is also important to mention that many works on the molecular epidemiology of HIV-1 remove resistance-associated positions from the analysis. This is usually done because the selection of resistant variants caused by the selective pressures imposed by antivirals may produce spurious clustering by convergent evolution. In our analysis we did

not remove these positions, because their impact on the detection of transmission clusters has been demonstrated to be irrelevant (Hué et al. 2004). Furthermore, a great majority of our sequences were known to derive from treatment-naïve patients (only 66 patients were known to have been treated), and none of the large transmission clusters was defined by a shared resistant mutation.

In conclusion, our results evidence that the HIV-1 epidemic in the CV is dominated by subtype B, especially among local MSM. Although there were an important number of non-B cases, they occurred mostly among immigrants. This suggests that non-B infections are not well established in the local population. However, the detection of transmission clusters of non-B variants associated to a higher pathogenicity and affecting Spanish patients, urge to increase efforts on HIV testing and prevention campaigns to prevent their further expansion.

**References**


Abecasis AB, Wensing AMJ, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DAMC, Albert J, Asjö B, Balotta C, Beshkov D, et al. 2013. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 10:7.

Aiewsakun P, Katzourakis A. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol. Biol. 15:119.

Aiewsakun P, Katzourakis A. 2016. Time-Dependent Rate Phenomenon in Viruses. J. Virol. 90:7184–7195.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Bártolo I, Abecasis AB, Borrego P, Barroso H, McCutchan F, Gomes P, Camacho R, Taveira N. 2011. Origin and Epidemiological History of HIV-1 CRF14_BG.Martin DP, editor. PLoS One 6:e24130.

Bello G, Afonso JM, Morgado MG. 2012. Phylodynamics of HIV-1 subtype F1 in Angola, Brazil and Romania. Infect. Genet. Evol. 12:1079–1086.

Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, Gras L, Rodrigues Faria N, van den Hengel R, Duits AJ, et al. 2015. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLoS Med. 12:e1001898; discussion e1001898.

Buonaguro L, Tornesello ML, Buonaguro FM. 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J. Virol. 81:10209–10219.

Croissant Y. 2015. Package mlogit. CRAN. Available at: https://cran.r-project.org/web/packages/mlogit/index.html

Cuevas MT, Muñoz-Nieto M, Thomson MM, Delgado E, Iribarren JA, Cilla G, Fernández-García A, Santamaría JM, Lezaun MJ, Jiménez L, et al. 2009. HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. J. Acquir. Immune Defic. Syndr. 51:99–103.

DGSP. 2014. Vigilancia epidemiológica del VIH/SIDA enEspaña: Sistema de información sobre nuevos diagnósticos del VIH y registro nacional de casos de SIDA.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

ECDC. 2013. Thematic report : Men who have sex with men. Monitoring implementation of the Dublin Declaration on Partnership to fight HIV/AIDS in Europe and Central Asia: 2012 progress report.

González-Alba JM, Holguín A, Garcia R, García-Bujalance S, Alonso R, Suárez A, Delgado R, Cardeñoso L, González R, García-Bermejo I, et al. 2011. Molecular surveillance of HIV-1 in Madrid, Spain: a phylogeographic analysis. J. Virol. 85:10755–10763.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. 2000. AIDS as a zoonosis: scientific and public health implications. Science 287:607–614.

Hemelaar J. 2012. The origin and diversity of the HIV-1 pandemic. Trends Mol. Med. 18:182–192.

Ho SYW, Duchêne S, Molak M, Shapiro B. 2015. Time-dependent estimates of molecular evolutionary rates: evidence and causes. Mol. Ecol. 24:6007–6012.

Holguín A, Álvarez A, Soriano V. 2005. Heterogeneous nature of HIV-1 recombinants spreading in Spain. J. Med. Virol. 75:374–380.

Hué S, Clewley JP, Cane PA, Pillay D. 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS 18:719–728.

Hué S, Pillay D, Clewley JP, Pybus OG. 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. U. S. A. 102:4425–4429.

Kaleebu P, French N, Mahe C, Yirrell D, Watera C, Lyagoba F, Nakiyingi J, Rutebemberwa A, Morgan D, Weber J, et al. 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J. Infect. Dis. 185:1244–1250.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kouri V, Khouri R, Alemán Y, Abrahantes Y, Vercauteren J, Pineda-Peña A-C, Theys K, Megens S, Moutschen M, Pfeifer N, et al. 2015. CRF19_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba. EBioMedicine 2:244–254.

Kouyos RD, von Wyl V, Yerly S, Böni J, Taffé P, Shah C, Bürgisser P, Klimkait T, Weber R, Hirschel B, et al. 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J. Infect. Dis. 201:1488–1497.

Kuiken CL, Foley B, Leitner T, Apetrei C, Mizrachi Y, Mullins JI, Rambaut A, Wolinsky SM, Korber B. 2012. HIV Sequence Compendium 2012. New Mexico: Theoretical Biology and Niophysics Group. Los Alamos national Laboratory

Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT, UK HIV Drug Resistance Collaboration. 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J. Infect. Dis. 204:1463–1469.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 5:e50.

Nickle DC, Rolland M, Jensen MA, Pond SLK, Deng W, Seligman M, Heckerman D, Mullins JI, Jojic N. 2007. Coping with viral diversity in HIV vaccine design. PLoS Comput. Biol. 3:e75.

Niculescu I, Paraschiv S, Paraskevis D, Abagiu A, Batan I, Banica L, Otelea D. 2015. Recent HIV-1 Outbreak Among Intravenous Drug Users in Romania: Evidence for Cocirculation of CRF14_BG and Subtype F1 Strains. AIDS Res. Hum. Retroviruses 31:488–495.

Paraskevis D, Wensing AMJ, Vercauteren J, Vijver DA, Albert J, Asjo B, . on behalf of the SP. 2006. Prevalence of HIV-1 subtypes among newly HIV-1 diagnosed individuals during 2002–2003 in Europe: Evidence for a continuous introduction of non-B subtypes. In: 1st International Workshop on HIV Transmission; Toronto, Canada 200. p. p.31. Abstract N$^{o}$ 34.

Patiño Galindo JA, Torres-Puente M, Gimeno C, Ortega E, Navarro D, Galindo MJ, Navarro L, Navarro V, Juan A, Belda J, et al. 2015. Expansion of the CRF19_cpx Variant in Spain. J. Clin. Virol. 69:146–149.

Patiño-Galindo JA, Thomson MM, Pérez-Álvarez L, Delgado E, Cuevas MT, Fernández-García A, Nájera R, Iribarren JA, Cilla G, López-Soria L, et al. 2016. Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain. Infect. Genet. Evol. 40:91–97.

Pérez-Álvarez L, Delgado E, Vega Y, Montero V, Cuevas T, Fernández-García A, García-Riart B, Pérez-Castro S, Rodríguez-Real R, López-Álvarez MJ, et al. 2014. Predominance of CXCR4 tropism in HIV-1 CRF14_BG strains from newly diagnosed infections. J. Antimicrob. Chemother. 69:246–253.

Pineda-Peña A-C, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, Gómez-López A, Camacho RJ, de Oliveira T, Vandamme A-M. 2013. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect. Genet. Evol. 19:337–348.

Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F. 2009. A new human immunodeficiency virus derived from

gorillas. Nat. Med. 15:871–872.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical computing.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2:vew007.

Rousseeuw PJ, Ruts I, Tukey JW. 1999. The Bagplot: A Bivariate Boxplot. Am. Stat. 53:382–387.

UNAIDS. 2016. Fact sheet – Latest global and regional statistics on the status of the AIDS epidemic. GENEVA

UNAIDS. 2002. Report on the global HIV/AIDS epidemic. GENEVA.

Yebra G, Delgado R, Pulido F, Rubio R, Galán JC, Moreno S, Holguín Á. 2014. Different trends of transmitted HIV-1 drug resistance in Madrid, Spain, among risk groups in the last decade. Arch. Virol. 159:1079–1087.

Yebra G, Holguín A, Pillay D, Hué S. 2013. Phylogenetic and demographic characterization of HIV-1 transmission in Madrid, Spain. Infect. Genet. Evol. 14:232–239.

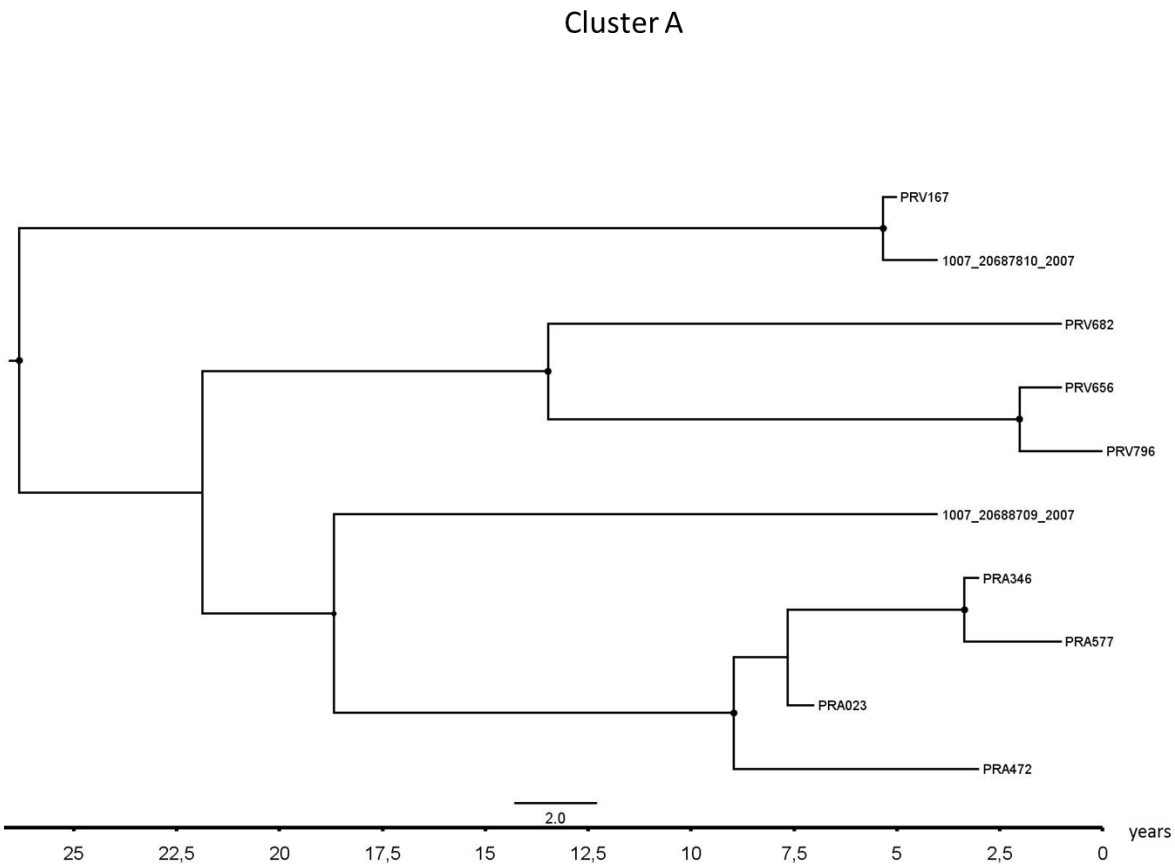
Yebra G, de Mulder M, Martín L, Rodríguez C, Labarga P, Viciana I, Berenguer J, Alemán MR, Pineda JA, García F, et al. 2012. Most HIV type 1 non-B infections in the Spanish cohort of antiretroviral treatment-naïve HIV-infected patients (CoRIS) are due to recombinant viruses. J. Clin. Microbiol. 50:407–413.

Zehender G, Ebranati E, Lai A, Santoro MM, Alteri C, Giuliani M, Palamara G, Perno CF, Galli M, Lo Presti A, et al. 2010. Population dynamics of HIV-1 subtype B in a cohort of men-having-sex-with-men in Rome, Italy. J. Acquir. Immune Defic. Syndr. 55:156–160.

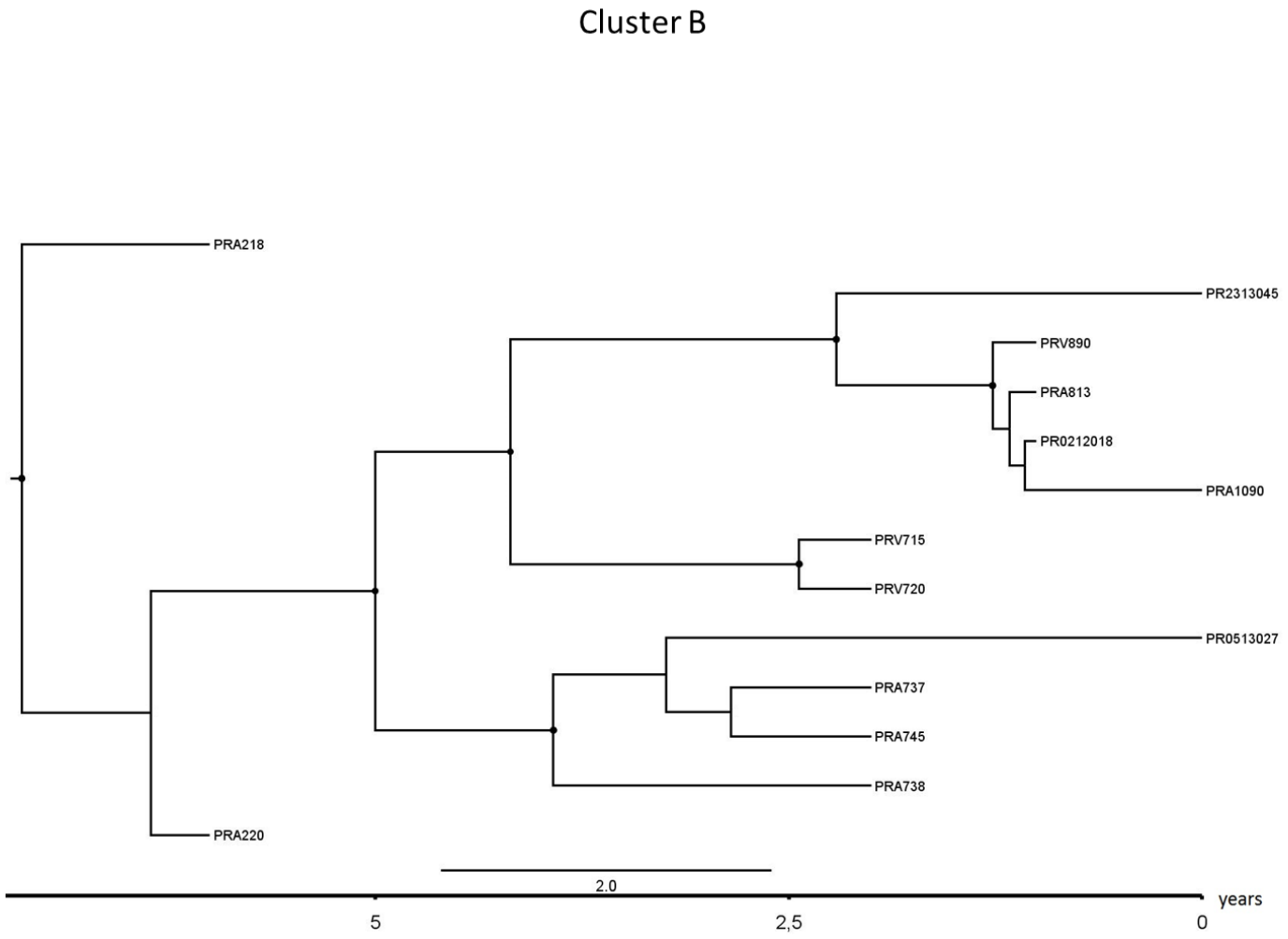
## Supplementary material

**Supplementary Table S.1**. Distribution of HIV cases in the dataset subjected to multinomial analysis (n=907), regarding the variables subtype, gender, age, nationality, risk group and clustering status

| | B (n=753) | A1 (n=14) | F1 (n=13) | G (n=11) | CRF02_AG (n=42) | CRF14_BG (n=15) | Others (n=59) | Total (n=907) |
|---|---|---|---|---|---|---|---|---|
| **Gender** | | | | | | | | |
| Male | 673 | 9 | 8 | 4 | 22 | 9 | 41 | 766 |
| Female | 80 | 5 | 5 | 7 | 20 | 6 | 18 | 141 |
| **Age** | | | | | | | | |
| <=20 | 30 | 1 | 2 | 0 | 1 | 0 | 1 | 35 |
| 21-29 | 246 | 5 | 3 | 9 | 24 | 2 | 25 | 314 |
| 30-35 | 195 | 4 | 2 | 1 | 9 | 6 | 14 | 231 |
| 36-50 | 252 | 3 | 6 | 1 | 8 | 4 | 18 | 292 |
| >50 | 30 | 1 | 0 | 0 | 0 | 3 | 1 | 35 |
| **Nationality** | | | | | | | | |
| Spain | 556 | 6 | 4 | 3 | 11 | 8 | 20 | 608 |
| W.Europe and N America | 27 | 0 | 0 | 0 | 1 | 0 | 1 | 29 |
| Eastern Europe | 15 | 5 | 3 | 0 | 0 | 4 | 6 | 33 |
| Africa and M. East | 13 | 2 | 3 | 8 | 23 | 0 | 11 | 60 |
| Latin America | 141 | 1 | 3 | 0 | 7 | 3 | 20 | 175 |
| Others | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2 |
| **Risk group** | | | | | | | | |
| HT | 119 | 8 | 7 | 9 | 29 | 2 | 24 | 198 |
| MSM | 565 | 3 | 6 | 0 | 11 | 1 | 30 | 616 |
| IDU | 67 | 3 | 0 | 2 | 2 | 12 | 5 | 91 |
| Other | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 |
| **Clustering** | | | | | | | | |
| No | 278 | 10 | 10 | 5 | 23 | 3 | 34 | 363 |
| Small cluster (2-3) | 170 | 4 | 1 | 6 | 14 | 2 | 18 | 215 |
| Medium cluster (4-9) | 142 | 0 | 2 | 0 | 5 | 0 | 7 | 156 |
| Large cluster (>=10) | 163 | 0 | 0 | 0 | 0 | 10 | 0 | 173 |

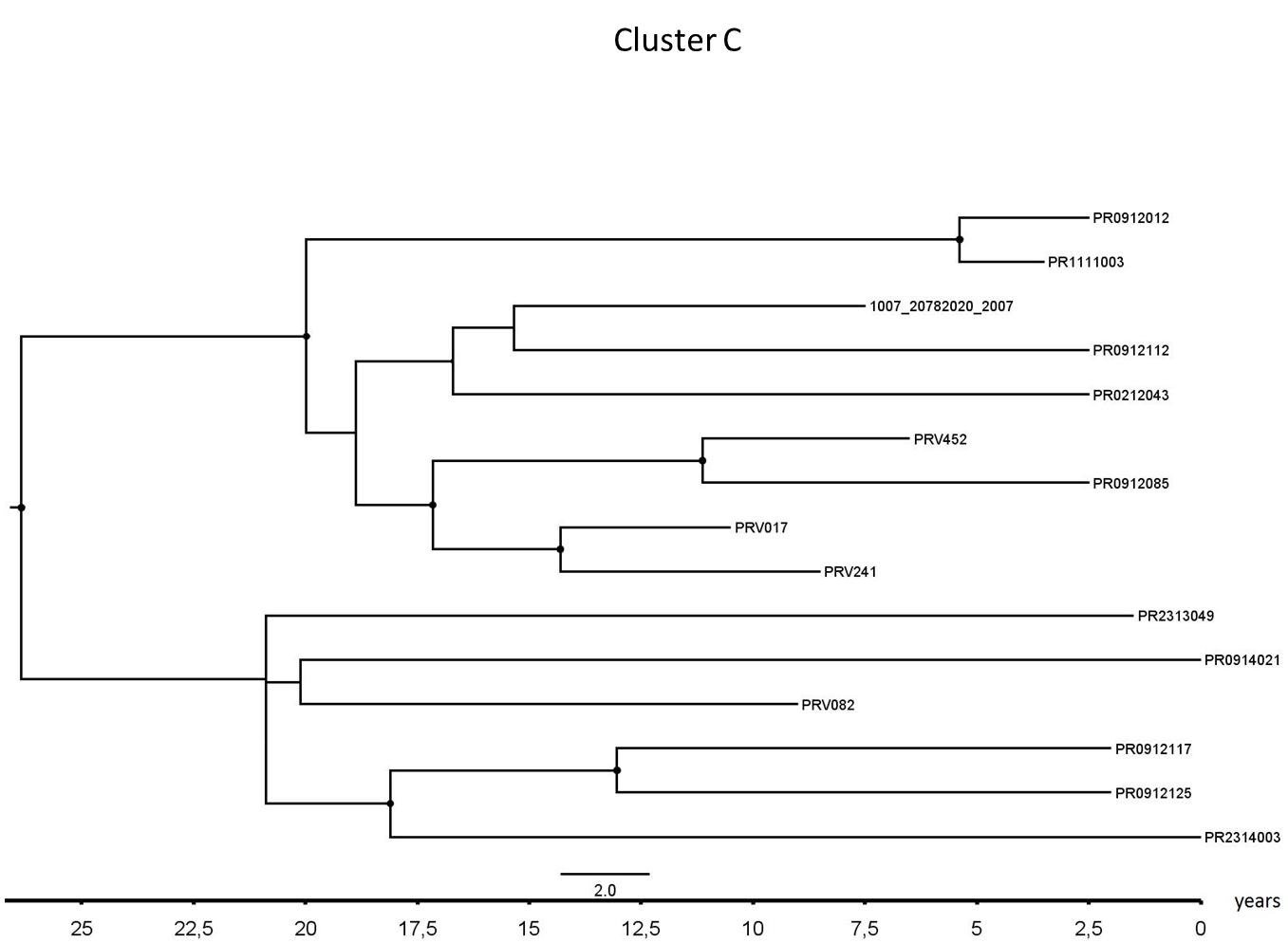



**Supplementary Figure S1.** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years. Nodes with Posterior Probabilities ≥0.90 are represented with black dots.



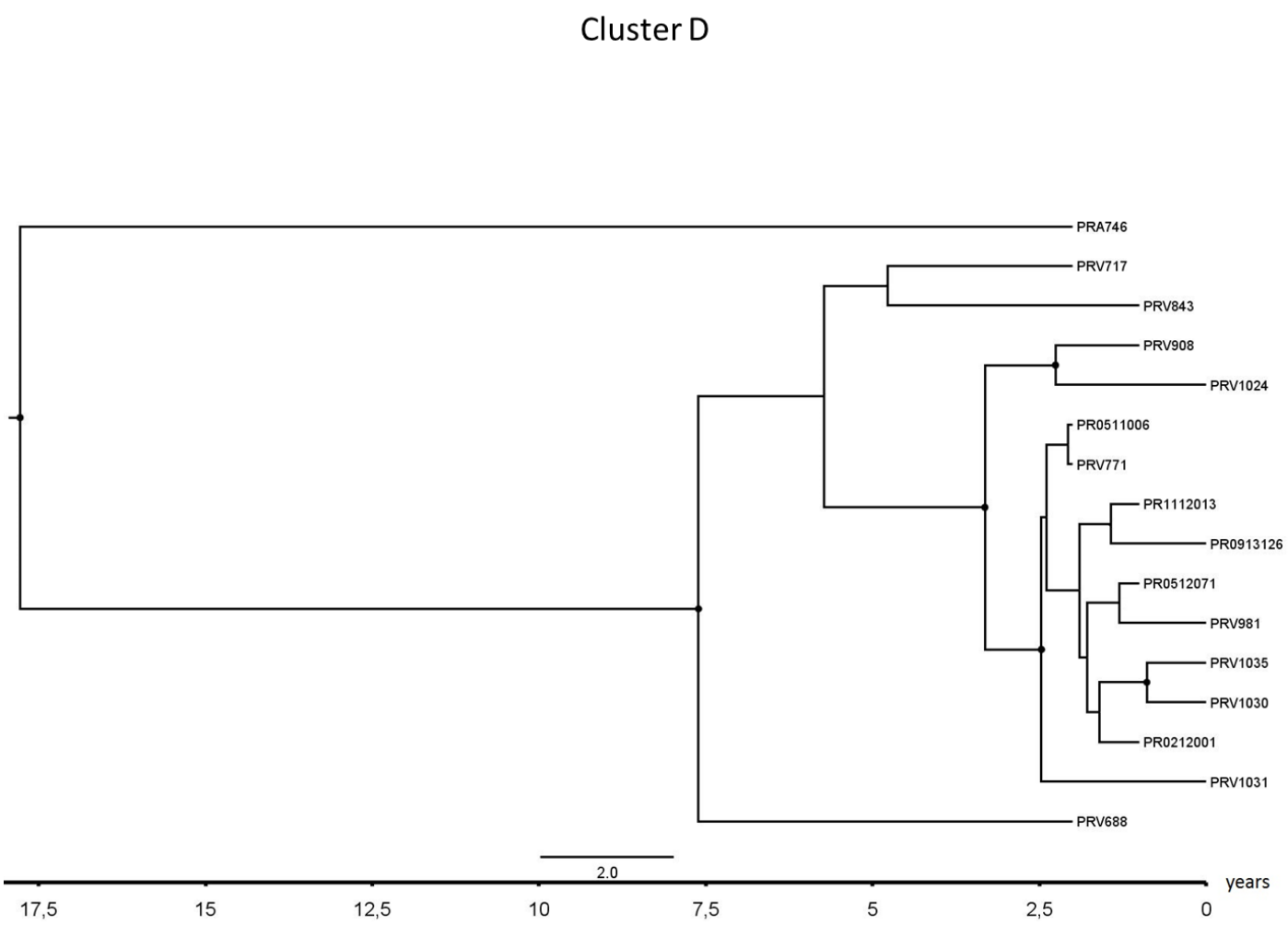

**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.



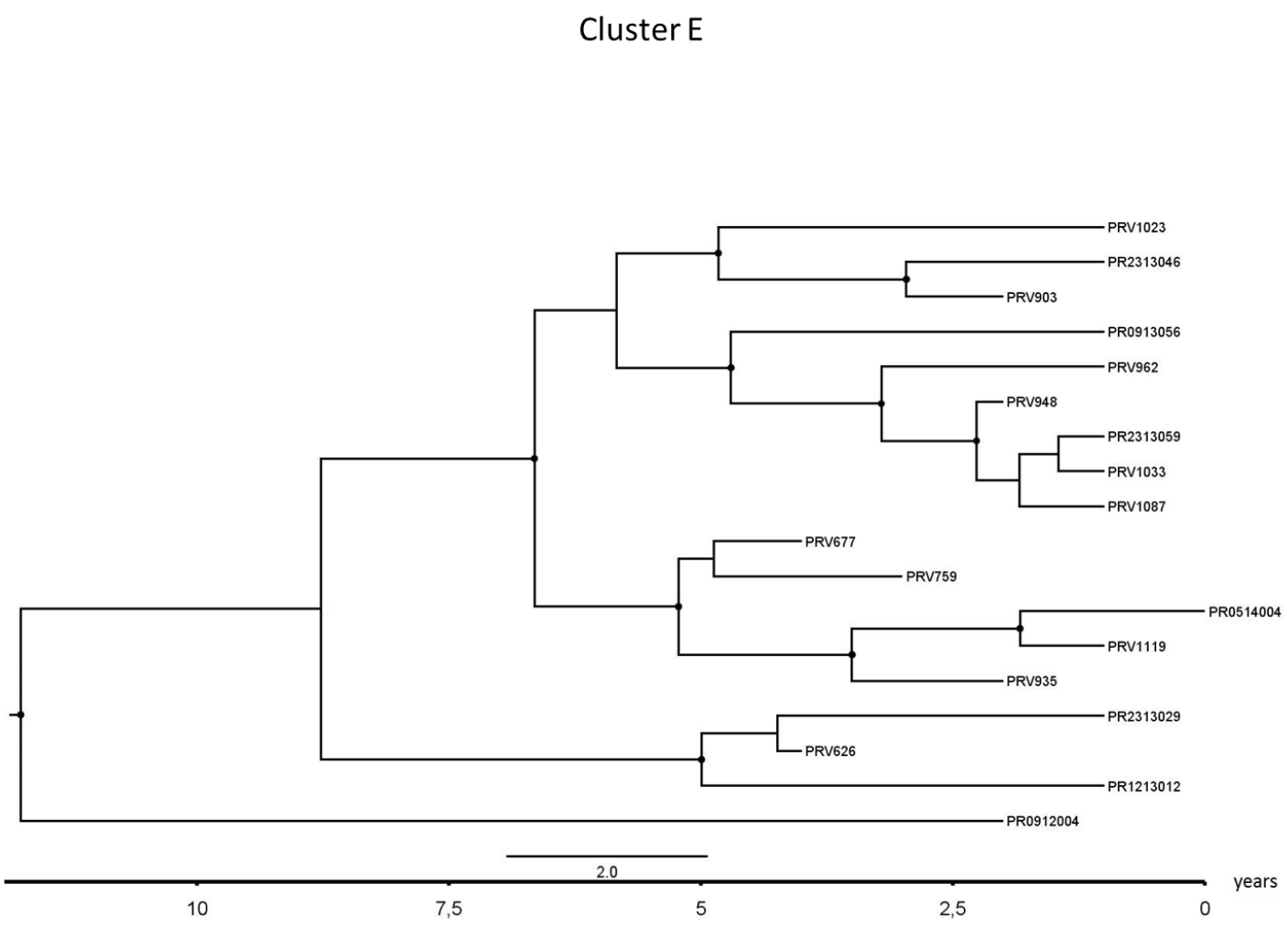

**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.



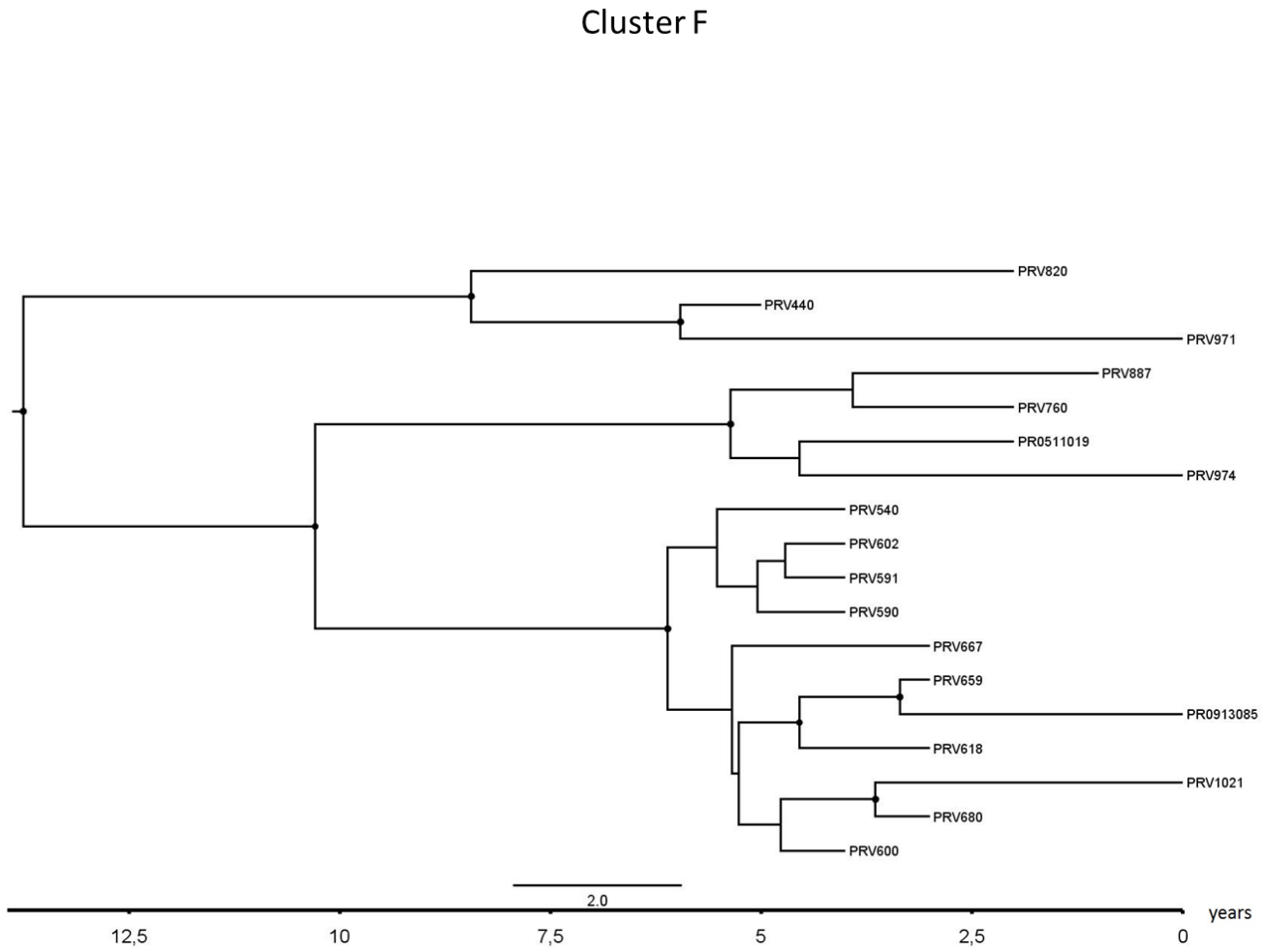

**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.

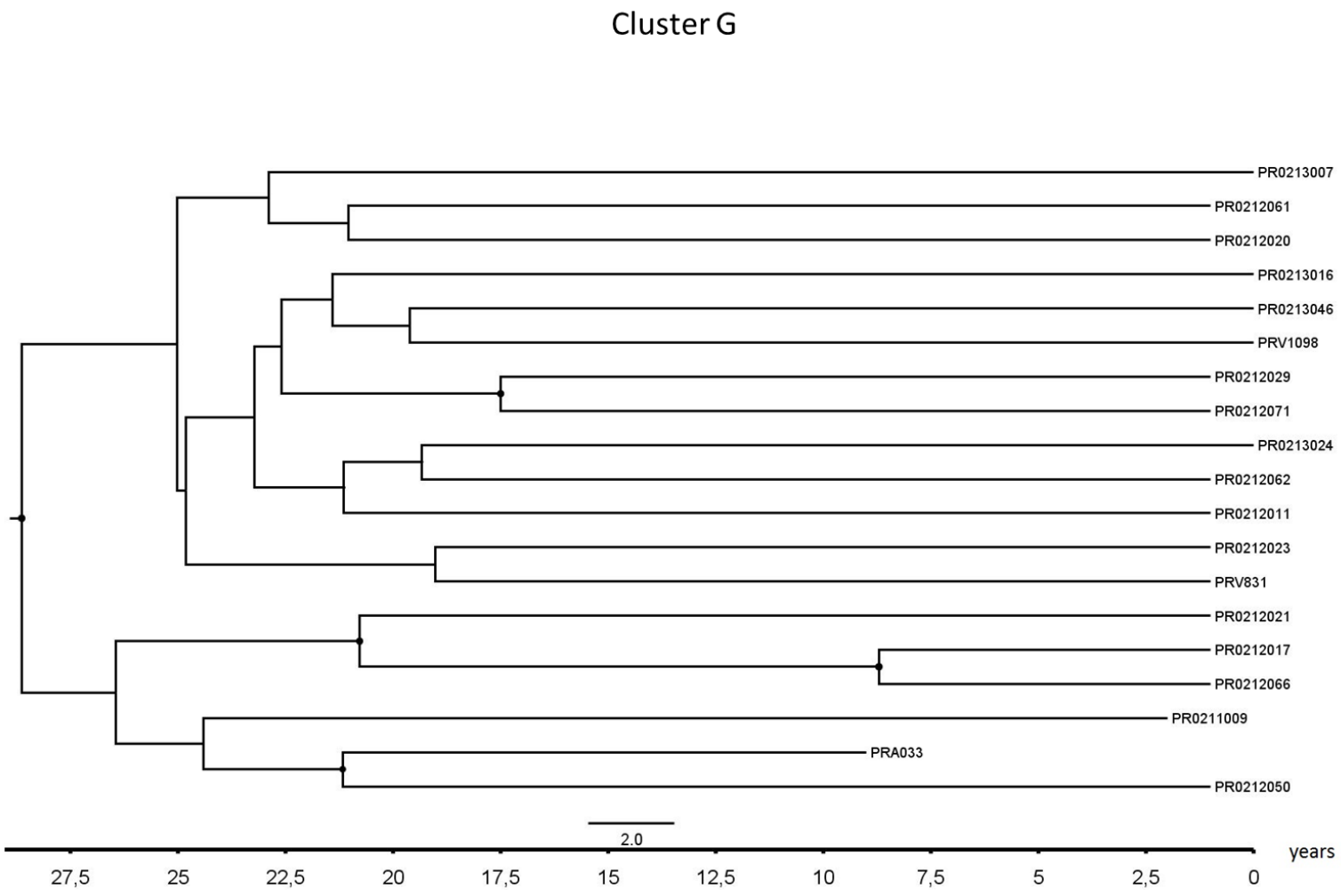



**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.

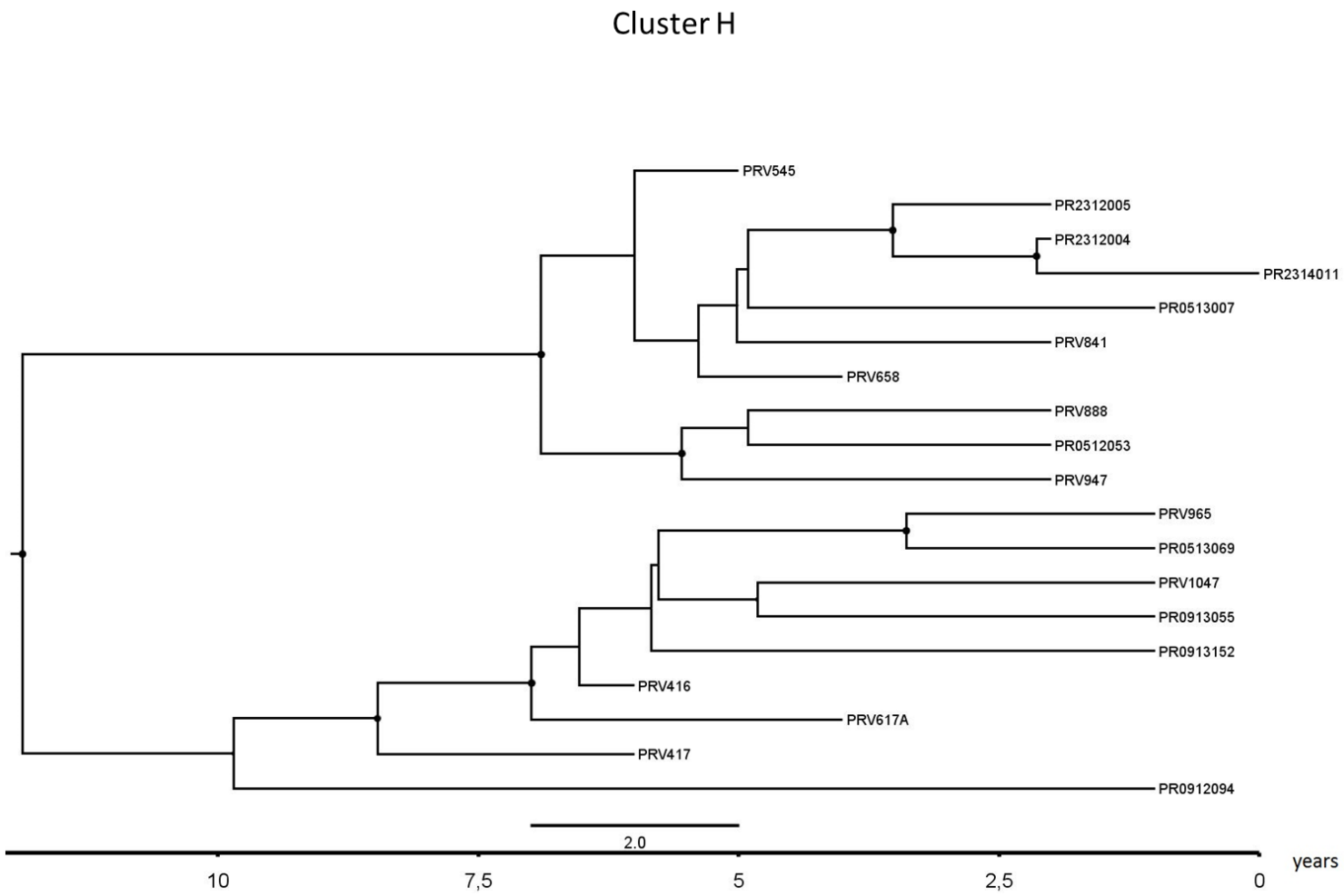



**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years. Nodes with Posterior Probabilities ≥0.90 are represented with black dots.

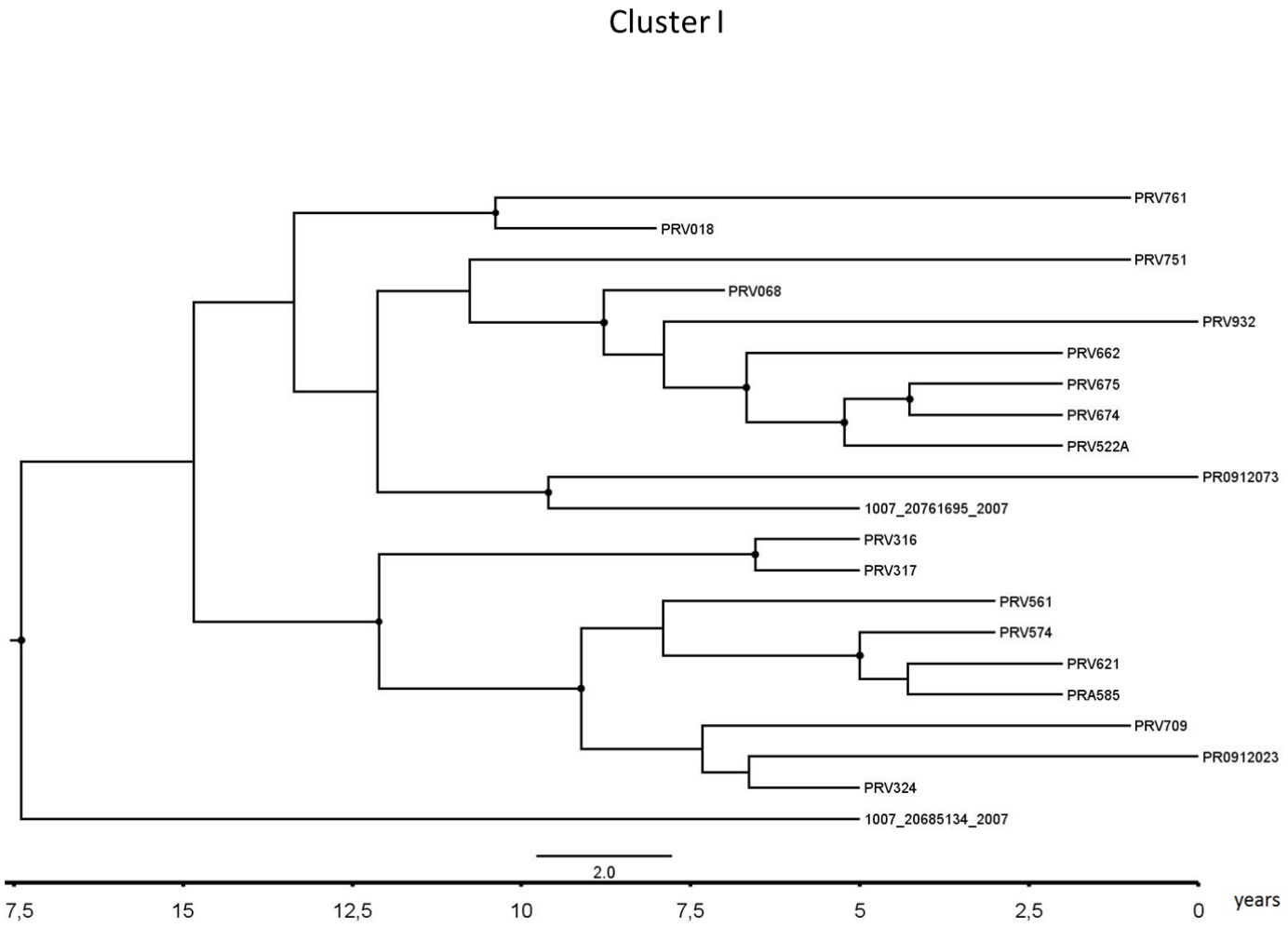



**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.

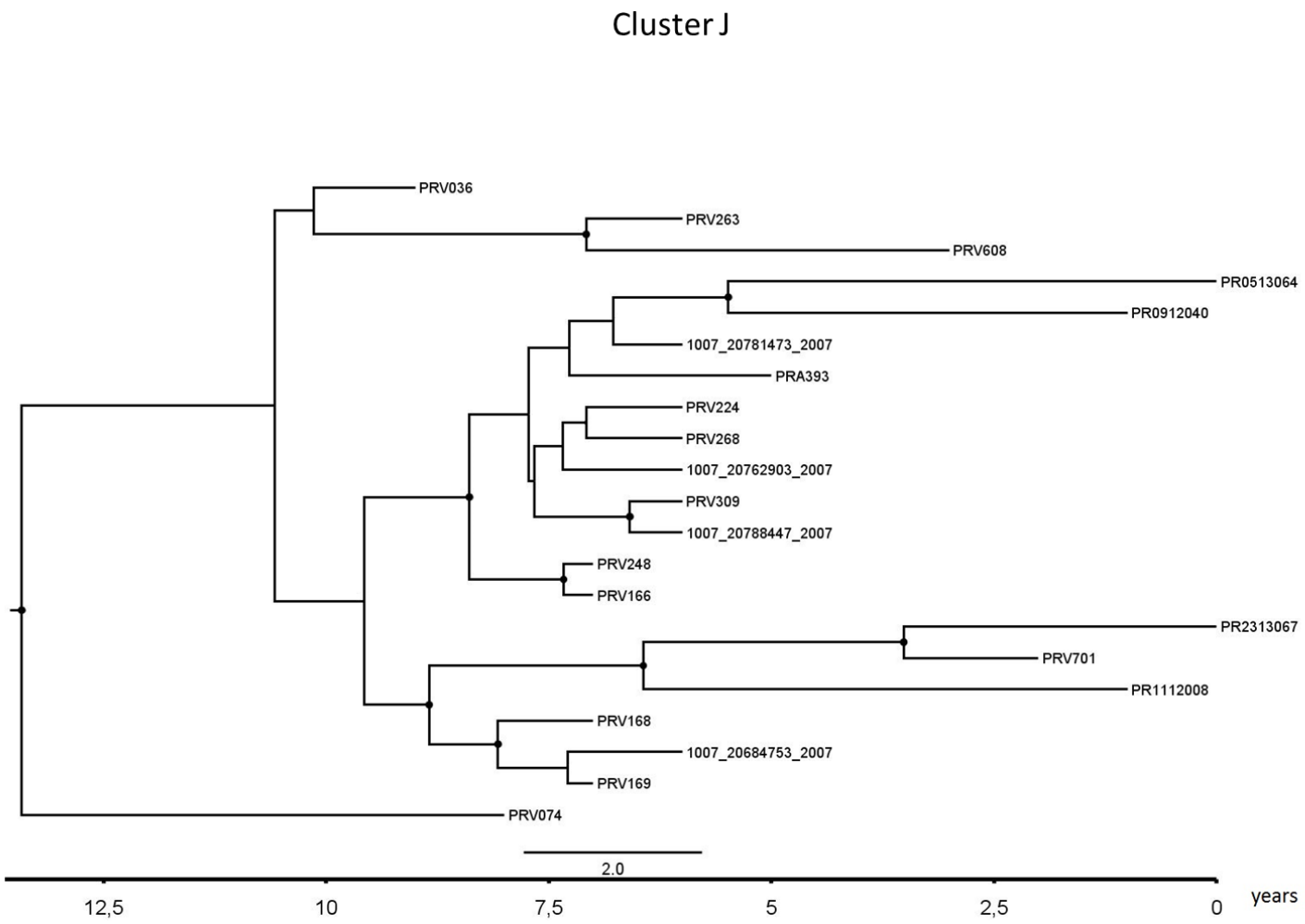



**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years. Nodes with Posterior Probabilities ≥0.90 are represented with black dots.




**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

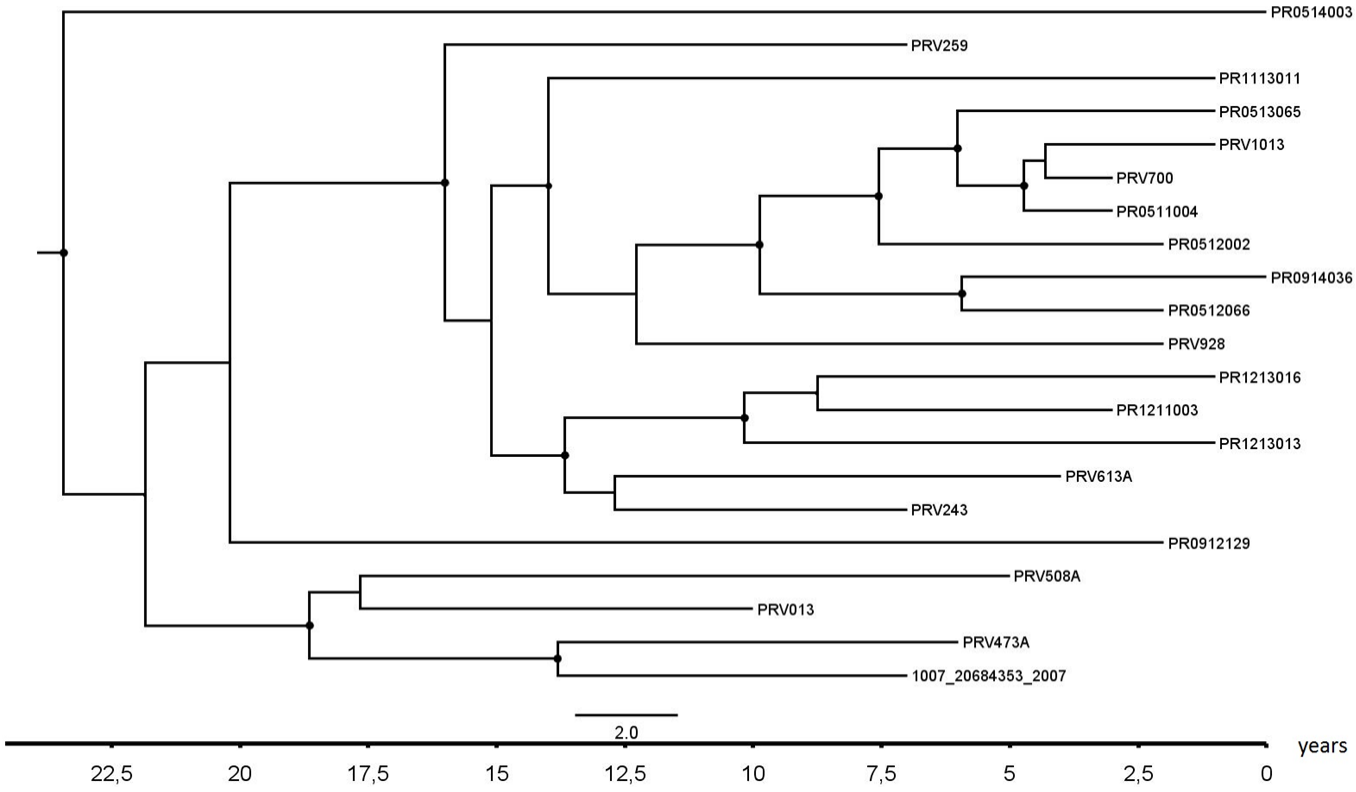
Nodes with Posterior Probabilities ≥0.90 are represented with black dots.




**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

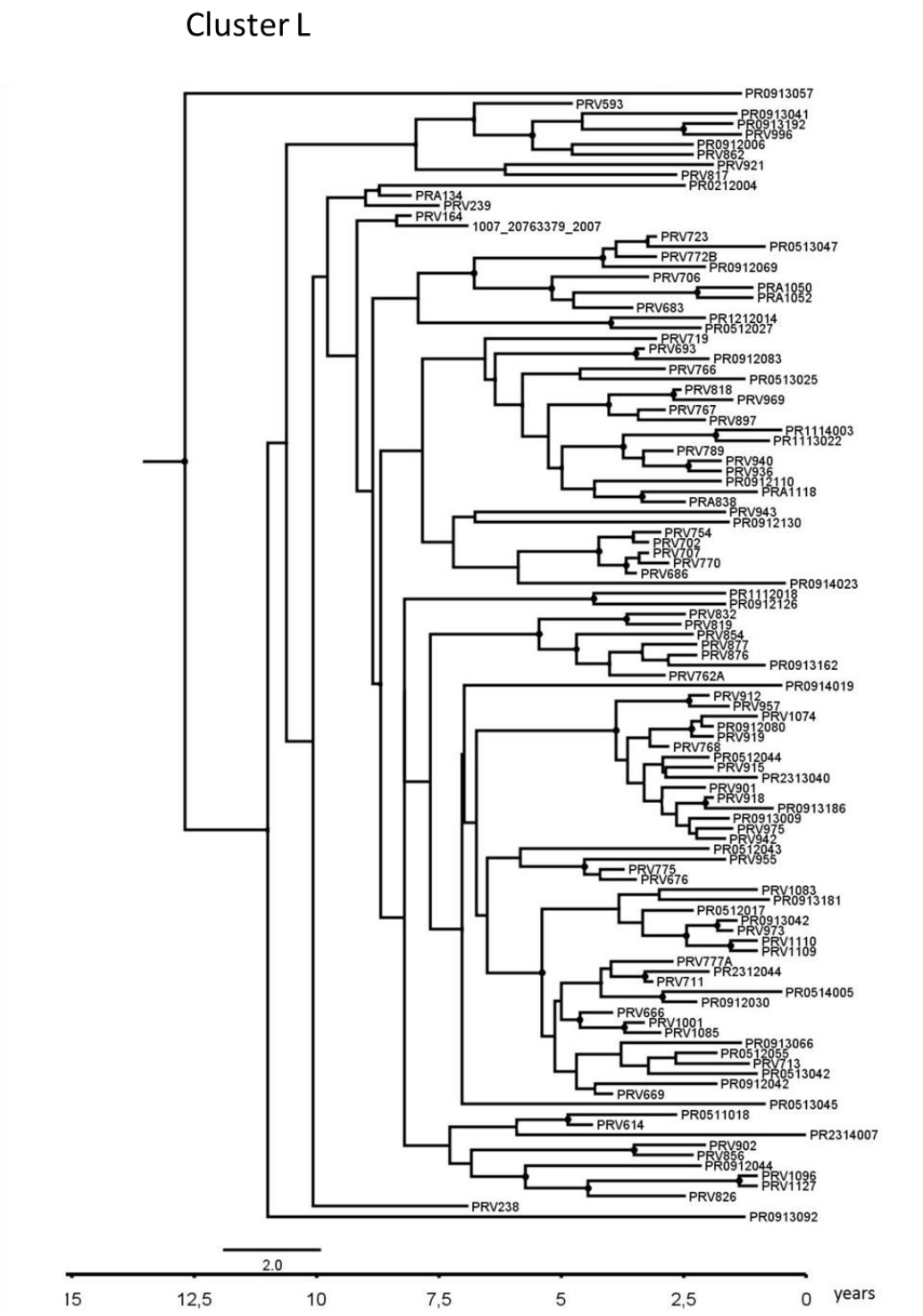
Nodes with Posterior Probabilities ≥0.90 are represented with black dots.

Cluster K




**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years.

Nodes with Posterior Probabilities ≥0.90 are represented with black dots.




**Supplementary Figure S1 (cont).** Dated phylogenetic trees of the 12 largest transmission clusters (A to L) analyzed with BEAST, Branch lengths represent years. Nodes with Posterior Probabilities ≥0.90 are represented with black dot

## 2.3- Chapter 3: Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain

***PLOS One. (2017) 12(2):e0171062.***

PLOS ONE

RESEARCH ARTICLE

# Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain

**Juan Ángel Patiño-Galindo[1], Manoli Torres-Puente[1], María Alma Bracho[1], Ignacio Alastrué[2], Amparo Juan[2], David Navarro[3], María José Galindo[3], Concepción Gimeno[4], Enrique Ortega[4], Fernando González-Candelas[1]***

**1** Unidad Mixta Infección y Salud Pública FISABIO-CSISP / Universidad de Valencia and CIBER Epidemiología y Salud Pública, Valencia, Spain, **2** Unidad Prevención del SIDA y otras ITS, Valencia, Spain, **3** Hospital Clínico Universitario-Universidad de Valencia, Valencia, Spain, **4** Hospital General Universitario, Valencia, Spain

* fernando.gonzalez@uv.es

Check for updates

OPEN ACCESS

**Citation:** Patiño-Galindo JÁ, Torres-Puente M, Bracho MA, Alastrué I, Juan A, Navarro D, et al. (2017) Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain. PLoS ONE 12(2): e0171062. doi:10.1371/journal.pone.0171062

**Editor:** Zhefeng Meng, Fudan University, CHINA

**Received:** August 6, 2016

**Accepted:** January 16, 2017

**Published:** February 2, 2017

**Copyright:** © 2017 Patiño-Galindo et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** Multiple sequence alignments and relevant sampling dates information are provided as Supporting Information in S1 and S2 Files.

**Funding:** This work was partly supported by project BFU2014- 58565-R from Ministerio de Economía y Competitividad (Spanish Government) to FGC, project CP07/00078 from Ministerio de Sanidad y Consumo (Spanish Government) to MAB, and Gilead Science SL. The funders had no role in study design, data collection and analysis,

## Abstract

We describe and characterize an exceptionally large HIV-1 subtype B transmission cluster occurring in the Comunidad Valenciana (CV, Spain). A total of 1806 HIV-1 protease-reverse transcriptase (PR/RT) sequences from different patients were obtained in the CV between 2004 and 2014. After subtyping and generating a phylogenetic tree with additional HIV-1 subtype B sequences, a very large transmission cluster which included almost exclusively sequences from the CV was detected (n = 143 patients). This cluster was then validated and characterized with further maximum-likelihood phylogenetic analyses and Bayesian coalescent reconstructions. With these analyses, the CV cluster was delimited to 113 patients, predominately men who have sex with men (MSM). Although it was significantly located in the city of Valencia (n = 105), phylogenetic analyses suggested this cluster derives from a larger HIV lineage affecting other Spanish localities (n = 194). Coalescent analyses estimated its expansion in Valencia to have started between 1998 and 2004. From 2004 to 2009, members of this cluster represented only 1.46% of the HIV-1 subtype B samples studied in Valencia (n = 5/143), whereas from 2010 onwards its prevalence raised to 12.64% (n = 100/791). In conclusion, we have detected a very large transmission cluster in the CV where it has experienced a very fast growth in the recent years in the city of Valencia, thus contributing significantly to the HIV epidemic in this locality. Its transmission efficiency evidences shortcomings in HIV control measures in Spain and particularly in Valencia.

## Introduction

Contrarily to intravenous drug users (IDUs) and heterosexual people (HT), the number of new HIV diagnosis among MSM in the European Union and European Economic Area (EU/EEA) has increased in the last years [1]. This trend is evident in the particular case of Spain, were IDU was consider the main transmission risk during the late 90s. However, in 2013,

decision to publish, or preparation of the manuscript.

**Competing Interests:** This study was partly funded by Gilead Science. There are no patents, products in development or marketed products to declare. This does not alter our adherence to all the PLOS ONE policies on sharing data and materials.

51.2% of the 3278 new HIV diagnoses reported in this country occurred among MSMs [2,3]. One of the factors contributing to this resurgence of HIV infections is the continued increase in unprotected anal sex among MSM that occurs since the highly active antiretroviral therapy (HAART) was introduced in 1996 [4,5]. Molecular epidemiology analyses have revealed the vulnerability of MSM to HIV infection in different ways, such as the frequent detection of transmission clusters affecting this risk group [6–10], and the estimation of shorter times between infections compared to those of HTs and IDUs [11]. In Spain, MSMs have been associated with significantly higher levels of local clustering than other risk groups [11,12]. Also, Delgado et al. [13] recently detected a large HIV-1 subtype F cluster affecting tens of MSM from different Spanish regions, indicating a fast and uncontrolled transmission among recently infected MSM who were unaware of their HIV status.

With approximately 5 million inhabitants, the Comunidad Valenciana (CV) is the fourth most populated region in Spain. Genotypic tests of resistance to antiviral drugs, by sequencing portions of the protease and retrotranscriptase (PR/RT) regions, are performed routinely in the CV. These tests produce large data sets of HIV-1 sequences that can be subjected to evolutionary analyses to better understand the local epidemics. The only molecular epidemiology analyses of HIV in the CV published so far were aimed at reporting the emergence of different CRFs among local MSM [14,15].

By means of phylogenetic analysis, we have identified an exceptionally large HIV-1 transmission cluster mainly localized in the city of Valencia (third largest city in Spain and the capital of the CV, with a metropolitan area of >1,500,000 inhabitants) and which is characterized by a rapid and recent expansion among MSM.

## Materials and methods

### Dataset

In order to assess the presence of resistance-associated mutations, 1806 PR/RT sequences were obtained from different newly HIV diagnosed people at seven different hospitals and two HIV counseling and testing centers (CIPS) from the three provinces in the CV between 2004 and 2014: six hospitals (Hospital Clínico de Valencia, Hospital General de Valencia, Hospital Universitario Doctor Peset, Hospital de Manises, Hospital La Ribera, Hospital Francesc de Borja) and one CIPS in Valencia, one CIPS in Alicante and one hospital (Hospital General de Castellón) in Castellón) (Fig 1). The sequences comprised the complete PR and the first 1005 nucleotides (335 amino acids) of the RT (1302 nt in total), and were obtained through viral RNA extraction followed by RT-PCR and direct sequencing using amplification and procedures described previously [16]. All the 1806 newly obtained sequences are available in S1 File. Sequences were subtyped using the REGAv3 [17] and COMET HIV-1 Subtyping tools (http://comet.retrovirology.lu/), and by examination of an initial phylogenetic tree obtained with FastTree 2.1 [18] under the GTR + Γ (4 CAT) model, which included 169 reference sequences downloaded from the Los Alamos HIV Database (LANL; http://www.hiv.lanl.gov), and represented the diversity of HIV-1 group M. Only those sequences classified as subtype B were considered in subsequent analyses. Nucleotide alignments were obtained with MAFFT version 7 [19]. This analysis was part of the surveillance program of communicable diseases by the General Directorate of Public Health of the Comunidad Valenciana and, as such, falls outside the mandate of the corresponding Ethics Committee for Biomedical Research. All personal information was anonymized and no data allowing individual identification was retained.

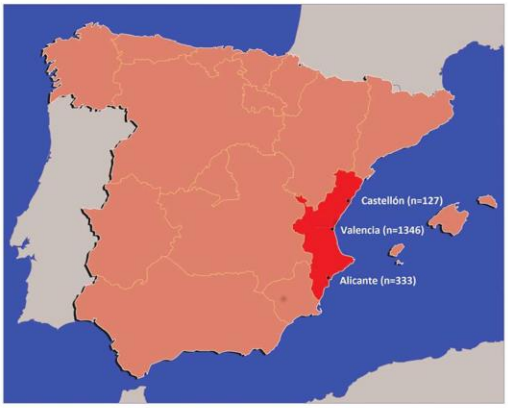



**Fig 1. Location of the Comunidad Valenciana in Spain.** The Comunidad Valenciana is highlighted in red. The locations of Valencia, Castellón and Alicante, and the number of HIV-1 sequences collected in each of the three localities (from a total of 1806 sequences), are specified.

doi:10.1371/journal.pone.0171062.g001

## Delimitation of the transmission cluster

An initial HIV-1 B phylogenetic tree was reconstructed with FastTree2.1, using the GTR + Γ (4 CAT) model. In addition to the 1514 HIV-1 subtype B sequences from the CV, 133 non-B reference sequences and 1787 HIV-1 subtype B sequences were downloaded from LANL and were included in this analysis. In this tree, a large and highly supported cluster of 143 sequences from the CV was found and further analyzed, given its potential interest within the local epidemic occurring in the CV. A BLAST search for the 100 sequences deposited in GenBank with the highest similarity to each of the sequences in the cluster was performed in October, 2015. This resulted in a non-redundant dataset of 587 BLAST-derived sequences which, along with those from the proposed transmission cluster, 6 subtype B references and 40 additional Spanish sequences (kindly provided by Dr. JC Galán and Dr. M Thomson), were used to reconstruct a maximum-likelihood (ML) tree with PhyML [20]. The resulting tree was used to confirm that these sequences conformed a true transmission cluster (defined as a group of epidemiologically related sequences which share a common, recent ancestor; [21,22], excluding those sequences that were more closely related to those from the BLAST search or additional controls but falling outside the CV clade. Consequently, the criterion used to confirm and delimitate the potential transmission cluster was finding a clade in which more than 90% of its sequences were from the CV and grouped with aLRT support $\geq$ 0.99.

## Dated phylogeny

The molecular clock signal of the transmission cluster was assessed by performing a linear regression analysis between the parameters "root-to-tip divergence" and "sampling date" with TempEST [23]). We used as input the subtree that included the 143 sequences potentially belonging to the CV transmission cluster, as extracted from the HIV-1 subtype B tree obtained with FastTree. All the 143 sequences from the CV that grouped initially were included in the analysis. The multiple alignment, including time of sampling information, is available in the S2 File.

The most recent common ancestor (tMRCA) of the transmission cluster was dated by means of Bayesian coalescent analysis as implemented in BEAST 1.8.1 [24]. All the 143 sequences from the CV were included in the analysis. For the coalescence analysis, a $GTR_{112}$ + $CP_{112}$ + $\Gamma_{112}$ (4 CAT) evolutionary model was used and combined with an uncorrelated log-normal relaxed molecular clock model and three different demographic models (Bayesian Sky-line Plot, and exponential or logistic demographic change). The best demographic model was chosen using Akaike's Information Criterion (AICM, [25]). For each demographic model, two independent runs of Bayesian MCMC, with chain lengths of at least 30 million states were performed, and sampled regularly every 3000 generations. These runs were then combined after discarding 10% as burn-in. The evolutionary parameters were estimated from an effective sampling size >200. Trees generated were then summarized using TreeAnnotator (http://beast.bio.ed.ac.uk/).

The internal branch lengths of a transmission cluster can be used as estimates of the time between transmission events [7]. We obtained the distribution of times between transmissions in the cluster from its internal branch lengths, using the summarized BEAST tree.

### Detection of drug resistance mutations

The presence of mutations associated with resistance to PR and RT inhibitors in the transmission cluster was assessed using the Stanford HIV Resistance database (http://sierra2.stanford.edu/sierra/servlet/JSierra; [26]). Only major mutations were taken into account.

### Statistical tests

Univariate analyses (Fisher's exact tests) were perform in R [27] to check whether the transmission cluster presented a significantly more affected gender (male vs female), transmission risk (MSM vs HT) and/or sampling locality (Valencia vs other localities). Patients from the transmission cluster were compared with the other HIV-1 B patients from the whole dataset.

## Results

Among the 1806 HIV-1 *pol* sequences obtained from different patients in the CV between 2004 and 2014, 1514 were classified as subtype B (prevalence = 83.83%). A potential transmission cluster was found in the initial HIV-1B tree obtained with FastTree (Fig 2), and it was further validated with a ML tree obtained with 633 additional, non-redundant sequences retrieved in a BLAST search and additional controls as detailed in Material and Methods.

The ML tree is shown in Fig 3 and it revealed that 111 of the 143 sequences from the CV initially detected as a potential transmission cluster retained monophyly with very high statistical support (approximate Likelihood-Ratio Test, aLRT = 0.99). This reduced cluster, including those 111 sequences, will be referred to as "CV-cluster". More specifically, the CV-cluster included 105 sequences from patients living in the city of Valencia and its metropolitan area and 6 from patients living in other localities from the CV. Only 2 sequences sampled in other Spanish cities outside the CV were included in this cluster, thus giving a total size for the CV-cluster of 113 patients.

Compared to the full dataset of HIV-1B sequences, the transmission cluster included a disproportionally large number of men [Fisher's exact test, FET: odds-ratio (OR) = 7.41 (95% CI: 1.94–62.95), $P < 0.001$], and MSM compared to HTs [FET: OR = 5.63 (1.80–28.40), $P < 0.001$]. Geographically, the transmission cluster included a very large number of sequences of persons living in the city of Valencia and surroundings compared to other localities [OR = 4.74 (2.29–11.38), $P<0.001$, Table 1]. The average pairwise genetic distance for

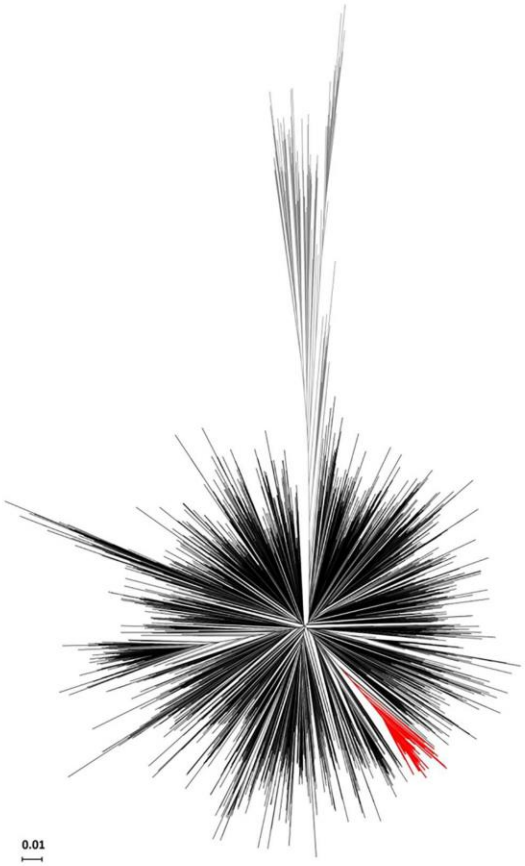



**Fig 2. Phylogenetic tree of the HIV-1 subtype B dataset.** The potential transmission cluster (n = 143) is highlighted in red, other subtype B sequences are colored in black and the reference sequences from other subtypes/CRFs are colored in grey.

doi:10.1371/journal.pone.0171062.g002

sequences in the CV-cluster was 0.0179 substitutions/site (max = 0.0392 s/s; 99 percentile = 0.0308 s/s; standard deviation = 0.0057 s/s).

The 111 Valencian sequences from the CV-cluster were sampled between 2006 and 2014, and 105 of them were sampled from 2010 onwards. Considering only those HIV-1 subtype B sequences from the city of Valencia (n = 1134), the transmission cluster accounted for 1.46% of all HIV-1 subtype B sequences sampled in this city between 2004 and 2009 (n = 5/343) but they represented 12.64% of the samples obtained between 2010 and 2014 (n = 100/791) (Fig 4).

The clock-like signal present in the analyzed dataset (n = 143) was evaluated by calculating the correlation coefficient (R) between the root-to-tip divergence and sampling date, obtaining an R value equal to 0.61, which was considered high enough as to proceed to estimating the Bayesian dated phylogenies with BEAST [28]. The exponential demographic model, combined with a lognormal relaxed molecular clock yielded the lowest AICM value. The Bayesian coalescent analysis performed under such model estimated the tMRCA of the CV-cluster to have occurred in 2001 (95% HPD = 1998–2004) (Table 2, Fig 5). The median of the internal branch

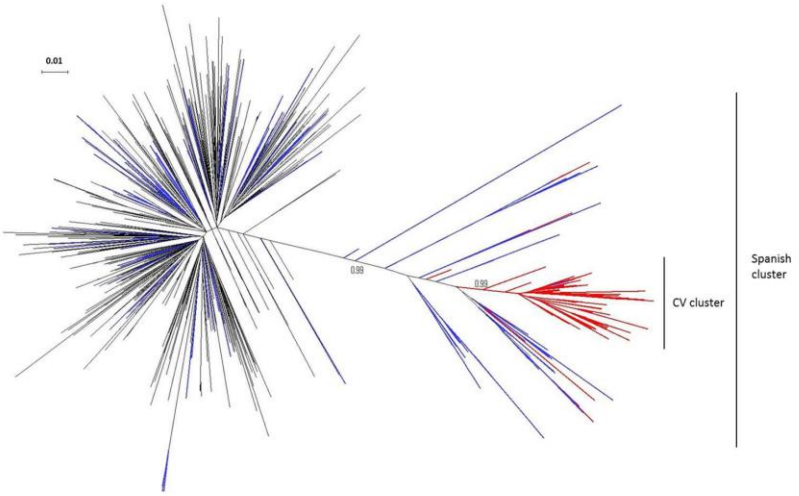



**Fig 3. Maximum likelihood tree of the CV-cluster.** It includes sequences from the city of Valencia (red), from other Spanish cities (blue) and sequences from other countries retrieved by BLAST analysis (grey). aLRT support values defining the CV-cluster (n = 113) and the larger Spanish clade it derives from (n = 194) are shown.

doi:10.1371/journal.pone.0171062.g003

lengths in the Bayesian tree for the CV-cluster was 0.68 years (95% CI = 0.14–3.08). The demographic reconstruction of the CV-cluster shows no indication of deceleration in its growth rate until 2014, the last year of sampling (Fig 6).

Major resistant mutations were present in very low frequency in the transmission cluster. Major Protease Inhibitors (PIs) resistance mutations were present in only one patient (V82F); major Nucleoside and Non-Nucleoside Reverse-transcriptase Inhibitors (NRTIs, NNRTIs) resistance mutations were present in only another patient (NRTI: M184V + K219E; NNRTI: K103N).

**Table 1. Epidemiological characteristics of the patients.**

| | Cluster (n = 111) | HIV-1 B outside the cluster (n = 1403) |
|---|---|---|
| **Sampling location** | | |
| **Valencia** | 105 | 1029 |
| **Alicante or Castellón** | 6 | 374 |
| **Gender** | | |
| **Male** | 95 | 724 |
| **Female** | 2 | 113 |
| **Unknown** | 14 | 566 |
| **Transmission risk** [a] | | |
| **MSM** | 72 | 515 |
| **HT** | 3 | 121 |
| **IDU** | 0 | 85 |
| **Unknown** | 36 | 682 |

[a] MSM:Men who have sex with men; HT: Heterosexual; IDU: Intravenous drug users.

doi:10.1371/journal.pone.0171062.t001

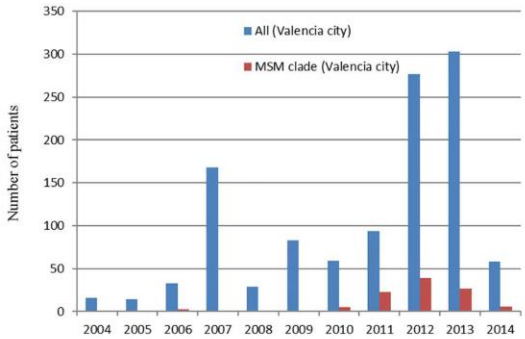



**Fig 4. Sampling distribution of HIV-1 subtype B sequences.** Sequences from patients inhabiting the city of Valencia or its metropolitan area are represented in blue (n = 1134) and those belonging to the transmission cluster sampled in this city are shown in red (n = 105).

doi:10.1371/journal.pone.0171062.g004

## Discussion

We have detected and characterized an HIV-1 subtype B transmission cluster which, affecting 105 patients solely in the city of Valencia, represents one of the largest local HIV-1 transmission clusters described so far in the HIV pandemic history. The report of clusters of similar or larger size is very rare [8,10,13,29,30], especially those that, like the CV-cluster, affect so many people in a single location in such a short time span.

Although this cluster corresponds to a highly localized HIV-1 outbreak of fast expansion among MSM in the city of Valencia (Spain), phylogenetic analyses revealed that this local cluster is related to sequences sampled in other Spanish cities. Hence, it is likely that, before its introduction and fast expansion in Valencia, this lineage has been circulating in other Spanish localities.

This transmission cluster started its expansion in the city of Valencia around 2001, five years after the introduction of HAART. However, most infections appear to have occurred after 2010, accounting for more than 12% of the HIV-1 subtype B sequences sampled in Valencia since then. This dynamic parallels the increase of HIV diagnosis among MSM in Europe in the last few years [1]. The transmission efficiency of this HIV-1 lineage in the Valencia population, whose growth rate had not reached a steady state by 2014 (last year in our sampling), reveals shortcomings in the HIV control measures in Spain, and particularly in the CV, at least for some specific, vulnerable groups such as MSM.

Previous works have found that recently infected MSMs, who are usually unaware of their HIV status, are a significant source of onward transmissions [31,32] and that, within the MSM

**Table 2. Akaike's Information Criterion (AICM) values and tMRCA of the CV cluster for three demographic models.**

| Demographic model | AICM [a] | tMRCA [b] |
|---|---|---|
| **Bayesian Skyline Plot** | 17577.24 +/- 0.58 | 2001.6 (1998.2–2004.6) |
| **Exponential** | 17559.72 +/- 0.38 | 2001.8 (1998.6–2004.8) |
| **Logistic** | 17585.04 +/- 0.69 | 2001.9 (1998.8–2004.8) |

[a] (value +/- SE)

[b] Median and 95% HPD limits

doi:10.1371/journal.pone.0171062.t002

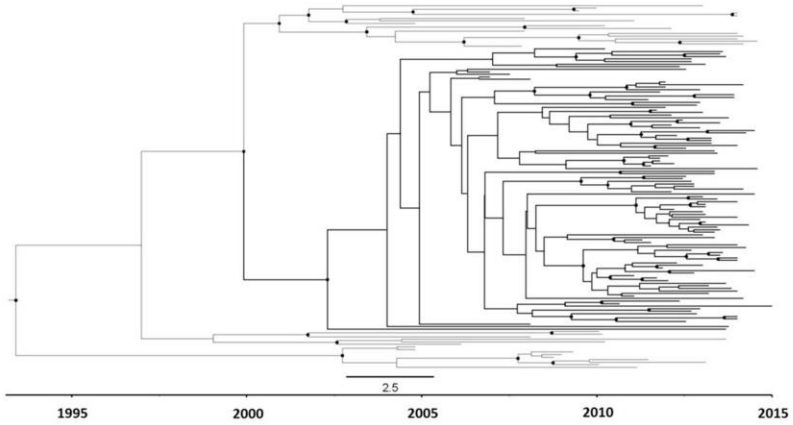



**Fig 5. Dated phylogeny of the 143 sequences from the initial CV cluster.** Branch lengths represent years. The CV cluster is highlighted in black. Dots on nodes represent posterior probabilities $\geq$ 0.90.

doi:10.1371/journal.pone.0171062.g005

collective, the increasing high-risk sexual behavior rates occurring in the last years may hamper the epidemiological benefits of HAART on controlling HIV incidence [4]. In Spain, where there is a high HAART coverage (approximately 74% of the estimated number of persons living with HIV receive treatment; Spain Country factsheet 2015, available at http://aidsinfo.unaids.org), the HIV incidence rate by year of diagnosis for MSMs increased from 2 seroconversions per 100 persons per year (p-y) in 2000 to 2.5/100 p-y in 2009, being the only risk group with a consistently increasing incidence rate along this time-span [33].

Usually, the analysis of transmission clusters is performed after removing major resistance-associated positions from the multiple alignment, in order to prevent spurious clusters resulting from convergent evolution. In this work, these positions were not removed in the identification and characterization of the CV-cluster because most of the sequences included in the analyses derive from treatment-naïve patients. Specifically, none of the 143 patients clustering in the initial group was reported to be or have been under antiretroviral treatment. Furthermore, only two sequences in the CV-cluster presented major resistance mutations. For this

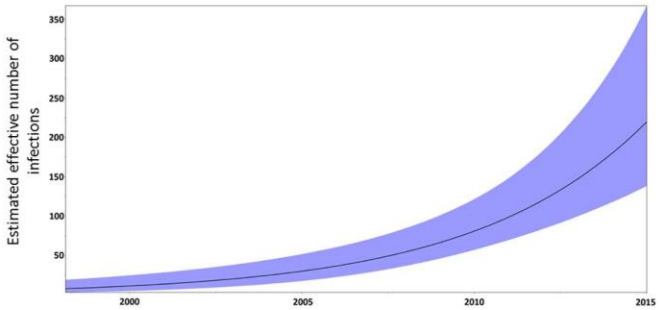



**Fig 6. Population dynamics of the 143 sequences from the initial CV cluster.** The dynamics was inferred with BEAST, based on the exponential growth model. The black line represents the median estimate of the effective number of infections, and the shaded area represents the 95% HPD.

doi:10.1371/journal.pone.0171062.g006

reason, we can discard artifactual clustering as a possible explanation for the identification of this cluster.

The criteria used for defining transmission clusters vary among studies, with many of them combining high phylogenetic support with genetic distance thresholds. Although in this work we did not consider large pairwise genetic distances as an exclusion criterion, the distribution of pairwise genetic distances obtained fell within the inclusion criteria of most studies [34]. Furthermore, the maximum distance between any two sequences was lower than 0.045 s/s, a threshold value used previously to consider the exclusion of epidemiologically unlinked sequences [8,9].

The phylogenetic analysis of sequences obtained during routine evaluation of resistance mutations for antiretroviral drugs can provide crucial information about the detailed local, regional and global HIV epidemics [7,9,11]. The identification of transmission clusters is just one of the many possible benefits of the molecular surveillance of infectious diseases, particularly HIV, in which traditional epidemiological analysis based on contact tracing or direct interviews with newly diagnosed individuals are hampered by social and personal attitudes, resulting in lack of useful information for further prevention. This cluster, which was not noticed by local or regional public health officials, was detected and characterized from basic patient information, completely anonymous for the researchers. It also illustrates the need to enhance prevention and information campaigns in a specific risk group.

In conclusion, this work has reported the existence of a large and fast expanding HIV-1 transmission cluster affecting newly diagnosed MSM in the city of Valencia. Given that factors such as high-risk sexual behaviors and unawareness of HIV status may hamper the control of the HIV epidemic in MSM, it is necessary to reinforce the campaigns for HIV prevention, such as condom distribution programs and HIV testing in the Valencian MSM community. The results obtained also stress the importance and interest of implementing surveillance strategies that use viral sequencing information derived from the genotypic analysis of resistance mutations in HIV-infected patients.

## Supporting information

**S1 File. Fasta alignment of the 1806 HIV-1 sequences obtained from the CV.**
(RAR)

**S2 File. Fasta alignment of the 143 HIV subtype B sequences from the CV.** These were the sequences initially detected to conform a well-supported cluster and information of their sampling date used in the BEAST analysis is included.
(RAR)

## Acknowledgments

We thank Dr. M Thomson and Dr. JC Galán-Montemayor for providing access to some unpublished HIV-1 sequences included in the analysis. We also thank all the patients who accepted to provide personal information on their risk of infection factors.

## Author Contributions

**Conceptualization:** JAPG MTP MAB FGC.

**Data curation:** JAPG MTP MAB.

**Formal analysis:** JAPG FGC.

**Funding acquisition:** FGC MAB.

**Investigation:** JAPG MTP MAB FGC.

**Methodology:** JAPG FGC.

**Project administration:** FGC.

**Resources:** MTP MAB IA AJ DN MJG CG EO.

**Software:** JAPG MTP.

**Supervision:** FGC.

**Validation:** JAPG MTP MAB FGC.

**Visualization:** JAPG FGC.

**Writing – original draft:** JAPG.

**Writing – review & editing:** JAPG MTP MAB IA AJ DN MJG CG EO FGC.

## References


1. ECDC (2013) Men who have sex with men. Monitoring implementation of the Dublin Declaration on Partnership to Fight HIV/AIDS in Europe and Central Asia: 2012 progress report. www.ecdc.europa.eu.
2. UNAIDS/WHO (2002) Report on the global HIV/AIDS epidemic–July 2002.
3. DGSP (2014) Vigilancia epidemiológica del VIH/SIDA enEspaña: Sistema de información sobre nuevos diagnósticos del VIH y registro nacional de casos de SIDA.
4. Bezemer D, de Wolf F, Boerlijst MC, van Sighem A, Hollingsworth TD, Prins M, et al. (2008) A resurgent HIV-1 epidemic among men who have sex with men in the era of potent antiretroviral therapy. AIDS 22: 1071–1077. doi: 10.1097/QAD.0b013e3282fd167c PMID: 18520351
5. Phillips AN, Cambiano V, Nakagawa F, Brown AE, Lampe F, Rodger A, et al. (2013) Increased HIV incidence in men who have sex with men despite high levels of ART-induced viral suppression: analysis of an extensively documented epidemic. PLoS ONE 8: e55312. doi: 10.1371/journal.pone.0055312 PMID: 23457467
6. Hue S, Clewley JP, Cane PA, Pillay D (2005) Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. AIDS 19: 449–450. PMID: 15750402
7. Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ (2008) Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Medicine 5: e50. doi: 10.1371/journal.pmed.0050050 PMID: 18351795
8. Kouyos RD, von Wyl V, Yerly S, Böni J, Taffé P, Shah C, et al. (2010) Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. Journal of Infectious Diseases 201: 1488–1497. doi: 10.1086/651951 PMID: 20384495
9. Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT (2011) Transmission network parameters estimated from HIV sequences for a nationwide epidemic. Journal of Infectious Diseases 204: 1463–1469. doi: 10.1093/infdis/jir550 PMID: 21921202
10. Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, et al. (2015) Dispersion of the HIV-1 epidemic in men who have xex with men in the Netherlands: A combined mathematical model and phylogenetic analysis. PLoS Medicine 12: e1001898. doi: 10.1371/journal.pmed.1001898 PMID: 26529093
11. Patiño Galindo JA, Thomson MM, Pérez-Álvarez L, Delgado E, Cuevas MT, Fernández-García A, et al. (2016) Transmission dynamics of HIV-1 subtype B in the Basque country, Spain. Infection, Genetics and Evolution 90: 91–97.
12. Yebra G, Holguín Á, Pillay D, Hué S (2013) Phylogenetic and demographic characterization of HIV-1 transmission in Madrid, Spain. Infection, Genetics and Evolution 14: 232–239. doi: 10.1016/j.meegid.2012.12.006 PMID: 23291408
13. Delgado E, Cuevas MT, Domínguez F, Vega Y, Cabello M, Fernández-García A, et al. (2015) Phylogeny and phylogeography of a recent HIV-1 subtype F outbreak among men who have sex with men in Spain deriving from a cluster with a wide geographic circulation in Western Europe. PLoS ONE 10: e0143325. doi: 10.1371/journal.pone.0143325 PMID: 26599410

14. Bracho MA, Sentandreu V, Alastrué I, Belda J, Juan A, Fernández-García E, et al. (2014) Emerging Trends in CRF02_AG Variants Transmission Among Men Who Have Sex With Men in Spain. JAIDS Journal of Acquired Immune Deficiency Syndromes 65: e130–e133. doi: 10.1097/01.qai.0000435602.73469.56 PMID: 24091696

15. Patiño Galindo JA, Torres-Puente M, Gimeno C, Ortega E, Navarro D, Galindo MJ, et al. (2015) Expansion of the CRF19_cpx variant in Spain. Journal of Clinical Virology 69: 146–149. doi: 10.1016/j.jcv.2015.06.094 PMID: 26209397

16. Holguin A, Alvarez A, Soriano V (2005) Heterogeneous nature of HIV-1 recombinants spreading in Spain. Journal of Medical Virology 75: 374–380. doi: 10.1002/jmv.20280 PMID: 15648070

17. Pineda-Peña AC, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, et al. (2013) Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools. Infection, Genetics and Evolution 19: 337–348. doi: 10.1016/j.meegid.2013.04.032 PMID: 23660484

18. Price MN, Dehal PS, Arkin AP (2010) FastTree 2—Approximately maximum-likelihood trees for large alignments. PLoS ONE 5: e9490. doi: 10.1371/journal.pone.0009490 PMID: 20224823

19. Katoh K, Standley DM (2013) MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772–780. doi: 10.1093/molbev/mst010 PMID: 23329690

20. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704. PMID: 14530136

21. Hué S, Clewley JP, Cane PA, Pillay D (2004) HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS 18: 719–728. PMID: 15075506

22. Hué S, Pillay D, Clewley JP, Pybus OG (2005) Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proceedings of the National Academy of Sciences USA 102: 4425–4429.

23. Rambaut A, Lam TT, Max Carvalho L, Pybus OG (2016) Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution 2.

24. Drummond AJ, Suchard MA, Xie D, Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29: 1969–1973. doi: 10.1093/molbev/mss075 PMID: 22367748

25. Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko AV (2012) Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Molecular Biology and Evolution 29: 2157–2167. doi: 10.1093/molbev/mss084 PMID: 22403239

26. Liu TF, Shafer RW (2006) Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical Infectious Diseases 42: 1608–1618. doi: 10.1086/503914 PMID: 16652319

27. R Development Core Team (2011) R: A language and environment for statistical computing., version Vienna, Austria.

28. Alizon S, Fraser C (2013) Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology 10: 49. doi: 10.1186/1742-4690-10-49 PMID: 23639104

29. Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, et al.(2014) The global transmission network of HIV-1. Journal of Infectious Diseases 209: 304–313. doi: 10.1093/infdis/jit524 PMID: 24151309

30. Peters PJ, Pontones P, Hoover KW, Patel MR, Galang RR, Shields J, et al. (2016) HIV infection linked to injection use of oxymorphone in Indiana, 2014–2015. New England Journal of Medicine 375: 229–239. doi: 10.1056/NEJMoa1515195 PMID: 27468059

31. Frange P, Meyer L, Deveau C, Tran L, Goujard C, Ghosn J, et al. (2012) Recent HIV-1 infection contributes to the viral diffusion over the French territory with a recent increasing frequency. PLoS ONE 7: e31695. doi: 10.1371/journal.pone.0031695 PMID: 22348121

32. Ambrosioni J, Junier T, Delhumeau C, Calmy A, Hirschel B, Zdobnov E, et al. (2012) Impact of highly active antiretroviral therapy on the molecular epidemiology of newly diagnosed HIV infections. AIDS 26: 2079–2086. doi: 10.1097/QAD.0b013e32835805b6 PMID: 23052354

33. Diez M, Bleda MJ, Varela JR, Ordonana J, Azpiri MA, Vall M, et al. (2014) Trends in HIV testing, prevalence among first-time testers, and incidence in most-at-risk populations in Spain: the EPI-VIH Study, 2000 to 2009. Euro Surveillance 19: 20971. PMID: 25443036

34. Grabowski MK, Redd AD (2014) Molecular tools for studying HIV transmission in sexual networks. Current Opinion in HIV and AIDS 9: 126–133. doi: 10.1097/COH.0000000000000040 PMID: 24384502

**Supplementary material**

Supplementary material (Supplementary Files S1 and S2) is freely available at

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0171062

# 2.4- Chapter 4: Expansion of the CRF19_cpx variant in Spain

*J Clin Virol (2015) 69: 146-149.*

Journal of Clinical Virology 69 (2015) 146–149

Contents lists available at ScienceDirect

Journal of Clinical Virology

journal homepage: www.elsevier.com/locate/jcv

Short communication

# Expansion of the CRF19_cpx Variant in Spain

Juan Angel Patiño Galindo[a,b], Manoli Torres-Puente[a,b], Concepción Gimeno[c,d], Enrique Ortega[e], David Navarro[d,f], María José Galindo[g], Laura Navarro[h], Vicente Navarro[h], Amparo Juan[i], Josefina Belda[j], María Alma Bracho[a,b], Fernando González- Candelas[a,b,k,*], On the behalf of: CRIVIH

[a] Unidad Mixta Infección y Salud FISABIO-Universidad de Valencia/Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Valencia, Spain
[b] CIBER en Epidemiología y Salud Pública (CIBERESP), Valencia, Spain
[c] Servicio de Microbiología, Hospital General Universitario, Valencia, Spain
[d] Dpto. Microbiología, Universidad de Valencia, Valencia, Spain
[e] Unidad de Enfermedades Infecciosas, Hospital General Universitario, Valencia, Spain
[f] Servicio de Microbiología, Hospital Clínico Universitario, Valencia, Spain
[g] Unidad de Enfermedades Infecciosas, Hospital Clínico Universitario, Valencia, Spain
[h] Servicio de Microbiología, Hospital de Manises, Manises, Spain
[i] Centro de Información y Prevención del SIDA, Valencia, Spain
[j] Centro de Información y Prevención del SIDA, Alicante, Spain
[k] Consorcio para la Investigación en VIH y SIDA de la Comunidad Valenciana, Spain

ARTICLE INFO

*Article history:*
Received 6 May 2015
Received in revised form 6 June 2015
Accepted 18 June 2015

*Keywords:*
HIV-1
Molecular epidemiology
CRF19_cpx

ABSTRACT

*Background:* HIV-1 CRF19_cpx, is a recombinant variant found almost exclusively in Cuba and recently associated to a faster AIDS onset. Infection with this variant leads to higher viral loads and levels of RANTES and CXCR4 co-receptor use.
*Objectives:* The goal of this study was to assess the presence of CRF19_cpx in the Spanish province of Valencia, given its high pathogenicity.
*Study design:* 1294 HIV-1 protease-reverse transcriptase (PR/RT) sequences were obtained in Valencia (Spain), between 2005 and 2014. After subtyping, the detected CRF19_cpx sequences were aligned with 201 CRF19_cpx and 66 subtype D sequences retrieved from LANL, and subjected to maximum-likelihood phylogenetic analyses and Bayesian coalescent reconstructions. The presence of resistance mutations in the PR/RT region of these sequences was also analyzed.
*Results:* Among the 9 CRF19_cpx sequences from different patients found (prevalence <0.1%), 7 grouped in two well-supported clades (groups A, $n = 4$, and B, $n = 3$), suggesting the existence of at least two independent introductions which subsequently started to expand in the studied Spanish region. Unprotected sex between men was the only known transmission route. Coalescent analyses suggested that the introductions in Valencia occurred between 2008 and 2010. Resistance mutations in the RT region were found in all sequences from group A (V139D) and in two sequences from group B (E138A).
*Conclusions:* This study reports for the first time the recent expansion of CRF19_cpx outside Cuba. Our results suggest that CRF19_cpx might become an emerging HIV variant in Spain, affecting Spanish native MSM and not only Cuban migrants.

© 2015 Elsevier B.V. All rights reserved.

## 1. Background

HIV-1CRF19_cpx was first described in 2005 as a circulating A, D and G intersubtype recombinant form infecting Cuban patients [1]. Although phylogeographic analyses suggest an African origin for this HIV-1 CRF, the vast majority of infections occur in Cuba, where prevalence is >15% of the HIV-infected population [2–5]. However, the reported incidence of CRF19_cpx cases in other countries is very low, and the existence of transmission groups outside

*Abbreviations:* PT/RT, protease-retrotranscriptase; CAT, category; ML, maximum-likelihood; tMRCA, time for the most recent common ancestor; AICM, Akaike's information criterion; MCMC, Markov Chain Monte Carlo; MSM, men who have sex with men.

* Corresponding author at: Unidad Mixta Infección y Salud FISABIO-Universidad de Valencia/Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Catedrático José Beltrán 2, 46980-Paterna, Valencia, Spain. Fax: +34 963 543 670.
*E-mail address:* fernando.gonzalez@uv.es (F. González- Candelas).

http://dx.doi.org/10.1016/j.jcv.2015.06.094
1386-6532/© 2015 Elsevier B.V. All rights reserved.

Cuba has never been reported. Of the 268 entries for this CRF in the Los Alamos HIV Database (LANL, http://www.hiv.lanl.gov/content/sequence/HIV/mainpage.html, last accessed on February 2015), only 19 come from other countries: six from Africa (4 from Congo and one each from Central African Republic and Nigeria), three from the United States of America, and ten from Europe (one isolate each from Greece and the United Kingdom, two from France and six from Spain). An increase in the number of cases with rapid progression to AIDS among Cuban patients has been recently linked to CRF19_cpx [6]. Infection with this variant is also associated to a higher viral load and higher levels of RANTES and CXCR4 co- receptor use. These results suggest that CRF19_cpx is a more pathogenic virus.

## 2. Objectives

Given the higher pathogenicity of CRF19_cpx, the identification of cases outside Cuba may be of public health concern. The main goal of this study was to check, retrospectively, for the presence of CRF19_cpx in the Valencia region (Spain) and, if detected, to describe its epidemic scenario by means of phylogenetic tools.

## 3. Study design

In order to assess the presence of resistance-associated mutations, 1294 HIV-1 protease-reverse transcriptase (PR/RT) sequences were obtained from newly HIV-1 diagnosed people in different hospitals and two first-line anonymous HIV counseling and testing centers (CIPS) in the Valencia region between 2009 and 2014. Information about country of origin and likely route of HIV transmission was recorded with patients' consent. Sequences included the complete PR and the first 1005 nucleotides (335 amino acids) of the RT, and were obtained through viral RNA extraction followed by RT-PCR and direct sequencing using amplification and sequencing procedures as described previously [7,8]. Sequences were subtyped using REGAv3 [9] and COMET (http://comet.retrovirology.lu/). Subtyping was further confirmed with a phylogenetic analysis using FastTree with the GTR + GAMMA(4 categories, CAT) [10] evolution model on a multiple sequence alignment produced with MUSCLE [11] which included 362 additional HIV-1 sequences retrieved from LANL and used as subtype references.

In order to further characterize the relationships among the CRF19_cpx sequences found, a Maximum-Likelihood (ML) phylogenetic tree was reconstructed with PhyML [12] using the GTR + GAMMA(4CAT) model. In addition to the newly detected Spanish sequences, the dataset included 201 CRF19_cpx and 66 subtype D sequences retrieved from LANL, which were used as outgroup.

In order to date the most recent common ancestor (tMRCA) of the transmission groups detected (defined as Valencia sequences grouping with support >0.90), a coalescent analysis was performed, as implemented in BEAST 1.8.1 [13]. For that goal, only the 38 closest Cuban reference sequences were added to those in the Valencian clusters. The molecular clock signal of each transmission group was assessed by performing linear regression analyses between the parameters "root-to-tip divergence" and "sampling date" with the software Path-*O*-Gen v1.4 (http://tree.bio.ed.ac.uk/software/pathogen), using the ML subtree that included the sequences used in the coalescent analyses. The analyses were performed with the ($GTR_{112} + CP_{112} + \Gamma_{112}$ (4CAT)) substitution model. A log-normal prior was chosen for ucld.mean (median = $3.35 \times 10^{-3}$, 95% HPD upper limit = $5.00 \times 10^{-3}$ substitutions per site and year [5]). Under an uncorrelated lognormal relaxed molecular clock model, the most appropriate demographic model (Bayesian Skyline Plot-BSP, constant, exponential or logistic demographic change) was chosen as the one with the lowest Akaike's Information Criterion (AICM, [14]) value.

For each demographic model, at least two independent runs of Bayesian MCMC with chain lengths of 10 million states were performed. All the parameters were estimated from an effective sampling size >200 using the software Tracer 1.5 (http://tree.bio.ed.ac.uk/software/tracer/). Trees generated from the two BEAST runs were combined and summarized after discarding a 10% burn-in using TreeAnnotator (http://beast.bio.ed.ac.uk/).

Detection of mutations associated with resistance to PR and RT inhibitors in these Valencian sequences was performed with the Stanford University HIV Drug Resistance Database (http://sierra2.stanford.edu/sierra/servlet/JSierra, [15]).

## 4. Results

Nine CRF19_cpx sequences from different patients, sampled between 2011 and 2014, were found among all the Valencian sequences analyzed (prevalence = 0.007, 95% CI: 0.003–0.013). All CRF19_cpx sequences considered ($n = 201$) conformed a well-supported group within the subtype D clade in the FastTree tree (Supplementary Fig. 1). Although CRF19_cpx does not have breakpoints in *pol*, subtyping with the aforementioned tools was consistent. Only one sequence (PR0914037), which clustered within the CRF19_cpx clade, was reported by Comet and REGA to belong to subtype D. However, a Bootscan analysis with REGA reported all windows with support >70% to be associated with CRF19. Furthermore, a partial *env* sequence was available for this sample and it matched with this recombinant form. One sequence from Valencia, which belonged to a Cuban man who reported having unprotected sex with men (MSM), was found to be an independent introduction from Cuba. The other eight sequences clustered together. However, they did not group with a high enough support as to be considered as a single transmission cluster (<0.90 support by aLRT [12]). Two well- supported subgroups of Valencian sequences (A and B, with aLRT support values of 0.99 and 1.00 respectively) were identified, thus suggesting the existence of two transmission clusters (Fig. 1, Supplementary Fig. 1). Cluster A included four patients, two of whom were known to be native Spanish MSM. Cluster B included three patients, two of whom were confirmed native Spanish MSM. No information about country of origin nor about the transmission route was available for the other Spanish sequences.

A relatively high root-to-tip vs divergence correlation ($R = 0.69$), calculated with Path-*O*-Gen, confirmed the presence of enough molecular clock signal for estimating node dates. We selected the exponential growth with a lognormal relaxed molecular clock model for estimating the tMRCA because it yielded the lowest AICM value (Table 1). The results obtained suggest two recent, independent introductions of CRF19_cpx in the Valencia region (median age for tMRCA of group A:2008.2, 95% HPD = 2005.4–2010.3; median age for tMRCA of group B:2010.6, 95% HPD = 2008.4–2012.1).

The screening for the presence of antiviral resistance mutations showed that there is a mutation (V179D), conferring low-level resistance to non-nucleoside retrotranscriptase inhibitors (NNRTI), which is present in all sequences from cluster A. On the other hand, two sequences from cluster B (PR0913158 and PRV1121) presented the mutation E138A, which is also associated with low-level resistance to some NNRTIs. While the existence of resistance mutations in these sequences may have caused spurious clustering due to the effect of convergent evolution, a phylogenetic tree with only third codon positions resolved clusters A and B, with the same topology (data not shown).

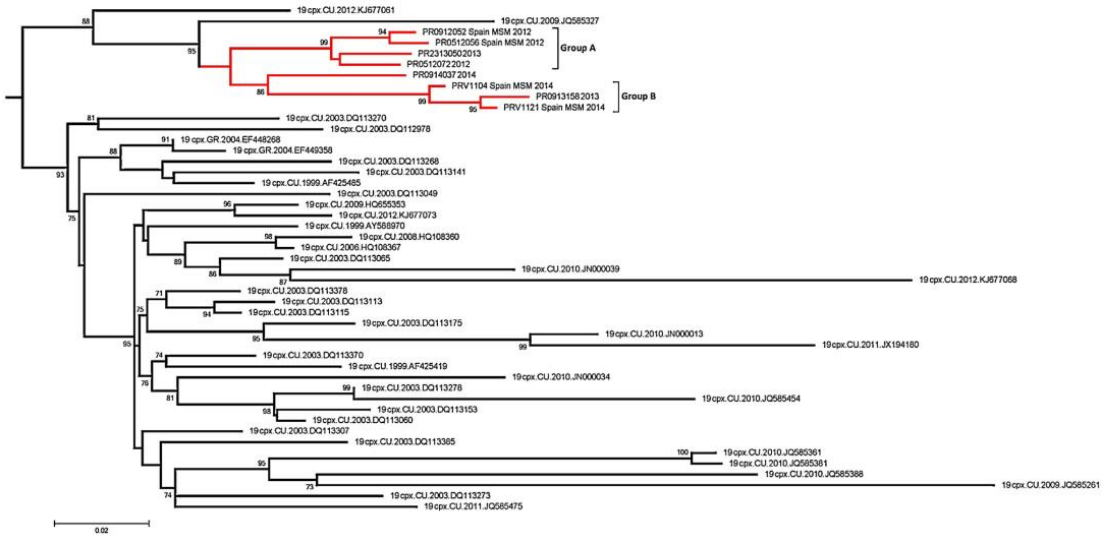



**Fig. 1.** Subtree containing the 8 Spanish CRF19_cpx sequences which group together (groups A and B), and the 38 nearest reference sequences as reconstructed from the full ML tree (Supplementary Fig. 1). Valencian sequences are highlighted in red. For each patient, sampling year and, if available, transmission route and country of origin are given. Only support values (aLRT) ≥70% are shown.

**Table 1**
Akaike's Information Criteria (AICM) values and median tMRCA for the transmission groups described, using four different demographic models.

| Demographic Model | AICM (value +/− SE) | median tMRCA-group A | median tMRCA-group B |
|---|---|---|---|
| Bayesian skyline Plot | 14333.60 +/− 0.78 | 2008.7 | 2010.9 |
| Constant | 14335.90 +/− 0.63 | 2009.2 | 2011.2 |
| Exponential | 14331.52 +/− 0.42 | 2008.4 | 2010.6 |
| Logistic | 14334.57 +/− 0.28 | 2008.0 | 2010.4 |

## 5. Discussion

This study reports, for the first time, the detection and apparent initial expansion of CRF19_cpx outside Cuba. Our results suggest that there have been at least three independent introductions of this HIV variant in the Valencian region, two of which have started to expand recently. Low-level resistance mutations were found in most sequences included in the detected transmission groups. Although CRF19_cpx is still a low prevalent variant in the Spanish population, the presence of transmission groups of recent origin suggests it as an emerging HIV variant in Spain, affecting Spanish native MSM and not only Cuban migrants. In light of the recent characterization of this HIV-1 CRF as a highly fit and pathogenic variant [6], the information obtained in this work should be taken into account in the design and implementation of public health control measures against HIV infections, especially among MSM, to prevent further expansion of this pathogenic variant in Spain and possibly in other countries.

## Conflict of interest declarations

We declare not to have any conflict of interest.

## Funding

Article partially funded by a contract with Gilead Sciences SL and project BFU2011-24112 from Ministerio de Economía y Competitividad (Spain).

## Competing interests

Funders played no role in the design, analysis and conclusions from this work. The authors declare not to have any conflict of interests.

## Ethical approval

The study was approved by the Committee for Ethics and Research (CEI) for DGSP/FISABIO.

## Acknowledgements

The authors would like to thank the patients for their participation in the study.

## Appendix A. Supplementary data

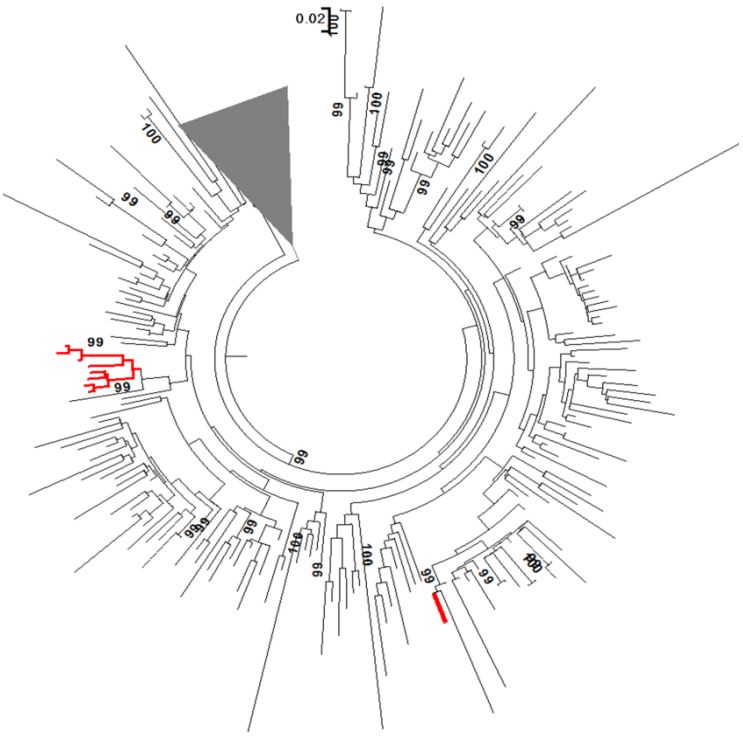
Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jcv.2015.06.094

## References

[1] G. Casado, M.M. Thomson, M. Sierra, R. Najera, Identification of a novel HIV-1 circulating ADG intersubtype recombinant form (CRF19_cpx) in Cuba, J. Acquir. Immune. Defic. Syndr. 40 (5) (2005) 532–537.
[2] L. Perez, M.M. Thomson, M.J. Bleda, C. Aragones, Z. Gonzalez, J. Perez, et al., HIV Type 1 molecular epidemiology in Cuba: high genetic diversity, frequent mosaicism, and recent expansion of BG intersubtype recombinant forms, AIDS Res. Hum. Retroviruses 22 (8) (2006) 724–733.

[3] V. Kourí, Y. Alemán, L. Pérez, J. Pérez, C. Fonseca, C. Correa, et al., High frequency of antiviral drug resistance and non-B subtypes in HIV-1 patients failing antiviral therapy in Cuba, J. Clin. Virol. 55 (4) (2012) 348–355.
[4] L. Pérez, V. Kourí, Y. Alemán, Y. Abrahantes, C. Correa, C. Aragonés, et al., Antiretroviral drug resistance in HIV-1 therapy-naive patients in Cuba, Infect. Genet. Evol. 16 (0) (2013) 144–150.
[5] E. Delatorre, G. Bello, Phylodynamics of the HIV-1 epidemic in Cuba, PLoS One 8 (9) (2013) e72448.
[6] Kourí V, Khouri R, Alemán Y, Abrahantes Y, Vercauteren J, Pineda-Peña AC, et al. CRF19_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba. EBioMedicine 2015. In press.
[7] Holguin A, Alvarez A, Soriano V. Heterogeneous nature of HIV-1 recombinants spreading in Spain. J Med Virol. 2005;75 (3):374-380.75 (3):374- 80.
[8] M.A. Bracho, V. Sentandreu, I. Alastrué, J. Belda, A. Juan, E. Fernández- García, et al., Emerging Trends in CRF02_AG Variants Transmission Among Men Who Have Sex With Men in Spain, J. Acquir. Immune. Defic. Synd. 65 (3) (2014) e130–e133.
[9] A.C. Pineda- Peña, N.R. Faria, S. Imbrechts, P. Libin, A.B. Abecasis, K. Deforche, et al., Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: Performance evaluation of the new REGA version 3 and seven other tools, Infect. Gen. Evol. 19 (2013) 337–348.
[10] M.N. Price, P.S. Dehal, A.P. Arkin, FastTree 2-Approximately maximum-likelihood trees for large alignments, PLoS One 5 (3) (2010) e9490.
[11] R.C. Edgar, MUSCLE. A multiple sequence alignment method with reduced time and space complexity, BMC Bioinformat. 5 (1) (2004) 113.
[12] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3. 0, Syst. Biol. 59 (3) (2010) 307–321.
[13] A.J. Drummond, A. Rambaut, BEAST. Bayesian evolutionary analysis by sampling trees, BMC Evol. Biol. 7 (2007) 214.
[14] G. Baele, P. Lemey, T. Bedford, A. Rambaut, M.A. Suchard, A.V. Alekseyenko, Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty, Mol. Biol. Evol. 29 (9) (2012) 2157–2167.
[15] T.F. Liu, R.W. Shafer, Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation, Clin Infect Dis. 42 (11) (2006) 1608–1618.

## Supplementary material




Supplementary Figure 1: ML tree obtained with PhyML, which includes the Valencian CRF19_cpx sequences (red), and subtype D (grey) and CRF19_cpx (black) reference sequences.

# 2.5- Chapter 5: The evolutionary rate of HIV-1 subtypes: a genomic approach

***This work is in the final stage of preparation.***

## The evolutionary rate of HIV-1 subtypes: a genomic approach

Juan Ángel Patiño Galindo[1] and Fernando González-Candelas[1]

[1]Unidad Mixta Infección y Salud Pública FISABIO-Salud Pública / Universitat de València.

CIBERESP, Valencia. SPAIN.

**Abstract**

**Background:** HIV-1 M causes most infections in the AIDS pandemic. This viral group presents a huge genetic diversity, with at least 10 pure subtypes and about 60 recombinant forms. We have performed a comparative analysis of the evolutionary rate of five pure subtypes (A1, B, C, D, and G) and two recombinant forms (CRF01_AE and CRF02 AG) using data obtained from nearly complete genome coding sequences.

**Materials and methods:** tMRCAs and evolutionary rates of these HIV genomes, and their genes, were estimated by Bayesian coalescent analyses. Genomic evolutionary rate estimates were compared between the HIV-1 datasets analyzed by means of randomization tests.

**Results:** Significant differences in the rate of evolution were found between subtypes, with subtypes C, A1 and CRF01_AE displaying the higest rates. On the other hand, CRF02_AG and subtype D had the lowest rates. Using a different molecular clock model at each genomic partition led to more accurate tMRCA estimates than when linking a same clock model along the HIV genome. Overall, the earliest tMRCA corresponded to subtype A1 (median=1941, 95%HPD= 1943-1955) whereas the most recent tMRCAs corresponded subtype G and CRF01_AE subset 3 (median =1971, 95% HPD= 1967-1975 and median = 1972, 95%HPD= 1970- 1975 respectively).

**Discussion:** These results suggest that both biological and epidemiological differences among HIV-1 M subtypes are reflected in their evolutionary dynamics. The obtained tMRCA and evolutionary rate estimates provide information that can be used as prior distributions in future Bayesian coalescent analyses of specific HIV -1 subtypes/CRFs and genes.

**Background**

HIV is a retrovirus of the genus *Lentivirus* and is characterized by a very high genetic diversity. There exist two types of HIV: HIV-1 and HIV-2. The former causes the AIDS pandemics and comprises four phylogenetically distinct groups: M, N, O, and P. Groups N, O are found almost exclusively in West-Central Africa (Hahn et al. 2000). Only two strains from group P have been reported so far, both in Cameroon (Plantier et al. 2009; Vallari et al. 2011). HIV-1 group M is the main driver of the HIV pandemics. Within this group, nine subtypes exist (denoted A, B, C, D, F, G, H, J, and K) and at least 61 circulating recombinant forms (Kuiken et al. 2012).

High mutation and evolutionary rates favor the genetic diversity of HIV. These are due to three main causes: (i) polymerization errors of the reverse transcriptase (Roberts et al. 1988); (ii) genetic recombination that produces viral chimeras (Temin 1993); and (iii) an explosive within-host proliferation and a large, and still growing, number of infected persons that lead to very large population sizes (Pennings et al. 2014). These factors facilitate the action of natural selection, favoring those mutations that increase the biological fitness of the virus and the elimination of disadvantageous alleles (Moya et al. 2004). Other factors act in the opposite direction. For instance, Simon-Loriere et al. (2013) showed that gene overlapping, which affects to all genes in the HIV genome, is inversely correlated to the rate of RNA virus evolution due to a reduction in the number of synonymous substitutions, although it would be less relevant in cases of terminal gene overlaps, which are the predominant type of overlap in HIV. Genetic bottlenecks during transmission also act slowing the pace of evolution in this virus, because many mutations accumulated within a host are lost after

transmission. The fact that adaptive changes at the within-host level are lost or reverted after transmission explains the consistently reported higher intra-host than between-host evolutionary rates of HIV-1 (Alizon & Fraser 2013; Duchêne et al. 2014; Lin et al. 2015). The speed of spread of HIV in an epidemic also influences its evolutionary rate (Maljkovic Berry et al. 2007). Hence, differences in selective pressures, mutation rates, replication capacity and/or epidemics dynamics may explain differences in the evolutionary rate among subtypes.

There are important differences in the prevalence of the different subtypes around the world. Subtype C is the most prevalent variant of HIV-1, occurring mainly in Africa (which presents the highest diversity of HIV-1) and Asia and accounting for almost 50% of the infections, but subtype B is the most widespread one, mainly affecting developed countries. Subtypes A, D, F, G, H, J and K display their highest prevalence in Sub-Saharan Africa. It is important to mention the increasing prevalence of circulating recombinant forms, especially CRF01_AE and CRF02_AG, which cause most of the infections in South-East Asia and Western Africa, respectively (Buonaguro et al. 2007).

Differences among subtypes in the intensity of selection have also been reported (Choisy et al. 2004). Its importance on the differential pace at which HIV-1 M subtypes evolve has been addressed by analyses of partial genes (Abecasis et al. 2009; Wertheim et al. 2012), thus ignoring the differences in the mutation rate and/or selective constraints that are known to exist between genomic regions (Geller et al. 2015).

Here, we present a comparative analysis of the evolution of the main HIV-1 subtypes using Bayesian coalescent reconstructions. The primary goal of our study was to compare the evolutionary rates of the main HIV-1 subtypes from a genomic perspective, by using near-full viral coding sequences, which should be more informative for the inference of the evolutionary rates and diversification dates than the individual genomic regions used so far.

## Materials and methods

### *Datasets*

Full coding-region sequences (CDS) were retrieved from the Los Alamos HIV Sequence Database, LANL (http://www.hiv.lanl.gov/) on October, 2015. Independent datasets were obtained for subtypes A1, B, C, D and G, and the CRF01_AE and CRF02_AG circulating recombinant forms. Although subtype F1 was initially considered, it was excluded from the study due to the low number of sequences retrieved. The first criterion for the selection of these sequences was the absence of problematic sequences (defined in LANL as sequences with a high proportion of non-ACTG characters or stop codons, presenting hypermutations, deletions or being either contaminants, synthetic constructs or reverse complements), the presence of only one sequence per patient and the exclusion of sequences with large deletions or undetermined regions (>5% of the sequence length). Sequences without known sampling date were also excluded. In order to exclude recombinant or incorrectly subtyped sequences the retained sequences were re-subtyped with the Comet HIV-1 (http://comet.retrovirology.lu) and REGAv3 HIV-1 subtyping tools (Pineda-Peña et al.

2013). All the sequences were also analyzed with five methods of recombination detection implemented in the RDP4 software, RDP, Geneconv, Bootscan, Maxchi and Chimaera (Smith 1992; Posada 2002; Padidam et al. 1999; Martin & Rybicki 2000; Martin et al. 2005; Martin et al. 2015). Sequences in which at least one method suggested recombination, with a P value <0.05, were considered for exclusion. In order to remove redundant sequences, alignments of the concatenated sequences were processed with CD-HIT (Huang et al. 2010) using a similarity threshold at 0.98. One sequence from each of the clusters found at this level was retained for further analysis.

Independent alignments of the non-overlapping regions from all genes were obtained, including the region spanning vpr to vpu, using MAFFT version 7 ("auto" strategy; Katoh & Standley 2013) Subsequently, regions of poor homology (gappy sites) were trimmed with trimAl (Capella-Gutiérrez et al. 2009). The final alignment lengths were: *gag*- 1295 nts*, pol* – 2746 nts, *vif* – 464 nts, *vpr*-to-*vpu*- 743 nts, *env* – 2316 nts, *nef* – 609 nts. Consequently, up to 8173 nts of the 8627 bp spanning the HIV-1 CDS were analyzed.

Due to the high number of B, C and CRF01_AE sequences that accomplished the selection criteria, and the computational limitations associated with the analysis of large genomic datasets, three different random subsets (n=100) for each of these HIV-1groups were generated with replacement. These subsets also allowed to check the robustness of the obtained estimates for these subtypes.

*Molecular clock signal analysis*

We checked the clock-likeness of our datasets by performing linear regression analyses between the parameters "root-to-tip divergence" and "sampling date" with

TempEST (Rambaut et al. 2016). For each subtype and CRF, a tree reconstructed with Fasttree2.1 (Price et al. 2010) was used as input, and the root was chosen as the branch that maximized the coefficient of correlation (R), under the assumption of a strict molecular clock.

*Evolutionary analyses*

tMRCA and genomic evolutionary rate estimates of each HIV-1 subtype and CRF were obtained by independent Bayesian Markov Chain Monte Carlo (MCMC) coalescent analyses, as implemented in BEAST v1.8.1 (Drummond & Rambaut 2007). Initially, the same partition tree and clock models were applied to all gene regions. All the analyses were performed with the HKY + Γ (4 categories) substitution model, combined with either an uncorrelated lognormal relaxed or the strict molecular clock model and three different demographic models (Bayesian Skyline Plot, and exponential or logistic demographic change). The best demographic model was chosen as that with the lowest Akaike's Information Criterion value (AICM) (Baele et al. 2012). We repeated the coalescent analyses using the GTR + Γ model, obtaining identical tMRCA and evolutionary rates results (data not shown).

For each viral group, we also estimated its tMRCA and the evolutionary rate of each gene partition by repeating the BEAST analyses by assigning a different molecular clock model to each gene partition.

At least two independent runs of BEAST were performed for each alignment, with MCMC chain lengths ranging between $30x10^6$ and $20x10^7$ states. Convergence of the estimated parameters was confirmed with Tracer (http://tree.bio.ed.ac.uk/software/tracer/) checking that effective sample sizes (ESS)

were larger than 200 for the estimated parameters.

Because substitutions in external branches may include recent, deleterious mutations leading to overestimates of the actual evolutionary rates, we compared the genomic evolutionary rates in internal and external branches for each subtype/CRF, using a Perl script (available upon request) to estimate the evolutionary rate ("ucld.mean") parameter independently for internal and external branches. As the estimates for internal and external branches were almost identical, tMRCA and evolutionary rate estimates reported in this work correspond to values estimated from both external and internal branches (data not shown).

*Pairwise comparisons of genomic evolutionary rates*

We tested whether the genomic evolutionary rate distributions estimated from BEAST were significantly different among subtypes/CRFs comparing pairwise posterior distributions by means of randomization tests, in which p-values were calculated by counting the number times that one evolutionary rate was lower than the other, considering 5000 tree states chosen randomly (with replacement) of each subtype/CRF. The obtained value (v) divided by 5000 (number of comparisons) was considered as the probability that the compared values belong to different distributions (Abecasis et al. 2009). P-values were obtained as "P=1-v", and were adjusted with the false discovery rate method (FDR; Benjamini & Hochberg 1995). These comparisons were performed using an in-home R script available upon request (R Core Team 2014).

## Results

*Datasets*

Of the 2399 HIV-1 M sequences initially retrieved from LANL, 1441 (A1: 96, B: 580, C: 450, D: 45, G: 32, AE: 208, AG: 30) were kept for further analyses. As mentioned in Material and Methods, three different random subsets (n=100) were obtained with replacement for subtypes B and C and CRF01_AE. In total, 248 subtype B, 234 subtype C and 177 01_AE sequences were included in the analyses. Consequently, 862 different sequences from the seven HIV-1 subtypes/CRFs were analyzed. Information on HIV-1 subtype/CRF, country of origin, sampling year and accession number of the sequences used in each dataset is shown in Supplementary Table S1.

*Molecular clock signal analyses*

The clock-like signal present in the analyzed datasets was evaluated by calculating the correlation coefficients (R) between the root-to-tip divergence and sampling date. R ranged between 0.50 (CRF02_AG) and 0.90 (subtype G) (Table 1). The possible existence of over dispersion of the HIV-1 molecular clock, which could be a major limitation for our comparisons, was rejected by ensuring that plots produced in the linear regression analyses of root-to-tip divergence vs sampling date for the concatenates did not present large dispersed clouds of points around the regression line. Residual mean squared values, which estimate the variance of the rates, for all the subtypes were lower than $2x10^{-4}$ (Table 1).

**Table 1**. Molecular clock signal of each HIV-1 dataset analyzed: correlation coefficient (R) and residual mean squared value obtained in the root-to-tip divergence vs sampling date correlation analyses

| Dataset | R | Residual mean squared values |
|---|---|---|
| A1 | 0.65 | 8.50E-05 |
| B-1 | 0.80 | 6.60E-05 |
| B-2 | 0.70 | 6.80E-05 |
| B-3 | 0.70 | 1.50E-04 |
| C-1 | 0.62 | 7.80E-05 |
| C-2 | 0.55 | 6.10E-05 |
| C-3 | 0.59 | 6.30E-05 |
| D | 0.77 | 2.00E-05 |
| G | 0.90 | 1.40E-05 |
| 01_AE-1 | 0.89 | 2.50E-05 |
| 01_AE-2 | 0.82 | 3.00E-05 |
| 01_AE-3 | 0.86 | 4.60E-05 |
| 02_AG | 0.50 | 1.00E-04 |

*Evolutionary rate and tMRCA estimates and comparisons*

Genomic evolutionary rate estimates of each HIV-1 subtype and CRF were obtained by Bayesian Markov Chain Monte Carlo (MCMC) coalescent analyses, as implemented in BEAST. The best-fitting demographic and molecular clock models for each HIV-1 subtype/CRF is shown in Table 2, and dated phylogenetic trees obtained from the near full CDS of each dataset under the best-fitting demographic and clock model are shown in Supplementary Figure S1.

**Table 2.** Akaike's Information Criterion values (AICM) obtained with the three demographic models (under a relaxed molecular clock model) and the strict lock model for each HIV-1 subtype/CRF. Values in brackets represent standard deviation values. The best fitting-model is highlighted in black

| | A1 | B# | C# | D | G | CRF01_AE# | CRF02_AG |
|---|---|---|---|---|---|---|---|
| BSP | 273384.5 (0.4) | 333487.0 (0.1) | **323525.1 (2.3)** | **155726.5 (0.2)** | **123944.8 (0.2)** | **230898.6 (0.6)** | **110222.6 (0.2)** |
| Expo | 273353.9 (0.8) | **333463.2 (0.7)** | 323525.9 (1.5) | 155744.4 (0.2) | 123948.2 (0.2) | 230907.8 (0.6) | 110224.7 (0.2) |
| Logistic | **273338.5 (0.5)** | 333466.4 (0.9) | 323528.4 (0.7) | 155733.4 (0.2) | 123946.7 (0.4) | 230925.9 (0.6) | 110224.4 (0.2) |
| Strict* | 273612.1 (0.3) | 333785.6 (0.3) | 323748.2 (0.2) | 155882.3 (0.1) | 123992.4 (0.1) | 231026.6 (0.7) | 110283.0 (0.2) |

*AICM value obtained using the best-fitting demographic model.

# For subtypes B, C and CRF01_AE, only the subdataset with highest molecular clock signal was subjected to model comparison.

Genomic HIV-1 evolutionary rates ranged between $1.3x10^{-3}$ s/s/y (95% HPD = 0.7 -$1.8x10^{-3}$ s/s/y) for CRF02_AG and $3.5{\cdot}10^{-3}$ s/s/y (95% HPD = 2.9 – $4.1x10^{-3}$ s/s/y) for subtype C dataset 2 (Figue 1A; Table 3). Randomization tests revealed significant inter-subtype differences, with subtypes A, C and CRF01_AE displaying significantly higher genomic evolutionary rates than CRF02_AG and subtypes B, D and G. Importantly, no significant intrasubtype differences were found between the random subsets of subtypes B, C and CRF01_AE (Figure 1B).

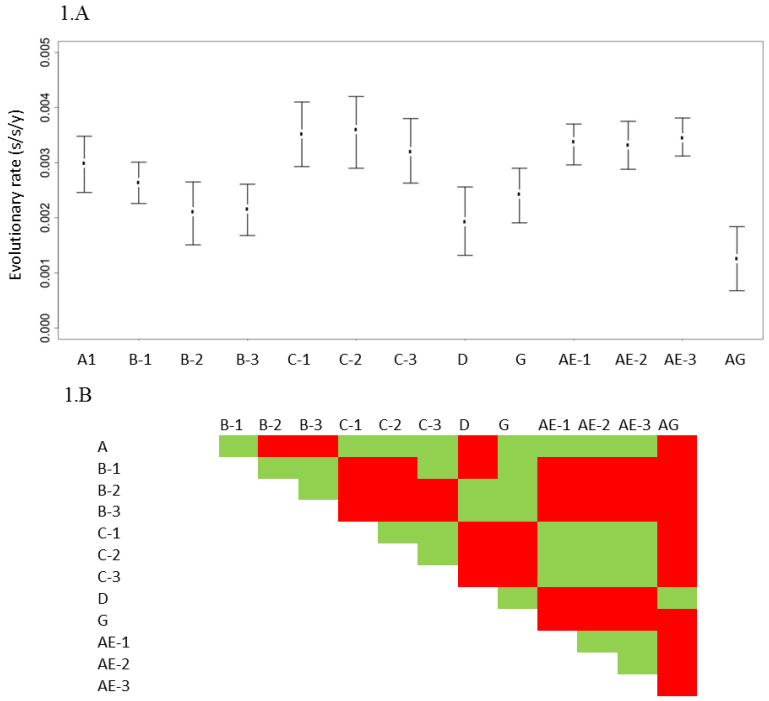



**Figure 1.** Comparison of the genomic evolutionary rate and tMRCA estimates of HIV-1 subtypes, as obtained with BEAST with the best-fitting demographic and molecular clock models: 1A) Plots of the median and 95% HPD intervals for the evolutionary rate (ucld.mean parameter); 1B) pairwise comparisons of the posterior distributions estimated for the evolutionary rate of each HIV-1 subtype, as obtained with a randomization test. Red: significantly different intervals (P value <0.05 after FDR correction). Green: not significantly different intervals**.**

**Table 3**. Estimates (median and 95% HPD lower and upper limit) for the tMRCA, and for the evolutionary rates (value x $10^{-3}$) of each genomic partition of the HIV-1 datasets analyzed, as obtained using the best-fitting demographic and molecular clock models

| Dataset (time-span) | tMRCA (unlinked clock model) | tMRCA (linked clock model) | Rate (CDS) | Rate (gag) | Rate (pol) | Rate (vif) | Rate (vpr-to-vpu) | Rate (env) | Rate (nef) |
|---|---|---|---|---|---|---|---|---|---|
| A1 (1985 - 2011) | 1949(1943-1955) | 1953(1944-1960) | 3.0(2.5-3.5) | 2.3(2.0-2.6) | 1.5(1.4-1.7) | 2.5(2.1-2.9) | 2.8(2.4-3.1) | 4.1(3.6-4.5) | 3.9(3.4-4.4) |
| B-1 (1983 - 2014) | 1955(1950-1959) | 1953(1945-1960) | 2.6(2.3-3.0) | 2.4(2.1-2.7) | 1.5(1.3-1.7) | 2.4(2.1-2.8) | 3.0(2.6-3.5) | 3.8(3.4-4.3) | 3.2(2.8-3.6) |
| B-2 (1984 - 2014) | 1949(1939-1957) | 1944(1923-1958) | 2.1(1.5-2.7) | 1.7(1.5-2.0) | 1.2(1.0-1.4) | 2.0(1.7-2.3) | 2.1(2.3-3.1) | 3.0(2.6-3.5) | 3.3(2.8-3.8) |
| B-3 (1983 - 2014) | 1950(1944-1955) | 1944(1930-1956) | 2.2(1.7-2.6) | 1.7(1.5-1.9) | 1.3(1.2-1.5) | 2.2(1.9-2.5) | 2.9(2.6-3.2) | 3.2(2.8-3.5) | 2.8(2.5-3.2) |
| C-1 (1989 - 2011) | 1965(1960-1969) | 1965(1958-1970) | 3.5(2.9-4.1) | 3.5(3.0-4.0) | 1.7(1.5-1.9) | 3.0(2.5-3.5) | 3.5(3.0-4.0) | 4.8(4.1-5.4) | 3.8(3.3-4.4) |
| C-2 (1989 - 2012) | 1964(1959-1969) | 1965(1957-1971) | 3.6(2.9-4.2) | 3.1(2.7-3.6) | 1.7(1.5-2.0) | 2.5(2.1-3.0) | 3.5(3.0-4.0) | 5.1(4.4-5.8) | 4.0(4.0-4.7) |
| C-3 (1989 - 2011) | 1961(1955-1966) | 1962(1955-1969) | 3.2(2.6-3.8) | 3.0(2.6-3.4) | 1.6(1.4-1.8) | 2.3(1.9-2.7) | 3.0(2.6-3.6) | 4.6(4.0-5.3) | 3.5(3.0-4.1) |
| D (1983 - 2011) | 1953(1946-1960) | 1956-1927-1961) | 1.9(1.3-2.6) | 1.9(1.5-2.2) | 1.2(1.0-1.4) | 1.9(1.5-2.4) | 2.9(2.3-3.6) | 3.4(2.8-4.1) | 2.8(2.3-3.4) |
| G (1992 - 2014) | 1971(1967-1975) | 1969(1961-1974) | 2.4(1.9-2.9) | 2.4(2.0-2.7) | 1.6(1.3-1.8) | 2.6(2.0-3.1) | 2.6(2.2-3.1) | 4.2(3.6-4.9) | 3.1(2.6-3.6) |
| AE-1 (1990 - 2012) | 1971(1969-1974) | 1970(1967-1974) | 3.4(3.0-3.7) | 3.1(2.7-3.4) | 1.4(1.3-1.6) | 3.4(2.9-3.8) | 4.0(3.5-4.5) | 4.6(4.1-5.0) | 4.2(3.7-4.8) |
| AE-2 (1990 - 2012) | 1971(1967-1974) | 1971(1965-1975) | 3.3(2.9-3.8) | 3.0(2.7-3.5) | 1.5(1.3-1.6) | 3.4(2.9-3.9) | 4.0(3.5-4.6) | 4.2(3.7-4.6) | 4.4(3.8-5.0) |
| AE-3 (1990 - 2011) | 1972(1970-1975) | 1972(1968-1975) | 3.5(3.1-3.8) | 3.2(2.9-3.6) | 1.5(1.3-1.6) | 4.0(3.5-4.6) | 4.3(3.7-4.8) | 4.6(4.1-5.0) | 4.3(3.8-4.9) |
| AG (1991 - 2012) | 1954(1939-1965) | 1948(1913-1969) | 1.3(0.7-1.8) | 1.2(0.9-1.5) | 0.8(0.6-1.1) | 1.5(1.0-2.0) | 1.6(1.2-2.1) | 2.1(1.6-2.7) | 2.0(1.4-2.6) |

Bayesian coalescent analyses were also performed, unlinking the molecular clock models of the different genomic partitions. Median tMRCAs estimated from this approach were very similar to those obtained when an only clock model was used (largest difference = 6 years, for subtype B subset 3 and CRF02_AG). However, 95% HPDs were more accurate (narrower) than when applying the same molecular clock model along the whole CDS (Table 3). 95%HPDs of tMRCAs estimated unlinking the molecular clock models were narrower than 15 years for all datasets, with subtype B dataset 2 (19 years) and CRF02_AG (27 years) as the only exceptions. However, when applying the same clock model along the whole CDS, 95% HPDs narrower than 15 years were only obtained for subtype C (all three datasets), G and CRF1_AE (all three datasets). Regarding the tMRCAs estimated obtained from the different subsets of subtypes B and C and CRF01_AE, the largest difference between medians of a same HIV-1 variant was found that between B0 and B3 subsets (6 years).

Overall, the earliest tMRCA corresponded to subtype A1 (median=1941, 95%HPD= 1943-1955) whereas the most recent tMRCAs corresponded subtype G and CRF01_AE subset 3 (median =1971, 95% HPD= 1967-1975 and median = 1972, 95%HPD= 1970- 1975 respectively).

In all cases, the evolutionary rates of the 5' half of HIV-1 genome (*gag*, *pol* and *vif*) were lower than the 3' half (*vpr*-to-*vpu*, *env* and *nef*). Concretely, *pol* presented the lowest evolutionary rate and *env* the highest in all HIV-1 subtypes/CRFs (Table 3).

**Discussion**

We have estimated, and compared, the genomic evolutionary rates of different HIV-1 subtypes (A1, B, C, D and G) and CRFs (CRF01_AE and CRF02_AG). To obtain representative datasets of the publicly available genomes for each HIV-1 variant we included sequences from the most complete geographical, temporal and genetic range as possible and removed epidemiologically related variants, including those obtained from the same patient.

Our analyses, performed by using tip-dates of heterochronous samples as the only calibration method, have revealed differences in the evolutionary rates between the analyzed subtypes and CRFs. Subtypes C and A1, and CRF01_AE presented the fastest evolutionary rates among the studied HIV-1 datasets, with CRF02_AG and subtype D being the slowest-evolving groups.

As expected, evolutionary rates estimates for the different subtypes and CRFs differed from previous analyses working with partial *pol* and/or *env* regions (Abecasis et al. 2009; Wertheim et al. 2012). Abecasis et al. (2009) analyzed partial *pol* and *env* sequences of up to 799 and 931 bp respectively and found that subtype G and CRF02_AG had the highest evolutionary rate, and subtype D the lowest, for these two genes. On the other hand, Wertheim et al. (2012) analyzed complete pol sequences and found subtype B to be evolving faster than subtypes D and C. These incongruences between different studies can be explained by the different genomic regions that they analyzed, focusing on the evolutionary rates of short genomic regions, but ignoring differences in selective constraints or mutation rates that exist along the whole HIV CDS (Geller et al. 2015). These discrepancies could also be

explained by the different dataset sizes: Although, compared to our datasets, larger datasets have been previously analyzed for CFR02_AG and subtype G (Abecasis et al. 2009), and for subtype D (Wertheim et al. 2012), we analyzed larger datasets than previous works for the remaining HIV-1 groups.

All the analyzed HIV-1 subtypes and CRFs, displayed a similar pattern, regarding the evolutionary rates estimated along their genomes: genes *gag*, *pol* and *vif* presented lower rates than the *vpr*-to-*vpu* segment and *env* and *nef* genes, with *pol* and *env* presenting the slowest and fastest evolutionary rates, respectively. Li et al. (2015) found higher levels of amino acid diversity in the proteins encoded by *tat*, *rev*, *vpu*, *env* and *nef* genes, and associated their higher levels of variability to different factors: firstly, it could be associated with the presence of CD4 T cell and antibody epitopes, which would favor diversifying selective pressures. Secondly, these proteins were found to present higher numbers of HIV-human associations, which may lead these proteins to present higher structural flexibility.

Using nearly complete genome coding regions, the 95% HPD intervals obtained for the tMRCA of each subtype and gene were in most cases in agreement with previous estimates obtained (Abecasis et al. 2009; Gray et al. 2009; Yebra et al. 2016). However, tMRCA estimates for subtypes B, C and were discordant with respect those obtained by Faria et al. (2014), who estimated the tMRCA of subtype B to have occurred in the 40s, and that of subtype C in the late 30s. The most plausible reasons for such discrepancies is that Faria et al. (2014) disposed of older sequences, obtained from samples existing in the late 50s-early 60s in Kinsasha (Democratic Republic of the Congo), while our oldest sequences were obtained in the 80s, at the

beginning of the AIDS pandemic. Indeed, the tMRCA of our subtype C datasets are more similar to that obtained by (Wilkinson et al. 2015) for the origin of the southern African epidemic (median: year 1960), the geographic region from which most of our sequences were obtained. In this way, such discrepancies can be explained by the different time-spans sampled between these previous estimates and ours.

tMRCA estimated for CRF02_AG presented the broadest 95%HPD among all the HIV subtypes/CRFs analyzed in this work, probably because it was also the dataset with the lowest molecular clock signal. However, the median tMRCA obtained for this CRF was only 8 years older than that estimated by Yebra et al. (2016) in West Africa (median tMRCA between 1962 and 1963 as estimated from PR and gp41, respectively). These tMRCA estimates obtained for CRF02_AG suggest an earlier origin than that of HIV-1 subtype G, which was supposed to be one of its parental subtypes. These estimates support previous results revealing that, indeed, CRF02_AG is the parent of subtype G, which is not an actual pure subtype although it remains classified as one (Abecasis et al. 2007) .

Overall, analyzing nearly complete coding regions has produced more accurate tMRCA and evolutionary rate estimates than others obtained previously (Abecasis et al. 2009; Gray et al. 2009; Wertheim et al. 2012), especially when different molecular clock models are used for the different gene partitions comprising the CDS. However, is noteworthy that for some HIV-1 groups (subtypes D, G and CRF02_AG), the number of available genomes was much lower than for the others. This work aimed to obtain the most representative CDS datasets as possible, and the analyzed datasets represent the genome availability at public databases for

each HIV-1 subtype/CRF. However, it is possible that the limited number of complete CDS sequences available for subtypes D, G and CRF02_AG may have affected our estimates. Despite this potential caveat, estimating tMRCA and evolutionary rate estimates from independent datasets for each genome partition would lack the statistical power that confers the information present in the different genome partitions in BEAST. Furthermore, it would introduce other bias, such as sampling differences between genomic regions from a same HIV subtype/CRF.

It is also important to mention other possible bias in the estimates, such as those caused by the time-dependency of the evolutionary rates. Rates can be very different when analyzing viral datasets from different timescales, with shorter timescales associated with higher estimates (Duchêne et al. 2014; Aiewsakun & Katzourakis 2015; Aiewsakun & Katzourakis 2016). Meyer et al. (2015) assessed the effect of time dependence of the evolutionary rate of influenza during the 2009 pandemic outbreak, and found that at least 9 months of temporal divergence was needed for precise estimates for long term values. In our work, we have used datasets with similar scales for the sampling time-spans, which ranged between 21 and 31 years. For this reason, such phenomenon should not bias our comparisons.

In conclusion, we have estimated and compared the tMRCAs and genomic evolutionary rates of the main HIV-1 M subtypes and CRFs from a genomic perspective, using the longest non-overlapping coding regions as possible. The results obtained show that evolutionary rates differ significantly among HIV-1 subtypes and CRFs, and that the accuracy of the estimated evolutionary parameters increases when independent molecular clock models are applied at each genomic

partition. The obtained results provide information that can be used as prior distributions in future Bayesian coalescent analyses of specific HIV -1 subtypes/CRFs and genes, given that the evolutionary rates of HIV-1 varies between subtypes/CRFs and genomic regions.

**Acknowledgements**

This work was supported by Ministerio de Economía y Competitividad (MINECO), Spanish Government [projects BFU2011-24112 and BFU2014-58656-R] and Dirección General de Salud Pública (Generalitat Valenciana, Spain). JAPG was recipient of a Formación de Profesorado Universitario (FPU) predoctoral fellowship from Ministerio de Educación, Spanish Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**References**


Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme A-M. 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81:8543–8551.

Abecasis AB, Vandamme A-M, Lemey P. 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J. Virol. 83:12917–12924.

Aiewsakun P, Katzourakis A. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol. Biol. 15:119.

Aiewsakun P, Katzourakis A. 2016. Time-Dependent Rate Phenomenon in Viruses. J. Virol. 90:7184–7195.

Alizon S, Fraser C. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology 10:49.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57:289–300.

Buonaguro L, Tornesello ML, Buonaguro FM. 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J. Virol. 81:10209–10219.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Choisy M, Woelk CH, Guégan J-F, Robertson DL. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J. Virol. 78:1962–1970.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Duchêne S, Holmes EC, Ho SYW. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. 281.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, et al. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 346:56–61.

Geller R, Domingo-Calap P, Cuevas JM, Rossolillo P, Negroni M, Sanjuán R. 2015. The external domains of the HIV-1 envelope are a mutational cold spot. Nat.

Commun. 6:8571.

Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, Serwadda D, Sewankambo N, Gray RH, Wawer M, Quinn TC, et al. 2009. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. AIDS 23:F9–F17.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. 2000. AIDS as a zoonosis: scientific and public health implications. Science 287:607–614.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kuiken CL, Foley B, Leitner T, Apetrei C, Mizrachi Y, Mullins JI, Rambaut A, Wolinsky SM, Korber B. 2012. HIV Sequence Compendium 2012. New Mexico: Theoretical Biology and Niophysics Group. Los Alamos national Laboratory

Li G, Piampongsant S, Faria NR, Voet A, Pineda-Peña A-C, Khouri R, Lemey P, Vandamme A-M, Theys K. 2015. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology 12:18.

Lin Y-Y, Liu C, Chien W-H, Wu L-L, Tao Y, Wu D, Lu X, Hsieh C-H, Chen P-J, Wang H-Y, et al. 2015. New insights into the evolutionary rate of hepatitis B virus at different biological scales. J. Virol. 89:3512–3522.

Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, Bruno W, Leitner T. 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J. Virol. 81:10625–10635.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562–563.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1:vev003.

Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retroviruses 21:98–102.

Meyer AG, Spielman SJ, Bedford T, Wilke CO. 2015. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evol. 1.

Moya A, Holmes EC, González-Candelas F. 2004. The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2:279–288.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–225.

Pennings PS, Kryazhimskiy S, Wakeley J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 10:e1004000.

Pineda-Peña A-C, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, Gómez-López A, Camacho RJ, de Oliveira T, Vandamme A-M. 2013. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect. Genet. Evol. 19:337–348.

Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F. 2009. A new human immunodeficiency virus derived from gorillas. Nat. Med. 15:871–872.

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19:708–717.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical computing.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2:vew007.

Roberts JD, Bebenek K, Kunkel TA. 1988. The accuracy of reverse transcriptase from HIV-1. Science 242:1171–1173.

Simon-Loriere E, Holmes EC, Pagán I. 2013. The effect of gene overlapping on the rate of RNA virus evolution. Mol. Biol. Evol. 30:1916–1928.

Smith JM. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34:126–129.

Temin HM. 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc. Natl. Acad. Sci. U. S. A. 90:6900–6903.

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, Mbanya D, Kaptué L, Ndembi N, Gürtler L, et al. 2011. Confirmation of putative HIV-1 group P in Cameroon. J. Virol. 85:1403–1407.

Wertheim JO, Fourment M, Kosakovsky Pond SL. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol. Biol. Evol. 29:451–456.

Wilkinson E, Engelbrecht S, de Oliveira T. 2015. History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. Sci. Rep. 5:16897.

Yebra G, Kalish ML, Leigh Brown AJ. 2016. Reconstructing the HIV-1 CRF02_AG and CRF06_cpx epidemics in Burkina Faso and West Africa using early samples.

Infect. Genet. Evol. In Press .

**Supplementary material**

**Supplementary Table S1.** Subtype, country of origin, isolation date and accession number of the HIV-1 sequences from each dataset analyzed

| Subtype/CRF | Country | Isolation date | Accession number |
|---|---|---|---|
| A1 | Uganda | 1985 | M62320 |
| A1 | Kenya | 1986 | AF539405 |
| A1 | Uganda | 1992 | AB098332 |
| A1 | Uganda | 1992 | AB253428 |
| A1 | Rwanda | 1992 | AB253421 |
| A1 | Rwanda | 1992 | AB287376 |
| A1 | Rwanda | 1993 | AB287378 |
| A1 | Rwanda | 1993 | AY713406 |
| A1 | Sweden | 1994 | AF069670 |
| A1 | Kenya | 1994 | AF004885 |
| A1 | Sweden | 1994 | AF069671 |
| A1 | Sweden | 1995 | AF069669 |
| A1 | Kenya | 1997 | AY322193 |
| A1 | Tanzania | 1997 | AF361873 |
| A1 | Tanzania | 1997 | AF361872 |
| A1 | DR Congo | 1997 | AM000054 |
| A1 | DR Congo | 1997 | AM000053 |
| A1 | Uganda | 1998 | AF484507 |
| A1 | Uganda | 1998 | AF484509 |
| A1 | Uganda | 1998 | AF484512 |
| A1 | Kenya | 1999 | AF457063 |
| A1 | Kenya | 1999 | AF457075 |
| A1 | India | 1999 | KT152841 |
| A1 | Kenya | 1999 | AF457065 |
| A1 | Kenya | 2000 | AF457089 |
| A1 | Kenya | 2000 | AF457055 |
| A1 | Kenya | 2000 | AF457068 |
| A1 | Kenya | 2000 | AF457069 |
| A1 | Kenya | 2000 | AF457086 |
| A1 | Kenya | 2000 | AF457070 |
| A1 | Kenya | 2000 | AF457066 |
| A1 | India | 2000 | KT152846 |
| A1 | Kenya | 2000 | AF457081 |
| A1 | Kenya | 2000 | AF457080 |
| A1 | Kenya | 2000 | AF457053 |
| A1 | Ukraine | 2000 | AF413987 |
| A1 | Russia | 2000 | EF545108 |
| A1 | Kenya | 2001 | EU110087 |
| A1 | Tanzania | 2001 | AY253305 |
| A1 | Tanzania | 2001 | AY253314 |
| A1 | Senegal | 2001 | AY521629 |
| A1 | Ukraine | 2001 | DQ823358 |
| A1 | Ukraine | 2001 | DQ823359 |
| A1 | Ukraine | 2001 | DQ823365 |
| A1 | Ukraine | 2001 | DQ823361 |

| | | | |
|---|---|---|---|
| A1 | Ukraine | 2001 | DQ823366 |
| A1 | Ukraine | 2001 | DQ823357 |
| A1 | Kenya | 2002 | EU110092 |
| A1 | Russia | 2002 | JQ292892 |
| A1 | Kazakhstan | 2002 | EF589043 |
| A1 | Kazakhstan | 2002 | EF589042 |
| A1 | Uzbekistan | 2002 | AY829210 |
| A1 | Italy | 2002 | EU861977 |
| A1 | Australia | 2003 | DQ676872 |
| A1 | Switzerland | 2003 | JQ403028 |
| A1 | Russia | 2003 | AY500393 |
| A1 | South Africa | 2004 | KT183312 |
| A1 | Russia | 2005 | JQ292895 |
| A1 | Spain | 2005 | FJ670519 |
| A1 | Kenya | 2006 | FJ623480 |
| A1 | Kenya | 2006 | FJ623477 |
| A1 | Kenya | 2006 | FJ623485 |
| A1 | Kenya | 2006 | FJ623486 |
| A1 | Kenya | 2006 | FJ623478 |
| A1 | Kenya | 2006 | FJ623475 |
| A1 | Kenya | 2006 | FJ623484 |
| A1 | Kenya | 2006 | FJ623481 |
| A1 | Kenya | 2006 | FJ623476 |
| A1 | Kenya | 2006 | FJ623487 |
| A1 | Kenya | 2006 | FJ623479 |
| A1 | Kenya | 2006 | FJ623483 |
| A1 | Kenya | 2006 | FJ623488 |
| A1 | Russia | 2006 | JQ292900 |
| A1 | Russia | 2006 | JQ292899 |
| A1 | Russia | 2006 | JQ292897 |
| A1 | Russia | 2006 | JQ292896 |
| A1 | Russia | 2006 | JQ292898 |
| A1 | Uganda | 2007 | JX236676 |
| A1 | Rwanda | 2007 | KP223844 |
| A1 | Cameroon | 2007 | KP718918 |
| A1 | Uganda | 2007 | JX236669 |
| A1 | Uganda | 2007 | JX236671 |
| A1 | Russia | 2007 | JQ292891 |
| A1 | Cameroon | 2008 | KP718928 |
| A1 | Russia | 2008 | KF716492 |
| A1 | Russia | 2008 | JQ292894 |
| A1 | Russia | 2008 | JQ292893 |
| A1 | Russia | 2008 | KF716491 |
| A1 | Russia | 2008 | FJ864679 |
| A1 | Uganda | 2009 | KF716478 |
| A1 | Russia | 2010 | JX500695 |
| A1 | Russia | 2010 | JX500696 |
| A1 | Uganda | 2011 | KF859745 |
| A1 | Kenya | 2011 | KF716475 |

| | | | |
|---|---|---|---|
| A1 | Russia | 2011 | JX500694 |
| A1 | Uganda | 2011 | KF716486 |
| B-1 | Great Britain | 1983 | D10112 |
| B-1 | USA | 1983 | M17451 |
| B-1 | USA | 1983 | K02007 |
| B-1 | South Africa | 1985 | FJ647145 |
| B-1 | USA | 1990 | AY173954 |
| B-1 | USA | 1991 | AB485638 |
| B-1 | Soth Korea | 1992 | KJ140247 |
| B-1 | Soth Korea | 1992 | KJ140267 |
| B-1 | Soth Korea | 1992 | KJ140266 |
| B-1 | Australia | 1995 | AF538304 |
| B-1 | Thailand | 1996 | DQ354118 |
| B-1 | Georgia | 1998 | DQ207943 |
| B-1 | USA | 1998 | AY331294 |
| B-1 | USA | 1998 | EF175209 |
| B-1 | CA | 1998 | AY779550 |
| B-1 | USA | 1998 | AY560108 |
| B-1 | Australia | 1999 | AF538303 |
| B-1 | Cuba | 1999 | AY586542 |
| B-1 | Cuba | 1999 | AY586543 |
| B-1 | Japan | 2000 | AB289587 |
| B-1 | USA | 2000 | JN944930 |
| B-1 | USA | 2000 | EF363122 |
| B-1 | USA | 2000 | FJ496167 |
| B-1 | Switzerland | 2000 | KC797171 |
| B-1 | USA | 2001 | FJ496151 |
| B-1 | UA | 2001 | DQ823364 |
| B-1 | Denmark | 2001 | EF514711 |
| B-1 | Denmark | 2001 | EF514707 |
| B-1 | Colombia | 2001 | AY561237 |
| B-1 | Brazil | 2002 | JN692431 |
| B-1 | Brazil | 2002 | JN692444 |
| B-1 | Brazil | 2002 | JN692432 |
| B-1 | USA | 2002 | FJ469741 |
| B-1 | USA | 2002 | FJ496072 |
| B-1 | Brazil | 2002 | DQ358805 |
| B-1 | Soth Korea | 2003 | JQ316131 |
| B-1 | Brazil | 2003 | EF637048 |
| B-1 | Switzerland | 2003 | JQ403045 |
| B-1 | USA | 2003 | FJ469731 |
| B-1 | Brazil | 2003 | EF637056 |
| B-1 | Georgia | 2003 | DQ207942 |
| B-1 | Georgia | 2003 | DQ207940 |
| B-1 | Brazil | 2004 | FJ195090 |
| B-1 | Soth Korea | 2004 | DQ295196 |
| B-1 | Germany | 2004 | JQ403038 |
| B-1 | Argentina | 2004 | DQ383750 |
| B-1 | Japan | 2004 | AB480692 |

| | | | |
|---|---|---|---|
| B-1 | Denmark | 2004 | EF514701 |
| B-1 | Russia | 2004 | AY682547 |
| B-1 | Australia | 2004 | AY818644 |
| B-1 | USA | 2004 | JN024203 |
| B-1 | Japan | 2004 | AB221125 |
| B-1 | Spain | 2005 | KT200351 |
| B-1 | Japan | 2005 | AB428556 |
| B-1 | Brazil | 2005 | JN692474 |
| B-1 | Thailand | 2005 | JN248347 |
| B-1 | USA | 2005 | FJ496081 |
| B-1 | USA | 2005 | JF320361 |
| B-1 | USA | 2006 | JQ403095 |
| B-1 | USA | 2006 | JF320120 |
| B-1 | USA | 2006 | JF320615 |
| B-1 | USA | 2006 | FJ495937 |
| B-1 | USA | 2006 | FJ469696 |
| B-1 | Spain | 2006 | KT200350 |
| B-1 | Peru | 2006 | JF320241 |
| B-1 | USA | 2006 | JF320564 |
| B-1 | USA | 2006 | JF320036 |
| B-1 | USA | 2006 | JF689865 |
| B-1 | China | 2007 | JF932469 |
| B-1 | China | 2007 | JF932490 |
| B-1 | USA | 2007 | JF689885 |
| B-1 | USA | 2007 | JQ403056 |
| B-1 | China | 2007 | JF932493 |
| B-1 | USA | 2007 | JQ403060 |
| B-1 | China | 2007 | JF932487 |
| B-1 | China | 2007 | JF932489 |
| B-1 | China | 2007 | JF932470 |
| B-1 | Soth Korea | 2007 | JQ341411 |
| B-1 | Spain | 2008 | FJ853620 |
| B-1 | Spain | 2008 | KT200358 |
| B-1 | USA | 2008 | JF689896 |
| B-1 | USA | 2008 | JQ403088 |
| B-1 | Spain | 2009 | GU362885 |
| B-1 | Spain | 2009 | KC473841 |
| B-1 | USA | 2010 | KC473826 |
| B-1 | USA | 2010 | JN397365 |
| B-1 | Brazil | 2010 | KJ849808 |
| B-1 | Brazil | 2010 | KJ849784 |
| B-1 | China | 2010 | JX140658 |
| B-1 | Russia | 2010 | JX500707 |
| B-1 | Russia | 2011 | JX500708 |
| B-1 | USA | 2011 | KF384806 |
| B-1 | USA | 2011 | KF384805 |
| B-1 | USA | 2011 | KF384804 |
| B-1 | USA | 2011 | KF384807 |
| B-1 | USA | 2011 | KF384799 |

| | | | |
|---|---|---|---|
| B-1 | China | 2012 | KP109511 |
| B-1 | Spain | 2014 | KT276262 |
| B-1 | Spain | 2014 | KT276263 |
| B-1 | Spain | 2014 | KT276268 |
| B-2 | USA | 1984 | AY835779 |
| B-2 | USA | 1986 | M38429 |
| B-2 | Great Britain | 1986 | AJ271445 |
| B-2 | Thailand | 1990 | AY173951 |
| B-2 | Brazil | 1990 | AB485641 |
| B-2 | USA | 1990 | AY173954 |
| B-2 | Soth Korea | 1992 | KJ140257 |
| B-2 | USA | 1993 | DQ487188 |
| B-2 | USA | 1994 | AY713410 |
| B-2 | USA | 1995 | JN024274 |
| B-2 | Australia | 1995 | AF538307 |
| B-2 | Soth Korea | 1995 | KJ140250 |
| B-2 | USA | 1997 | AY331289 |
| B-2 | Argentina | 1998 | AY037268 |
| B-2 | Uruguay | 1999 | JN235959 |
| B-2 | Japan | 2000 | AB289587 |
| B-2 | USA | 2000 | JN944917 |
| B-2 | USA | 2000 | FJ496167 |
| B-2 | Switzerland | 2000 | KC797171 |
| B-2 | Denmark | 2001 | EF514711 |
| B-2 | China | 2002 | DQ007901 |
| B-2 | Paraguay | 2002 | JN251901 |
| B-2 | Japan | 2002 | AB428553 |
| B-2 | Brazil | 2002 | JN692470 |
| B-2 | Australia | 2003 | DQ676880 |
| B-2 | Brazil | 2003 | JN692445 |
| B-2 | Australia | 2003 | DQ676877 |
| B-2 | USA | 2003 | FJ469764 |
| B-2 | Switzerland | 2003 | JQ403045 |
| B-2 | Brazil | 2003 | EF637056 |
| B-2 | China | 2003 | JF932492 |
| B-2 | USA | 2003 | FJ469731 |
| B-2 | Paraguay | 2003 | JN251906 |
| B-2 | Brazil | 2004 | JN692457 |
| B-2 | Thailand | 2004 | JN248329 |
| B-2 | Russia | 2004 | AY751407 |
| B-2 | Soth Korea | 2004 | DQ295195 |
| B-2 | USA | 2004 | JQ403107 |
| B-2 | Great Britain | 2004 | HM586198 |
| B-2 | Russia | 2004 | AY682547 |
| B-2 | USA | 2004 | FJ469695 |
| B-2 | Denmark | 2004 | EF514698 |
| B-2 | Switzerland | 2004 | JQ403042 |
| B-2 | Brazil | 2004 | FJ195090 |
| B-2 | USA | 2004 | JN024100 |

| | | | |
|---|---|---|---|
| B-2 | Brazil | 2004 | FJ195086 |
| B-2 | Germany | 2004 | JQ403038 |
| B-2 | USA | 2005 | JF320054 |
| B-2 | Spain | 2005 | KT200351 |
| B-2 | USA | 2005 | DQ886037 |
| B-2 | USA | 2005 | JF689854 |
| B-2 | Thailand | 2005 | JN248353 |
| B-2 | USA | 2005 | JF689857 |
| B-2 | Japan | 2005 | AB287363 |
| B-2 | China | 2005 | DQ990880 |
| B-2 | USA | 2005 | FJ469724 |
| B-2 | Spain | 2005 | EU786672 |
| B-2 | USA | 2005 | JF689852 |
| B-2 | USA | 2006 | JF320044 |
| B-2 | USA | 2006 | JF320169 |
| B-2 | USA | 2006 | JF689871 |
| B-2 | USA | 2006 | JF689874 |
| B-2 | USA | 2006 | FJ495818 |
| B-2 | USA | 2006 | FJ495937 |
| B-2 | Spain | 2006 | EU786675 |
| B-2 | USA | 2006 | JF320145 |
| B-2 | Spain | 2006 | EU786674 |
| B-2 | Peru | 2006 | JF320186 |
| B-2 | Peru | 2006 | JF320196 |
| B-2 | China | 2006 | JF932482 |
| B-2 | USA | 2006 | JF689864 |
| B-2 | USA | 2006 | JF689872 |
| B-2 | Hong Kong | 2006 | FJ460501 |
| B-2 | Peru | 2006 | JF320008 |
| B-2 | USA | 2007 | JQ403078 |
| B-2 | USA | 2007 | JF320150 |
| B-2 | USA | 2007 | JF689890 |
| B-2 | China | 2007 | JF932493 |
| B-2 | USA | 2007 | FJ469689 |
| B-2 | USA | 2007 | JF689889 |
| B-2 | Spain | 2008 | FJ670531 |
| B-2 | China | 2008 | JF932479 |
| B-2 | USA | 2008 | JQ403087 |
| B-2 | USA | 2008 | JQ403035 |
| B-2 | USA | 2008 | JQ403062 |
| B-2 | USA | 2008 | JF689896 |
| B-2 | China | 2008 | JF932478 |
| B-2 | France | 2009 | KF716494 |
| B-2 | Spain | 2010 | JX140659 |
| B-2 | Brazil | 2010 | KJ849808 |
| B-2 | Brazil | 2010 | KJ849780 |
| B-2 | Spain | 2010 | KT200356 |
| B-2 | Japan | 2011 | KF716497 |
| B-2 | USA | 2011 | KF384806 |

| | | | |
|---|---|---|---|
| B-2 | USA | 2011 | KF384798 |
| B-2 | USA | 2011 | KF384807 |
| B-2 | USA | 2011 | KF526174 |
| B-2 | Japan | 2012 | KF716498 |
| B-2 | Spain | 2013 | KT200348 |
| B-2 | Spain | 2014 | KT276267 |
| B-3 | USA | 1983 | K02007 |
| B-3 | Great Britain | 1983 | D10112 |
| B-3 | USA | 1984 | M17449 |
| B-3 | South Africa | 1985 | FJ647145 |
| B-3 | USA | 1985 | AY835769 |
| B-3 | USA | 1986 | M93258 |
| B-3 | Germany | 1986 | U43141 |
| B-3 | USA | 1988 | AF286365 |
| B-3 | Thailand | 1990 | AY173951 |
| B-3 | USA | 1991 | AF069140 |
| B-3 | Soth Korea | 1992 | KJ140266 |
| B-3 | Soth Korea | 1992 | KJ140261 |
| B-3 | Great Britain | 1994 | KJ019215 |
| B-3 | USA | 1996 | AY331292 |
| B-3 | USA | 1999 | JQ403100 |
| B-3 | Australia | 1999 | AF538303 |
| B-3 | USA | 1999 | AY331296 |
| B-3 | USA | 2000 | JN944930 |
| B-3 | Japan | 2001 | AB289589 |
| B-3 | Uruguay | 2001 | AY781127 |
| B-3 | USA | 2002 | FJ469760 |
| B-3 | USA | 2002 | FJ469758 |
| B-3 | USA | 2002 | FJ469687 |
| B-3 | USA | 2002 | FJ496072 |
| B-3 | Australia | 2002 | DQ676883 |
| B-3 | Brazil | 2002 | JN692433 |
| B-3 | Brazil | 2002 | JN692432 |
| B-3 | South Africa | 2003 | DQ396398 |
| B-3 | USA | 2003 | FJ469764 |
| B-3 | Switzerland | 2003 | JQ403045 |
| B-3 | Brazil | 2003 | EF637056 |
| B-3 | Brazil | 2003 | JN692445 |
| B-3 | USA | 2004 | FJ469733 |
| B-3 | Germany | 2004 | JQ403050 |
| B-3 | Thailand | 2004 | JN248337 |
| B-3 | Soth Korea | 2004 | DQ295195 |
| B-3 | Japan | 2005 | AB287364 |
| B-3 | Thailand | 2005 | JN248346 |
| B-3 | USA | 2005 | FJ469682 |
| B-3 | USA | 2005 | JF689854 |
| B-3 | USA | 2005 | JF689859 |
| B-3 | USA | 2005 | KF990605 |
| B-3 | Spain | 2005 | EU786672 |

| | | | |
|---|---|---|---|
| B-3 | USA | 2005 | JF320185 |
| B-3 | USA | 2005 | FJ469683 |
| B-3 | Brazil | 2005 | JN692463 |
| B-3 | Thailand | 2005 | JN248344 |
| B-3 | Brazil | 2005 | JN692473 |
| B-3 | Brazil | 2005 | JN692475 |
| B-3 | USA | 2006 | JF320615 |
| B-3 | CA | 2006 | JF320427 |
| B-3 | USA | 2006 | KF990608 |
| B-3 | USA | 2006 | JF320120 |
| B-3 | China | 2006 | JF932482 |
| B-3 | USA | 2006 | JF689862 |
| B-3 | USA | 2006 | JN944897 |
| B-3 | Brazil | 2006 | JN692479 |
| B-3 | Peru | 2007 | JF320189 |
| B-3 | USA | 2007 | JF689892 |
| B-3 | Peru | 2007 | JF320018 |
| B-3 | USA | 2007 | JQ403060 |
| B-3 | USA | 2007 | JF320559 |
| B-3 | USA | 2007 | JF320150 |
| B-3 | China | 2007 | JF932485 |
| B-3 | China | 2007 | JF932493 |
| B-3 | USA | 2007 | JF689879 |
| B-3 | Soth Korea | 2007 | JQ341411 |
| B-3 | USA | 2007 | JF320197 |
| B-3 | USA | 2008 | JQ403061 |
| B-3 | USA | 2008 | JQ403069 |
| B-3 | USA | 2008 | JQ403035 |
| B-3 | France | 2008 | JX140654 |
| B-3 | Thailand | 2008 | JN860769 |
| B-3 | USA | 2008 | JQ403062 |
| B-3 | Spain | 2008 | GQ372988 |
| B-3 | USA | 2009 | JX140657 |
| B-3 | France | 2009 | KF716494 |
| B-3 | China | 2009 | KC899011 |
| B-3 | China | 2009 | KC596066 |
| B-3 | Spain | 2010 | KT200356 |
| B-3 | Brazil | 2010 | KJ849784 |
| B-3 | USA | 2010 | JN397365 |
| B-3 | USA | 2010 | KC473825 |
| B-3 | USA | 2010 | KC473829 |
| B-3 | USA | 2010 | KC473826 |
| B-3 | USA | 2010 | KC473830 |
| B-3 | Spain | 2010 | KT200353 |
| B-3 | Brazil | 2010 | KJ849814 |
| B-3 | Brazil | 2010 | KJ849812 |
| B-3 | Brazil | 2010 | KJ849767 |
| B-3 | USA | 2011 | KF384798 |
| B-3 | USA | 2011 | KF384805 |

| | | | |
|---|---|---|---|
| B-3 | USA | 2011 | KC473831 |
| B-3 | USA | 2011 | KF526261 |
| B-3 | USA | 2011 | KF384803 |
| B-3 | USA | 2011 | JN397362 |
| B-3 | China | 2012 | KP109512 |
| B-3 | China | 2012 | KP109511 |
| B-3 | Spain | 2014 | KT276256 |
| B-3 | Spain | 2014 | KT276267 |
| C-1 | Somalia | 1989 | AY713415 |
| C-1 | South Africa | 1990 | JN188292 |
| C-1 | Malawi | 1993 | AY713413 |
| C-1 | India | 1993 | AB023804 |
| C-1 | Botswana | 1996 | AF443075 |
| C-1 | Tanzania | 1997 | AF361875 |
| C-1 | South Africa | 1998 | AX455929 |
| C-1 | South Africa | 1998 | AY043176 |
| C-1 | Tanzania | 1998 | AF286234 |
| C-1 | Botswana | 1998 | AF443076 |
| C-1 | Botswana | 1999 | AF443083 |
| C-1 | Botswana | 1999 | AF443087 |
| C-1 | South Africa | 1999 | AF411967 |
| C-1 | India | 1999 | KP109487 |
| C-1 | Botswana | 2000 | AF443097 |
| C-1 | South Africa | 2000 | AY463233 |
| C-1 | South Africa | 2000 | AY463232 |
| C-1 | South Africa | 2000 | AY463217 |
| C-1 | South Africa | 2000 | AY463231 |
| C-1 | India | 2000 | KP109483 |
| C-1 | Botswana | 2000 | AF443113 |
| C-1 | South Africa | 2000 | AY585265 |
| C-1 | Botswana | 2000 | AF443109 |
| C-1 | Botswana | 2000 | AF443094 |
| C-1 | Botswana | 2000 | AF443090 |
| C-1 | Kenya | 2000 | AF457054 |
| C-1 | South Africa | 2000 | AY463234 |
| C-1 | Tanzania | 2001 | AY253308 |
| C-1 | Tanzania | 2001 | AY253317 |
| C-1 | Tanzania | 2001 | AY253313 |
| C-1 | Brazil | 2002 | JN692434 |
| C-1 | Zambia | 2002 | AB254143 |
| C-1 | Zambia | 2002 | AB254149 |
| C-1 | South Africa | 2003 | DQ351216 |
| C-1 | South Africa | 2003 | DQ396369 |
| C-1 | South Africa | 2003 | DQ351223 |
| C-1 | South Africa | 2003 | DQ396395 |
| C-1 | South Africa | 2003 | DQ396374 |
| C-1 | South Africa | 2003 | AY878061 |
| C-1 | South Africa | 2003 | DQ396377 |
| C-1 | India | 2003 | EF469243 |

| | | | |
|---|---|---|---|
| C-1 | South Africa | 2003 | DQ396380 |
| C-1 | South Africa | 2003 | DQ396386 |
| C-1 | South Africa | 2003 | DQ275656 |
| C-1 | South Africa | 2003 | DQ275653 |
| C-1 | South Africa | 2003 | AY901981 |
| C-1 | South Africa | 2003 | DQ369985 |
| C-1 | South Africa | 2003 | KT183301 |
| C-1 | South Africa | 2003 | DQ396391 |
| C-1 | South Africa | 2003 | DQ275660 |
| C-1 | South Africa | 2003 | DQ093592 |
| C-1 | South Africa | 2003 | DQ396375 |
| C-1 | South Africa | 2004 | DQ056415 |
| C-1 | South Africa | 2004 | DQ093604 |
| C-1 | South Africa | 2004 | DQ011180 |
| C-1 | Brazil | 2004 | AY727523 |
| C-1 | South Africa | 2004 | DQ093590 |
| C-1 | South Africa | 2004 | AY878059 |
| C-1 | Brazil | 2004 | AY727522 |
| C-1 | South Africa | 2004 | DQ164126 |
| C-1 | South Africa | 2004 | AY772699 |
| C-1 | Brazil | 2004 | AY727525 |
| C-1 | South Africa | 2004 | AY901978 |
| C-1 | Brazil | 2004 | AY727524 |
| C-1 | South Africa | 2004 | DQ056406 |
| C-1 | South Africa | 2004 | DQ164119 |
| C-1 | South Africa | 2004 | DQ056416 |
| C-1 | South Africa | 2004 | AY878072 |
| C-1 | South Africa | 2004 | DQ011173 |
| C-1 | South Africa | 2004 | DQ164110 |
| C-1 | South Africa | 2004 | DQ011174 |
| C-1 | South Africa | 2005 | GQ999991 |
| C-1 | South Africa | 2005 | GQ999982 |
| C-1 | South Africa | 2005 | GQ999988 |
| C-1 | South Africa | 2005 | GQ999989 |
| C-1 | South Africa | 2005 | GQ999979 |
| C-1 | South Africa | 2007 | JX974245 |
| C-1 | South Africa | 2007 | KT183089 |
| C-1 | South Africa | 2007 | KT183264 |
| C-1 | Malawi | 2007 | KC156213 |
| C-1 | South Africa | 2007 | KT183208 |
| C-1 | China | 2007 | KF835522 |
| C-1 | South Africa | 2007 | KC156127 |
| C-1 | Sweden | 2007 | KP411832 |
| C-1 | South Africa | 2008 | KT183218 |
| C-1 | South Africa | 2008 | KT183135 |
| C-1 | South Africa | 2008 | KT183250 |
| C-1 | South Africa | 2008 | KT183155 |
| C-1 | Spain | 2008 | EU786681 |
| C-1 | South Africa | 2008 | KT183274 |

| | | | |
|---|---|---|---|
| C-1 | Tanzania | 2008 | KC156220 |
| C-1 | South Africa | 2008 | KT183188 |
| C-1 | China | 2009 | KC898996 |
| C-1 | Zambia | 2009 | KR820367 |
| C-1 | Malawi | 2009 | KP109527 |
| C-1 | Sweden | 2009 | KP411836 |
| C-1 | Zambia | 2009 | KR820326 |
| C-1 | China | 2009 | KC898995 |
| C-1 | Zambia | 2011 | KP109494 |
| C-1 | Zambia | 2011 | KP109495 |
| C-2 | Somalia | 1989 | AY713415 |
| C-2 | Zambia | 1989 | AB485645 |
| C-2 | India | 1993 | AF067158 |
| C-2 | India | 1993 | AB023804 |
| C-2 | India | 1994 | AF067159 |
| C-2 | Botswana | 1996 | AF110972 |
| C-2 | Botswana | 1996 | AF443074 |
| C-2 | Botswana | 1996 | AF110969 |
| C-2 | Botswana | 1996 | AF110967 |
| C-2 | South Africa | 1997 | AY118166 |
| C-2 | South Africa | 1997 | AF286227 |
| C-2 | Tanzania | 1997 | AF361874 |
| C-2 | South Africa | 1998 | AY043173 |
| C-2 | South Africa | 1998 | AX455929 |
| C-2 | South Africa | 1998 | AY162225 |
| C-2 | South Africa | 1998 | AY158535 |
| C-2 | South Africa | 1999 | EU293445 |
| C-2 | India | 1999 | KP109487 |
| C-2 | Botswana | 2000 | AF443105 |
| C-2 | Botswana | 2000 | AF443114 |
| C-2 | Botswana | 2000 | AF443092 |
| C-2 | Botswana | 2000 | AF443094 |
| C-2 | Botswana | 2000 | AF443093 |
| C-2 | South Africa | 2001 | AY463237 |
| C-2 | Tanzania | 2001 | AY253312 |
| C-2 | Denmark | 2001 | EF514713 |
| C-2 | South Africa | 2002 | DQ351235 |
| C-2 | Zambia | 2002 | AB254148 |
| C-2 | Zambia | 2002 | AB254141 |
| C-2 | Tanzania | 2002 | AY734551 |
| C-2 | South Africa | 2003 | DQ351224 |
| C-2 | South Africa | 2003 | DQ093592 |
| C-2 | South Africa | 2003 | DQ093593 |
| C-2 | South Africa | 2003 | AY878068 |
| C-2 | South Africa | 2003 | DQ445635 |
| C-2 | South Africa | 2003 | DQ011175 |
| C-2 | South Africa | 2003 | DQ396390 |
| C-2 | South Africa | 2003 | DQ275656 |
| C-2 | South Africa | 2003 | DQ351228 |

| | | | |
|---|---|---|---|
| C-2 | South Africa | 2003 | DQ275660 |
| C-2 | South Africa | 2003 | DQ369987 |
| C-2 | South Africa | 2003 | DQ396368 |
| C-2 | South Africa | 2003 | DQ275657 |
| C-2 | South Africa | 2003 | AY772695 |
| C-2 | South Africa | 2003 | DQ011176 |
| C-2 | South Africa | 2003 | DQ396375 |
| C-2 | South Africa | 2003 | DQ351216 |
| C-2 | South Africa | 2003 | DQ275646 |
| C-2 | South Africa | 2003 | DQ369990 |
| C-2 | South Africa | 2003 | DQ011169 |
| C-2 | South Africa | 2003 | DQ275645 |
| C-2 | South Africa | 2003 | DQ351226 |
| C-2 | South Africa | 2003 | DQ351219 |
| C-2 | South Africa | 2003 | DQ396369 |
| C-2 | South Africa | 2003 | DQ351230 |
| C-2 | South Africa | 2003 | DQ275655 |
| C-2 | South Africa | 2003 | AY878060 |
| C-2 | South Africa | 2003 | AY878061 |
| C-2 | South Africa | 2003 | DQ396380 |
| C-2 | South Africa | 2003 | DQ369980 |
| C-2 | South Africa | 2003 | DQ275650 |
| C-2 | South Africa | 2003 | AY772700 |
| C-2 | South Africa | 2003 | DQ275664 |
| C-2 | Zambia | 2003 | FJ496195 |
| C-2 | South Africa | 2003 | AY878065 |
| C-2 | South Africa | 2003 | DQ056404 |
| C-2 | South Africa | 2003 | DQ396386 |
| C-2 | South Africa | 2004 | DQ445637 |
| C-2 | South Africa | 2004 | DQ056413 |
| C-2 | South Africa | 2004 | DQ093600 |
| C-2 | South Africa | 2004 | AY878059 |
| C-2 | South Africa | 2004 | DQ011170 |
| C-2 | South Africa | 2004 | DQ093587 |
| C-2 | South Africa | 2004 | DQ275659 |
| C-2 | South Africa | 2004 | DQ164115 |
| C-2 | South Africa | 2004 | DQ351217 |
| C-2 | South Africa | 2004 | DQ093595 |
| C-2 | South Africa | 2004 | DQ011177 |
| C-2 | South Africa | 2004 | AY703911 |
| C-2 | South Africa | 2004 | DQ093605 |
| C-2 | South Africa | 2004 | DQ011166 |
| C-2 | South Africa | 2004 | AY901973 |
| C-2 | South Africa | 2004 | DQ093604 |
| C-2 | South Africa | 2004 | AY703909 |
| C-2 | South Africa | 2004 | AY878062 |
| C-2 | South Africa | 2005 | GQ999990 |
| C-2 | South Africa | 2005 | GQ999981 |
| C-2 | India | 2005 | KF766540 |

| | | | |
|---|---|---|---|
| C-2 | South Africa | 2007 | KT183094 |
| C-2 | South Africa | 2007 | KT183208 |
| C-2 | South Africa | 2007 | KT183089 |
| C-2 | South Africa | 2007 | KT183196 |
| C-2 | Malawi | 2008 | KF527172 |
| C-2 | South Africa | 2008 | KT183201 |
| C-2 | Malawi | 2008 | KC156216 |
| C-2 | South Africa | 2008 | KT183172 |
| C-2 | South Africa | 2009 | JX140668 |
| C-2 | China | 2009 | KC898995 |
| C-2 | India | 2009 | KC156210 |
| C-2 | South Africa | 2012 | KP109516 |
| C-3 | Somalia | 1989 | AY713415 |
| C-3 | South Africa | 1990 | JN188292 |
| C-3 | India | 1993 | AF067157 |
| C-3 | India | 1994 | AF067159 |
| C-3 | India | 1994 | AF286223 |
| C-3 | India | 1995 | AF067155 |
| C-3 | Botswana | 1996 | AF443075 |
| C-3 | Tanzania | 1997 | AF361874 |
| C-3 | South Africa | 1997 | AF286227 |
| C-3 | Tanzania | 1998 | AF286235 |
| C-3 | India | 1998 | AF286232 |
| C-3 | Botswana | 1998 | AF443076 |
| C-3 | South Africa | 1998 | AY162225 |
| C-3 | South Africa | 1999 | EU293444 |
| C-3 | South Africa | 1999 | EU293445 |
| C-3 | South Africa | 1999 | EU293448 |
| C-3 | Botswana | 1999 | AF443086 |
| C-3 | South Africa | 2000 | AY463225 |
| C-3 | Botswana | 2000 | AF443103 |
| C-3 | South Africa | 2000 | AY463234 |
| C-3 | Botswana | 2000 | AF443102 |
| C-3 | Botswana | 2000 | AF443115 |
| C-3 | Botswana | 2000 | AF443105 |
| C-3 | South Africa | 2000 | AY585264 |
| C-3 | Kenya | 2000 | AF457054 |
| C-3 | Botswana | 2000 | AF443097 |
| C-3 | Tanzania | 2001 | AY253317 |
| C-3 | Tanzania | 2001 | AY253303 |
| C-3 | South Africa | 2002 | DQ351220 |
| C-3 | Zambia | 2002 | AB254142 |
| C-3 | South Africa | 2003 | DQ011175 |
| C-3 | South Africa | 2003 | AY772700 |
| C-3 | South Africa | 2003 | DQ093593 |
| C-3 | South Africa | 2003 | DQ369980 |
| C-3 | South Africa | 2003 | DQ369981 |
| C-3 | Georgia | 2003 | DQ207941 |
| C-3 | South Africa | 2003 | DQ351237 |

| | | | |
|---|---|---|---|
| C-3 | South Africa | 2003 | DQ093591 |
| C-3 | South Africa | 2003 | DQ351224 |
| C-3 | South Africa | 2003 | AY901966 |
| C-3 | South Africa | 2003 | DQ396367 |
| C-3 | South Africa | 2003 | DQ164113 |
| C-3 | South Africa | 2003 | DQ093597 |
| C-3 | South Africa | 2003 | DQ396368 |
| C-3 | South Africa | 2003 | DQ396376 |
| C-3 | South Africa | 2003 | DQ369986 |
| C-3 | Zambia | 2003 | FJ496185 |
| C-3 | South Africa | 2003 | DQ164104 |
| C-3 | South Africa | 2003 | DQ056404 |
| C-3 | South Africa | 2004 | DQ164126 |
| C-3 | South Africa | 2004 | DQ093590 |
| C-3 | South Africa | 2004 | DQ056413 |
| C-3 | Brazil | 2004 | AY727523 |
| C-3 | South Africa | 2004 | DQ011178 |
| C-3 | South Africa | 2004 | AY901979 |
| C-3 | South Africa | 2004 | DQ056409 |
| C-3 | South Africa | 2004 | DQ445637 |
| C-3 | South Africa | 2004 | DQ164110 |
| C-3 | South Africa | 2004 | DQ164117 |
| C-3 | South Africa | 2004 | DQ093600 |
| C-3 | Brazil | 2004 | AY727524 |
| C-3 | South Africa | 2004 | AY703911 |
| C-3 | Brazil | 2004 | AY727525 |
| C-3 | South Africa | 2004 | DQ164114 |
| C-3 | South Africa | 2004 | DQ351217 |
| C-3 | South Africa | 2004 | DQ164129 |
| C-3 | South Africa | 2004 | DQ093604 |
| C-3 | South Africa | 2004 | AY901974 |
| C-3 | South Africa | 2004 | DQ351232 |
| C-3 | South Africa | 2004 | DQ011172 |
| C-3 | South Africa | 2004 | AY901977 |
| C-3 | South Africa | 2004 | DQ056405 |
| C-3 | Brazil | 2004 | AY727522 |
| C-3 | South Africa | 2005 | GQ999987 |
| C-3 | South Africa | 2005 | GQ999984 |
| C-3 | South Africa | 2005 | GQ999976 |
| C-3 | South Africa | 2005 | GQ999989 |
| C-3 | India | 2005 | KF766540 |
| C-3 | South Africa | 2005 | GQ999982 |
| C-3 | South Africa | 2005 | GQ999991 |
| C-3 | Spain | 2006 | EU786673 |
| C-3 | South Africa | 2007 | KT183264 |
| C-3 | Malawi | 2007 | KC156122 |
| C-3 | South Africa | 2007 | KC156127 |
| C-3 | South Africa | 2007 | JX974245 |
| C-3 | Malawi | 2007 | KC156213 |

| | | | |
|---|---|---|---|
| C-3 | South Africa | 2008 | KT183258 |
| C-3 | Malawi | 2008 | KC156218 |
| C-3 | South Africa | 2008 | KC156221 |
| C-3 | South Africa | 2008 | JX140666 |
| C-3 | South Africa | 2008 | KT183201 |
| C-3 | Malawi | 2008 | KF527172 |
| C-3 | Spain | 2008 | EU786681 |
| C-3 | South Africa | 2008 | KT183336 |
| C-3 | South Africa | 2008 | KT183128 |
| C-3 | South Africa | 2008 | KT183078 |
| C-3 | South Africa | 2009 | JX140668 |
| C-3 | Malawi | 2009 | KP109523 |
| C-3 | South Africa | 2009 | KT183289 |
| C-3 | Zambia | 2011 | KP109495 |
| D | DR Congo | 1983 | A07108 |
| D | DR Congo | 1984 | U88822 |
| D | Senegal | 1990 | AB485648 |
| D | Uganda | 1991 | AB485650 |
| D | Uganda | 1993 | AY713418 |
| D | Uganda | 1994 | U88824 |
| D | Uganda | 1998 | AF484505 |
| D | Uganda | 1998 | AF484516 |
| D | Uganda | 1998 | AF484502 |
| D | Uganda | 1998 | AF484511 |
| D | Uganda | 1998 | AF484506 |
| D | Uganda | 1998 | AF484513 |
| D | Uganda | 1998 | AF484514 |
| D | Uganda | 1998 | AF484504 |
| D | Uganda | 1999 | AF484499 |
| D | Uganda | 1999 | AF484518 |
| D | Uganda | 1999 | AF484489 |
| D | Uganda | 1999 | AF484481 |
| D | Uganda | 1999 | AF484490 |
| D | Uganda | 1999 | AF484487 |
| D | Uganda | 1999 | AF484498 |
| D | Uganda | 1999 | AF484519 |
| D | Uganda | 1999 | AF484494 |
| D | Uganda | 1999 | AF484483 |
| D | Uganda | 1999 | AF484497 |
| D | Uganda | 1999 | AY304496 |
| D | Uganda | 1999 | AF484485 |
| D | Uganda | 1999 | AF484480 |
| D | Uganda | 1999 | AF484477 |
| D | Uganda | 1999 | AF484495 |
| D | Uganda | 1999 | AF484486 |
| D | Chad | 1999 | AJ488926 |
| D | Chad | 1999 | AJ488927 |
| D | Kenya | 2001 | AF457090 |
| D | Yemen | 2001 | AY795903 |

| | | | |
|---|---|---|---|
| D | Yemen | 2002 | AY795907 |
| D | Uganda | 2005 | JX236668 |
| D | Uganda | 2007 | JX236670 |
| D | Uganda | 2007 | JX236673 |
| D | Uganda | 2008 | JX236672 |
| D | Uganda | 2010 | KF716479 |
| D | Brazil | 2010 | KJ787684 |
| D | Cameroon | 2010 | JX140670 |
| D | Kenya | 2011 | KF716476 |
| D | Uganda | 2011 | KF716480 |
| G | Nigeria | 1992 | U88826 |
| G | Kenya | 1993 | AB485662 |
| G | Sweden | 1993 | AF061642 |
| G | Cameroon | 1996 | AY772535 |
| G | Belgium | 1996 | AF084936 |
| G | Cuba | 1999 | AY586548 |
| G | Cuba | 1999 | AY586549 |
| G | Spain | 2000 | AF423760 |
| G | Nigeria | 2001 | DQ168575 |
| G | Nigeria | 2001 | DQ168579 |
| G | Nigeria | 2001 | DQ168573 |
| G | Nigeria | 2001 | DQ168576 |
| G | Cameroon | 2001 | FJ389367 |
| G | Ghana | 2003 | AB231893 |
| G | Cameroon | 2004 | FJ389363 |
| G | Spain | 2005 | EU786670 |
| G | Spain | 2005 | FJ670520 |
| G | Cameroon | 2006 | KP718915 |
| G | Cameroon | 2007 | KP718923 |
| G | Nigeria | 2008 | JN248582 |
| G | China | 2008 | JN106043 |
| G | Cameroon | 2008 | KP718925 |
| G | Spain | 2008 | FJ670530 |
| G | Nigeria | 2009 | JN248591 |
| G | Nigeria | 2009 | JN248586 |
| G | Nigeria | 2009 | JN248593 |
| G | Kenya | 2009 | KF716477 |
| G | Spain | 2009 | GU362882 |
| G | Nigeria | 2009 | JN248584 |
| G | Cameroon | 2010 | KP109502 |
| G | Cameroon | 2010 | JX140676 |
| G | Spain | 2014 | KT276261 |
| 01_AE-1 | Central African Rep | 1990 | AF197341 |
| 01_AE-1 | Central African Rep | 1990 | U51188 |
| 01_AE-1 | Central African Rep | 1990 | AF197340 |
| 01_AE-1 | Thailand | 1993 | AB220944 |
| 01_AE-1 | Thailand | 1993 | U51189 |
| 01_AE-1 | Indonesia | 1993 | AB485652 |
| 01_AE-1 | Thailand | 1995 | AB032741 |

| | | | |
|---|---|---|---|
| 01_AE-1 | Thailand | 1997 | AY713419 |
| 01_AE-1 | Viet Nam | 1997 | FJ185239 |
| 01_AE-1 | Viet Nam | 1997 | FJ185249 |
| 01_AE-1 | Viet Nam | 1997 | FJ185258 |
| 01_AE-1 | Viet Nam | 1997 | FJ185241 |
| 01_AE-1 | Viet Nam | 1997 | FJ185260 |
| 01_AE-1 | Thailand | 1998 | AY713422 |
| 01_AE-1 | Viet Nam | 1998 | FJ185236 |
| 01_AE-1 | Thailand | 2000 | DQ789392 |
| 01_AE-1 | China | 2002 | JX112866 |
| 01_AE-1 | China | 2002 | JX112863 |
| 01_AE-1 | China | 2002 | JX112862 |
| 01_AE-1 | China | 2002 | JX112861 |
| 01_AE-1 | Thailand | 2004 | JN248328 |
| 01_AE-1 | Hong Kong | 2004 | DQ234790 |
| 01_AE-1 | China | 2005 | GU564223 |
| 01_AE-1 | China | 2005 | EF036535 |
| 01_AE-1 | China | 2005 | EF036530 |
| 01_AE-1 | Thailand | 2005 | JN248342 |
| 01_AE-1 | Thailand | 2005 | JX447618 |
| 01_AE-1 | China | 2005 | GQ845125 |
| 01_AE-1 | China | 2005 | DQ859178 |
| 01_AE-1 | China | 2005 | GU564221 |
| 01_AE-1 | Thailand | 2005 | JX447077 |
| 01_AE-1 | China | 2005 | EF036527 |
| 01_AE-1 | Thailand | 2005 | JX447268 |
| 01_AE-1 | Thailand | 2005 | JX447589 |
| 01_AE-1 | Thailand | 2005 | JN248341 |
| 01_AE-1 | China | 2005 | GU564229 |
| 01_AE-1 | Thailand | 2006 | JX447315 |
| 01_AE-1 | Thailand | 2006 | JX447884 |
| 01_AE-1 | Thailand | 2006 | JX446853 |
| 01_AE-1 | Thailand | 2006 | JX447132 |
| 01_AE-1 | Thailand | 2006 | JX447638 |
| 01_AE-1 | Thailand | 2006 | JX447403 |
| 01_AE-1 | Thailand | 2006 | JX447678 |
| 01_AE-1 | China | 2006 | JX112858 |
| 01_AE-1 | China | 2007 | JX112840 |
| 01_AE-1 | Thailand | 2007 | JX447054 |
| 01_AE-1 | China | 2007 | KF835499 |
| 01_AE-1 | China | 2007 | JX112836 |
| 01_AE-1 | Thailand | 2007 | JN860766 |
| 01_AE-1 | China | 2007 | JX112813 |
| 01_AE-1 | China | 2007 | JX112855 |
| 01_AE-1 | Thailand | 2007 | JN860767 |
| 01_AE-1 | China | 2007 | JX112830 |
| 01_AE-1 | China | 2007 | JX112822 |
| 01_AE-1 | China | 2007 | JX112833 |
| 01_AE-1 | China | 2007 | KF835519 |

| | | | |
|---|---|---|---|
| 01_AE-1 | China | 2007 | JX112826 |
| 01_AE-1 | Thailand | 2007 | JX446877 |
| 01_AE-1 | China | 2007 | JX112851 |
| 01_AE-1 | China | 2007 | JX112828 |
| 01_AE-1 | China | 2007 | JX112834 |
| 01_AE-1 | China | 2007 | JX112821 |
| 01_AE-1 | China | 2007 | JX112839 |
| 01_AE-1 | Thailand | 2007 | JX447028 |
| 01_AE-1 | China | 2007 | JX112820 |
| 01_AE-1 | Thailand | 2007 | JN860761 |
| 01_AE-1 | China | 2007 | JX112844 |
| 01_AE-1 | China | 2007 | KF835527 |
| 01_AE-1 | China | 2007 | JX112859 |
| 01_AE-1 | China | 2007 | JX112854 |
| 01_AE-1 | Thailand | 2007 | JN860763 |
| 01_AE-1 | China | 2007 | JX112856 |
| 01_AE-1 | China | 2007 | JX112831 |
| 01_AE-1 | China | 2007 | KF835502 |
| 01_AE-1 | China | 2007 | KF835513 |
| 01_AE-1 | China | 2007 | JX112811 |
| 01_AE-1 | Thailand | 2007 | JN860764 |
| 01_AE-1 | Thailand | 2008 | JX446755 |
| 01_AE-1 | Thailand | 2008 | JX447305 |
| 01_AE-1 | Thailand | 2008 | JX446848 |
| 01_AE-1 | Thailand | 2008 | JX447122 |
| 01_AE-1 | Thailand | 2008 | JX447359 |
| 01_AE-1 | China | 2009 | HQ215555 |
| 01_AE-1 | China | 2009 | HQ215553 |
| 01_AE-1 | Thailand | 2009 | JX448057 |
| 01_AE-1 | China | 2010 | JX112848 |
| 01_AE-1 | China | 2010 | JX112800 |
| 01_AE-1 | China | 2010 | KC870029 |
| 01_AE-1 | China | 2010 | KP109507 |
| 01_AE-1 | China | 2010 | JX112849 |
| 01_AE-1 | China | 2010 | JX112846 |
| 01_AE-1 | China | 2010 | JX112801 |
| 01_AE-1 | China | 2010 | JX112805 |
| 01_AE-1 | China | 2010 | JX112798 |
| 01_AE-1 | China | 2010 | JX112799 |
| 01_AE-1 | China | 2010 | JX112847 |
| 01_AE-1 | China | 2010 | JX112802 |
| 01_AE-1 | China | 2010 | KP109506 |
| 01_AE-1 | Sweden | 2011 | KP411842 |
| 01_AE-1 | China | 2012 | KP109508 |
| 01_AE-2 | Central African Rep | 1990 | AF197341 |
| 01_AE-2 | Thailand | 1993 | AF164485 |
| 01_AE-2 | Thailand | 1993 | U51189 |
| 01_AE-2 | Indonesia | 1993 | AB485652 |
| 01_AE-2 | Japan | 1993 | AB052995 |

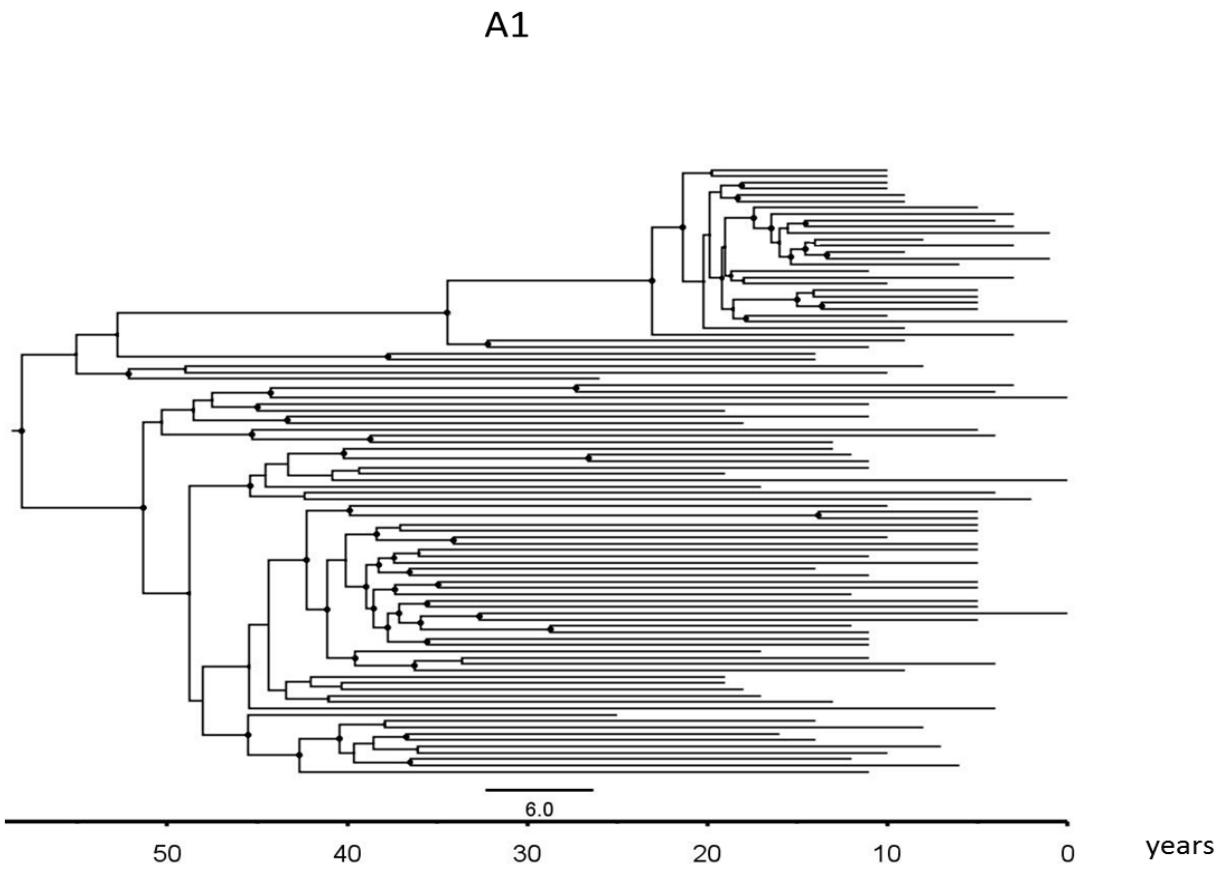
| | | | |
|---|---|---|---|
| 01_AE-2 | Thailand | 1995 | AB032741 |
| 01_AE-2 | Thailand | 1996 | AY713424 |
| 01_AE-2 | Viet Nam | 1997 | FJ185260 |
| 01_AE-2 | Viet Nam | 1997 | FJ185242 |
| 01_AE-2 | Viet Nam | 1997 | FJ185239 |
| 01_AE-2 | Viet Nam | 1997 | FJ185241 |
| 01_AE-2 | Viet Nam | 1997 | FJ185248 |
| 01_AE-2 | Viet Nam | 1997 | FJ185251 |
| 01_AE-2 | Viet Nam | 1997 | FJ185250 |
| 01_AE-2 | Viet Nam | 1997 | FJ185259 |
| 01_AE-2 | Thailand | 1997 | AY125894 |
| 01_AE-2 | Thailand | 1999 | AY713423 |
| 01_AE-2 | Thailand | 2000 | DQ789392 |
| 01_AE-2 | China | 2002 | JX112865 |
| 01_AE-2 | China | 2002 | JX112862 |
| 01_AE-2 | Thailand | 2004 | JN248327 |
| 01_AE-2 | Thailand | 2004 | JN248334 |
| 01_AE-2 | Thailand | 2004 | JN248328 |
| 01_AE-2 | Hong Kong | 2004 | DQ234790 |
| 01_AE-2 | Thailand | 2004 | JN248336 |
| 01_AE-2 | Thailand | 2004 | JN248330 |
| 01_AE-2 | Thailand | 2005 | JX447589 |
| 01_AE-2 | China | 2005 | EF036527 |
| 01_AE-2 | Thailand | 2005 | JX447268 |
| 01_AE-2 | China | 2005 | GU564227 |
| 01_AE-2 | China | 2005 | GU564228 |
| 01_AE-2 | Thailand | 2005 | JX447936 |
| 01_AE-2 | China | 2005 | EF036534 |
| 01_AE-2 | Thailand | 2005 | JN248339 |
| 01_AE-2 | China | 2005 | GU564223 |
| 01_AE-2 | China | 2005 | GU564229 |
| 01_AE-2 | China | 2005 | EF036530 |
| 01_AE-2 | Thailand | 2005 | JN248355 |
| 01_AE-2 | Thailand | 2005 | JX447618 |
| 01_AE-2 | Thailand | 2006 | JX447638 |
| 01_AE-2 | Thailand | 2006 | JX447678 |
| 01_AE-2 | Thailand | 2006 | JX446899 |
| 01_AE-2 | Thailand | 2006 | JX447403 |
| 01_AE-2 | Thailand | 2006 | JX447315 |
| 01_AE-2 | China | 2006 | GU564230 |
| 01_AE-2 | Thailand | 2006 | JX446736 |
| 01_AE-2 | China | 2006 | JX112857 |
| 01_AE-2 | Thailand | 2006 | JX447884 |
| 01_AE-2 | China | 2007 | JX112821 |
| 01_AE-2 | China | 2007 | JX112845 |
| 01_AE-2 | China | 2007 | JX112826 |
| 01_AE-2 | China | 2007 | JX112820 |
| 01_AE-2 | Thailand | 2007 | JX446987 |
| 01_AE-2 | China | 2007 | KF835503 |

| | | | |
|---|---|---|---|
| 01_AE-2 | China | 2007 | KF835516 |
| 01_AE-2 | China | 2007 | JX112840 |
| 01_AE-2 | Thailand | 2007 | JN860761 |
| 01_AE-2 | China | 2007 | JX112825 |
| 01_AE-2 | China | 2007 | JX112856 |
| 01_AE-2 | China | 2007 | JX112841 |
| 01_AE-2 | China | 2007 | JX112810 |
| 01_AE-2 | China | 2007 | JX112855 |
| 01_AE-2 | China | 2007 | JX112813 |
| 01_AE-2 | China | 2007 | JX112839 |
| 01_AE-2 | China | 2007 | JX112828 |
| 01_AE-2 | China | 2007 | JX112827 |
| 01_AE-2 | China | 2007 | JX112833 |
| 01_AE-2 | Thailand | 2007 | JX447283 |
| 01_AE-2 | Thailand | 2007 | JN860766 |
| 01_AE-2 | China | 2007 | KF835499 |
| 01_AE-2 | China | 2007 | KF835542 |
| 01_AE-2 | China | 2007 | JX112814 |
| 01_AE-2 | China | 2007 | JX112837 |
| 01_AE-2 | Thailand | 2007 | JN860767 |
| 01_AE-2 | China | 2007 | JX112818 |
| 01_AE-2 | China | 2007 | JX112817 |
| 01_AE-2 | Thailand | 2007 | JX447054 |
| 01_AE-2 | China | 2007 | JX112823 |
| 01_AE-2 | Thailand | 2007 | JX447499 |
| 01_AE-2 | Thailand | 2007 | JN860764 |
| 01_AE-2 | Thailand | 2008 | JX447072 |
| 01_AE-2 | Thailand | 2008 | JX447454 |
| 01_AE-2 | Thailand | 2008 | JX447300 |
| 01_AE-2 | Thailand | 2008 | JX447122 |
| 01_AE-2 | China | 2009 | HQ215555 |
| 01_AE-2 | China | 2010 | JX112799 |
| 01_AE-2 | China | 2010 | JX112800 |
| 01_AE-2 | China | 2010 | JX112802 |
| 01_AE-2 | China | 2010 | KC870029 |
| 01_AE-2 | China | 2010 | JX112805 |
| 01_AE-2 | China | 2010 | JX112803 |
| 01_AE-2 | China | 2010 | KP109506 |
| 01_AE-2 | China | 2010 | JX112801 |
| 01_AE-2 | China | 2010 | JX112804 |
| 01_AE-2 | China | 2010 | JX112848 |
| 01_AE-2 | China | 2010 | JX112807 |
| 01_AE-2 | Sweden | 2011 | KP411842 |
| 01_AE-2 | Cameroon | 2011 | KP718930 |
| 01_AE-2 | China | 2011 | KC596065 |
| 01_AE-2 | China | 2012 | KP109508 |
| 01_AE-3 | Thailand | 1990 | AF259954 |
| 01_AE-3 | Central African Rep | 1990 | AF197340 |
| 01_AE-3 | Central African Rep | 1990 | U51188 |

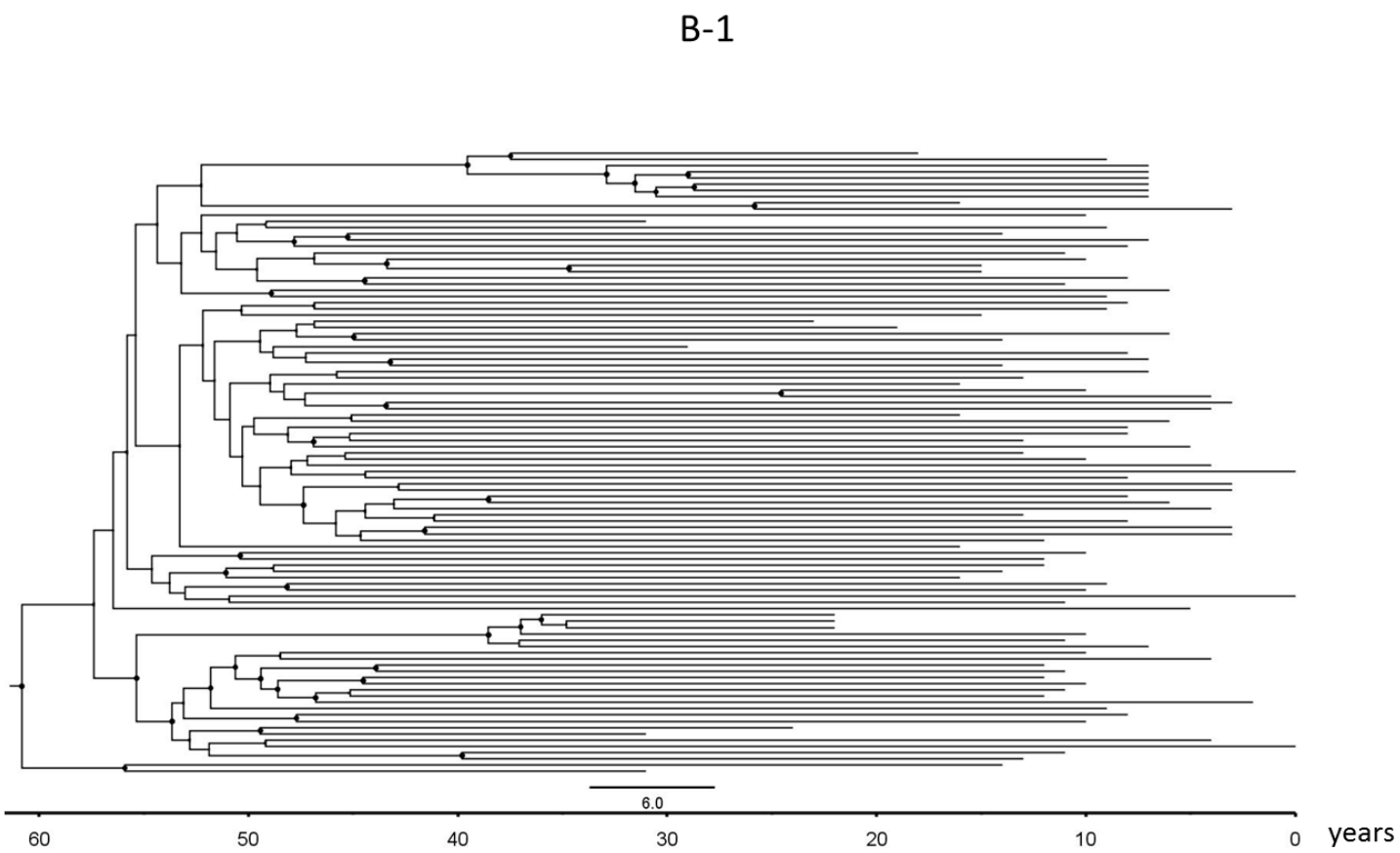
| | | | |
|---|---|---|---|
| 01_AE-3 | Japan | 1993 | AB070352 |
| 01_AE-3 | Thailand | 1993 | AB220946 |
| 01_AE-3 | Thailand | 1993 | U51189 |
| 01_AE-3 | Thailand | 1995 | AB032741 |
| 01_AE-3 | Viet Nam | 1997 | FJ185258 |
| 01_AE-3 | Viet Nam | 1997 | FJ185247 |
| 01_AE-3 | Viet Nam | 1997 | FJ185237 |
| 01_AE-3 | Viet Nam | 1997 | FJ185260 |
| 01_AE-3 | Viet Nam | 1997 | FJ185248 |
| 01_AE-3 | Viet Nam | 1997 | FJ185242 |
| 01_AE-3 | Viet Nam | 1997 | FJ185241 |
| 01_AE-3 | Viet Nam | 1997 | FJ185250 |
| 01_AE-3 | Thailand | 1998 | AY713422 |
| 01_AE-3 | Thailand | 1999 | AY713423 |
| 01_AE-3 | Thailand | 2000 | DQ789392 |
| 01_AE-3 | China | 2002 | JX112861 |
| 01_AE-3 | China | 2002 | JX112865 |
| 01_AE-3 | Thailand | 2004 | JN248334 |
| 01_AE-3 | Hong Kong | 2004 | DQ234790 |
| 01_AE-3 | Thailand | 2004 | JN248336 |
| 01_AE-3 | Thailand | 2004 | DQ314732 |
| 01_AE-3 | China | 2005 | GU564221 |
| 01_AE-3 | China | 2005 | GU564223 |
| 01_AE-3 | China | 2005 | EF036530 |
| 01_AE-3 | Thailand | 2005 | JX447936 |
| 01_AE-3 | Thailand | 2005 | JN248356 |
| 01_AE-3 | China | 2005 | GU564222 |
| 01_AE-3 | Thailand | 2005 | JN248341 |
| 01_AE-3 | China | 2005 | GU564227 |
| 01_AE-3 | Thailand | 2005 | JN248339 |
| 01_AE-3 | China | 2005 | EF036528 |
| 01_AE-3 | Thailand | 2005 | JX447618 |
| 01_AE-3 | China | 2005 | GU564224 |
| 01_AE-3 | China | 2005 | GQ845124 |
| 01_AE-3 | China | 2005 | DQ859178 |
| 01_AE-3 | Thailand | 2005 | JX447268 |
| 01_AE-3 | Thailand | 2005 | JX448019 |
| 01_AE-3 | Thailand | 2006 | JX447403 |
| 01_AE-3 | Thailand | 2006 | JX447638 |
| 01_AE-3 | China | 2006 | GU564230 |
| 01_AE-3 | China | 2006 | JX112857 |
| 01_AE-3 | Thailand | 2006 | JX447465 |
| 01_AE-3 | Thailand | 2006 | JX446853 |
| 01_AE-3 | China | 2007 | JX112828 |
| 01_AE-3 | China | 2007 | KF835543 |
| 01_AE-3 | China | 2007 | JX112839 |
| 01_AE-3 | China | 2007 | KF835503 |
| 01_AE-3 | China | 2007 | JX112837 |
| 01_AE-3 | China | 2007 | JX112831 |

| | | | |
|---|---|---|---|
| 01_AE-3 | China | 2007 | JX112850 |
| 01_AE-3 | China | 2007 | JX112811 |
| 01_AE-3 | China | 2007 | KF835542 |
| 01_AE-3 | Thailand | 2007 | JX447283 |
| 01_AE-3 | China | 2007 | KF835502 |
| 01_AE-3 | China | 2007 | JX112859 |
| 01_AE-3 | Thailand | 2007 | JN860761 |
| 01_AE-3 | China | 2007 | KF835527 |
| 01_AE-3 | China | 2007 | JX112810 |
| 01_AE-3 | Thailand | 2007 | JX447054 |
| 01_AE-3 | China | 2007 | JX112823 |
| 01_AE-3 | China | 2007 | KF835499 |
| 01_AE-3 | China | 2007 | JX112845 |
| 01_AE-3 | China | 2007 | JX112853 |
| 01_AE-3 | Thailand | 2007 | JN860764 |
| 01_AE-3 | China | 2007 | KF835519 |
| 01_AE-3 | China | 2007 | JX112829 |
| 01_AE-3 | China | 2007 | JX112841 |
| 01_AE-3 | China | 2007 | JX112836 |
| 01_AE-3 | Afghanistan | 2007 | GQ477441 |
| 01_AE-3 | China | 2007 | JX112820 |
| 01_AE-3 | Thailand | 2008 | JX447072 |
| 01_AE-3 | Thailand | 2008 | JX447454 |
| 01_AE-3 | Thailand | 2008 | JN860768 |
| 01_AE-3 | Thailand | 2008 | JX446755 |
| 01_AE-3 | Thailand | 2008 | JX447305 |
| 01_AE-3 | Thailand | 2008 | JX447359 |
| 01_AE-3 | China | 2009 | KP109505 |
| 01_AE-3 | China | 2009 | JX112870 |
| 01_AE-3 | China | 2009 | JX112868 |
| 01_AE-3 | Thailand | 2009 | JX448057 |
| 01_AE-3 | China | 2009 | JX112867 |
| 01_AE-3 | Thailand | 2009 | JX447726 |
| 01_AE-3 | China | 2009 | HQ215555 |
| 01_AE-3 | China | 2010 | JX112805 |
| 01_AE-3 | China | 2010 | JX112801 |
| 01_AE-3 | Thailand | 2010 | KP109513 |
| 01_AE-3 | China | 2010 | JX112807 |
| 01_AE-3 | China | 2010 | KP109506 |
| 01_AE-3 | China | 2010 | JX112798 |
| 01_AE-3 | China | 2010 | JX112806 |
| 01_AE-3 | China | 2010 | JX112800 |
| 01_AE-3 | China | 2010 | JX112848 |
| 01_AE-3 | China | 2010 | KC870041 |
| 01_AE-3 | China | 2010 | JX112804 |
| 01_AE-3 | Sweden | 2011 | KP411842 |
| 01_AE-3 | China | 2011 | KC596065 |
| 01_AE-3 | Japan | 2011 | KF859741 |
| 02_AG | France | 1991 | AF063224 |

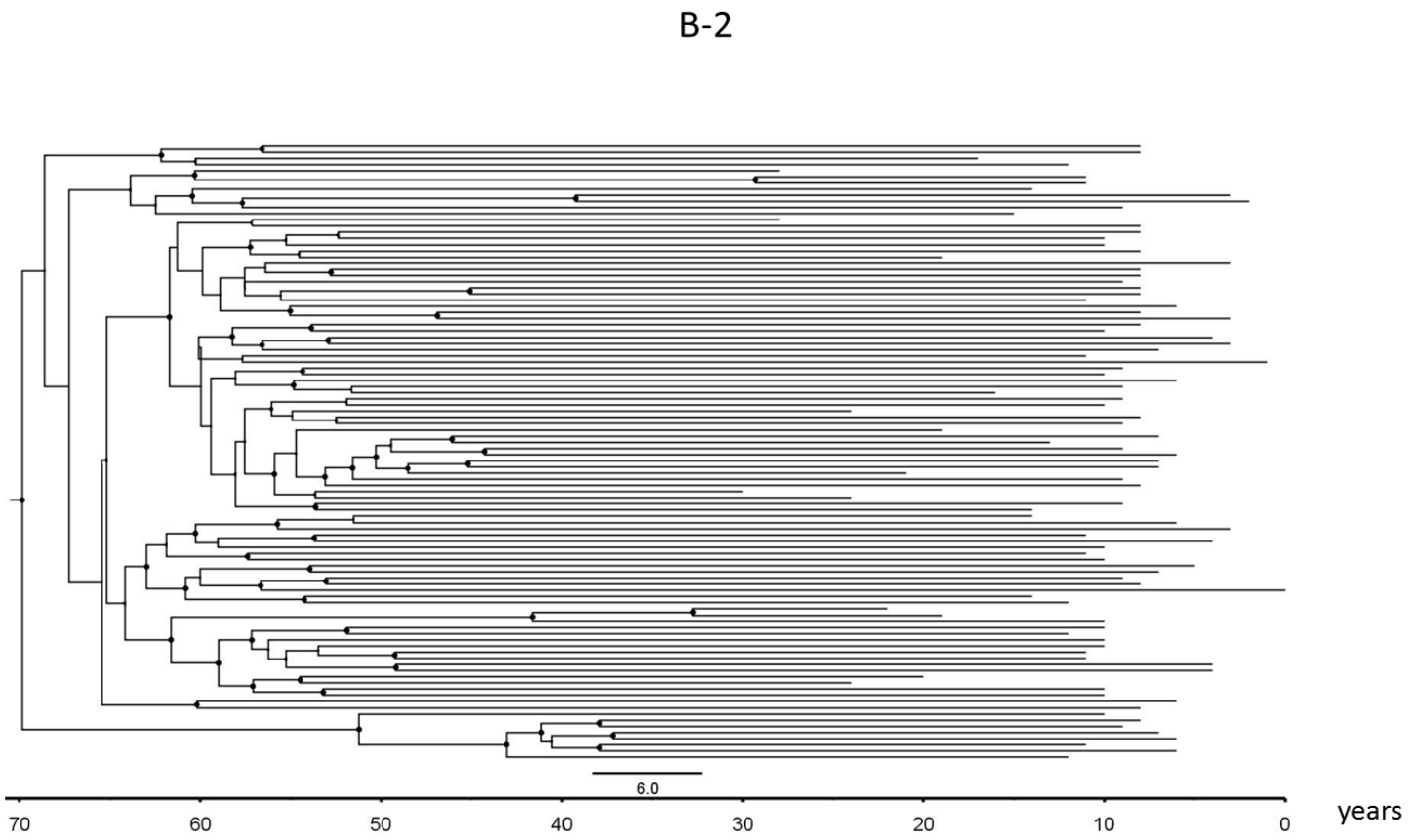
| | | | |
|---|---|---|---|
| 02_AG | Sweden | 1994 | AF107770 |
| 02_AG | Ghana | 1997 | AB049811 |
| 02_AG | Cameroon | 1997 | AF377955 |
| 02_AG | Cameroon | 1997 | AF377954 |
| 02_AG | Senegal | 1998 | AJ251056 |
| 02_AG | Cameroon | 1999 | AY271690 |
| 02_AG | Cameroon | 2001 | GU201499 |
| 02_AG | Nigeria | 2001 | DQ168578 |
| 02_AG | Cameroon | 2001 | GU201498 |
| 02_AG | Cameroon | 2001 | GU201495 |
| 02_AG | Cameroon | 2002 | GU201514 |
| 02_AG | Cameroon | 2002 | GU201512 |
| 02_AG | Cameroon | 2002 | GU201513 |
| 02_AG | Ghana | 2003 | AB231898 |
| 02_AG | Ghana | 2003 | AB286862 |
| 02_AG | Ghana | 2003 | AB231896 |
| 02_AG | Guinea-Bissau | 2004 | FJ694791 |
| 02_AG | Spain | 2006 | EU786671 |
| 02_AG | Nigeria | 2006 | JN248580 |
| 02_AG | USA | 2006 | JF320297 |
| 02_AG | Spain | 2006 | EU884501 |
| 02_AG | South Korea | 2007 | JQ316136 |
| 02_AG | Cameroon | 2008 | JX140647 |
| 02_AG | Cameroon | 2008 | JX140646 |
| 02_AG | Nigeria | 2009 | JN248589 |
| 02_AG | Nigeria | 2009 | JN248585 |
| 02_AG | Nigeria | 2009 | JN248590 |
| 02_AG | Nigeria | 2009 | JN248592 |
| 02_AG | South Korea | 2012 | KF561435 |



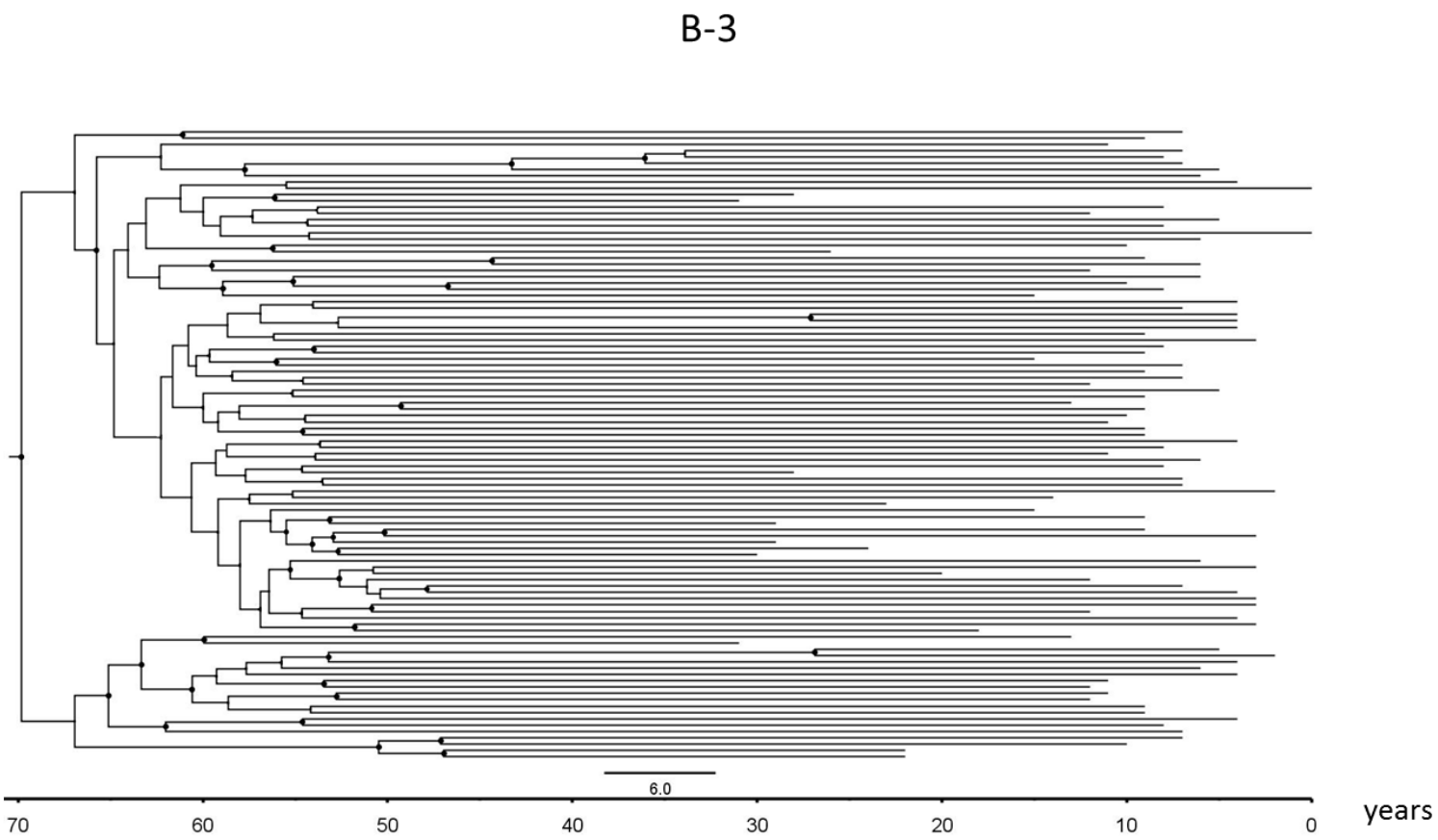

**Supplementary Figure S1**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



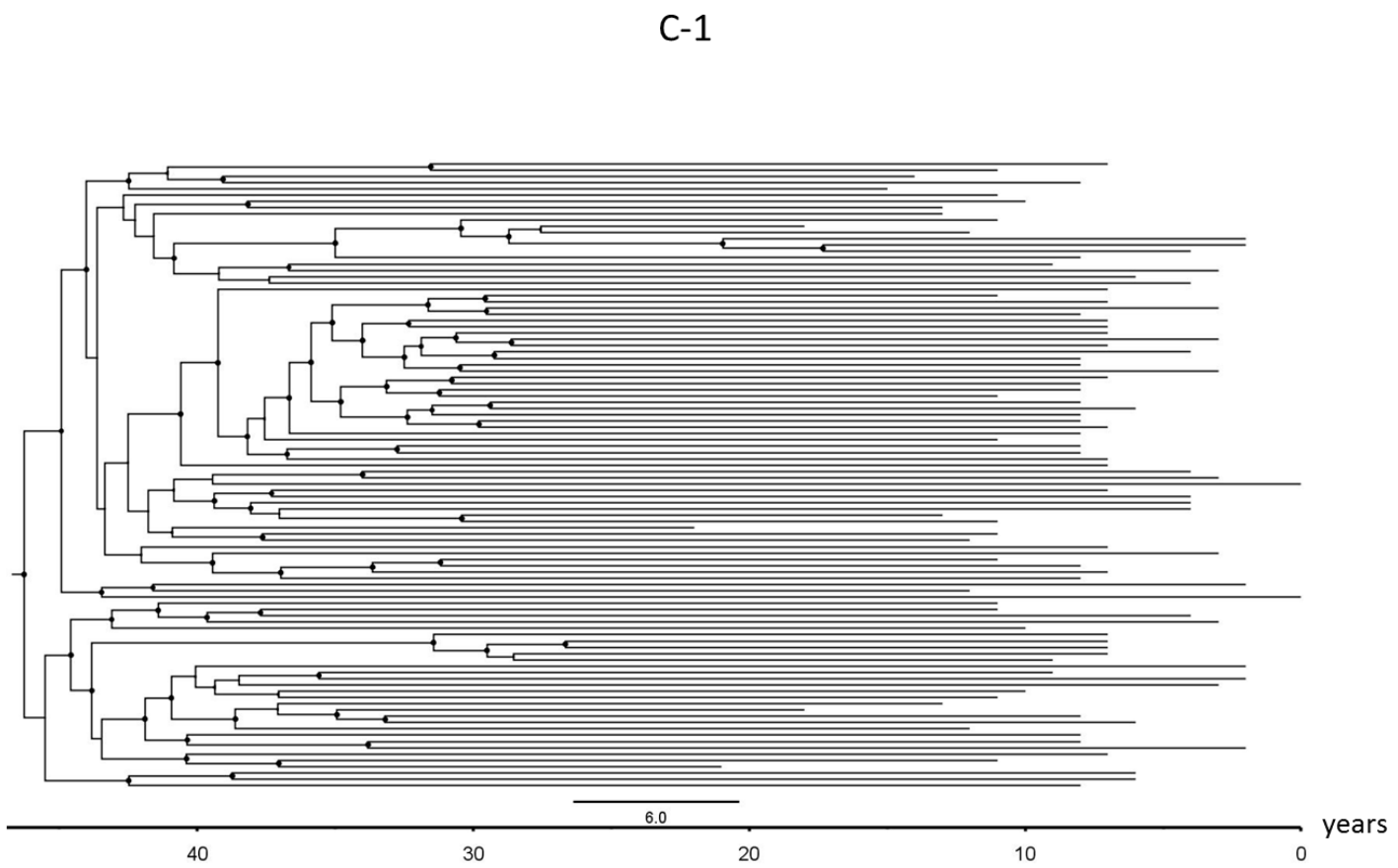

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



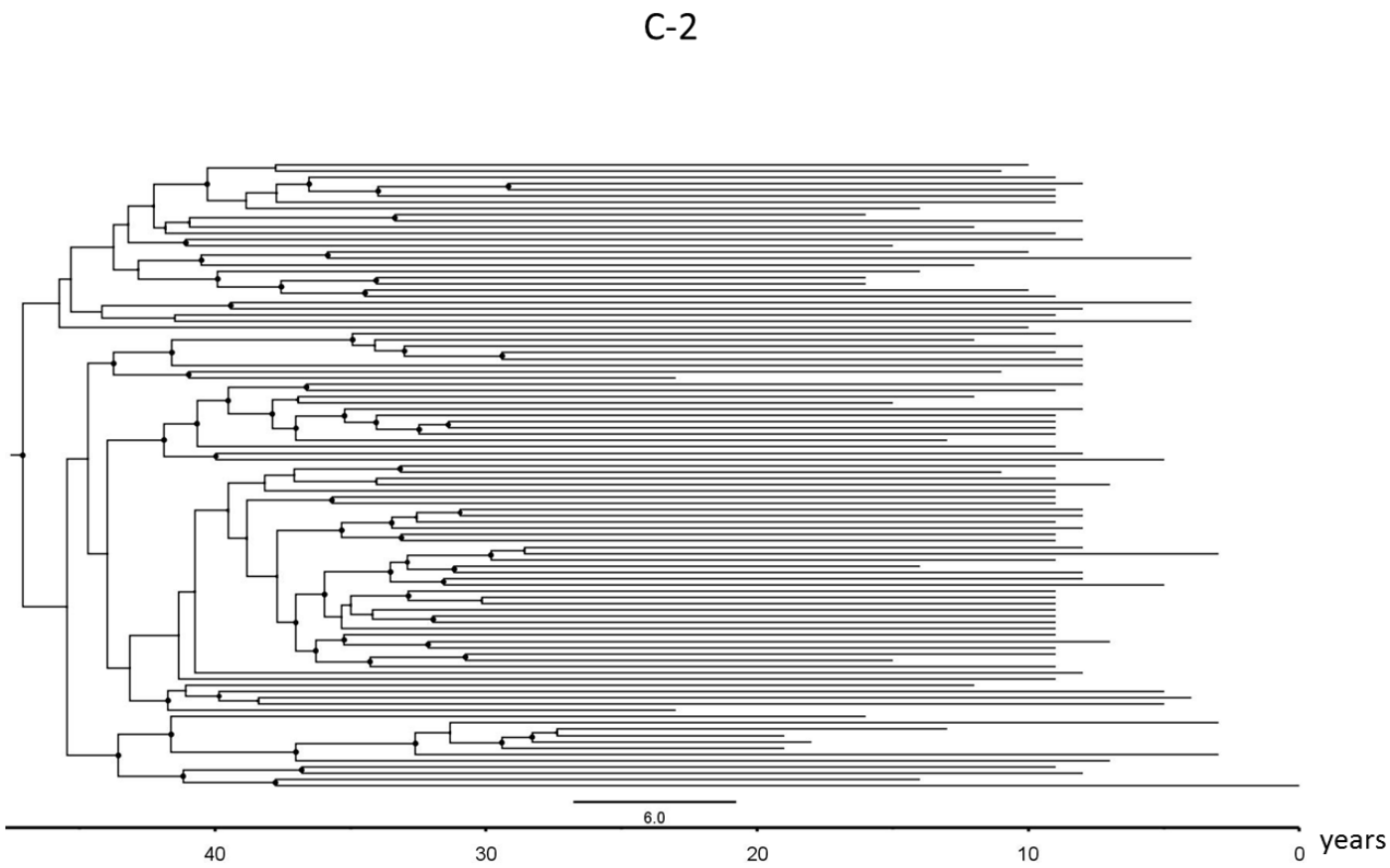

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



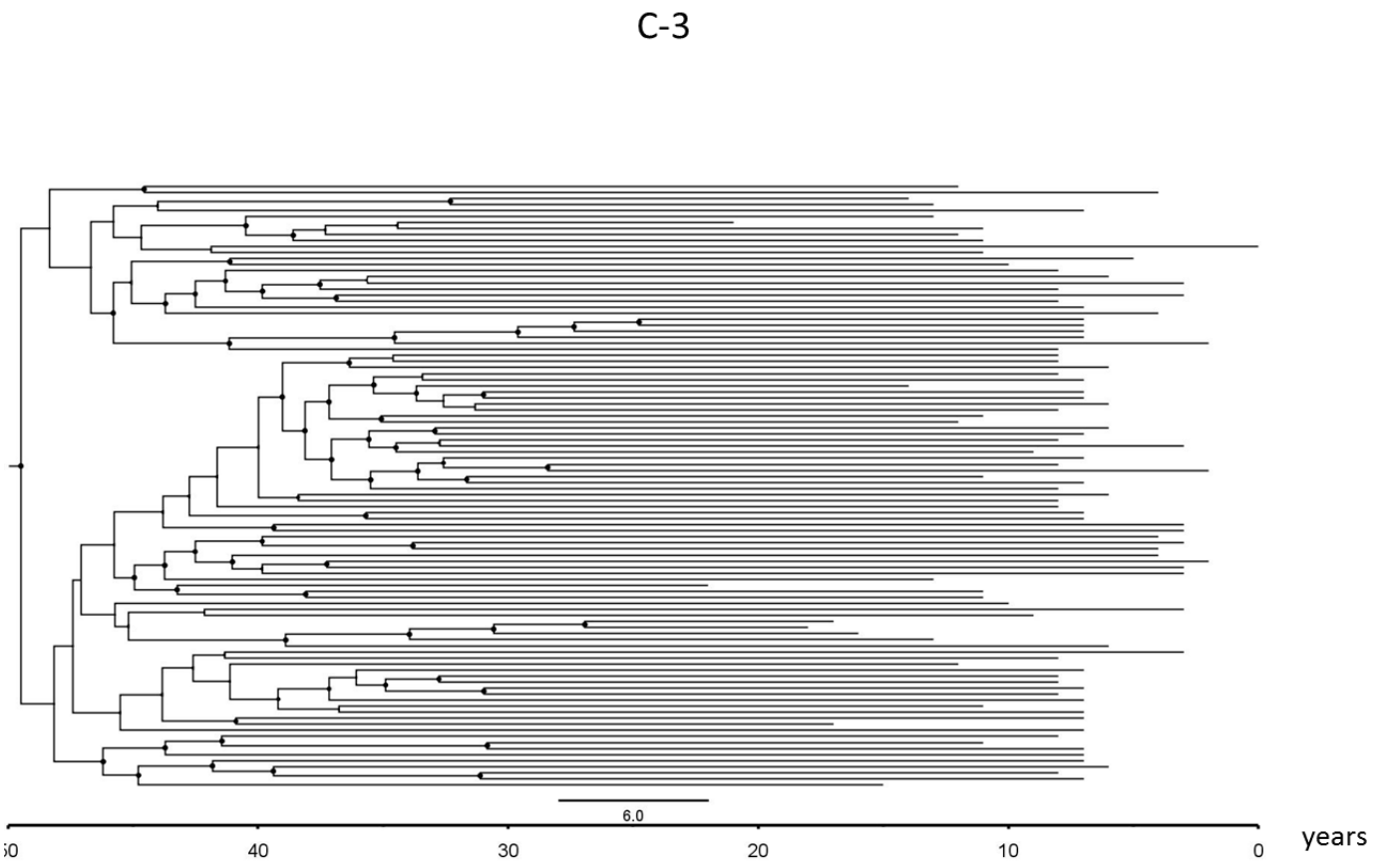

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



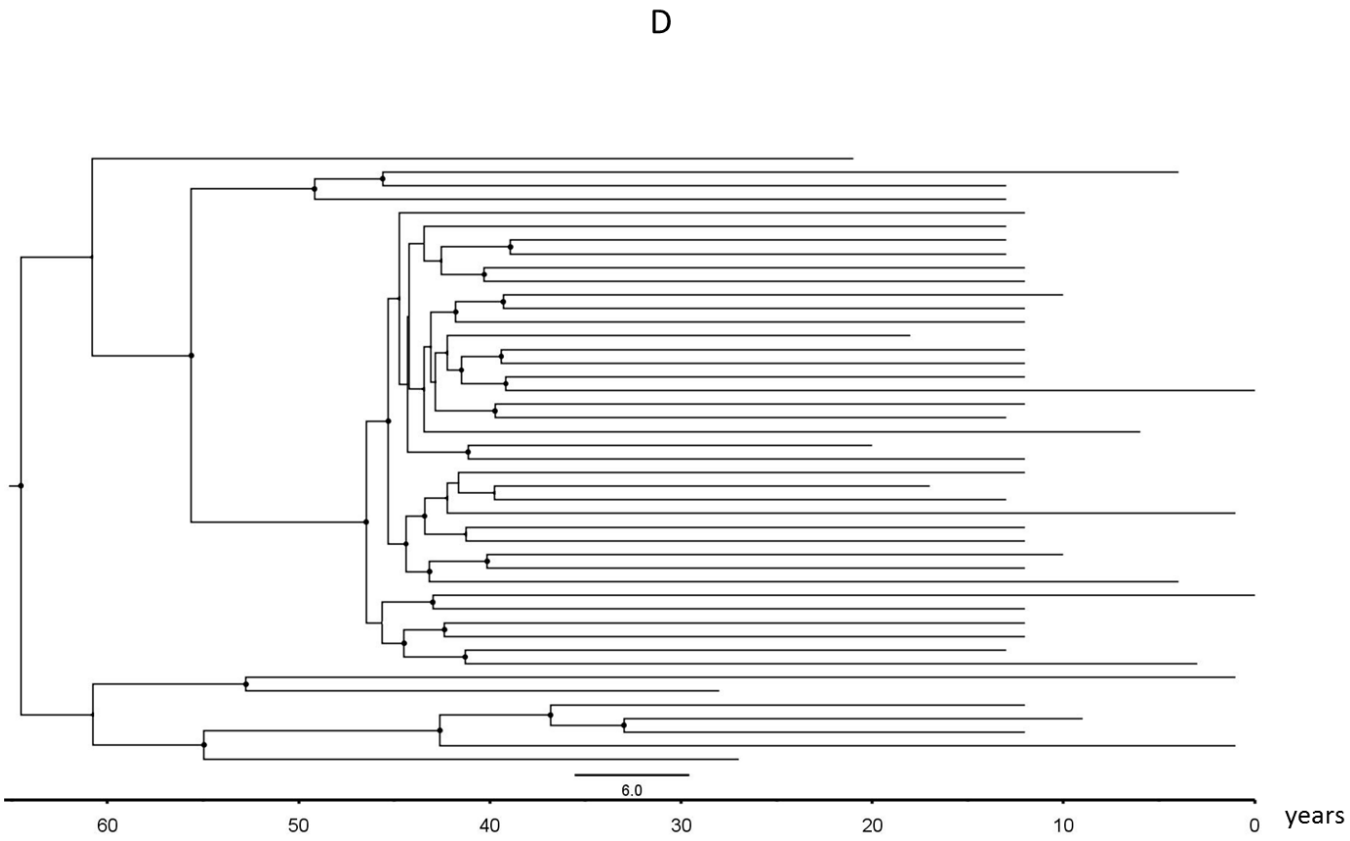

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



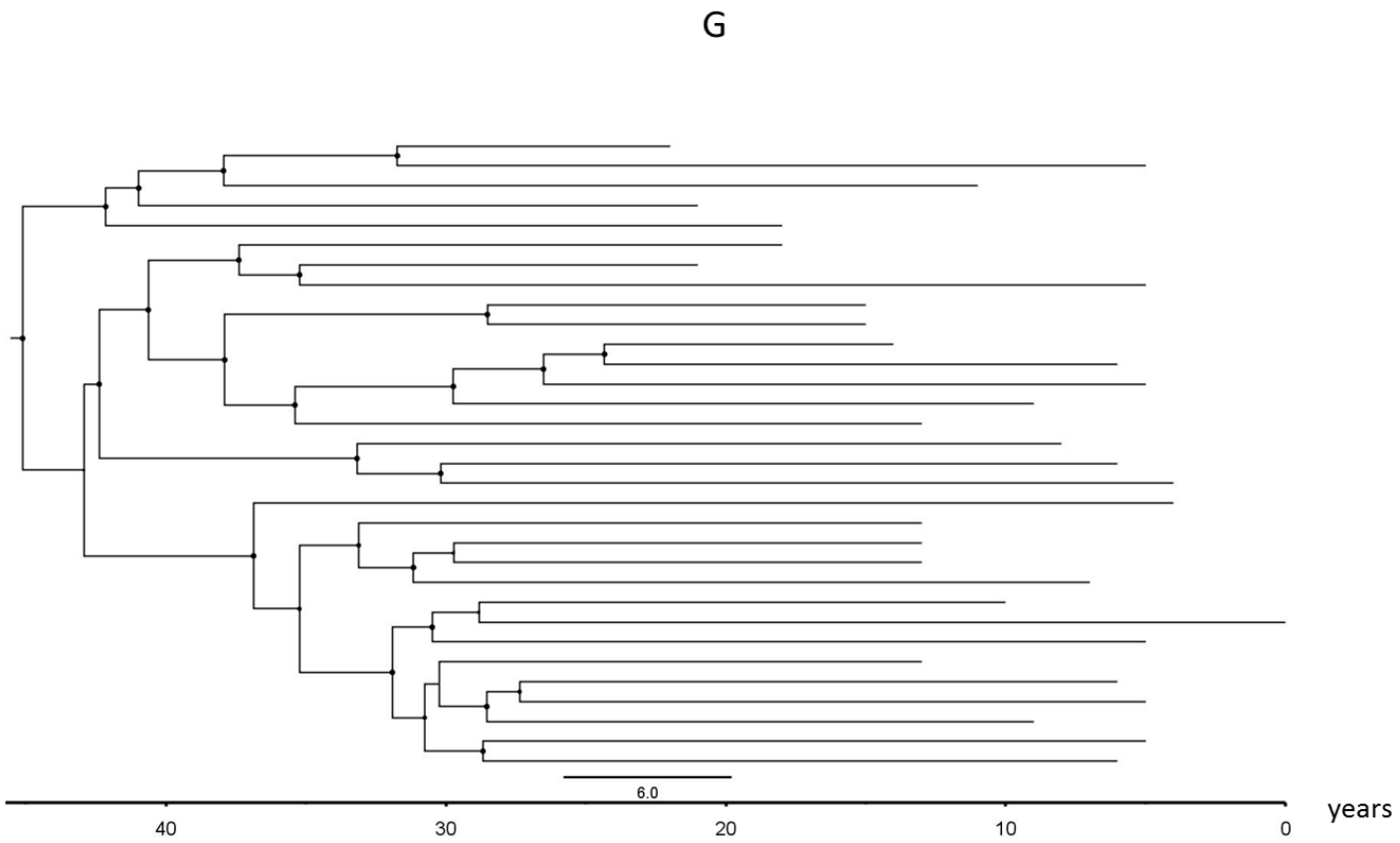

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



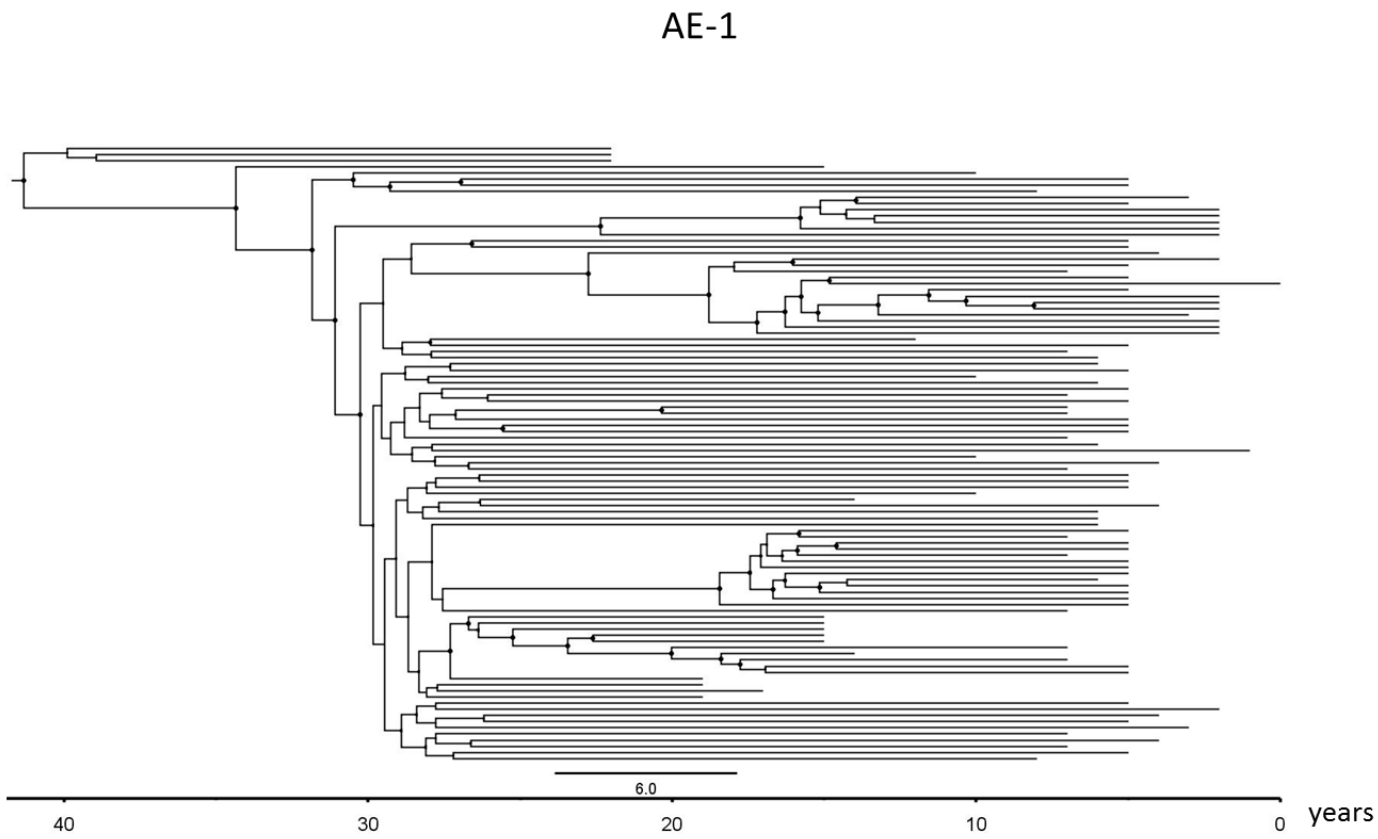

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



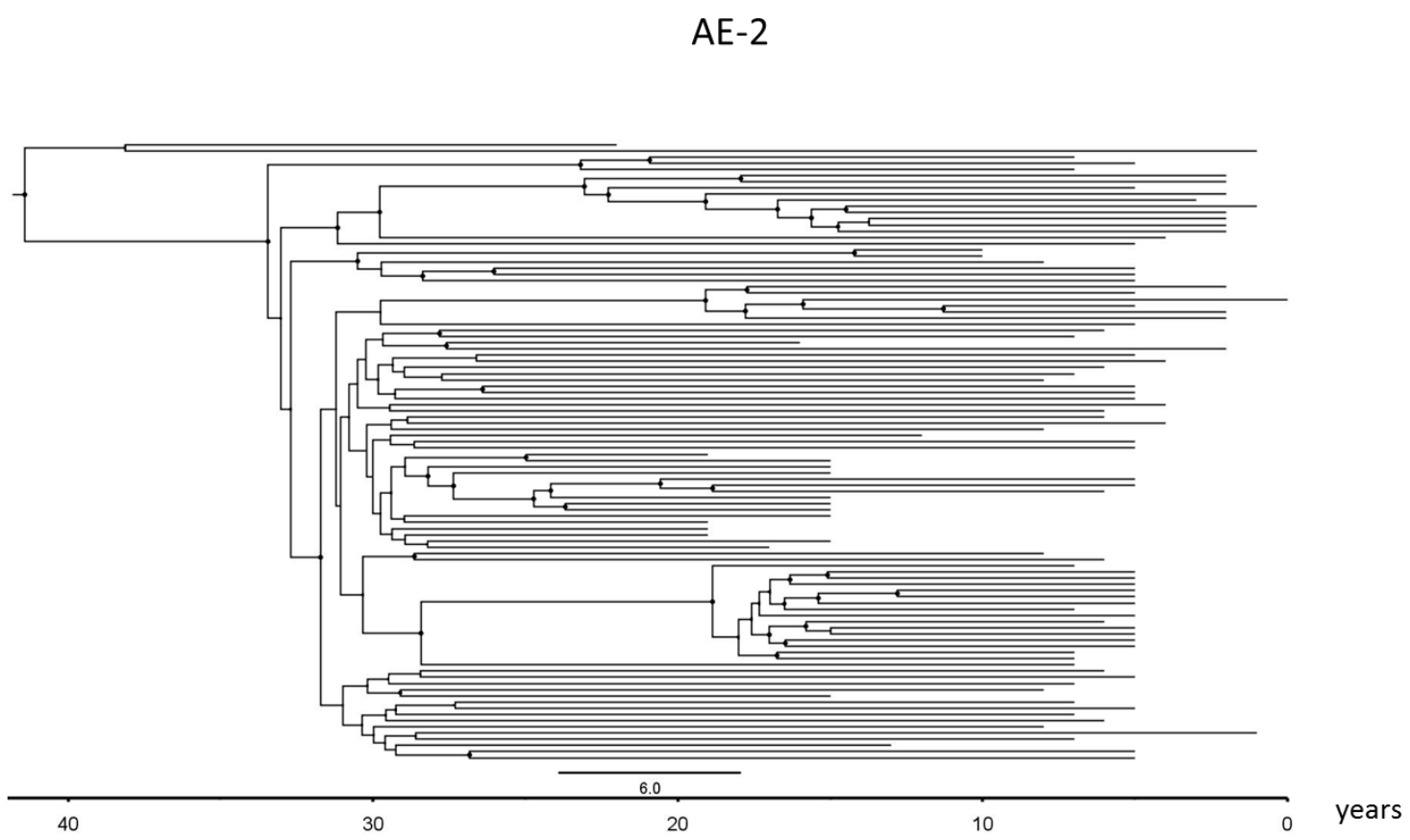

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



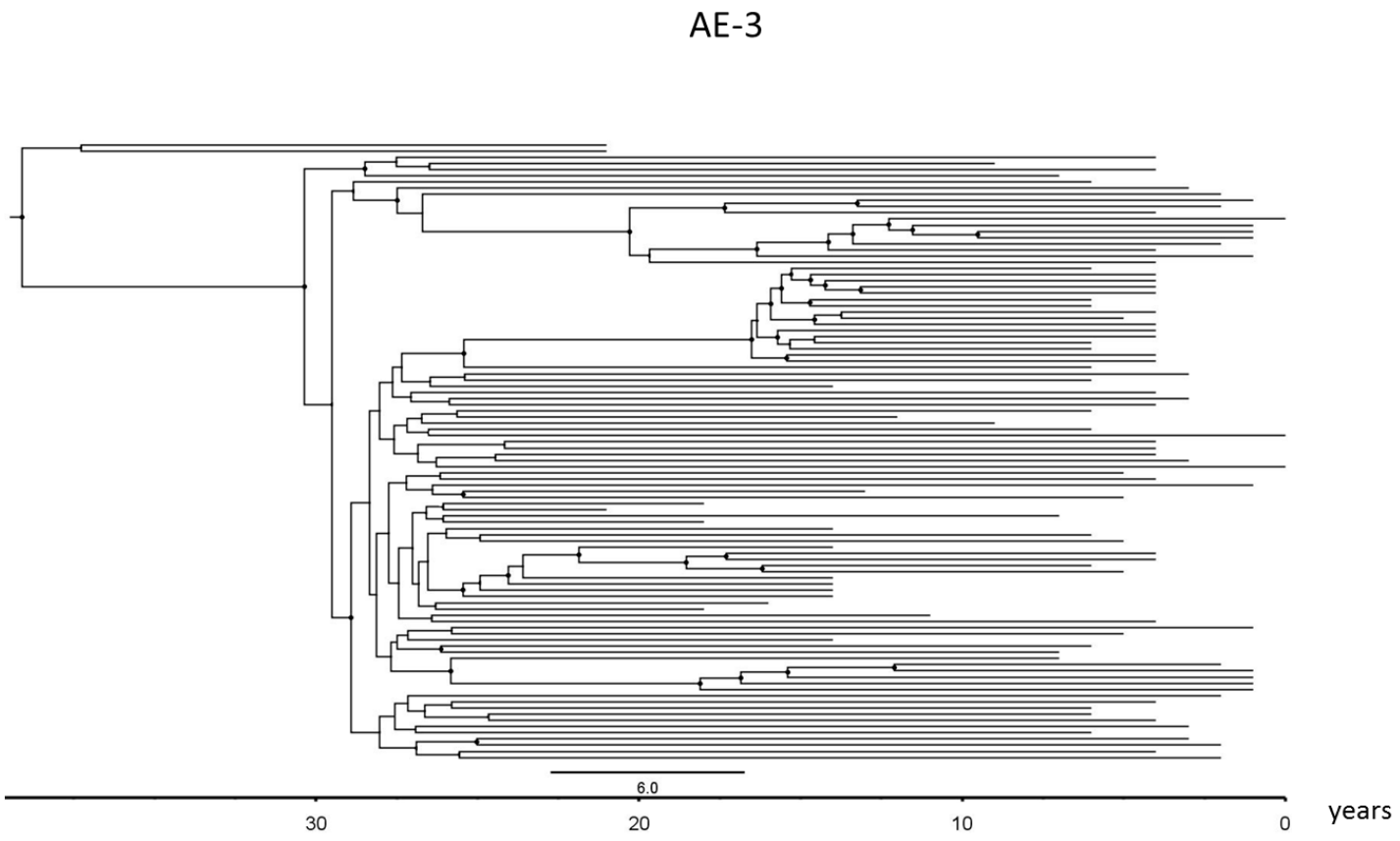

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.



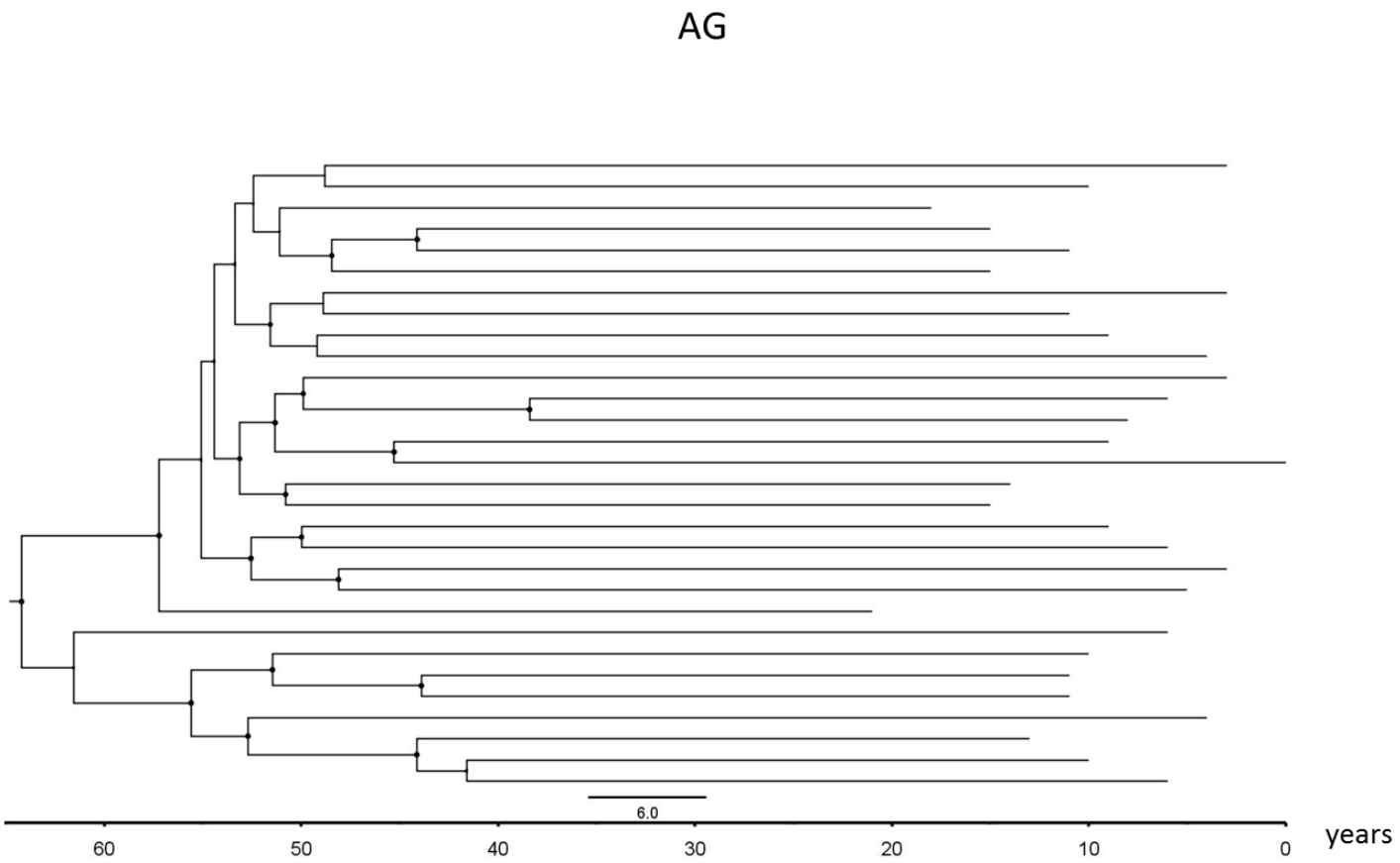

**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.




**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.




**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.




**Supplementary Figure S1 (cont)**. Dated phylogenetic trees obtained from the concatenate of all non-overlapping genomic regions of each subtype analyzed with BEAST, as obtained using the best-fitting demographic and molecular clock models. Node circles represent posterior probabilities >0.90. Time scale: years from present.

## 2.6- Chapter 6: Comprehensive screening for naturally occurring hepatitis C virus resistance to direct-acting antivirals in the NS3, NS5A, and NS5B genes in worldwide isolates of viral genotypes 1 to 6

*Antimicrob Agents Chemother (2016) 60(4):2402-2416.*

AMERICAN SOCIETY FOR MICROBIOLOGY | Antimicrobial Agents and Chemotherapy

# Comprehensive Screening for Naturally Occurring Hepatitis C Virus Resistance to Direct-Acting Antivirals in the NS3, NS5A, and NS5B Genes in Worldwide Isolates of Viral Genotypes 1 to 6

**Juan Ángel Patiño-Galindo,[a,d] Karina Salvatierra,[b,c*] Fernando González-Candelas,[a,d] F. Xavier López-Labrador[a,b,c,d]**

Joint Units in Infection and Health[a] and Genomics and Health,[b] FISABIO-Public Health/Cavanilles Institute for Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain; Virology Laboratory, Genomics and Health Area, FISABIO-Public Health, Generalitat Valenciana, Valencia, Spain[c]; CIBER-ESP (Centro de Investigación Biomédica en Epidemiología y Salud Publica), Instituto de Salud Carlos III, Madrid, Spain[d]

**There is no comprehensive study available on the natural hepatitis C virus (HCV) polymorphism in sites associated with resistance including all viral genotypes which may present variable susceptibilities to particular direct-acting antivirals (DAAs). This study aimed to analyze the frequencies, genetic barriers, and evolutionary histories of naturally occurring resistance-associated variants (RAVs) in the six main HCV genotypes. A comprehensive analysis of up to 103 RAVs was performed in 2,901, 2,216, and 1,344 HCV isolates for the NS3, NS5A, and NS5B genes, respectively. We report significant intergenotypic differences in the frequencies of natural RAVs for these three HCV genes. In addition, we found a low genetic barrier for the generation of new RAVs, irrespective of the viral genotype. Furthermore, in 1,126 HCV genomes, including sequences spanning the three genes, haplotype analysis revealed a remarkably high frequency of viruses carrying more than one natural RAV to DAAs (53% of HCV-1a, 28.5% of HCV-1b, 67.1% of HCV-6, and 100% of genotype 2, 3, 4, and 5 haplotypes). With the exception of HCV-1a, the most prevalent haplotypes showed RAVs in at least two different viral genes. Finally, evolutionary analyses revealed that, while most natural RAVs appeared recently, others have been efficiently transmitted over time and cluster in well-supported clades. In summary, and despite the observed high efficacy of DAA-based regimens, we show that naturally occurring RAVs are common in all HCV genotypes and that there is an overall low genetic barrier for the selection of resistance mutations. There is a need for natural DAA resistance profiling specific for each HCV genotype.**

Hepatitis C virus (HCV) infection is considered a major public health problem. More than 170 million people are chronically infected worldwide, with the consequent risk of developing liver diseases such as cirrhosis and liver cancer, which can eventually cause death (1, 2). HCV is a highly variable RNA virus which has been classified into 7 known genotypes (3). Genetic distances among genotypes reach up to 30% (4). Such diversity can be explained by an evolutionary rate of a magnitude of $10^{-3}$ substitutions per site and year (5). These differences at the genomic level also appear to be relevant at the clinical level. Treatment of chronic HCV infection with peginterferon-ribavirin combination therapy (P/R) shows variable sustained virological response (SVR) rates depending on the infecting HCV genotype (GT), with average SVR rates of 46%, 80%, 66%, and 60% for GTs 1, 2, 3, and 4, respectively (6, 7). Even within HCV GT1, a significant difference in SVR between subtypes 1a and 1b has been reported (8).

In recent years, the field of HCV therapy is blooming because of the clinical development of direct-acting antiviral drugs (DAAs) that are more effective than P/R (SVR up to >90%) and can be given in interferon (IFN)-free regimens with reduced toxicity (9, 10). DAAs that have advanced to clinical trials target three essential proteins for the HCV life cycle, encoded by the nonstructural (NS) protein genes: inhibitors of the NS3 serine protease (protease inhibitors [PIs]), inhibitors of the NS5A protein (NS5A inhibitors), and inhibitors of the NS5B RNA polymerase, either nucleos(t)idic (NI) or non-nucleos(t)idic (NNI) (11). More than 20 different DAA compounds targeting any of these proteins have been approved or are being investigated in advanced clinical trials. These regimens result in an increase in SVR rates to above 90% and reduce the duration of treatment to 12 weeks or less (12). Despite the high SVR rates obtained with these antivirals in HCV GT1 infection, the high variability and epidemic history of HCV may condition the real-world effectiveness of DAAs in a significant proportion of patients infected with other viral genotypes. Natural HCV variation is generated and transmitted over time even in the absence of DAAs. Indeed, naturally occurring DAA resistance-associated HCV variants (RAVs) have already been reported in DAA-naive patients, with some RAVs showing differential prevalence between genotypes and subtypes (13–21). A relevant example is the NS3-Q80K variant associated with resistance to simeprevir (SMV), which was previously found to be present in >30% of GT1a sequences but

**Received** 16 November 2015 **Returned for modification** 1 December 2015
**Accepted** 1 February 2016
**Accepted manuscript posted online** 8 February 2016
**Citation** Patiño-Galindo JÁ, Salvatierra K, González-Candelas F, López-Labrador FX. 2016. Comprehensive screening for naturally occurring hepatitis C virus resistance to direct-acting antivirals in the NS3, NS5A, and NS5B genes in worldwide isolates of viral genotypes 1 to 6. Antimicrob Agents Chemother 60:2402–2416. doi:10.1128/AAC.02776-15.
Address correspondence to F. Xavier López-Labrador, F.Xavier.Lopez@uv.es.
* Present address: Karina Salvatierra, Facultad de Ciencias Exactas, Químicas y Naturales, Universidad Nacional de Misiones, Posadas, Argentina.
Supplemental material for this article may be found at http://dx.doi.org/10.1128/AAC.02776-15.
Copyright © 2016, American Society for Microbiology. All Rights Reserved.

almost absent in GTs 1b, 2, 3, 4, and 5 (13, 16, 21). Current clinical guidelines include NS3-Q80K sequencing to evaluate treatment with SMV for GT1a-infected patients (9, 10). Other studies noted different rates of naturally occurring RAVs in NS5B. The NS5B-C316N variant seems highly prevalent in GT1b but not in GT1a isolates (17, 20), and NS5B-414L and -423I seem highly prevalent in GTs 4 and 5, respectively, but almost nonexistent in other HCV genotypes (20). Another relevant aspect rarely as yet explored is whether, for every resistance-associated site, differences in the genetic barrier (the minimal number of nucleotide mutations required to generate a given RAV) exist between HCV genotypes. This issue has been explored only for the emergence of substitutions causing drug resistance to some DAAs and for some particular, but not all, viral genotypes (20, 22).

To our knowledge, to date there is no study comprehensively describing the natural HCV variation and polymorphism in all three targets of approved DAAs (NS3, NS5A, and NS5B) and comparing the frequencies of natural RAVs between different viral genotypes and within isolates from the same genotype. Here, we report the analysis of an extensive data set of HCV sequences from DAA-naive patients available to date, trying to answer the following questions: (i) how prevalent natural RAVs are, (ii) whether there are inter- and/or intragenotypic differences in their frequencies, (iii) whether naturally occurring multiple resistance to different DAA classes exists, (iv) whether the genetic barrier for resistance differs between HCV genotypes, and (v) whether natural RAVs are generated *de novo* or are characteristic of a subset of viruses being already transmitted over time.

## MATERIALS AND METHODS

**HCV data sets.** In April 2013, data mining was performed on three different databases: the European HCV database (EuHCVdb; http://euhcvdb.ibcp.fr/), the Los Alamos HCV database (LANL; http://hcv.lanl.gov/components/sequence/HCV/search/searchi.html), and The Virus Pathogen Database and Analysis Resource (VIPRBC; www.viprbrc.org/) in order to retrieve any nucleotide sequence including HCV nucleotide positions 3420 to 9000 (H77 reference, GenBank accession no. NC_004102) (23–25). A total of 2,931 sequences were retrieved: 1,328 sequences from LANL, 847 sequences from EuHCVdb, and 756 sequences from VIPRBC. A nonredundant data set of 1,407 sequences spanning the three NS3, NS5A, and NS5B genes was created using an in-house Perl script. To complement this data set with independent sequences from NS3, NS5A, and NS5B, in December 2013 additional searches were performed in LANL, EuHCVdb, and GenBank (www.ncbi.nlm.nih.gov/GenBank/) databases, expanding the initial data set to 4,551 NS3, 7,385 NS5A, and 2,182 NS5B nonredundant sequences. Low-quality sequences (those containing stop codons and/or with >1.2% of the amino acid positions consisting of indeterminate residues) and non-human-derived sequences (accession numbers were searched according to the keyword "HCV" and each of the following: "recombinant," "replicon," "mouse," "*Mus musculus*," "chimpanzee," "chimp," "*Pan troglodytes*," "replicon," "plasmid," "cell culture," "construct," and "chimera") were excluded. Only those HCV sequences derived from DAA-naive patients were retained, excluding all sequences from DAA-treated patients described in the literature up to January 2015. Reference isolates for the different HCV genotypes/subtypes were chosen in accordance with reference 3 and included in the data sets (available upon request).

Nucleotide sequence alignments were obtained with MUSCLE (26, 27) and manually edited with MEGA5.2 (28). In order to identify very closely related sequences (i.e., clones or sequential samples from the same individual), a clustering was performed for each data set using CD-HIT (29), setting the threshold of similarity to 0.95. Further exclusion of sequences originating from the same patient at different time points was done after retrieving additional information from GenBank.

The final data comprised four data sets: (i) NS3-NS5A-NS5B (concatenated set; 1,143 sequences), (ii) NS3 set (2,936 sequences), (iii) NS5A set (2,242 sequences), and (iv) NS5B set (1,376 sequences). All sequences from the concatenated data set were present in the corresponding individual gene data set. The genotype of the sequences was confirmed by using two different methods: (i) the CRP COMET subtyping tool (http://comet.retrovirology.lu/hcv/) and (ii) maximum likelihood (ML) phylogenetic analyses performed with PhyML using the GTR+GAMMA model (30). Recombinant and incorrectly genotyped sequences were excluded from the final data sets ($n$ = 17, 35, 26, and 32 for NS3-NS5A-NS5B, NS3, NS5A, and NS5B sets, respectively).

**Naturally occurring polymorphism at positions associated with resistance to DAA (natural RAVs).** A compilation of 99 RAVs previously described in the literature in NS3, NS5A, and NS5B (31) (see Table S1 in the supplemental material) were used for computing the number and type of amino acid variants at the corresponding positions with BioEdit (32), first for HCV genotypes 1 to 6 and thereafter for subtypes 1a and 1b separately. To calculate the allele frequencies for each RAV, the BioEdit output table was further analyzed with R (33) at each position associated with resistance to DAA. Natural RAVs from all the analyzed positions associated with major resistance to approved DAAs were identified, according to their published phenotypic fold change in 50% inhibitory concentrations ($IC_{50}$s) of >10 on HCV replicon culture (see Table S1 in the supplemental material). Major mutations compiled in references 20 and 22 for which no >10-fold change in $IC_{50}$ was obtained in our bibliographic search were also considered.

**Amino acid signature patterns of resistance to DAA for HCV genotypes 1 to 6.** Amino acid signature patterns for each position in the NS3, NS5A, and NS5B proteins associated with resistance to DAAs were inferred using VESPA (34), The HCV Con1 replicon or all GT1 amino acid sequences were used as background sets, which were compared with each of the query HCV genotypes/subtypes. For a given HCV GT, two analyses were performed: first to determine the consensus amino acid at each position (threshold = 0.50) and thereafter to detect the signature amino acids at each position (threshold = 1.00).

**Searching for natural multiple resistance to different DAA classes.** For each HCV genotype, the data set containing the concatenated NS3-NS5A-NS5B sequences was analyzed to determine the frequency of viral haplotypes with multiple RAVs. A data frame was created using in-house Perl and R scripts (33), with every row representing a sequence and each column representing a resistance-associated position. We computed the haplotype frequencies considering only resistance positions. The frequencies of resistant haplotypes for each analyzed HCV genotype/subtype were inferred with Arlequin v3.1 (35). Given the large number of possible haplotype combinations, the dimensions of the data frames were reduced by considering only 37 positions consistently associated with resistance to approved DAAs (see Table S2 in the supplemental material).

**Calculation of the genetic barrier.** The genetic barrier for the evolution of DAA-susceptible codons into RAVs was calculated according to a previous model applied to human immunodeficiency virus and HBV (36, 37) and recently to HCV (20). Given a position related to RAVs, each nonresistance codon present with a prevalence of >1% in the data sets was compared with all codons encoding a resistance amino acid, assigning a score of 1 to transitions and a score of 2.5 to transversions, because transitions occur 2.5 times more frequently than transversions (20, 36). The genetic barrier was calculated as the addition of the values assigned to the transitions and/or transversions needed to convert a nonresistance codon into a RAV. For each nonresistance codon, only the RAV codon with the minimal genetic barrier was chosen (the most parsimonious one), together with its minimal score. Due to the high number of possible RAVs in the HCV genes

analyzed, the genetic barrier was calculated only for the most clinically significant ones: NS3-36M/A, 54A/S, 55A, 80K/L/R, 155K, 156S/T/V, 168A/E/T/V/Y, and 170A; NS5A-28T, 30R/H, 31V/M, and 93H/C; and NS5B-159F, 282T, 316N/Y, 414T/V, 448H/C, 495L/Q/S, 554S, 556G, and 559G (reviewed in reference 38).

**Evolutionary history of naturally occurring RAVs.** For each data set, an ML phylogenetic tree was obtained with PhyML (30) with the GTR+GAMMA model for nucleotide substitutions and rooted with representative sequences from HCV genotypes 1 to 6. The evolutionary history of natural RAVs for each GT was analyzed with Mesquite (version 2.74; available at http://mesquiteproject.org/mesquite/download/download.html) using the phylogenetic tree obtained and the complete amino acid alignment of each GT as input. The evolution for each amino acid position of interest along the lineages of the phylogeny was inferred using a parsimony approach. Clades were defined as groups of clustered sequences with approximate likelihood ratio test (aLRT) values higher than 0.85.

## RESULTS

**Frequency of natural RAVs in HCV genotypes 1 to 6.** A total of 2,901, 2,216, and 1,344 NS3, NS5A, and NS5B curated sequences, respectively, were included in the analyses of HCV GTs 1a, 1b, 2, 3, 4, 5, and 6. These data sets included 1,126 isolates with sequences spanning all of the complete NS3, NS5A, and NS5B genes.

Tables 1 to 4 (also see Table S3 in the supplemental material) summarize the amino acid frequencies at the most relevant resistance-associated positions in NS3, NS5A, and NS5B, by HCV genotype and the breakdown by subtypes. Among NS3 RAVs to approved protease inhibitors (PIs), V36L (low-level resistance to boceprevir [BOC]/telaprevir [TPV]/SMV/paritaprevir) was highly frequent in GT2 (99.3%), GT3 (100%), GT4 (100%), GT5 (100%), and GT6 (18.9%) but present in only 1.4% and 0.8% of GT1a and GT1b isolates, respectively. Variants in T54 associated with resistance to BOC/TPV showed low frequencies in GT1 (2.29% and 2.38% for GT1a and GT1b, respectively) and were rare in other GTs (found only in one GT2 and two GT3 isolates). V55A/I variants (associated with resistance to BOC) were concentrated in GT1a sequences (48 isolates, 3.91%). Only six non-GT1a isolates carried these variants (five GT1b and one GT3). The NS3-Q80K variant (resistance to SMV) was, as expected, highly frequent in GT1a (36.6%) but also in GT5 (100%) and GT6 (24.3%), whereas Q80G was a signature of GT2 (100% of isolates). R117H (resistance *in vitro* to BOC/TPV) was rare, present only in GT1 and with a very low prevalence (0.67% and 3.02% of GT1a and -1b sequences, respectively). Mutations at position NS3-122 (resistance to SMV) were common and differentially distributed among genotypes (122G/N was present in 7.24% of GT1a and in 13% of GT1b sequences; 122N in 33.8% of GT6 sequences; and 122T in 85.9% of GT4, 83.8% of GT5, and 48.6% of GT6 sequences, respectively). While NS3-122R was present only in GT2 (71.89%), 122A was unique to GT5 (10.8%). The major RAVs for PIs are located at NS3 positions 155, 156, and 168. R155K was very rare, present in only 13 GT1a, one GT1b, and two GT3 isolates (frequencies of 0.88%, 0.10%, and 1.71%, respectively). There was only one A156 variant in the data set: one GT6 isolate with A156V. In contrast, whereas D168 variants were uncommon in GT1 (0.20% and 0.91% of GT1a and GT1b sequences, respectively), D168Q was fixed in GT3 (100%), and 168E (resistance to most noncovalent PIs) was found in 48.6% of GT5 and 2.7% of GT6 sequences. Only three isolates (GTs 1a, 1b, and 6) showed V170A/T (resistance to BOC/SMV/asunaprevir), and I170 was dominant in GTs 1a, 2, 3, and 5 (97.3%, 84.61%, 98.52%, and 89.74%, respectively) and common in GTs 1b and 5 (30.94% and 31.08%, respectively). Finally, M175L (low-level resistance to BOC) was characteristic of HCV GTs 1a, 2 (99% of isolates), 3, 4, and 5 (100% of isolates) but present in only 1.4% of GT1b sequences.

For mutations occurring in NS5A, L28M (low-level resistance to daclatasvir [DCV]/ledipasvir [LDV]/ombitasvir) was present in 98.7% of GT3, 95.7% of GT1a, and 15% of GT4 sequences but in only 2% of GT1b sequences. On the other hand, L28V was common in GT6 isolates (43.8%) but rare in GT1a or GT4 sequences (3.3%). R30Q/L is a RAV studied in GT1b (Con1 replicon) which also causes resistance to DCV/LDV/ombitasvir. Although 30Q was found in only 4.8% of HCV-1b and 3.3% of GT3 sequences, it was present in 100% of GT5 sequences, and 30L was present in 28.3% of GT4 sequences. For variations in Q30 (Q30H/R/E) characterized in subtype 1a replicons, NS5A-30H was present in only 1.6% of GT1a sequences and 30R was present in 3.4% of GT2 sequences, but 30R was common for GT4 and GT6 (53.3% and 32.3%, respectively). Further, H58T (low-level resistance to LDV/ombitasvir) had a low prevalence of 1.2% in GT1b and 10% in GT4 sequences but of 40.6% in GT6 sequences. H58P (resistance in GT1a replicons) was present in 45.8% and 100% of GT6 and GT5 sequences, respectively, compared to only 2.7% of HCV-1a isolates. Finally, 153L (compensatory for resistance to thiazole analogues such as BP008 and DBPR110) was characteristic for GT5 (100%) and GT1a (96.9%) but less common in GT1b (18.1%), GT3 (12.4%), and GT4 (1.6%). The most important RAVs associated with clinical resistance to NS5A inhibitors are L31V/M and Y93H/N. L31M (low-level resistance to DCV/LDV) was highly represented in GT2 (81.2%) and GT4 (88.3%) but rare in GTs 1a, 1b, and 3 (0.9%, 4.6%, and 6.6%, respectively). Remarkably, within GT3, L31M variants were concentrated in non-GT3a isolates (see Table S3 in the supplemental material). NS5A-93H (resistance to DCV/LDV/ombitasvir) had a prevalence of 4.8% in GT1b, 3.1% in GT3, and 3.3% in GT4.

The prevalence of natural RAVs to NS5B NI and NNI was also differentially distributed among HCV genotypes. Among those variations affecting NIs, clinical resistance to sofosbuvir (SOF)/mericitabine (MCB) is associated with mutations in NS5B-S283, and L159F facilitates resistance (39, 40). S282T/R were present in just one isolate each for GTs 1a, 1b, 3, and 4 (frequencies of 0.17%, 0.24%, 1.24%, and 1.63%, respectively), and L159F was identified in only 4.5% of GT1b sequences. No other RAVs in these two NS5B positions were found for any genotype. In contrast to NIs, natural RAVs to NNIs were common and varied by virus subtype. Resistance patterns to NNIs depend on the binding site of the drug (38). NS5B-C316N/Y, M414T/V, Y448H/C, G554S, S556G, and D559G are associated with resistance to ABT-072 and dasabuvir, whereas A421V and P495L/S are associated with resistance to beclabuvir. Interestingly, C316N was found in up to 30% of GT1b and in 8.1% of GT4 sequences, all from non-GT4a isolates. It is worth noting that, while A421V represents a minority of GT1a (12.56%) or GT1b (5.03%) sequences, this variant is highly dominant in the other genotypes (80.32%, 98.77%, 91.94%, 37.5%, and 98.63% for GTs 2, 3, 4, 5, and 6, respectively). In contrast, only one sequence (GT6) carried P495 variants. At NS5B position 414, only 2.5% of GT1 sequences showed variants, while almost all GT2 (414Q/L, 97.5% and 2.5%, respectively) and GT4 (45.2%,

**TABLE 1** Amino acid frequencies in HCV genotypes 1 to 6 at sites associated with resistance to NS3 inhibitors

| NS3 position | Amino acid variant(s) (no. of isolates) for genotype[a]: | | | | | | |
|---|---|---|---|---|---|---|---|
| | GT1a ($n$ = 1,482) | GT1b ($n$ = 992) | GT2 ($n$ = 135) | GT3 ($n$ = 117) | GT4 (64) | GT5 ($n$ = 37) | GT6 ($n$ = 74) |
| C16S | C (1,476), **S (1),** T (1) | C (984), T (2), **S (1),** Y (1) | A (111), T (18), **S (4),** P (2) | T (117) | T (64) | A (37) | T (74) |
| V36A/M/L/G/I/C | V (1,450), **L (21), M (7), I (1)** | V (978), **L (8), I (3)** | **L (134), M (1)** | **L (117)** | **L (64)** | **L (37)** | V (57), **L (14), I (3)** |
| A39V | A (1,461), T (8), G (7), S (3), **V (1)** | A (979), T (4), S (1) | **V (134),** I (1) | A (102), T (10), S (5) | A (60), T (4) | A (37) | A (72), T (1), D (1) |
| Q41R/K/P/H | Q (1,467), **H (12)** | Q (988), **H (2)** | Q (135) | Q (116) | Q (64) | Q (36), **H (1)** | Q (74) |
| F43S/C/Y/V/I/L | F (1,481), **Y (1)** | F (992) | F (135) | F (117) | F (63) | F (36) | F (74) |
| I48V | I (1,443), **V (33)** | I (382), **V (597),** L (4), F (2), T (1) | I (132), **V (3)** | I (15), **V (98),** L (4) | I (25), **V (39)** | I (37) | I (42), **V (28),** L (4) |
| T54A/S/V/G/C | T (1,446), **S (34)** | T (966), **S (23)** | T (133), **A (1)** | T (117) | T (59), **S (2)** | T (37) | T (73) |
| V55A/F/I/K/T | V (1,410), **I (36), A (32)** | V (984), **I (3), A (2)** | V (134), G (1) | V (116), **I (1)** | V (64) | V (36), L (1) | V (73) |
| D79E | D (1,478), **E (2),** N (1) | D (990), N (2) | D (1), **E (133),** K (1) | D (112), **E (4),** N (1) | D (64) | D (37) | D (71), **E (3)** |
| Q80K/L/N/R/H/G | Q (884), **K (542), L (28), R (9), N (7),** S (1), M (1) | Q (948), **L (33), R (5), K (4)** | **G (135)** | Q (116), **R (1)** | Q (63) | **K (37)** | Q (55), **K (18), L (1)** |
| A87T | A (1,435), S (22), **T (7),** C (5), I (1) | A (969), S (13), P (2), **T (2),** M (1), E (1) | A (1), S (127), **T (3),** N (3), G (1) | A (114), S (3) | A (56), S (5), C (1), V (1) | S (35), **T (2)** | A (54), S (19), **T (1)** |
| Y105C | Y (1,476), F (3) | Y (976), F (11) | Y (135) | Y (117) | Y (30), F (34) | Y (37) | Y (70), F (3), **C (1)** |
| R109K | R (1,482) | R (986), M (1), **K (1)** | R (135) | R (117) | R (63) | R (37) | R (73) |
| R117H | R (1,448), C (19), **H (10),** S (2), L (1) | R (929), **H (30),** C (28), Q (1), F (1) | R (135) | R (117) | R (64) | R (37) | R (74) |
| S122G/A/R/N/T | S (1,365), **G (96), N (11), T (2),** C (2) | S (798), **G (94), T (56), N (35),** C (2) | **R (110),** K (23), **T (1)** | S (117) | S (5), **T (55), N (4)** | S (2), **T (31), A (4)** | S (13), **T (36), N (25)** |
| R123H/K | R (1,477), **K (3)** | R (983), **K (8)** | R (135) | T (117) | R (61), **K (1)** | R (34), **H (1)** | R (74) |
| I132V | I (1,466), **V (5),** L (2) | I (282), **V (692),** L (10) | I (7), L (127) | I (8), L (109) | I (59), L (4), **V (1)** | I (35), L (1), **V (1)** | I (51), L (22) |
| S138T/D/P | S (1,479) | S (988), F (1) | S (135) | S (117) | S (63), F (1) | S (37) | S (71), F (1) |
| R155K/T/I/M/G/L/S/Q/P/N/W | R (1,463), **K (13)** | R (991), **P (1)** | R (135) | R (115), **K (2)** | R (63) | R (36) | R (73) |
| A156S/T/V/I/F/N/G/D | A (1,478) | A (992) | A (134) | A (117) | A (64) | A (37) | A (73), **V (1)** |
| V158I/M | V (1,480) | V (990), **I (1)** | V (130), **I (3), M (2)** | V (115), A (1), **I (1)** | V (62), L (1), **I (1)** | V (33), **M (2),** L (1) | V (71), **I (2)** |
| V163L | V (1,478) | V (984), I (4) | V (133), I (1), A (1) | V (114), E (1) | V (62) | V (36), A (1) | V (74) |
| D168Q/A/Y/V/E/T/N/P/I/H/G/F/S/K | D (1,471), **E (3), G (1)** | D (979), **E (9)** | D (135) | **Q (116)** | D (64) | D (19), **E (18)** | D (72), **E (2)** |
| V170A/T/G/L/M | V (55), I (1,422), **T (1)** | V (682), I (307), **T (1)** | V (2), I (133) | V (11), I (105) | V (63), I (1) | V (3), I (33) | V (50), I (23), **A (1)** |
| E173G | E (1,481) | E (992) | E (133), D (1), **G (1)** | E (116), D (1) | E (63) | E (36) | E (74) |
| S174F/P | S (584), N (827), G (59), D (2), A (1), C (1) | S (960), A (19), **F (6),** T (4), H (1), L (1), C (1) | S (106), T (18), A (11) | S (3), T (106), A (7), I (1) | S (59), A (3) | S (4), N (32) | S (27), N (44), A (2), G (1) |
| M175L | **L (1,476)** | M (977), **L (14)** | **L (134),** I (1) | **L (117)** | **L (64)** | **L (37)** | M (74) |
| E176G | E (1,466), **G (7),** D (4), A (1), S (1) | E (986), Q (2), D (1), **G (1)** | D (110), A (17), N (8) | S (93), N (18), A (5), H (1) | E (62), A (1) | E (37) | E (46), Q (26), D (2) |

[a] RAVs are indicated in bold.

**TABLE 2** Amino acid frequencies in HCV genotypes 1 to 6 at sites associated with resistance to NS5a inhibitors

| NS5a position | Amino acid variant(s) (no. of isolates) for genotype[a]: | | | | | | |
|---|---|---|---|---|---|---|---|
| | GT1a ($n$ = 861) | GT1b ($n$ = 810) | GT2 ($n$ = 149) | GT3 ($n$ = 226) | GT4 ($n$ = 60) | GT5 ($n$ = 14) | GT6 ($n$ = 96) |
| Reference GT1b (Con1) | | | | | | | |
| Q24L | Q (2), K (844), R (5), E (4), S (1) | Q (788), K (11), R (10) | S (113), T (35), A (1) | S (218), **L (5)**, A (1), G (1), T (1) | K (60) | Q (14) | K (65), Q (29), R (2) |
| L28M/V | L (1), **M (824)**, **V (28)**, T (6), I (2) | L (789), **M (16)**, **V (3)**, P (1) | L (90), F (56), C (3) | L (1), **M (223)**, I (2) | L (48), **M (9)**, **V (2)**, I (1) | L (14) | L (22), **V (42)**, F (11), T (11), A (5), **M (3)**, G (1) |
| R30Q/L | R (4), **Q (840)**, H (14), **L (2)** | R (756), **Q (39)**, K (7), **L (3)**, M (2), H (1) | R (5), K (143) | R (2), A (180), K (30), T (6), **L (3)**, S (2), V (2), M (1) | R (32), **L (17)**, S (4), T (4), **Q (2)**, A (1) | **Q (14)** | R (3), S (41), A (22), N (1) |
| L31M/V/F | L (853), **M (8)** | L (766), **M (37)**, I (3), **V (1)**, **F (1)**, P (1) | L (28), **M (121)** | L (208), **M (15)**, **V (2)** | L (7), **M (53)** | L (14) | L (95), I (1) |
| P32L | P (860) | P 810 | P (149) | P (226) | P (70) | P (14) | P (96) |
| Q54H/N/L/Y | **H (847)**, **Y (11)**, C (1), F (1) | Q (524), **H (253)**, **N (9)**, **L (6)**, **Y (15)**, C (1), E (1), T (1) | T (147), **N (1)**, S (1) | S (196), T (29) | **H (60)** | S (13), **Y (1)** | **H (69)**, R (10), **Y (7)**, T (3), C (2), **N (2)**, S (1) |
| P58S/A/L/T/H | P (23), **H (823)**, R (6), Y (6), Q (1), C (1), D (1) | P (750), **S (30)**, T (10), **L (7)**, Q (6), **A (6)** | P (141), **S (6)**, **H (1)**, **T (1)** | P (215), **S (8)**, R (1), **A (1)**, **T (1)** | P (52), **T (6)**, R (1) | P (14) | P (44), **T (39)**, G (5), **S (4)**, **A (2)**, **L (1)** |
| Q62E/R/A/P/S | **E (830)**, D (24), G (3), **A (1)** | Q (774), **E (16)**, **R (6)**, H (5), N (2), K (2), **S (1)**, D (1), G (1), L (1) | N (128), **A (6)**, S (4), T (4), V (3), H (1), I (1), **E (1)** | Q (1), **S (135)**, T (46), M (7), D (7), L (6), **E (6)**, V (5), **A (4)**, **P (4)**, I (4), N (1) | Q (4), **E (42)**, **S (4)**, N (4), D (3), **R (2)**, V (1) | T (13), **A (1)** | Q (4), V (34), D (15), N (11), T (8), **A (8)**, **S (5)**, M (4), **E (3)**, H (1), I (1), **R (1)**, K (1) |
| A92T | A (854), P (7) | A (783), **T (16)**, V (7), G (1) | A (1), C (142), S (6) | E (224), G (1) | A (57), **T (1)** | A (14) | A (96), P (1) |
| Y93H/N/C | Y (849), **H (5)**, **C (4)**, **N (1)**, F (1) | Y (769), **H (39)**, **C (2)** | Y (148), F (1) | Y (219), **H (7)** | Y (55), **H (2)**, T (1), R (1), S (1) | T (14) | T (79), S (14), I (2) |
| F149L | F (858), **L (1)** | F (809) | F (149) | F (226) | F (58), **L (1)**, S (1) | F (14) | F (96) |
| V153M/L/I | V (15), **L (834)**, **I (4)**, P (1) | V (657), **L (147)**, **I (5)**, T (1) | V (149) | V (197), **L (28)** | V (57), **L (1)** | **L (14)** | V (92), **I (4)** |
| M202L | M (856), **L (2)**, Q (2), I (1) | M (807), **L (3)** | M (148), V (1) | M (223), **L (1)** | M (60) | M (14) | M (94), **L (2)** |
| M265V/T | M (858), **V (2)**, L (1) | M (791), K (9), **V (5)**, **T (1)**, W (1) | M (148), L (1) | M (222), **V (2)**, I (1), L (1) | M (13) | M (14) | M (57), **V (32)**, L (3), A (2), I (1) |
| D320E | D (857), **E (2)**, N (1), G (1) | D (796), **E (6)**, S (3), N (2) | D (131), **E (7)**, N (4), S (3), G (2), Q (1) | D (226) | D (43), **E (11)**, A (4), S (1) | D (14) | D (94), N (1), G (1) |
| Y321N | Y (856) | Y (805), H (1) | Y (147) | Y (226) | Y (58) | Y (14) | Y (95), **N (1)** |
| Reference GT1a (H77) | | | | | | | |
| L23F | L (856), M (1) | L (807), P (2), I (1) | L (149) | L (220), P (5), I (1) | L (57) | L (14) | L (96) |
| M28T | M (824), V (28), **T (6)**, I (2), L (1) | M (16), L (789), V (3), P (1) | L (90), F (56), C (3) | M (223), I (2), L (1) | M (9), L (48), V (2), I (1) | L (14) | M (3), V (42), L (22), F (11), **T (11)**, A (5), G (1) |
| Q30H/R/E | Q (840), **H (14)**, **R (4)**, L (2) | Q (39), **R (755)**, K (7), L (3), M (2), **H (1)** | K (143), **R (5)** | A (180), K (30), T (6), L (3), V (2), S (2), **R (2)**, M (1) | Q (2), **R (32)**, L (17), S (4), T (4), A (1) | Q (14) | S (41), **R (31)**, A (22), N (1) |
| L31V/M | L (853), **M (8)** | L (766), **M (37)**, I (3), **V (1)**, F (1), P (1) | L (28), **M (121)** | L (208), **M (15)**, **V (2)** | L (7), **M (53)** | L (14) | L (95), I (1) |
| P32L | P (860) | P (810) | P (149) | P (226) | P (60) | P (14) | P (96) |
| H58D/P | H (823), **P (23)**, R (6), Y (6), Q (1), **D (1)**, C (1) | **P (750)**, S (30), T (10), L (7), Q (6), A (6) | H (1), **P (141)**, S (6), T (1) | **P (215)**, S (8), R (1), A (1), T (1) | **P (52)**, T (6), R (1) | **P (14)** | **P (44)**, T (39), G (5), S (4), A (2), L (1) |
| Y93N/C | Y (849), H (5), **C (4)**, **N (1)**, F (1) | Y (769), H (39), **C (2)** | Y (148), F (1) | Y (219), H (7) | Y (55), H (2), T (1), R (1), S (1) | T (14) | T (79), S (14), I (2) |
| D320E | D (857), **E (2)**, N (1), G (1) | D (796), **E (6)**, S (3), N (2) | D (131), **E (7)**, N (4), S (3), G (2), Q (1) | D (226) | D (43), **E (11)**, A (4), S (1) | D (14) | D (94), N (1), G (1) |

[a] RAVs are indicated in bold.

**TABLE 3** Amino acid frequencies in HCV genotypes 1 to 6 at sites associated with resistance to NS5B NIs

| NS5B NI resistance position | Amino acid variant(s) (no. of isolates) for genotype[a]: | | | | | | |
|---|---|---|---|---|---|---|---|
| | GT1a ($n$ = 581) | GT1b ($n$ = 417) | GT2 ($n$ = 122) | GT3 ($n$ = 81) | GT4 ($n$ = 62) | GT5 ($n$ = 8) | GT6 ($n$ = 73) |
| A15G | A (562), S (14), T (3), V (2) | A (379), S (32), T (3), V (2), **G (1)** | A (4), **G (80),** S (32), C (2), R (1), I (1) | A (2), S (75), T (2), N (2) | A (62) | S (8) | A (71), V (2) |
| K72M | K (579), **M (1)** | K (416), N (1) | K (122) | K (80), R (1) | K (61) | K (6), R (2) | K (72) |
| S96T | S (581) | S (417) | S (122) | S (81) | S (62) | S (8) | S (73) |
| N142T | N (581) | N (385), S (29), **T (1)** | N (120), **T (2)** | N (78), S (3) | N (58), S (3), **T (1)** | N (8) | N (73) |
| L159F | L (581) | L (397), **F (19)** | L (122) | L (81) | L (62) | L (8) | L (72) |
| R222Q | R (579), D (1) | R (417) | R (122) | R (81) | R (62) | R (8) | R (65), K (5), S (3) |
| C223H/Y | C (581) | C (416), R (10) | C (121), W (1) | C (81) | C (62) | C (8) | C (73) |
| I239V/L | I (580), **V (1)** | I (416) | I (120), **L (2)** | I (78), **V (3)** | I (10), **V (52)** | I (8) | I (62), **V (10)** |
| S282T/R | S (580), **R (1)** | S (416), **T (1)** | S (122) | S (79), **R (1)** | S (61), **T (1)** | S (8) | S (73) |
| A300T | Q (296), R (260), K (21), L (4) | A (18), S (300), **T (95),** V (1), R (1), F (1) | L (105), R (8), Q (8), K (1) | **T (67),** S (6), K (5), M (2), R (1) | **T (46),** S (8), V (5), N (1), M (1) | L (8) | A (1), Q (53), E (5), K (4), M (3), S (3), **T (2),** N 1 |
| L320I/F | L (579) | L (415), R (1) | L (122) | L (80), Q (1) | L (61) | L (8) | L (73) |
| V321I | V (580), **I (1)** | V (414), **I (3)** | V (120), F (1), **I (1)** | V (81) | V (59), **I (3)** | V (8) | V (73) |
| A396G | A (579) | A (417) | A (122) | A (81) | A (61), V (1) | A (8) | A (73) |
| Y586C | Y (544), **C (2)** | Y (399), N (1), F (1), **C (1)** | F (114), L (2), V (1) | F (57), V (1) | F (49) | F (7) | F (71), I (1) |

[a] RAVs are indicated in bold.

37.1%, 16.12%, and 1.6% with variants 414L/V/I/Q, respectively) isolates did. Although NS5B-556 RAVs were rare for GT1 (556G/N in 7.2% and 1.4% of HCV-1b and 1a isolates, respectively), 556G was dominant in all non-1 HCV genotypes except GT6 (95.9%, 93.5%, 82.2%, 100%, and 4.2% of GT2, GT3, GT4, GT5, and GT6 sequences, respectively). In contrast, G554S was uncommon (7.4% and 2.5% of GT2 and GT3 isolates, respectively). Other natural RAVs to NNIs identified included L392I (resistance to deleobuvir and TMC647055), which was more frequently found in GT2 (69.7%) than in HCV-1b and GT6 (3.8% and 1.6%, respectively). Finally, RAVs P495L/Q/S (resistance to deleobuvir and TMC647055) were identified in only one GT6 isolate (P495L), and Y448H/C (resistance to tegobuvir) was identified in one GT1b isolate (Y448C).

**Amino acid signature patterns and RAVs.** We used the VESPA program to determine the overall resistance profile to DAAs within and between HCV genotypes and to identify the predominant and signature amino acids in positions associated with resistance at the three NS3, NS5A, and NS5B genes. Tables 5 to 8 (see also Table S4 in the supplemental material) show the consensus amino acid pattern for each HCV genotype and the breakdown by subtypes. For a given position, one amino acid was considered a signature when present in 100% of the sequences from a given genotype. For HCV GTs 2, 3, 4, and 5, all the sequences presented more than one natural RAV as a signature amino acid. For GT2, signature RAVs were NS3-Q80G and NS5B-I424V, C445F, I482L, and V494A. For GT3, signature RAVs were NS3-V36L, D168Q, and M175L and NS5B-I424V, C445F, and I482L. For GT4, two NS3, one NS5A, and three NS5B RAVs were found as signature amino acids (NS3-V36L and M175L, NS5a Q54H and NS5B-L419I, I482L, and R531K). In GT5, signature RAVs were NS3-V36L, Q80K, and M175L and NS5B-I363V, M423I, I424V, C445F, R531K, and S556G. For GT6, three signature RAVs were found in NS5B (NS5B-C455F, I482L, and V499A). Finally, we found several positions with differential signature/majority amino acid patterns between subtypes of the same GTs (Tables 5 to 8; see also Table S4). These positions were NS3-48, 132, 170, 174, and 175, NS5A-24, 28, 30, 54, 58, 62, and 153, and NS5B-19, 71, 300, 338, and 499 for GT1; NS3-122, 173, and 174, NS5A-24 and 28, and NS5B-15, 71, 338, and 392 for GT2; NS3-39, 132, and 174, NS5A-30, 31, 54, 62, and 153, and NS5B-300, 442, and 494 for GT3; NS3-105 for GT4; and NS3-48, 80, 87, 174, and 176, NS5A-24, 28, 30, 58, 62, and 265, and NS5B-300, 494, 486, 555, 556, and 571 for GT6.

**Haplotypes with multiple RAVs to different DAA classes.** For the subset of HCV isolates for which all the NS3, NS5A, and NS5B gene sequences were available ($n$ = 1,126), the frequency of isolates carrying more than one RAV was calculated considering all positions associated with resistance. For each HCV genotype, the haplotypes with multiple RAVs are detailed in Table S5 in the supplemental material, including their frequencies. Haplotypes with more than one RAV accounted for 53% of GT1a sequences, with NS3-(Q80K+175L) being present in 32% of them. For GT1b, the frequency of sequences with more than one RAV was 28.5%, and 4.5% of GT1b sequences presented the haplotype NS3-(122G)+NS5B-(316N). It is worth noting that all (100%) the haplotypes of genotypes 2, 3, 4, and 5 carried more than one natural RAV. In GT2, the most prevalent haplotype with multiple RAVs (40%) was NS3-(36L+80G+122R+175L)+NS5A-(31M)+NS5B-(392I+414Q+556G). For GT3, NS3-(36L+168Q+175L)+NS5A-(28M) and NS3-(36L+168Q+175L)+NS5A-(28M)+NS5B-(556G) were present in 42.9% and 28.6% of the sequences, respectively. For GT4, there were two equally frequent haplotypes, NS3-(36L+122T+175L)+NS5A-(28M+31M)+NS5B-(414L+556G) and NS3-(36L+122T+175L)+NS5A-(31M)+NS5B-(414L+556G), both present in only 8.8% of the sequences. Up to 50% of the GT5 sequences showed the haplotype NS3-(36L+80K+122T+168E+170V+175L)+NS5B-(556G). Finally, for GT6, 67.1% of the

**TABLE 4** Amino acid frequencies in HCV genotypes 1 to 6 at sites associated with resistance to NS5B NNIs

| NS5B NNI resistance position | Amino acid variant(s) (no. of isolates) for genotype[a]: | | | | | | |
|---|---|---|---|---|---|---|---|
| | GT1a ($n$ = 581) | GT1b ($n$ = 417) | GT2 ($n$ = 122) | GT3 ($n$ = 81) | GT4 ($n$ = 62) | GT5 ($n$ = 8) | GT6 ($n$ = 73) |
| T19S/P | Q (574), E (5), D (1), **P (1)** | T (43), **S (368),** N (7), G (1) | E (118), G (2), D (1), R (1) | E (81) | T (17), **S (37),** A (4), **P (3),** N (1) | E (8) | E (73) |
| K50R | K (576), **R (5)** | K (403), **R (12),** Q (1) | K (122) | K (78), **R (3)** | K (61) | K (8) | K (69), **R (4)** |
| D55E | D (578), **E (3)** | D (517) | D (122) | D (81) | D (62) | D (8) | D (68), **E (2),** H (1), G (1) |
| M71V | **V (578),** I (2), A (1) | M (393), I (13), **V (10),** L (1) | **V (75),** I (47) | **V (71),** I (9) | I (59), T (1), **V (1)** | M (8) | M (3), **V (18),** I (45), T (6) |
| H95Q/R | H (575) | H (414), L (1), **Q (1)** | H (122) | H (78), N (2), K (1) | H (62) | H (8) | H (71), C (1), **R (1)** |
| V138I | V (1), **I (579),** L (1) | V (27), **I (390)** | V (10), **I (111),** A (1) | V (1), **I (80)** | **I (62)** | **I (8)** | V (2), **I (70)** |
| L314F | L (577), P (1), **F (1)** | L (417) | L (122) | L (81) | L (62) | L (8) | L (73) |
| C316Y/F/N/S | C (580) | C (287), **N (126),** H (1), R (1), **Y (1), S (1)** | C (121), W (1) | C (79), G (2) | C (54), **N (5),** H (3) | C (8) | C (73) |
| A338V | A (579) | A (31), **V (386)** | A (96), **V (24),** T (1) | A (75), **V (5),** T (1) | A (58), **V (3),** T (1) | A (8) | A (70), **V (1)** |
| I363V | I (578), **V (1)** | I (416), **V (1)** | I (122) | I (81) | I (62) | **V (8)** | I (73) |
| S365T/A/L/O/F | S (579) | S (416), **A (1)** | S (122) | S (81) | S (62) | S (8) | S (73) |
| S368A/T | S (577) | S (416), P (1) | S (122) | S (81) | S (62) | S (8) | S (70), **A (1)** |
| T389S/A | T (572), **A (3), S (2),** M (1), V (1) | T (400), **A (9),** I (4), **S (2),** M (2) | T (119), **A (1),** I (1), **S (1)** | T (58), E (22), D (1) | E (62) | Q (8) | T (52), V (13), I (5), L (1), **S (1), A (1)** |
| L392I | L (568), F (11), **I (3)** | L (396), **I (16),** F (4) | L (35), **I (85),** F (1) | L (81) | L (62) | L (8) | L (71), **I (1)** |
| N411S | N (578) | N (417) | N (120), T (2) | N (79), **S (2)** | N (62) | N (8) | N (73) |
| M414L/T/I/V/Q | M (577), **L (1), V (1)** | M (415), **L (1), I (1)** | **Q (119), L (3)** | M (81) | **L (28), V (23), I (10), Q (1)** | M (8) | M (71), A (1) |
| L419M/V/S/I | L (578) | L (411), **I (5)** | L (2), **I (117), V (3)** | L (3), **I (78)** | **I (61)** | L (8) | L (1), **I (72)** |
| A421V | A (505), **V (73),** M (1) | A (396), **V (21)** | A (23), **V (98)** | A (1), **V (80)** | A (5), **V (57)** | A (5), **V (3)** | A (1), **V (72)** |
| R422K | R (576), **K (3)** | R (417) | R (122) | R (81) | R (62) | R (8) | R (73) |
| M423T/V/I/A | M (561), **I (13), V (3), T (1), A (1)** | M (423) | M (122) | M (79), **I (2)** | M (62) | **I (8)** | M (73) |
| I424V | I (560), **V (19)** | I (381), **V (35)** | **V (122)** | **V (81)** | I (16), **V (54)** | **V (8)** | I (1), **V (72)** |
| M426T/V/I | M (531), L (41), F (1), **T (1), I (1)** | M (405), L (10), A (1), **T (1)** | M (120), L (2) | M (75), L (4), **V (1), I (1)** | M (61) | **T (1)** | M (63), C (10) |
| A442T | A (579) | A (405), **T (10),** V (2) | N (101), D (19), S (2) | A (23), P (54), S (4) | A (62) | A (7), **T (1)** | A (49), V (19), P (3) |
| C445F | C (573), G (1), Y (1) | C (411), **F (4),** W (1) | **F (122)** | **F (81)** | **F (62)** | **F (8)** | **F (73)** |
| I447F | I (578), **F (1)** | I (447) | M (118), L (4) | M (81) | M (62) | M (8) | I (39), M (32), L (1) |
| Y448H/C | Y (579) | Y (416), **C (1)** | Y (122) | Y (81) | Y (62) | Y (8) | Y (73) |
| C451R | C (572), Y (4), H (2), W (1) | C (333), T (46), Y (13), H (7), I (7), N (3), S (3), V (2), **R (1)** | V (120), A (1), I (1) | T (80), V (1) | T (62) | V (8) | T (73) |
| Y452H | Y (577), **H (2)** | Y (410), **H (7)** | Y (122) | Y (81) | Y (62) | Y (8) | Y (72), **H (1)** |
| C455F | E (567), Q (6), K (5) | E (412), Q (4), G (1) | C (1), N (87), S (30), T (4) | T (81) | T (62) | T (8) | C (1), T (43), S (24), N (3), A (2) |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M/I462T | I (578), L (1) | I (416), V (1) | I (119), V (3) | I (80) | I (62) | R (8) | I (72), V (1) |
| R465G | R (575), K (1) | R (416), L (1) | R (122) | R (80) | R (58), K (4) | R (8) | R (73) |
| I482L/V/T/S | I (564), **L (1)** | I (411), **L (1)** | **L (119)** | **L (75)** | **L (60)** | I (8) | **L (72)** |
| A486V/I/T/M | A (563), D (1) | A (411) | A (119) | A (73), S (2) | A (55), S (1) | A (7), S (19) | A (34), G (38) |
| V494A/I | V (562), **I (1)** | V (411), **A (1)** | **A (119)** | C (56), **I (14)**, **A (3)**, S (2) | V (56), **A (2)** | V (8) | V (14), **A (57)**, M (1) |
| P495L/A/S/T/Q | P (564) | P (412) | P (119) | P (75) | P (57) | P (8) | P (71), **L (1)** |
| P496A/T/S | P (563) | P (412) | P (118), **S (1)** | P (75) | P (58) | P (8) | P (72) |
| V499A | V (4), **A (545)**, T (15) | V (364), **A (41)**, T (6), I (1) | V (9), **A (106)**, T (3), M (1) | V (1), **A (74)** | V (1), **A (57)** | **A (8)** | **A (72)** |
| R531K | R (467), K (84) | R (275), **K (136)** | R (15), **K (102)** | R (59) | **K (53)** | **K (8)** | R (32), **K (40)** |
| G554D/S | G (548) | G (402), **D (1)** | G (109), **S (9)** | G (57), **S (2)** | G (52) | G (8) | G (72) |
| Y555C | Y (548) | Y (411) | A (111), S (7) | V (51), I (4), A (3), G (1) | A (44), S (4), G (3) | A (8) | Y (34), G (19), F (17), V (1) |
| S556G/N/C | S (543), **G (4)**, **N (1)** | S (373), **G (30)**, **N (6)**, D (1) | **G (117)**, **C (1)** | S (1), **G (58)** | **G (49)**, **N (2)**, A (1) | **G (8)** | S (41), D (18), R (11), **G (2)** |
| G558R | G (548) | G (405) | G (114), S (4) | G (3), N (55), S (1) | G (51), E (1) | G (8) | G (70), A (1), V (1) |
| D559G/S/N | D (548) | D (403), **N (1)** | D (117), H (1) | D (59) | D (52) | D (8) | D (71), **G (1)** |
| W571R | W (545), **R (2)** | W (402), L (1), **R (1)** | L (101), S (6), F (4), Y (3), I (3) | H (39), Y (16), L (1), Q (1) | Y (49), F (1), H (1) | Y (5), N (1), C (1) | M (24), L (21), I (12), T (6), V (3), F (3), N (1) |

[a] RAVs are indicated in bold.

sequences presented haplotypes with at least two natural RAVs, NS3-(Q80K+122N)+NS5A-(58T) and NS3-(36L+122T) (17.8% and 8.8% of the total haplotypes, respectively).

**Genetic barriers.** The genetic barrier for the evolution of DAA resistance was calculated for the analyzed data sets of each gene and HCV genotype. For each original codon with no resistance, the minimal genetic barrier was defined as the lowest score obtained from the total number of transitions (score = 1) and/or transversions (score = 2.5) needed to generate a given RAV. Genetic barrier scores were classified as "intermediate" for values lower than 5.0 but equal to or higher than 3.0, while values equal to or higher than 5.0 were classified as "high." Thus, both intermediate and high genetic barrier scores necessarily imply at least two nucleotide changes. The results are shown in Table S6 in the supplemental material. Considering the most frequent nonresistant codon for each position and GT, four positions in NS3, two in NS5A, and seven in NS5B presented very low genetic barriers. For NS3, the following codons had minimal scores of 1.0: V36(GTG) in GT1a; V55(GTC) in GTs 1a, 1b, 5, and 6; V55(GTT) in GT3; V55(GTA) in GT2; R155(AGA) in GT2; R155(AGG) in GTs 1a, 3, and 5; and V170(GTG) in GTs 4 and 6 or V170(GTA) in GT1b. Only for positions NS3-L36(CTT) in GT3, R155(CGC) in GTs 4 and 6, and Q168(CAG) in GT3 was a high genetic barrier evident (L36M score = 5; R155K score = 6; and Q168A/E/V score = 5), and not for all possible RAVs. Thus, most positions presented low or intermediate genetic barrier scores. A similar pattern was found among less frequent codons in all genotypes. For NS5A, two positions presented very low genetic barrier scores: M28(ATG) in GTs 1a and 3 and Y93(TAC) in GTs 1a, 1b, 2, 3, and 4. On the other hand, NS5A-T93(ACA) in GT5 and T93(ACC) in GT6 presented high genetic barriers toward all possible RAVs (all minimal scores were ≥5). Most codon variants found at lower frequencies also showed low or intermediate genetic barriers for all GTs. For NS5B, the following codons presented low genetic barrier scores for all possible RAVs: L159(CTT) in GT1b and L159(CTC) in GTs 1a, 2, and 3; L320(CTT) in GTs 1b and 5; M414(ATG) in GTs 1a, 1b, 3, 5, and 6; Y448(TAC) in GTs 1 to 6; Y448(TAT) in GT5; G554(GGT) in GT1b; G554(GGC) in GTs 1a, 2, 3, 4, 5, and 6; S556(AGC) in GTs 1a, 1b, and 6; and A559(GAC) in HCV GTs 1 to 6. In all HCV GTs, the generation of S282T was associated with a low genetic barrier score, requiring only one substitution but at the second base of the codon (see Table S4 in the supplemental material). Other highly frequent codons, such as C316(TGT) in GTs 1a, 2, and 5; P495(CCT) in GTs 2 and 6; and P495(CCC) in GTs 3, 4, and 5, presented high genetic barriers for change to some RAVs but low barriers for others. A high genetic barrier (score = 5) was found only for Q414(CAA), a polymorphism particularly frequent in genotype 2 isolates. Overall, most of the codon variants present at lower frequencies also showed low or intermediate genetic barriers to generate RAVs in all HCV GTs.

Finally, when analyzing each HCV genotype/subtype independently, the most frequent nonresistant codons from NS3, NS5A, and NS5B presented, on average, low genetic barrier scores. However, several differences in genetic barriers were observed between genotypes/subtypes. For NS3-36, the genetic barrier is very low in GT1a (V36A/M score = 1) but high in GT3 (L30A, score = 3.5; L36M, score = 5) isolates. Similarly, for NS3-R155, the genetic barrier is very low for GTs 1a (as previously known), 2, 3, and 5 (score = 1) but high for GTs 1b, 4, and 6 (score = 6). For NS5A

**TABLE 5** Majority amino acids and signatures of HCV genotypes 1 to 6 at sites associated with resistance to NS3 inhibitors

| NS3 position | Amino acid for sequence or genotype[a]: Con1 reference (1b) | GT1 consensus amino acid | All GT1 | 1a | 1b | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| C16S | C | C | ● | ● | ● | A | T* | T* | A* | T* |
| V36A/M/L/G/I/C | V | V | ● | ● | ● | L | L* | L* | L* | ● |
| A39V | A | A | ● | ● | ● | **V** | ● | ● | ● | ● |
| Q41R/K/P/H | Q | Q | ● | ● | ● | ● | ● | ● | ● | ● |
| F43S/C/Y/V/I/L | F | F | ● | ● | ● | ● | ● | ● | ● | ● |
| I48V | V | I | I | I | ● | I | ● | ● | I | I |
| T54A/S/V/G/C | T | T | ● | ● | ● | ● | ● | ● | ● | ● |
| V55A/F/I/K/T | V | V | ● | ● | ● | ● | ● | ● | ● | ● |
| D79E | D | D | ● | ● | ● | **E** | ● | ● | ● | ● |
| Q80K/L/N/R/H/G | Q | Q | ● | ● | ● | **G*** | ● | ● | **K*** | ● |
| A87T | A | A | ● | ● | ● | S | ● | ● | S | ● |
| Y105C | Y | Y | ● | ● | ● | ● | ● | **F** | ● | ● |
| R109K | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| R117H | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| S122G/A/R/N/T | S | S | ● | ● | ● | **R** | ● | T | T | T (<50%) |
| R123H/K | R | R | ● | ● | ● | ● | **T*** | ● | ● | ● |
| I132V | V | I | I | I | ● | L | L | I | I | I |
| S138T/D/P | S | S | ● | ● | ● | ● | ● | ● | ● | ● |
| R155K/T/I/M/G/L/S/Q/P/N/W | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| A156S/T/V/I/F/N/G/D | A | A | ● | ● | ● | ● | ● | ● | ● | ● |
| V158I/M | V | V | ● | ● | ● | ● | ● | ● | ● | ● |
| V163L | V | V | ● | ● | ● | ● | ● | ● | ● | ● |
| D168Q/A/Y/V/E/T/N/P/I/H/G/F/S/K | D | D | ● | ● | ● | ● | **Q*?** | ● | ● | ● |
| V170A/T/G/L/M | V | I | I | I | ● | I | I | ● | I | ● |
| E173G | E | E | ● | ● | ● | ● | ● | ● | ● | ● |
| S174F/P | S | S | ● | N | ● | ● | **T** | ● | N | N |
| M175L | M | L | L | L*? | ● | L | L* | L* | L* | ● |
| E176G | E | E | ● | ● | ● | **D** | **S** | ● | ● | ● |

[a] Symbols: ●, amino acid identical to the Con1 prototype replicon (HCV-1b); *, amino acid present in 100% of the sequences for the given HCV genotype; *?, amino acid present in 100% of the sequences for the given HCV genotype, with the exception of ambiguities. Amino acids without asterisks are present in the majority, but not all, of the sequences for the given HCV genotype. Amino acids unique for the given HCV genotype are presented in bold.

Y93H/C, GTs 1a and 1b showed a low genetic barrier in line with published clinical data (41), together with GT2, GT3, and GT4 isolates (score = 1). In contrast, GT5 and GT6 isolates showed scores of ≥5. Finally, for NS5B-M414, the genetic barrier was high only for GT2 isolates (score = 5) and low for GTs 1a, 1b, 3, 5, and 6 (score = 1). There were no differences in genetic barriers among HCV genotypes/subtypes for NS5B-S282 (all scores = 2.5).

**Evolutionary history of natural RAVs.** To find out whether naturally occurring RAVs are associated with ongoing random variation and/or founder effects in transmission clusters, the evolutionary history of these RAVs was traced by means of parsimony analyses, as implemented in Mesquite, using the phylogenetic trees obtained for each HCV variant and genomic region analyzed. Most natural RAVs seem to have occurred independently, although some grouped in well-supported clades (Fig. 1). The most significant cases (Fig. 1) are NS3-80K (GTs 1a and 6), NS5A-54H (GT1b) and 153L (GT1b), and NS5B-316N (GT1b) and 392I (GT2). Other RAVs forming well-supported clades are shown in Fig. S1 in the supplemental material: NS3-122G/T (GT1b) and 48V (GT6); NS5A-28M (GT1b), 31M (GT3), and 153L (GT3); and NS5B-414G/T (GT4). The same tree topologies were obtained when phylogenies were reconstructed from alignments without the codon positions implicated in resistance to DAAs (data not shown).

## DISCUSSION

This study provides a comprehensive view on naturally occurring RAVs in all sequences publicly available to date for HCV genotypes 1 to 6. We obtained 2,901, 2,216, and 1,344 HCV sequences from HCV isolates from DAA-naive patients for NS3, NS5A, and NS5B, respectively, and they were screened for 28 NS3, 17 NS5A, and 58 NS5B relevant RAVs to compare their relative frequencies in the different HCV genotypes/subtypes. In addition, for the 1,126 HCV isolates in which sequences from the three genomic regions were available, we estimated the frequency of haplotypes carrying multiple RAVs to the four different classes of approved DAAs. Hence, to our knowledge, the current study represents the most comprehensive data set and analysis of naturally occurring HCV resistance to DAAs.

One of the most important findings is the remarkable difference observed in the prevalence of natural RAVs among HCV genotypes, with some RAVs being found in all isolates of a given genotype. First, for NS3, we found differences between genotypes for positions 36, 80, 117, 122, 168, and 175, in line with our previous results with a smaller data set (13) and with those observed by others (15, 16, 21, 22). Second, for NS5A, we identified relevant differences in the distribution of RAVs at positions 28, 30, 31, 58, 62, 93, and 153, particularly between GT1a and GT1b. Variation

**TABLE 6** Majority amino acids and signatures of HCV genotypes 1 to 6 at sites associated with resistance to NS5a inhibitors

| NS5a position | Amino acid for sequence or genotype[a]: Con1 reference (1b) | GT1 consensus amino acid | All GT1 | 1a | 1b | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| L23F | L | L | ● | ● | ● | ● | ● | ● | ● | ● |
| Q24L | Q | K | K | K | ● | S | S | K* | ● | K |
| L28M/T | L | M | M | M | ● | ● | M | ● | ● | **V (<50%)** |
| R30Q/L | R | Q | Q | Q | ● | **K** | **A** | ● | Q | **S (<50%)** |
| L31M/V/F | L | L | ● | ● | ● | M | ● | M | ● | ● |
| P32L | P | P | ● | ● | ● | ● | ● | ● | ● | ● |
| Q54H/N/L/Y | Q | H | H | H | ● | **T** | S | H* | S | H |
| P58S/A/L/T/H/D/P | P | H | H | H (49.5%) | **H** | ● | ● | ● | ● | ● |
| Q62E/R/A/P/S | Q | E | E | E | ● | **N** | **S** | E | **T** | **V** |
| A92T | A | A | ● | ● | ● | **C** | **E** | ● | ● | ● |
| Y93H/N/C | Y | Y | ● | ● | ● | ● | ● | ● | T* | T |
| F149L | F | F | ● | ● | ● | ● | ● | ● | ● | ● |
| V153M/L/I | V | L | L | L | ● | ● | ● | ● | L | ● |
| M202L | M | M | ● | ● | ● | ● | ● | ● | ● | ● |
| M265V/T | M | M | ● | ● | ● | ● | ● | ● | ● | ● |
| D320E | D | D | ● | ● | ● | ● | ● | ● | **G*** | ● |
| Y321N | Y | Y | ● | ● | ● | ● | ● | ● | ● | ● |

[a] Symbols: ●, amino acid identical to the Con1 prototype replicon (HCV-1b); *, amino acid present in 100% of the sequences for the given HCV genotype. Amino acids without asterisks are present in the majority, but not all, of the sequences for the given HCV genotype. Amino acids unique for the given HCV genotype are presented in bold.

in NS5A has been previously analyzed in 31 GT1a and 30 GT1b clinical isolates (17), finding similar frequencies for L28V in GT1a and L28M/V in GT1b. However, we report here a much lower prevalence of L31M in worldwide GT1a isolates (0.93%) and a higher or lower prevalence of Y93H/C in GT1a or GT1b isolates (0.70% versus 5.06%, respectively). Furthermore, while Paolucci et al. (17) did not find natural RAVs for NS5A inhibitors, we found up to 5.1% of HCV-1b isolates harboring RAVs. Third, NS5B RAVs to NNIs were frequent in positions 316, 392, 414, and 556 for certain HCV genotypes. For positions NS5B-316 and 556, our results are similar to those obtained previously in a smaller data set (20). However, we identified higher frequencies of RAVs at NS5B-414 in GT2 and GT4 isolates, probably because we considered M414Q/L (highly prevalent in these genotypes) a RAV. Other previous studies, analyzing a very limited number of isolates, also reported M414Q as a very frequent polymorphism in HCV GTs 2 and 3 (42). Finally, the most clinically relevant RAVs to NIs (L159F and S282T) were very infrequent (4.80% of GT1b isolates and 0.17%, 0.24%, and 1.24% of isolates from GTs 1a, 3, and 4, respectively). This near-absence of intergenotypic variation of NS5B-S282 suggests that NIs may indeed warrant their efficacy across genotypes, but the natural presence of the compensatory RAV L159F in a small proportion of GT1b isolates deserves further investigation. For other HCV genotypes,

**TABLE 7** Majority amino acids and signatures of HCV genotypes 1 to 6 at sites associated with resistance to NS5B NIs

| NS5B NI resistance position | Amino acid for sequence or genotype[a]: Con1 reference (1b) | GT1 consensus amino acid | All GT1 | 1a | 1b | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| A15G | A | A | ● | ● | ● | **G** | S | ● | S* | ● |
| K72M | K | K | ● | ● | ● | ● | ● | ● | ● | ● |
| S96T | S | S | ● | ● | ● | ● | ● | ● | ● | ● |
| N142T | N | N | ● | ● | ● | ● | ● | ● | ● | ● |
| L159F | L | L | ● | ● | ● | ● | ● | ● | ● | ● |
| R222Q | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| C223H/Y | C | C | ● | ● | ● | ● | ● | ● | ● | ● |
| I239V/L | I | I | ● | ● | ● | ● | ● | **V** | ● | ● |
| S282T/R | S | S | ● | ● | ● | ● | ● | ● | ● | ● |
| A300T | A | S (<30%) | S (<30%) | Q | **S** | L | T | T | L* | Q |
| L320F/I | L | L | ● | ● | ● | ● | ● | ● | ● | ● |
| V321I | V | V | ● | ● | ● | ● | ● | ● | ● | ● |
| A396G | A | A | ● | ● | ● | ● | ● | ● | ● | ● |
| Y586C | Y | Y | ● | ● | ● | F | F | F*# | F*# | F |

[a] Symbols: ●, amino acid identical to the Con1 prototype replicon (HCV-1b); *, amino acid present in 100% of the sequences for the given HCV genotype; *#, amino acid present in 100% of the sequences for the given HCV genotype, with the exception of gaps. Amino acids without asterisks are present in the majority, but not all, of the sequences for the given HCV genotype. Amino acids unique for the given HCV genotype are presented in bold.

**TABLE 8** Majority amino acids and signatures of HCV genotypes 1 to 6 at sites associated with resistance to NS5B NNIs

| NS5B NNI resistance position | Amino acid for sequence or genotype[a]: Con1 reference (1b) | GT1 consensus amino acid | All GT1 | 1a | 1b | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|
| T19S/P | T | Q | Q | **Q** | S | E | E* | S | E* | E* |
| K50R | K | K | ● | ● | ● | ● | ● | ● | ● | ● |
| D55E | D | D | ● | ● | ● | ● | ● | ● | ● | ● |
| M71V | M | V | V | V | ● | V | V | I | ● | I |
| H95Q/R | H | H | ● | ● | ● | ● | ● | ● | ● | ● |
| 138I | I | I | ● | ● | ● | ● | ● | ● | ● | ● |
| L314F | L | L | ● | ● | ● | ● | ● | ● | ● | ● |
| C316Y/F/N/S | C | C | ● | ● | ● | ● | ● | ● | ● | ● |
| A338V | A | A | ● | ● | **V** | ● | ● | ● | ● | ● |
| I363V | I | I | ● | ● | ● | ● | ● | ● | **V*** | ● |
| S365T/A/L/O/F | S | S | ● | ● | ● | ● | ● | ● | ● | ● |
| S368A/T | S | S | ● | ● | ● | ● | ● | ● | ● | ● |
| T389S/A | T | T | ● | ● | ● | ● | ● | **E*** | **Q*** | ● |
| L392I | L | L | ● | ● | ● | **I** | ● | ● | ● | ● |
| N411S | N | N | ● | ● | ● | ● | ● | ● | ● | ● |
| M414L/T/I/V/Q | M | M | ● | ● | ● | **Q** | ● | **L** | ● | ● |
| L419M/V/S/I | L | L | ● | ● | ● | I | I | I*? | ● | I |
| A421V | A | A | ● | ● | ● | V | V | V | ● | V |
| R422K | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| M423T/V/I/A | M | M | ● | ● | ● | ● | ● | ● | **I*** | ● |
| I424V | I | I | ● | ● | ● | V* | V* | V | V* | V |
| M426T/V/I | M | M | ● | ● | ● | ● | ● | ● | ● | ● |
| A442T | A | A | ● | ● | ● | **N** | **P** | ● | ● | ● |
| C445F | C | C | ● | ● | ● | F* | F* | F* | F* | F* |
| I447F | I | I | ● | ● | ● | M | M* | M* | M* | ● |
| Y448H/C | Y | Y | ● | ● | ● | ● | ● | ● | ● | ● |
| C451R | C | C | ● | ● | ● | V | T | T* | V* | T* |
| Y452H | Y | Y | ● | ● | ● | ● | ● | ● | ● | ● |
| C455F | E | E | ● | ● | ● | **N** | T* | T* | T* | T |
| M/I462T | I | I | ● | ● | ● | ● | ● | ● | ● | ● |
| R465G | R | R | ● | ● | ● | ● | ● | ● | ● | ● |
| I482L/V/T/S | I | I | ● | ● | ● | L*# | L*# | L*# | ● | L*# |
| A486V/I/T/M | A | A | ● | ● | ● | ● | ● | ● | ● | **G** |
| V494A/I | V | V | ● | ● | ● | A*# | **C** | ● | ● | A |
| P495L/A/S/T/Q | P | P | ● | ● | ● | ● | ● | ● | ● | ● |
| P496A/T/S | P | P | ● | ● | ● | ● | ● | ● | ● | ● |
| V499A | V | A | A | A | ● | A | A | A | A | A*# |
| R531K | R | R | ● | ● | ● | K | ● | K*# | K* | K |
| G554D/S | G | G | ● | ● | ● | ● | ● | ● | ● | ● |
| Y555C | Y | Y | ● | ● | ● | A | **V** | A | A* | ● |
| S556G/N/C | S | S | ● | ● | ● | G | G | G | G* | ● |
| G558R | G | G | ● | ● | ● | ● | **N** | ● | ● | ● |
| D559G/S/N | D | D | ● | ● | ● | ● | ● | ● | ● | ● |
| W571R | W | W | ● | ● | ● | L | **H** | Y | Y | **M (<50%)** |

[a] Symbols: ●, amino acid identical to the Con1 prototype replicon (HCV-1b); *, amino acid present in 100% of the sequences for the given HCV genotype; *#, amino acid present in 100% of the sequences for the given HCV genotype, with the exception of gaps; *?, amino acid present in 100% of the sequences for the given HCV genotype, with the exception of ambiguities. Amino acids without asterisks are present in the majority, but not all, of the sequences for the given HCV genotype. Amino acids unique for the given HCV genotype are presented in bold.

NS5B RAVs identified in HCV replicon assays (A15G and I239V) showed a significant prevalence in GT2 and GTs 4 and 6, respectively, but their potential clinical significance is unclear.

A second important finding is the presence of multiple RAVs to several DAAs in HCV isolates from DAA-naive patients, irrespective of the viral genotype. Considering those haplotypes with more than one RAV, the most frequent for GT1a showed at least two RAVs at NS3, but for the rest of the HCV genotypes, finding three or more RAVs involving at least two different genes was very common. These results indicate intergenotypic differences in the frequency and combinations of natural RAVs that may have potential clinical implications. However, the real levels of resistance conferred by these natural RAVs should be determined with specific assays for each genotype/subtype and not only with HCV-1 replicons (see below). Despite such differences, the most common

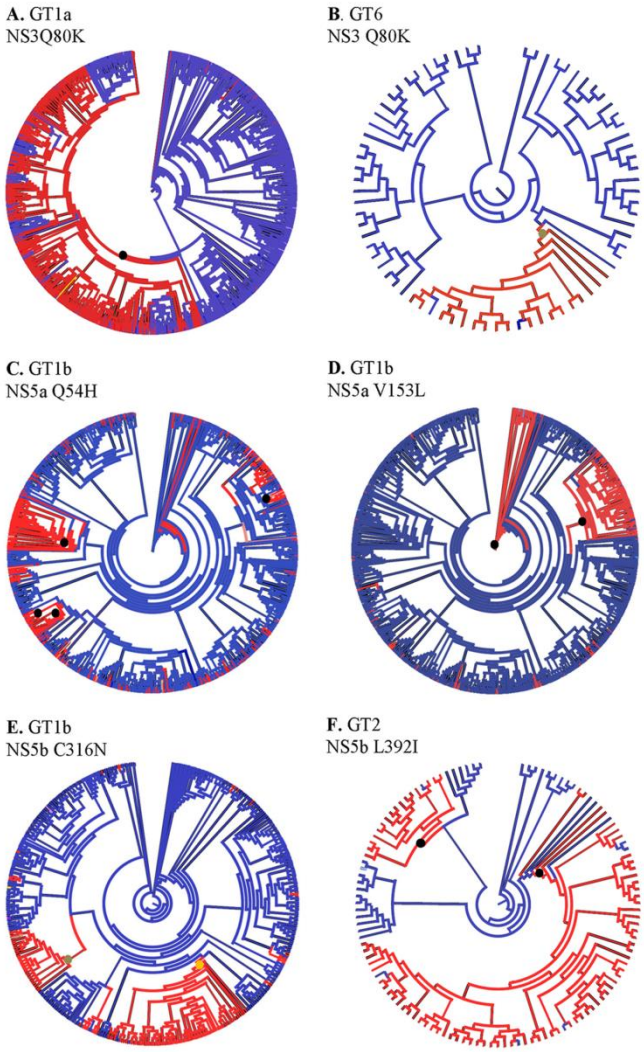



**FIG 1** Dendrograms representing the evolutionary history of the six most representative resistance mutations forming well-supported clades (A to F). Blue lines represent nonresistant lineages; red lines represent lineages carrying the resistance mutation; black dots represent the emergence of well-supported clades of resistance (aLRT, >0.85). In panel B, the node including the HCV-6a clade is indicated in green. In panel E, nodes including the Asian (yellow) and U.S. (green) clades are indicated.

combinations of natural RAVs included NS3-36L+175L and/or 80G/K (GTs 1a, 2, 3, 4, and 5), NS3-80G/K+122G/R/T/N (GTs 2, 5, and 6), and NS3-122T/R+NS5B-556G and/or 175L (GTs 2, 4, and 5). While the selection of several combinations of RAVs during DAA treatment has been reported previously for TPV (NS3-36M+155K), SMV (NS3-36M+155K; 80L+155K; 122R+155K; 155K+168E; 80R+168E), and MCB (NS5B-159F+320F) (38), none of these combinations was found in any of the HCV isolates from DAA-naive patients analyzed here. Collectively, these observations suggest that natural RAV profiles may be different from those induced during treatment with DAAs.

The third important finding of our study is that several

RAVs previously characterized for HCV GT1 (either in replicon assays or during DAA treatment failure) were present as majority or signature (unique) amino acids for other viral genotypes (detailed in Tables 5 to 8; see also Table S4 in the supplemental material). Further, we found several positions with differential signature/majority amino acid patterns between subtypes of the same GTs. The potential impact of these genotype/subtype-specific natural RAVs on the pan-genotypic efficacy of new interferon-free DAA regimens remains to be determined and deserves further research. An important limitation of our study is that the RAVs investigated here have been mainly characterized for HCV GT1 and might not necessarily confer resistance in other HCV genotypes. Previous work reported reduced DAA sensitivity in isolates from non-1 HCV genotypes carrying some of the consensus and/or signature amino acids identified here. Clear examples are NS3-Q80K for GT1a, NS3-Q80G for GT2, and NS3-D168Q for GT3 (13, 16, 43). However, this may not always be the case: Y93H confers several-thousandfold resistance to NS5A inhibitors in GT1 isolates, but GT5 and GT6 isolates with the naturally occurring variant Y93T do not show a great degree of natural resistance, and isolates with T93H show only low levels of resistance (44). Because the information available on specific RAVs for genotype 2 to 6 isolates (either *in vitro* in replicon assays or *in vivo*) is still limited, this issue remains beyond the scope of our study.

Nevertheless, we were able to identify several RAVs in non-GT1 isolates that are known to confer resistance in chimeric HCV replicons for the corresponding genotype (see Table S1 in the supplemental material), such as NS3-D/Q168E for GTs 5 and 6, NS5A-L28M/F/V for GTs 2 and 4, L/A30K/H and Y93H for GT3, and NS5B-L329I and V494 for GT2. Our results strongly suggest that DAA resistance profiles may be helpful for the choice of the most appropriate DAA combination for each particular HCV genotype/subtype. Thus, the identification of RAV profiles for GTs 2 to 6 *in vitro* using genotype/subtype-specific assays is urgently needed, together with their correlation with *in vivo* data from the increasing number of patients infected with HCV GTs 2 to 6 being treated with DAAs.

Given the high rates of evolution of HCV, the development of resistance during treatment is inevitable. Strikingly, our results show that most HCV isolates have a low genetic barrier toward mutations to generate RAVs to different DAAs, regardless of the HCV genotype. These results may seem difficult to reconcile with the high efficacy of DAA combinations but are in line with the clinical observations that SVR rates have been increasing when more DAAs are being included in treatment regimens. All the studies on DAAs administered as monotherapy showed the rapid selection of RAVs (45). While the combination of one DAA with P/R marked a milestone in achieving SVR rates higher than 70%, not until the use of two or more DAAs in combination did SVR rates reach 85 to 90%, presumably because of the continuous selection of RAVs in amino acid sites with low genetic barriers in single-DAA regimens. While this may hold true for PIs and NNIs, it is probably not that relevant for combinations with highly potent NS5A inhibitors and/or those with high clinical genetic barriers (NIs such as SOF or MCB). During treatment with these highly potent DAAs that immediately reduce viral replication, the virus probably has no capacity to generate and allow the spread of new RAVs emerging as minor variants in the viral population. However, a different question is whether DAA escape occurs because a natural RAV preexists as a predominant variant within an individual. Despite the high natural variability of HCV, SVR rates observed in pivotal studies are much higher than 90% for HCV GT1, but as different IFN-free DAA combinations are generalized for treatment of non-1 HCV genotypes, natural resistance/susceptibility profiling may still be useful for choosing the best drug regimens and for retreatment of relapses.

Finally, from a public health perspective it is important to investigate whether naturally occurring RAVs appear randomly without antiviral pressure, correspond to old transmission clusters, and/or share a geographical distribution pattern. Using phylogenetic analyses, we revealed that most natural RAVs appeared *de novo*, as they are found in external branches. However, in some cases such mutations were clearly differentiated in well-supported clades. It has been suggested recently that the majority of GT1a infections carrying the 80K variant (associated with resistance to SMV) were transmitted from a single origin (21). Our analyses further support their results and shed more light on the evolution of NS3-Q80K. We found that this variant in GT6 occurs only in subtype 6a, being present in 18 out of 20 of all GT-6a sequences, 13 of them from Hong Kong. In addition, we identified other well-differentiated clades grouping different RAVs, NS5A-153L and NS5B-316N in GT1b isolates. It is worth noting that our phylogenetic analyses support the existence of two distinct GT1b clades carrying NS5B-316N: one clade from Asian isolates (Japanese and Chinese sequences) and a second one from the United States. For these highly prevalent natural RAVs, surveillance may be useful to asses their potential impact on the efficacy of some of the new IFN-free DAA regimens containing NNIs. In addition, if proven clinically relevant, these natural RAVs also deserve attention for their potential to be transmitted in small local epidemics associated with risk behaviors. In addition, and despite the high efficacy of new DAA regimens shown in pivotal studies, in actual clinical practice a small but significant number of patients may fail DAA treatment and potentially transmit selected DAA-resistant HCV variants, which need to be identified and characterized.

In conclusion, our comprehensive analysis of natural HCV polymorphisms associated with resistance to DAAs indicates that natural RAVs are relatively common in viral isolates from treatment-naive patients, with frequencies clearly varying among HCV genotypes/subtypes. Importantly, for any HCV genotype, there is a significant proportion of viral isolates carrying at least two RAVs, and we identified distinct genotypic profiles for HCV natural resistance to DAAs. Furthermore, we observed that, for any HCV genotype, most of the more prevalent wild-type codons present low genetic barriers for the generation of new RAVs, emphasizing the importance of using at least one highly potent drug as a backbone in IFN-free DAA combinations. While our evolutionary analyses showed that the majority of naturally occurring RAVs are probably being generated at random and very recently, a few natural RAVs are associated with long-lasting, geographically delimited successful transmission with a high potential to keep being further transmitted over time.

## FUNDING INFORMATION

Ministerio de Economía y Competitividad (MINECO) | Instituto de Salud Carlos III (ISCIII) provided funding to F. Xavier Lopez-Labrador under grant number PI10/00512. Ministerio de Economía y Competitividad (MINECO) | Instituto de Salud Carlos III (ISCIII) provided funding to F. Xavier Lopez-Labrador under grant number PI10/01734. Ministerio de Economía y Competitividad (MINECO) provided funding to Fernando González-Candelas under grant number BFU2011-24112. Ministerio de Economía y Competitividad (MINECO) provided funding to Fernando González-Candelas under grant number BFU2014-58656R. Generalitat Valenciana (Regional Government of Valencia) provided funding to F. Xavier Lopez-Labrador under grant number FPA/2013/A/083. Generalitat Valenciana (Regional Government of Valencia) provided funding to F. Xavier Lopez-Labrador under grant number ACOMP/2013/086. Ministerio de Educación, Cultura y Deporte (MECD) provided funding to Juan Angel Patiño-Galindo under grant number FPU-AP2010-0561.

## REFERENCES

1. **Hajarizadeh B, Grebely J, Dore GJ.** 2013. Epidemiology and natural history of HCV infection. Nat Rev Gastroenterol Hepatol **10:**553–562. http://dx.doi.org/10.1038/nrgastro.2013.107.
2. **Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST.** 2013. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology **57:**1333–1342. http://dx.doi.org/10.1002/hep.26141.
3. **Smith DB, Bukh J, Kuiken C, Muerhoff AS, Rice CM, Stapleton JT, Simmonds P.** 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. Hepatology **59:**318–327. http://dx.doi.org/10.1002/hep.26744.
4. **Bukh J.** 1995. Genetic heterogeneity of hepatitis C virus: quasispecies and genotypes. Semin Liver Dis **15:**41–63. http://dx.doi.org/10.1055/s-2007-1007262.
5. **Okamoto H, Kojima M, Okada S, Yoshizwa H, Iizuka H, Tanaka T, Muchmore EE, Peterson DA, Ito Y, Mishiro S.** 1992. Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. Virology **190:**894–899. http://dx.doi.org/10.1016/0042-6822(92)90933-G.
6. **Mangia A, Santoro R, Minerva N, Ricci GL, Caretta V, Persico M, Vinelli F, Scotto G, Bacca D, Annese M, Romano M, Zechini F, Sogari F, Spirito F, Andriulli A.** 2005. Peginterferon alfa-2b and ribavirin for 12 vs. 24 weeks in HCV genotype 2 or 3. N Engl J Med **352:**2609–2617. http://dx.doi.org/10.1056/NEJMoa042608.
7. **Pang PS, Planet PJ, Glenn JS.** 2009. The evolution of the major hepatitis C genotypes correlates with clinical response to interferon therapy. PLoS One **4:**e6579. http://dx.doi.org/10.1371/journal.pone.0006579.
8. **Zein NN.** 2000. Clinical significance of hepatitis C virus genotypes. Clin Microbiol Rev **13:**223–235. http://dx.doi.org/10.1128/CMR.13.2.223-235.2000.
9. **AASLD/ISDA HCV Guidance Panel.** 2015. Hepatitis C guidance: AASLD-ISDA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology **62:**932–954. http://dx.doi.org/10.1002/hep.27950.
10. **European Association for the Study of the Liver.** 2015. EASL recommendations on treatment of hepatitis C 2015. J Hepatol **63:**199–236. http://dx.doi.org/10.1016/j.jhep.2015.03.025.
11. **Bartenschlager R, Lohmann V, Penin F.** 2013. The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection. Nat Rev Microbiol **11:**482–496. http://dx.doi.org/10.1038/nrmicro3046.
12. **Asselah T, Boyer N, Saadoun D, Martinot-Peignoux M, Marcellin P.** 2016. Direct-acting antivirals for the treatment of hepatitis C virus infection: optimizing current IFN-free treatment and future perspectives. Liver Int **36:**47–57. http://dx.doi.org/10.1111/liv.13027.
13. **López-Labrador FX, Moya A, González-Candelas F.** 2008. Mapping natural polymorphisms of hepatitis C virus NS3/4A protease and antiviral resistance to inhibitors in worldwide isolates. Antivir Ther **13:**481–494.
14. **Bartels DJ, Zhou Y, Zhang EZ, Marcial M, Byrn RA, Pfeiffer T, Tigges AM, Adiwijaya BS, Lin C, Kwong AD, Kieffer TL.** 2008. Natural prevalence of hepatitis C virus variants with decreased sensitivity to NS3-4A protease inhibitors in treatment-naive subjects. J Infect Dis **198:**800–807. http://dx.doi.org/10.1086/591141.
15. **Paolucci S, Fiorina L, Piralla A, Gulminetti R, Novati S, Barbarini G, Sacchi P, Gatti M, Dossena L, Baldanti F.** 2012. Naturally occurring mutations to HCV protease inhibitors in treatment-naïve patients. Virol J **9:**245. http://dx.doi.org/10.1186/1743-422X-9-245.
16. **Alves R, Queiro AT, Pessoa MG, da Silva EF, Mazo DF, Carrilho FJ, Carvalho-Filho RJ, de Carvalho IM.** 2013. The presence of resistance mutations to protease and polymerase inhibitors in hepatitis C virus sequences from the Los Alamos databank. J Viral Hepat **20:**414–421. http://dx.doi.org/10.1111/jvh.12051.
17. **Paolucci S, Fiorina L, Mariani B, Gulminetti R, Novati S, Barbarini G, Bruno R, Baldanti F.** 2013. Naturally occurring resistance mutations to inhibitors of HCV NS5A region and NS5B polymerase in DAA treatment-naïve patients. Virol J **10:**355. http://dx.doi.org/10.1186/1743-422X-10-355.
18. **de Carvalho IM, Alves R, de Souza PA, da Silva EF, Mazo D, Carrilho FJ, Queiroz AT, Pessoa MG.** 2014. Protease inhibitor resistance mutations in untreated Brazilian patients infected with HCV: novel insights about targeted genotyping approaches. J Med Virol **86:**1714–1721. http://dx.doi.org/10.1002/jmv.24015.
19. **Margeridon-Thermet S, Le Pogam S, Li L, Liu TF, Shulman N, Shafer RW, Najera I.** 2014. Similar prevalence of low-abundance drug-resistant variants in treatment-naive patients with genotype 1a and 1b hepatitis C virus infections as determined by ultradeep pyrosequencing. PLoS One **9:**e105569. http://dx.doi.org/10.1371/journal.pone.0105569.
20. **Di Maio VC, Cento V, Mirabelli C, Artese A, Costa G, Alcaro S, Perno CF, Ceccherini-Silberstein F.** 2014. Hepatitis C virus genetic variability and the presence of NS5B resistance-associated mutations as natural polymorphisms in selected genotypes could affect the response to NS5B inhibitors. Antimicrob Agents Chemother **58:**2781–2797. http://dx.doi.org/10.1128/AAC.02386-13.
21. **McCloskey RM, Liang RH, Joy JB, Krajden M, Montaner JS, Harrigan PR, Poon AF.** 2015. Global origin and transmission of hepatitis C virus nonstructural protein 3 Q80K polymorphism. J Infect Dis **211:**1288–1295. http://dx.doi.org/10.1093/infdis/jiu613.
22. **Cento V, Mirabelli C, Salpini R, Dimonte S, Artese A, Costa G, Mercurio F, Svicher V, Parrotta L, Bertoli A, Ciotti M, Di Paolo D, Sarrecchia C, Andreoni M, Alcaro S, Angelico M, Perno CF, Ceccherini-Silberstein F.** 2012. HCV genotypes are differently prone to the development of resistance to linear and macrocyclic protease inhibitors. PLoS One **7:**e39652. http://dx.doi.org/10.1371/journal.pone.0039652.
23. **Combet C, Garnier N, Charavay C, Grando D, Crisan D, Lopez J, Dehne-Garcia A, Geourjon C, Bettler E, Hulo C, Le Mercier P, Bartenschlager R, Diepolder H, Moradpour D, Pawlotsky JM, Rice CM, Trepo C, Penin F, Deleage G.** 2007. The European hepatitis C virus database. Nucleic Acids Res **35**(Database Issue)**:**D363–D366.
24. **Kuiken C, Yusim K, Boykin L, Richardson R.** 2005. The Los Alamos HCV sequence database. Bioinformatics **21:**379–384. http://dx.doi.org/10.1093/bioinformatics/bth485.
25. **Pickett B, Greer D, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen C, Jen W, Klem E, Scheuermann R.** 2012. Virus Pathogen Database and Analysis Resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses **4:**3209–3226. http://dx.doi.org/10.3390/v4113209.
26. **Edgar RC.** 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5:**113. http://dx.doi.org/10.1186/1471-2105-5-113.
27. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
28. **Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S.** 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony. Methods Mol Biol Evol **28:**2731–2739. http://dx.doi.org/10.1093/molbev/msr121.
29. **Huang Y, Niu B, Gao Y, Fu L, Li W.** 2010. CD-HIT Suite: a Web server for clustering and comparing biological sequences. Bioinformatics **26:**680–682. http://dx.doi.org/10.1093/bioinformatics/btq003.
30. **Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O.** 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **59:**307–321. http://dx.doi.org/10.1093/sysbio/syq010.
31. **Salvatierra KA.** 2014. Ph.D. thesis. University of Valencia, Valencia, Spain.
32. **Hall TA.** 1999. BioEdit: a user-friendly biological sequence alignment

editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp Ser **41:**95–98.
33. **R Core Team.** 2014. A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.
34. **Korber B, Myers G.** 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res Hum Retroviruses **8:**1549–1560. http://dx.doi.org/10.1089/aid.1992.8.1549.
35. **Excoffier L, Lischer HE.** 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resour **10:**564–567. http://dx.doi.org/10.1111/j.1755-0998.2010.02847.x.
36. **Van de Vijver DA, Wensing AM, Angarano G, Asjö B, Balotta C, Boeri E, Camacho R, Chaix ML, Costagliola D, De Luca A, Derdelinckx I, Grossman Z, Hamouda O, Hatzakis A, Hemmer R, Hoepelman A, Horban A, Korn K, Kücherer C, Leitner T, Loveday C, MacRae E, Maljkovic I, de Mendoza C, Meyer L, Nielsen C, Op de Coul EL, Ormaasen V, Paraskevis D, Perrin L, Puchhammer-Stöckl E, Ruiz L, Salminen M, Schmit JC, Schneider F, Schuurman R, Soriano V, Stanczak G, Stanojevic M, Vandamme AM, Van Laethem K, Violin M, Wilbe K, Yerly S, Zazzi M, Boucher CA.** 2006. The calculated genetic barrier for antiretroviral drug resistance substitutions is largely similar for different HIV-1 subtypes. J Acquir Immune Defic Syndr **41:**352–360. http://dx.doi.org/10.1097/01.qai.0000209899.05126.e4.
37. **Svicher V, Cento V, Salpini R, Mercurio F, Fraune M, Beggel B, Han Y, Gori C, Wittkop L, Bertoli A, Micheli V, Gubertini G, Longo R, Romano S, Visca M, Gallinaro V, Marino N, Mazzotta F, De Sanctis GM, Fleury H, Trimoulet P, Angelico M, Cappiello G, Zhang XX, Verheyen J, Ceccherini-Silberstein F, Perno CF.** 2011. Role of hepatitis B virus genetic barrier in drug-resistance and immune-escape development. Dig Liver Dis **43:**975–983. http://dx.doi.org/10.1016/j.dld.2011.07.002.
38. **Wyles DL, Gutierrez JA.** 2014. Importance of HCV genotype 1 subtypes for drug resistance and response to therapy. J Viral Hepat **21:**229–240. http://dx.doi.org/10.1111/jvh.12230.
39. **Svarovskaia ES, Dvory-Sobol H, Gontcharova V, Chiu S, Hebner CM, Hyland R, Kowdley K, Lawitz E, Gane E, Symonds WT, McHutchison JG, Miller MD, Mo H.** 2012. Comprehensive resistance testing in patients who relapsed after treatment with sofosbuvir (GS-7977)-containing regimens in phase 2 studies. Hepatology **56**(Suppl 1)**:**551A.
40. **Tong X, Li L, Haines K, Najera I.** 2014. Identification of the NS5B S282T resistant variant and two novel amino acid substitutions that affect replication capacity in hepatitis C virus-infected patients treated with mericitabine and danoprevir. Antimicrob Agents Chemother **58:**3105–3114. http://dx.doi.org/10.1128/AAC.02672-13.
41. **McPhee F, Hernandez D, Yu F, Ueland J, Monikowski A, Carifa A, Falk P, Wang C, Fridell R, Eley T, Zhou N, Gardiner D.** 2013. Resistance analysis of hepatitis C virus genotype 1 prior treatment null responders receiving daclatasvir and asunaprevir. Hepatology **58:**902–911. http://dx.doi.org/10.1002/hep.26388.
42. **Legrand-Abravanel F, Henquell C, Le Guillou-Guillemette H, Balan V, Mirand A, Dubois M, Lunel-Fabiani F, Payan C, Izopet J.** 2009. Naturally occurring substitutions conferring resistance to hepatitis C virus polymerase inhibitors in treatment-naive patients infected with genotypes 1-5. Antivir Ther **14:**723–730.
43. **Thibeault D, Bousquet C, Gingras R, Lagacé L, Maurice R, White PW, Lamarre D.** 2004. Sensitivity of NS3 serine proteases from hepatitis C virus genotypes 2 and 3 to the inhibitor BILN 2061. J Virol **78:**7352–7359. http://dx.doi.org/10.1128/JVI.78.14.7352-7359.2004.
44. **Scheel TK, Gottwein JM, Mikkelsen LS, Jensen TB, Bukh J.** 2011. Recombinant HCV variants with NS5A from genotypes 1-7 have different sensitivities to an NS5A inhibitor but not interferon-α. Gastroenterology **140:**1032–1042. http://dx.doi.org/10.1053/j.gastro.2010.11.036.
45. **Sarrazin C, Zeuzem S.** 2010. Resistance to direct antiviral agents in patients with hepatitis C virus infection. Gastroenterology **138:**447–462. http://dx.doi.org/10.1053/j.gastro.2009.11.055.

**Supplementary material**

Supplementary material (Fig S.1 and tables S1 to S6) is freely available at

http://aac.asm.org/content/suppl/2016/03/16/AAC.02776-

15.DCSupplemental/zac004165065so1.pdf

## 2.7- Chapter 7: Comparative analysis of variation and selection in the HCV genome

*Infect Genet Evol. (2017) 49:104-110.*

Infection, Genetics and Evolution 49 (2017) 104–110

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Research paper

# Comparative analysis of variation and selection in the HCV genome

Juan Ángel Patiño-Galindo, Fernando González-Candelas *

*Unidad Mixta Infección y Salud Pública FISABIO-CSISP/Universitat de València, CIBERESP, Valencia, Spain*

ARTICLE INFO

*Article history:*
Received 12 October 2016
Received in revised form 6 January 2017
Accepted 9 January 2017
Available online 10 January 2017

*Keywords:*
HCV
Genome
Selection
RNA secondary structure
Epitope

ABSTRACT

Genotype 1 of the hepatitis C virus (HCV) is the most prevalent of the variants of this virus. Its two main subtypes, HCV-1a and HCV-1b, are associated to differences in epidemic features and risk groups, despite sharing similar features in most biological properties. We have analyzed the impact of positive selection on the evolution of these variants using complete genome coding regions, and compared the levels of genetic variability and the distribution of positively selected sites. We have also compared the distributions of positively selected and conserved sites considering different factors such as RNA secondary structure, the presence of different epitopes (antibody, CD4 and CD8), and secondary protein structure. <10% of the genome was found to be under positive selection, and purifying selection was the main evolutionary process acting in both subtypes. We found differences in the number of positively selected sites between subtypes in several genes (*Core*, HVR2 in *E2*, *P7*, helicase in *NS3* and *NS4a*). Heterozygosity values in positively selected sites and the rate of non-synonymous substitutions were significantly higher in subtype HCV-1b. Logistic regression analyses revealed that similar selective forces act at the genome level in both subtypes: RNA secondary structure and CD4 T-cell epitopes are associated with conserved sites, while CD8 T-cell epitopes are associated with positive selection in both subtypes. These results indicate that similar selective constraints are acting along HCV-1a and HCV-1 b genomes, despite some differences in the distribution of positively selected sites at independent genes.

© 2017 Elsevier B.V. All rights reserved.

## 1. Introduction

Hepatitis C Virus (HCV) is the main causal agent of non-A non-B viral hepatitis, a pandemic with a global prevalence of 2.8%, affecting >185 million people worldwide (Mohd Hanafiah et al., 2013). HCV belongs to the genus *Hepacivirus* in the *Flaviviridae* family and its single, positive-sense, single-stranded RNA genome of 9.6 kb encodes a polyprotein of >3000 amino acids (Takamizawa et al., 1991). This polyprotein is cleaved into 3 structural (Core, Envelope 1 (E1), E2) and 7 non-structural (P7, NS2, NS3, NS4A, NS4b, NS5A and NS5b) proteins by means of both viral and human proteases (reviewed in Lindenbach and Rice, 2005).

Like most RNA viruses, HCV is highly variable genetically and is phylogenetically divided into seven genotypes (named from 1 to 7) (Smith et al., 2014) which present >30% divergence at the nucleotide level among them. Most genotypes are further divided into subtypes, with 20–25% divergence among them (Simmonds et al., 1993). Genotype 1 is the most prevalent variant (Messina et al., 2015) and also the one with the lowest sustained viral response (SVR) rate to treatment with pegylated interferon and ribavirin (Pang et al., 2009), which still remains as the most frequent treatment for hepatitis C infection. >10 HCV-1 subtypes have been reported so far (Kuiken et al., 2005). Among them, subtypes 1a and 1b are the most prevalent and cause about 40% of the total infections by the virus.

These two viral variants present some differences. There are epidemiological differences between their major transmission groups and these influence their phylodynamics. HCV-1a has been historically associated with transmission among intravenous drug users and HCV-1b with transfusions and other nosocomial and community transmissions (Shepard et al., 2005). Although subtype 1b apparently appeared around 10 years later than subtype 1a, it started an explosive growth phase 20 years earlier (in the 1940s), coincident with the start of widespread use of blood and blood derivatives in transfusions (Magiorkinis et al., 2009). Clinical differences between these two subtypes may also exist: sustained viral response (SVR) to treatment with pegylated interferon and ribavirin has been reported to be significantly higher in subtype 1a than in 1b (Pellicelli et al., 2012). Finally, at the genetic level, Torres-Puente et al. (2008) reported differences in genetic variability between these two subtypes, with HCV-1b showing higher nucleotide diversity in genes *E1*, *E2* and *NS5A*.

Most published studies detecting positively selected sites in this virus have analyzed only the *E1–E2* and/or *NS5A* genes (Cuevas et al., 2009; Cuevas et al., 2008; Humphreys et al., 2009; Sheridan et al., 2004). This is due to the relevance of the proteins that they encode in the viral response to the immune system and to treatment and, in

* Corresponding author at: Unidad Mixta Infección y Salud Pública FISABIO-CSISP/Universitat de València – Instituto de Biología Integrativa de Sistemas (I2SysBio), c/ Catedrático José Beltrán, 2, 46980 Paterna, Valencia, Spain.
*E-mail address:* fernando.gonzalez@uv.es (F. González-Candelas).

http://dx.doi.org/10.1016/j.meegid.2017.01.010
1567-1348/© 2017 Elsevier B.V. All rights reserved.

consequence, for the establishment of persistent infection. Similar studies using complete coding regions published so far (Campo et al., 2008; Cannon et al., 2008) have used methods for the inference of positive selection that lack power, as they are based on estimating the ratio of nonsynonymous to synonymous substitution rates (dN/dS) from the reconstruction of ancestral states (Nei and Gojobori, 1986). Hence, we currently lack a detailed view of the action of selection as a factor in the evolution of the HCV genome and how it may differentially affect major variants of this virus.

Our goal in this work is to analyze the impact of positive selection on the evolution of the genomes of these two viral variants, comparing the selective pressures in the different proteins of these two subtypes. For this, we present analyses with a comprehensive dataset of complete HCV-1a and 1b genomes and report a detailed comparative map of positively selected sites using, for the first time in HCV, the mixed effects model of evolution (MEME) (Murrell et al., 2012). We have also estimated and compared the synonymous and non-synonymous substitution rates (dS and dN, respectively) and the heterozygosity (H) at individual sites. Finally, we have performed multivariate analyses in order to test the effects of RNA and protein structures and the presence of epitopes on the distribution of positively and conserved sites along their genomes.

The results obtained can give information to better understand the evolution of the HCV genome, especially genotype 1, regarding the effects of the different levels at which virus-host interactions can occur on genetic variability. The results can also indicate whether the most prevalent HCV-1 subtypes interact with the host in a similar way and have similar levels of genetic variability, or present differences which may be of interest to consider for antiviral research.

## 2. Materials and methods

Full coding regions from HCV-1a and HCV-1b genomes were retrieved from the VIPRBRC dataset on May 2013 (Pickett et al., 2012) (amino acid positions 1 to 3011, according to the reference genome H77 - GenBank accession number AF011753). Only sequences derived from human hosts were included. In addition, we ensured that all HCV sequences derived from DAA-naïve patients, excluding all sequences from DAA-treated patients described in the literature up to January 2015. Additional inclusion criteria were:

1- We included only one sequence per patient.
2- Viral subtypes were confirmed using the COMET HIV-1/2 & HCV subtyping tool (Alcantara et al., 2009; De Oliveira et al., 2005, accessible at http://comet.retrovirology.lu).
3- Recombinant sequences were detected and subsequently removed using five different methods implemented in the RDP3 software: RDP, Geneconv, Bootscan, Maxchi and Chimera (Martin et al., 2005; Martin and Rybicki, 2000; Martin et al., 2010; Padidam et al., 1999; Posada and Crandall, 2001). The criterion to remove putative recombinant sequences was to obtain significant results with at least two different methods.

Multiple alignments were obtained for each HCV subtype independently using Muscle (Edgar, 2004), as implemented in MEGA 5 (Tamura et al., 2011).

The statistical power to detect sites under positive selection increases as the dataset contains more sequences (Kosakovsky Pond and Frost, 2005). To deal with the differences in size between the final datasets of HCV-1a and HCV-1b (393 and 179 sequences, respectively) we obtained 5 random subsets, each including half the total size of HCV-1a dataset ($n = 197$), and performed the same positive selection analyses.

For each dataset, a phylogenetic tree was obtained with FastTree2 (Price et al., 2009), using a GTR + GAMMA evolutionary model, as recommended by the AICM analyses implemented in jModeltest (Posada, 2008).

Considering diversifying selection as adaptive evolution in which increasing the global variability is favored in the population (Murrell et al., 2012), for each dataset and its corresponding tree, two different diversifying positive selection analyses were performed: (1) the two-rate Fixed Effects likelihood (FEL), a maximum-likelihood (ML) method used to find independent sites under positive, neutral or purifying selection which considers that both dN and dS can vary between sites, while dN/dS remains constant along the different lineages of a given phylogenetic tree (Kosakovsky Pond and Frost, 2005); and (2) Mixed Effects Model of Evolution (MEME), an extended version of FEL which considers that dN/dS can change across lineages. This method has been reported to be more efficient in finding sites under positive selection than FEL. Furthermore, it has been recommended for finding both episodic (that affects only to a subset of lineages) and pervasive (that affects to a large proportion of positively selected sites) selection, with a type I error probability not higher than 0.05 (Murrell et al., 2012). Both analyses were performed with Hyphy (Kosakovsky Pond and Muse, 2005) using the GTR model of nucleotide substitution and setting the significance level at 5%. Given that MEME and FEL are nested models, we compared their performance by calculating the ratio of their log-likelihoods (LRT) at each positively selected codon (Murrell et al., 2012). According to a Chi-square distribution with 2 degrees of freedom, LRT values larger than 5.99 were considered to be significant (P-value $< 0.05$).

A parsimony analysis was performed with MacClade 4.08 (Maddison and Maddison, 2005) in order to infer the number of changes accumulating in each branch of the HCV-1a and HCV-1b phylogenies, considering only the detected positively selected codons. This analysis allowed us also to detect sites identified as being under positive selection, but whose genetic variability was only due to singletons. Such sites were not considered for further analyses. This decision was taken in order to minimize the presence of false positives that could actually be sites in which a deleterious mutation had occurred but had not been removed from the viral population at the time of sampling. The remaining positively selected sites were mapped according to the H77 reference genome.

The number of sites under purifying selection was calculated from the list of negatively selected sites found in FEL. Then, we discarded those sites that actually were under positive selection, as detected by MEME.

For each subtype, a list of neutrally evolving sites was also obtained by means of the following procedure:

1- For each HCV subtype, a list of sites not evolving under positive nor negative selection was obtained from the results of FEL, as neutral evolution is considered the null hypothesis.
2- Potentially neutral sites that were actually positively selected, as detected by MEME, were excluded.
3- Completely conserved sites were also excluded, as they were considered to be under very strong purifying selection: it is possible that no variation is found in these positions, because mutations that have occur are deleterious, and are quickly removed from the viral populations. For this reason, we decided not to consider these sites as neutral. Importantly, when a site has no variation, then the LRT from FEL or MEME may not have enough statistical power to reject the null hypothesis of neutrality of these methods (Nei, 2005).

Gene diversity (H) at each amino acid and nucleotide site (first, second and third codon positions) was calculated using the expression $H = 1 - \sum p_i^2$ (where $p_i$ is the frequency of each allele at a given site) (Beebee and Rowe, 2008). dN, dS and H values were compared between subtypes using independent Mann-Whitney tests for: a) all genomic codons; b) positively selected codons and c) neutrally evolving codons. P-values obtained from non-independent statistical tests were corrected by means of False Discovery Rate corrections (FDR; (Benjamini and Hochberg, 1995).

Frequencies of base-pairing at each nucleotide position for each HCV subtype were estimated with the software STRUCTURE_DIST (Tuplin et

al., 2004), which analyses multiple RNA-folding patterns predicted by MFOLD (Zuker, 2003). Antibody, CD8 and CD4 T-cell epitope positions were retrieved from the Los Alamos National Laboratory website (http://hcv.lanl.gov/content/immuno/immuno-main.html) and the Immune Epitope Database and Analysis resource (http://www.iedb.org) on June 2015. Only human epitopes from HCV genotype 1 were included. Protein structures in both subtypes were inferred with JPred4 (Drozdetskiy et al., 2015), which predicts the location of secondary structures of proteins (alpha helixes, beta sheets) from multiple alignments of protein sequences. All these sites were mapped according to the H77 reference genome.

Logistic regression analyses (general linear model, GLM) were performed to compare, in each subtype, the distribution of positively selected and conserved positions. Several binary variables at each position were considered in the linear models: (1) RNA base-pairing (consensus, ≥0.50) at the subtype level (given that RNA structure applies to individual nucleotide positions, but selection acts at the codon level, we considered a codon as "structured" if at least 2 of the 3 positions were paired), (2) CD8 T-cell epitope, (3) CD4 T-cell epitope, (4) antibody epitope, (5) alpha helix, and (6) beta sheet. Positions were considered to be conserved if they had H = 0 at the amino acid level. An initial model, which included all the variables, was built. Stepwise model selection by Akaike Information Criterion (AIC) was performed with the R package "MASS" (Ripley et al., 2012) with the aim of including only relevant predictors. Stepwise search was performed in both directions (which tests at each step for variables to be included or excluded), and the best quality model for the GLM was chosen as the one with the lowest AIC value. P-values obtained in the logistic regression analyses were corrected by means of FDR. All statistical tests were performed as implemented in R (R Core Team, 2015).

## 3. Results

In total, 415 sequences from HCV-1a and 204 sequences from HCV-1b were retrieved. All sequences were correctly subtyped (no evidence of inter-subtype recombination was found) and, after removing duplicated and/or intra-subtype recombinants, the final data sets consisted of 393 HCV-1a and 179 HCV-1b sequences.

FEL analyses detected 95 and 62 positively selected sites along the HCV-1a and HCV-1b genomes, respectively. In contrast, MEME analyses detected 315 (total dataset of HCV-1a) and 248 (HCV-1b) sites under positive selection. From those, 282 sites (HCV-1a) and 211 (HCV-1b) were not singletons (found only in one external branch) and were retained for ensuing analyses. All positively selected sites found by FEL were also found by MEME, and all these sites represented 9.4% and 7.0% of the total length of 1a and 1b protein encoding genomes, respectively. Ninety-nine homologous sites were found to be positively selected in both subtypes. Additional details of these analyses are provided in Supplementary Table S1A and S1B (Supplementary material online), including information on the genome location of each selected position, the frequency at which each selected site found in the full set of HCV-1a was also detected in the five subsets and LRT comparisons between MEME and FEL.

The performance of MEME and FEL for the detection of positively selected sites was compared by performing LRTs at each positively selected position detected by MEME. The mixed effects model outperformed the fixed effects model in most positions (209 of 282 in HCV-1a and 159 of 211 in HCV-1b), and could never be considered as significantly worse than FEL for the inference of positive selection in the datasets analyzed (Supplementary Tables S1A and S1B, Supplementary material online).

Table 1 summarizes the distribution of positively selected sites along the different regions of the H77 reference polyprotein for the two HCV-1 subtypes and for the 5 random subsets of HCV-1a.

After considering those sites initially found to be under purifying selection by FEL (n = 2412 in HCV-1a; n = 2672 in HCV-1b), but that were actually under positive selection (as detected by MEME), there were 2256 and 2571 sites significantly associated with purifying selection in HCV-1a and 1b, respectively.

A list of neutral sites was obtained from those sites that were not found to be under positive or negative selection in FEL or MEME and that were not totally conserved. We found that 231 in HCV-1a (7.7% of the total genome) and 271 positions in HCV-1b (9.0%) were evolving under neutrality (dN = dS). Of them, only 74 homologous positions were coincident in both subtypes (Supplementary Table S1C, Supplementary material online).

Supplementary Figs. S1A and S1B (Supplementary material online) show the phylogenetic trees obtained from the HCV-1a and HCV-1b complete genome sequences, respectively. In these trees, the branches are colored according to the number of positively selected sites changing along each branch, as determined by parsimony using MacClade 4.08 (Maddison and Maddison, 2005). Changes at positively selected sites were observed in both internal and external branches, although most changes accumulated in external branches.

Differences in the number of positively selected sites between proteins were found. The protein with the highest proportion of sites under positive selection was E2, followed by NS2 and E1. The proteins with the lowest proportion of sites under positive selection were NS4A and Core (Table 1). HCV-1b tended to present more positively selected sites in the second hypervariable region (HVR2) of E2 (mean = 1.4 sites in the subsets of HCV-1a, vs 6 sites in HCV-1b), in P7 (1.8 sites in HCV-1a vs 5 sites in HCV-1b) and the NS3-helicase (13.8 sites in HCV-1a vs 21 in HCV-1b). HCV-1a presented more positively selected sites only in Core (5.8 sites in HCV-1a, vs 3 in HCV-1b) (Table 1).

Mean and standard error (SE) estimates of dS, dN and heterozygosity (for both amino acids and nucleotides) obtained for the whole genomes, positively selected sites and neutral sites are available in Supplementary Tables S2A–C (Supplementary material online), respectively, and in Fig. 1. After applying the FDR corrections, the heterozygosity at third codon positions was found to be significantly higher in HCV-1b than in HCV-1a at the genomic level (HCV-1a: 0.161 ± 0.002 – mean ± standard error, HCV-1b = 0.181 ± 0.003; $P < 0.001$). No significant differences were found in first or second codon positions nor amino acid sites (all P values > 0.05) (Fig. 1.A; Supplementary Table S.2A).

Mann-Whitney tests comparing heterozygosity between subtypes in all codon positions and amino acids and dS and dN of all positively

**Table 1**
Number positively selected sites, and its proportion (between brackets) at each gene of HCV-1a and HCV-1b. aa – number of amino acids encoded by each gene. n – number of sequences included in the dataset. Major domains are reported for the E2 (HVR = Hyper variable region) and NS3 (PR = protease, HE = helicase) genes.

| Dataset | n | CORE (191 aa) | E1 (192 aa) | E2 (363 aa) | P7 (63 aa) | NS2 (217 aa) | NS3 (631 aa) | NS4A 54 aa) | NS4B (261 aa) | NS5A (448 aa) | NS5B (591 aa) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HCV-1a | 393 | 10 (0.052) | 24 (0.125) | 70: 20 HVR1, 3 HVR2, 5 HVR3 (0.193) | 2 (0.032) | 28 (0.129) | 32: 9 PR, 23 HE (0.051) | 0 (0.000) | 15 (0.057) | 50 (0.112) | 51 (0.0.086) |
| Subsets HCV-1a[a] | 197 | 5.8 (0.030) | 14 (0.073) | 52: 17.2 HVR1, 1.4 HVR2, 4.2 HVR3 (0.143) | 1.8 (0.029) | 17.2 (0.079) | 21: 7.2 PR, 13.8 HE (0.033) | 0 (0.000) | 8.4 (0.032) | 36.8 (0.082) | 36.8 (0.062) |
| HCV-1b | 179 | 3 (h) (0.015) | 16 (0.083) | 61: 18 HVR1, 6 HVR2, 6 HVR3 (0.168) | 5 (0.079) | 17 (0.078) | 27: 6 PR, 21 HE (0.043) | 1 (0.019) | 8 (0.031) | 34 (0.076) | 39 (0.066) |

[a] Results are given as the mean number of positively selected sites found in the five HCV-1a subsets.

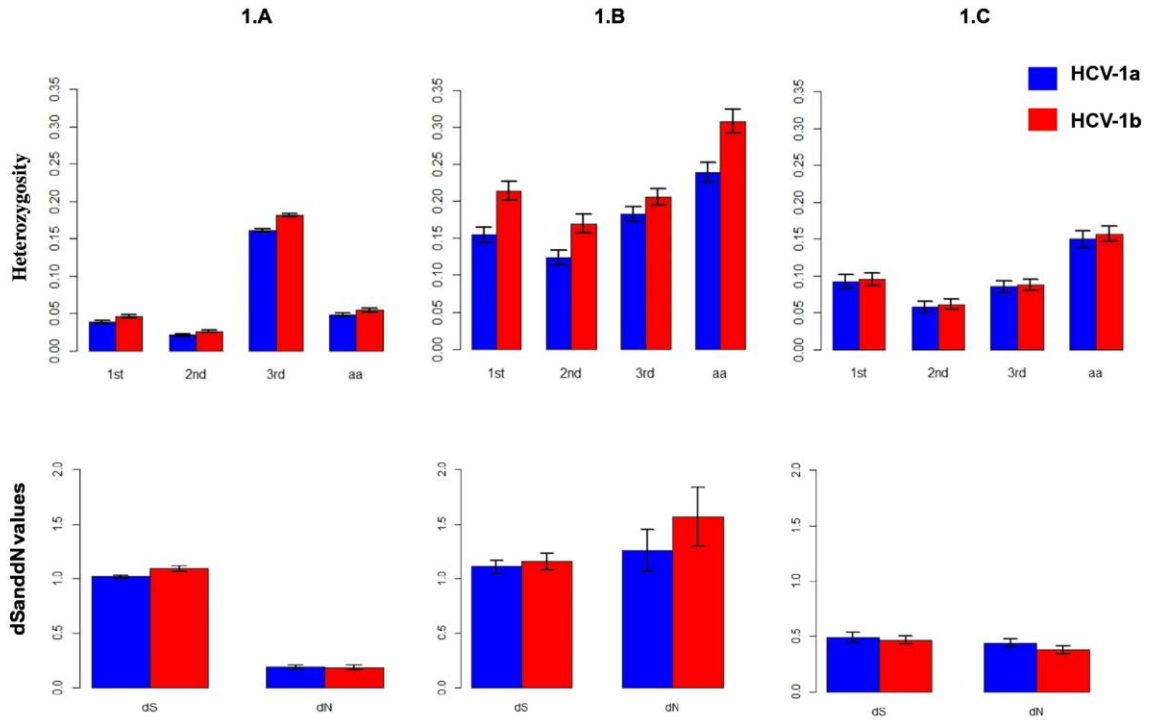



**Fig. 1.** Bar plots, with standard errors error bars, representing the mean heterozygosity, dS and dN values of the total genome (1.A), positively selected sites (1.B) and neutral sites (1.C) for both HCV-1a (blue) and HCV_1b (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

selected sites revealed that HCV-1b presents significantly higher dN (HCV-1a: 1.257 ± 0.192, HCV-1b: 1.563 ± 0.271; P = 0.040) and heterozygosity at amino acids (HCV-1a: 0.239 ± 0.013, HCV-1b: 0.308 ± 0.016; P < 0.001), first codon positions (HCV-1a: 0.155 ± 0.011, HCV-1b: 0.214 ± 0.013; <0.001), second codon positions (HCV-1a: 0.123 ± 0.010, HCV-1b: 0.169 ± 0.013; P = 0.003) and third codon positions (HCV-1a: 0.183 ± 0.010; HCV-1b: 0.206 ± 0.011; P = 0.023), but not significantly different dS (P = 0.90) (Fig. 1.B; Supplementary Table S.2B).

For neutral sites, no significant differences in dS, dN nor in heterozygosity between HCV-1a and HCV-1b were found, as concluded from the statistical tests (all P values > 0.05) (Fig. 1.C; Supplementary Table S.2C).

A map of the HCV-1a and HCV-1b genomes representing the different layers of data analyzed (conservation, positive selection, RNA structure, protein structure and epitopes for antibodies, CD8 and CD4 T cells) is shown in Fig. 2. In the logistic regression analyses, the models with the lowest AIC values were "RNA structure + CD4 epitope + CD8 epitope + Alpha helix" for HCV-1a and "RNA structure + CD4 epitope + CD8 epitope + Beta sheet" for HCV-1b. Comparing the distribution of positively and conserved sites at such layers of data revealed, for both HCV-1a and 1b, a positive association between conservation and secondary structure as well as with CD4 T cell epitopes. In contrast, although a positive association between selection and CD8 T cell epitopes was found in this case the P-values for both subtypes were >0.05 after the FDR correction (Supplementary Tables S3A and S3B, Supplementary material online).

The same analyses were performed for all genes in which at least 10 sites were found to be positively selected in both subtypes (E1, E2, NS2, NS3, NS5a and NS5b). In HCV-1a, a positive association between selection and the presence of CD8 epitopes was found in NS2, although the P-value increased to >0.05 after FDR correction. In subtype 1b, a similar association between conservation and the presence of CD4 epitopes was found in NS3. Although associations between selection and CD8 epitopes and between conservation and the presence of beta sheets were found in E2 of HCV-1b, the P-value increased to >0.05 after FDR correction (Supplementary Tables S.3A and S.3B).

## 4. Discussion

In this work we have performed a comparative analysis of the evolutionary forces that shape the genomes of HCV-1a and HCV-1b, using both a fixed effects (FEL) and a mixed effects model (MEME) to detect sites evolving under different selective pressures: positive, purifying or neutral. Our results reveal several differences as well as similarities between these two relevant subtypes of HCV and help us understand the underlying factors driving their evolution at the genome level.

Previous studies analyzing selection in the HCV genome are more conservative than this one because they are based on the Nei and Gojobori (1986) model for the calculation of dN/dS. Campo et al. (2008) found a very similar number (n = 60) of sites evolving under positive selection along the HCV-1b genome using the SLAC method to those we found with FEL (n = 62). However, our analysis with MEME detected >3 times more positively selected sites than these studies, in line with the expected increased power of this method (Murrell et al., 2012). In consequence, methods for the detection of positively selected sites that do not take into account that dN/dS can vary among lineages may underestimate the number of positively selected sites. This can occur when, for a given site, purifying selection prevails in most lineages, masking the detection of episodic positive selection occurring in a restricted number of lineages.

Amino acid changes in positively selected codons accumulated mainly on the external branches of the phylogenetic trees for both viral subtypes. This was expected, because the methods used to detect positive selection in fact identify sites under adaptive diversification

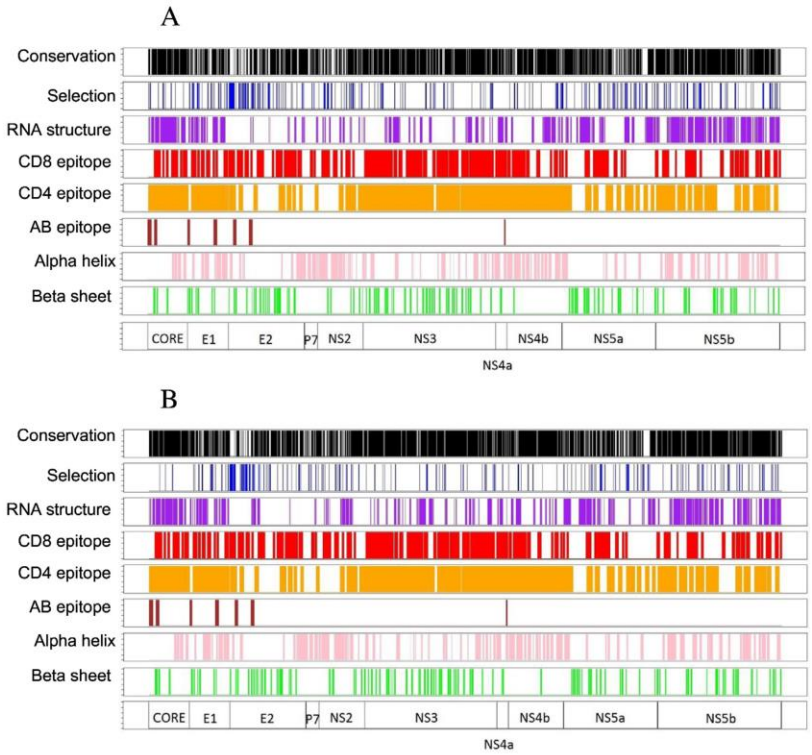



**Fig. 2.** Map of the HCV-1a and 1b genomes, indicating the location of totally conserved amino acids (black), positively selected sites (blue) RNA secondary structures present in at least 50% of the sequences of each dataset (purple), CD8 T cell epitopes (red), CD4 T cell epitopes (orange), antibody (AB) epitopes (brown), alpha helixes (pink), beta sheets (green). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

when applied to the population or within-species level. Diversifying selection is known to be of major importance in the adaptive evolution of HCV, in which increasing global variability would be favored (Cuevas et al., 2009). Another possible explanation is that recent deleterious mutations, most of which are expected to be nonsynonymous, may not have been purged by selection yet and they will likely map on the external branches of the phylogeny. In consequence, for a given position different nonsynonymous substitutions might be detected as being positively selected when analyzing viral sequences from different patients when they actually represent transient polymorphisms at the population level. However, we excluded singletons from the list of positively selected positions and analyzed only one sequence per patient, thus minimizing the presence of these false positives.

We also found that purifying selection plays a major role in the evolution of HCV, with >2200 codons estimated to be under purifying selection in both subtypes. However, the number of neutral sites inferred from our work differs markedly from those detected by Campo et al. (2008) (833 sites). This discrepancy might be explained by the different sampling sizes used in both studies (114 vs 179 HCV-1b sequences) but it is more likely due to differences in the methodology used in the two studies. The lower power of SLAC (Kosakovsky Pond and Frost, 2005) used by Campo et al. (2008) to reject the null hypothesis of neutrality would be reflected in a large number of undetected, negatively selected sites. Furthermore, our criteria to detect neutral sites were more stringent, including only sites that were considered not to be positively nor negatively selected in MEME and FEL and which presented $H > 0$ at the amino acid level, to avoid including codons under very strong, purifying selection for which no change could be detected.

We found positively selected sites in all HCV genes. Differences between subtypes in the distribution of positively selected sites were also detected. Whereas HCV-1b tended to present more positively selected sites in *E2* (HVR2), *P7* and *NS3* (helicase), HCV-1a presented more positively selected sites in the *Core* gene. This observation is based on the results obtained from the random subsets of HCV-1a, with a similar number of sequences to that obtained for HCV-1b. These subsets presented very similar levels of heterozygosity, dS, and dN when compared to the whole dataset of this subtype. Hence, differences between the two subtypes in these parameters are not likely due to their different sample sizes.

Differences in H between both subtypes were significant at the amino acid level and at first, second and third codon positions of positively selected sites and only at third codon positions at the complete genome level. In contrast, differences in dN between the two subtypes were significant only for positively selected sites, whereas no significant differences in dS were detected in any of the comparisons performed. The different phylodynamic histories of HCV-1a and HCV-1b (Magiorkinis et al., 2009) could explain the higher genetic variability in third codon positions of HCV-1b at the genomic scale. HCV-1b started its explosive growth 20 years earlier than HCV-1a and its current effective infection size remains higher. This implies that it may have accumulated more genetic variability. Although no significant differences were found at first and second codon positions, where most changes are nonsynonymous (a majority of these sites are totally conserved along the HCV genome), the significantly higher variability of HCV-1b with respect to HCV-1a was evident at third codon positions, in which most changes are synonymous. Interestingly, we did not find significantly higher dS values in HCV-1b. According to the neutral theory of evolution, dS depends only on the neutral mutation rate but H is dependent also on the effective population size (Kimura, 1983). In consequence, we would expect HCV-1a and HCV-1b to present similar dS if their

overall mutation rates are similar but to differ in H if they have different effective population sizes.

Despite the differences found in the distribution of positively selected sites, multivariate analyses at the genomic level revealed that HCV-1a and HCV-1b evolve under similar forces and constraints (RNA structure and CD4 T cell epitopes favoring conservation, and CD8 T cell epitopes favoring selection). Snoeck et al. (2011) mapped positively selected sites along the genome of HIV-1 and found that structured RNA and alpha helices in protein structure were associated with conservation whereas CD4 T-cell and antibody epitopes were associated with positive selection. However, they also found a significant association between CD8 and CD4 epitopes and conservation for several genes. Our results, regarding the overall association between secondary structure and conservation, are in line with those of Snoeck et al. (2011) and also with the results obtained by Mauger et al. (2015) with HCV. These authors suggested that conservation of secondary structures in this virus could facilitate persistent infection by masking the viral genome from degradation by RNase L and innate antiviral defenses (Li and Lemon, 2013; Washenberger et al., 2007).

In contrast to Snoeck et al. (2011) with HIV-1, we found no relevant association between antibody epitopes or protein structure and conservation or selection in HCV. Although the role of CD4 and CD8 T-cells in the immune response to HCV is well known, contrary to other viruses such as HIV or HBV (Koziel, 2005) there is not a clear pattern of antibodies response that distinguishes between recovery and chronic infection in HCV. Thus, further research to clarify their role in controlling HCV replication and to which extent such mechanisms influence on the evolution of HCV is certainly needed.

Recently, Geller et al. (2016) estimated the *per site* mutation rate along the HCV genome, and found a small reduction in sites predicted to form base pairs. We performed a univariate analysis to check whether there was an unequal distribution between sites with low and high mutation rates, considering conserved and selected codons, and found no significant differences (Fisher's exact test $P > 0.20$ in both subtypes).

The association between CD4 T-cell epitopes and conservation is remarkable. Given that epitopes are targets for the host immune system, it would be reasonable to expect epitopes to be under positive or diversifying selection (as for CD8), because an increased genetic variability would facilitate viral escape from the immune system. However, several studies have observed very conserved epitopes in HCV and other viruses (Lamonaca et al., 1999; Sanjuán et al., 2013; Sarobe et al., 2001; Snoeck et al., 2011). Sanjuán et al. (2013) suggested that HIV may take advantage of immune activation, thus favoring epitope conservation. Hence, if HCV also benefits from immune activation, the design of vaccines based on conserved epitopes would be detrimental.

After this manuscript was written, Cuypers et al. (2016) have published their analysis of the different distributions of positively selected and conserved sites in HCV considering variables such as protein and RNA secondary structure and B-cell, CD8 and CD4 epitopes. We were unaware of this paper during the development of our work. Despite using larger datasets, they found a similar proportion of sites under purifying selection (approximately 2500 positions) and <100 positively selected sites in each HCV subtype, probably because the positive selection analyses were performed using FEL, and not MEME. Indeed, these numbers are very similar to the positively selected sites that we found in HCV-1a with FEL, and is in agreement with previous observations which evidenced that methods for the detection of positive selection can be conservative for smaller sample sizes, but not for larger (Kosakovsky Pond and Frost, 2005; Murrell et al., 2012). In addition, the distribution of positively selected sites was similar in both works.

As expected, Cuypers et al. (2016) obtained similar results to ours for the association between conservation and RNA structure and CD4 epitopes. Such similarities were expected because, in both studies, the mapping of RNA secondary structure was performed computationally and the mapping of epitopes was a based on information available at public databases. However, some discrepancies were also obtained: they found an association between positive selection and CD8 epitopes only in HCV-1a, while we did not find inter-subtype differences. In addition, they obtained significant associations between different protein structures (alpha helix and B-sheets) and conservation, whilst we did not find any significant association for these two variables. Although Cuypers et al. (2016) used protein structure information derived from crystallized or nuclear magnetic resonance (NMS) when available, the extensive lack of such information for the HCV proteome led both studies to predict such structures by means of computational methods. Consequently, differences in the methodology used for mapping protein structures, including the use of different protein structure prediction programs, may have caused these incongruent results.

Is also important to point out that, unlike Cuypers et al. (2016) our work included a comparison of FEL and MEME, allowing to observe that most positively selected sites are under episodic/pervasive selection, hampering their detection when using more conservative, less powerful methods, such as FEL. Finally, our comparisons of the genetic variability (heterozygosity), and dN and dS, of HCV-1a and 1b along their genomes, as discussed above, have not been performed, or published, before.

In conclusion, we have produced a detailed map of positive selection along HCV-1a and 1b genomes and analyzed which variables can impose constraints or be associated to selection. We have shown that, despite purifying selection being the most extensive evolutionary process acting on HCV, positive selection affects all genes along the HCV genome. Although there are differences in variability and the distribution of positively selected sites, both viral subtypes share similar selective pressures along their genomes. The results obtained from this study give information about the effect of some of the interactions between HCV and its host on HCV variability, which may be useful for antiviral research against this virus.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.meegid.2017.01.010.

## Acknowledgements

This work was supported by projects BFU2011-24112 and BFU2014-58565-R from Ministerio de Economía y Competitividad (Spain). JAPG is recipient of a FPU fellowship from Ministerio de Educación y Ciencia (Spain).

## References

Alcantara, L.C.J., Cassol, S., Libin, P., Deforche, K., Pybus, O.G., Van Ranst, M., Galvao-Castro, B., Vandamme, A.M., De Oliveira, T., 2009. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res. 37, W634–W642.

Beebee, T., Rowe, G., 2008. An Introduction to Molecular Ecology. second ed. Oxford University Press.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Statist Soc* B. 57, 289–300.

Campo, D.S., Dimitrova, Z., Mitchell, R.J., Lara, J., Khudyakov, Y., 2008. Coordinated evolution of the hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 105, 9685–9690.

Cannon, N.A., Donlin, M.J., Fan, X., Aurora, R., Tavis, J.E., 2008. Hepatitis C virus diversity and evolution in the full open-reading frame during antiviral therapy. PLoS One 3, e2123.

Core Team, R., 2015. R: A language and environment for statistical computing. Available from:. http://www.r-project.org.

Cuevas, J.M., Torres-Puente, M., Jimenez-Hernandez, N., Bracho, M.A., Garcia-Robles, I., Carnicer, F., Del Olmo, J., Ortega, E., Moya, A., González-Candelas, F., 2008. Refined analysis of genetic variability parameters in hepatitis C virus and the ability to predict antiviral treatment response. J. Viral Hepat. 15, 578–590.

Cuevas, J.M., Gonzalez, M., Torres-Puente, M., Jiménez-Hernández, N., Bracho, M.A., García-Robles, I., González-Candelas, F., Moya, A., 2009. The role of positive selection in hepatitis C virus. Infect. Genet. Evol. 9, 860–866.

Cuypers, L., Li, G., Neuman-haefelin, C., Piampongstan, S., Libin, P., Vab Iaethem, K., Vandamme, A.-M., Theys, K., 2016. Mapping the genomic diversity of HCV subtypes 1a and 1b: implications of structural and immunological constraints for vaccine and drug development. Virus Evolution 2 (2) vew024.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., van Rensburg, E.J., Wensing, A.M.J., van de Vijver, D.A., et al., 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics 21, 3797–3800.

Drozdetskiy, A., Cole, C., Procter, J., Barton, G.J., 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43, W389–W394.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.
Geller, R., Estada, Ú., Peris, J.B., Andreu, I., Bou, J.V., Garijo, R., Cuevas, J.M., Sabariegos, R., Mas, A., Sanjuán, R., 2016. Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat Microbiol. 1, 16045.
Humphreys, I., Fleming, V., Fabris, P., Parker, J., Schulenberg, B., Brown, A., Demetriou, C., Gaudieri, S., Pfafferott, K., Lucas, M., et al., 2009. Full-length characterization of Hepatitis C Virus subtype 3a reveals novel hypervariable regions under positive selection during acute infection. J. Virol. 83, 11456–11466.
Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press.
Kosakovsky Pond, S.L., Frost, S.D.W., 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22, 1208–1222.
Kosakovsky Pond, S.L., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. In: Nielsen, R. (Ed.), Statistical Methods in Molecular Evolution. Springer, New York.
Koziel, M.J., 2005. Cellular immune responses against hepatitis C virus. Clin. Infect. Dis. 41, S25–S31.
Kuiken, C., Yusim, K., Boykin, L., Richardson, R., 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21, 379–384.
Lamonaca, V., Missale, G., Urbani, S., Pilli, M., Boni, C., Mori, C., Sette, A., Massari, M., Southwood, S., Bertoni, R., 1999. Conserved hepatitis C virus sequences are highly immunogenic for CD4+ T cells: implications for vaccine development. Hepatol. 30, 1088–1098.
Li, K., Lemon, S.M., 2013. Innate immune responses in hepatitis C virus infection. Semin. Immunopathol. 35, 53–72.
Lindenbach, B.D., Rice, C.M., 2005. Unravelling hepatitis C virus replication from genome to function. Nature 436, 933–938.
Maddison, D.R., Maddison, W.P., 2005. MacClade v. 4.08. Sinauer, Sunderland (MA).
Magiorkinis, G., Magiorkinis, E., Paraskevis, D., Ho, S.Y.W., Shapiro, B., Pybus, O.G., Allain, J.P., Hatzakis, A., 2009. The global spread of Hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. PLoS Med. 6, e1000198.
Martin, D., Rybicki, E., 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16, 562–563.
Martin, D.P., Posada, D., Crandall, K.A., Williamson, C., 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Human Retrovir. 21, 98–102.
Martin, D.P., Lemey, P., Lott, M., Moulton, V., Posada, D., Lefeuvre, P., 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26, 2462–2463.
Mauger, D.M., Golden, M., Yamane, D., Williford, S., Lemon, S.M., Martin, D.P., Weeks, K.M., 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc. Natl. Acad. Sci. U. S. A. 112, 3692–3697.
Messina, J.P., Humphreys, I., Flaxman, A., Brown, A., Cooke, G.S., Pybus, O.G., Barnes, E., 2015. Global distribution and prevalence of hepatitis C virus genotypes. Hepatol. 61, 77–87.
Mohd Hanafiah, K., Groeger, J., Flaxman, A.D., Wiersma, S.T., 2013. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatol. 57, 1333–1342.
Murrell, B., Wertheim, J.O., Moola, S., Weighill, T., Scheffler, K., Kosakovsky Pond, S.L., 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8, e1002764.
Nei, M., 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22, 2318–2342.
Nei, M., Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3, 418–423.
Padidam, M., Sawyer, S., Fauquet, C.M., 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265, 218–225.
Pang, P.S., Planet, P.J., Glenn, J.S., 2009. The evolution of the major Hepatitis C genotypes correlates with clinical response to interferon therapy. PLoS One 4, e6579.
Pellicelli, A.M., Romano, M., Stroffolini, T., Mazzoni, E., Mecenate, F., Monarca, R., Picardi, A., Bonaventura, M., Mastropietro, C., Vignally, P., et al., 2012. HCV genotype 1a shows a better virological response to antiviral therapy than HCV genotype 1b. BMC Gastroenterol. 12, 1–7.
Pickett, B.E., Greer, D.S., Zhang, Y., Stewart, L., Zhou, L., Sun, G., Gu, Z., Kumar, S., Zaremba, S., Larsen, C.N., 2012. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 4, 3209–3226.
Posada, D., 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.
Posada, D., Crandall, K.A., 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc. Natl. Acad. Sci. U. S. A. 98, 13757–13762.
Price, M.N., Dehal, P.S., Arkin, A.P., 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26, 1641–1650.
Ripley, B., Hornik, K., Gebhardt, A., Firth, D., 2012. Functions and datasets to support Venables and Ripley, 'Modern Applied Statistics with S' (2002): package "MASS". Accessed through. http://cran.r-project.org.
Sanjuán, R., Nebot, M.R., Peris, J.B., Alcamí, J., 2013. Immune activation promotes evolutionary conservation of T-cell epitopes in HIV-1. PLoS Biol. 11, e1001523.
Sarobe, P., Huarte, E., Lasarte, J.J., Lopez-Diaz de Cerio, A., Garcia, N., Borras-Cuesta, F., Prieto, J., 2001. Characterization of an immunologically conserved epitope from hepatitis C virus E2 glycoprotein recognized by HLA-A2 restricted cytotoxic T lymphocytes. J. Hepatol. 34, 321–329.
Shepard, C.W., Finelli, L., Alter, M.J., 2005. Global epidemiology of hepatitis C virus infection. Lancet Infect. Dis. 5, 558–567.
Sheridan, I., Pybus, O.G., Holmes, E.C., Klenerman, P., 2004. High resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. J. Virol. 78, 3447–3454.
Simmonds, P., Holmes, E.C., Cha, T.A., Chan, S.W., McOmish, F., Irvine, B., Beall, E., Yap, P.L., Kolberg, J., Urdea, M.S., 1993. Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. J Gen Virol. 74, 2391–2399.
Smith, D.B., Bukh, J., Kuiken, C., Muerhoff, A.S., Rice, C.M., Stapleton, J.T., Simmonds, P., 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and assignment web resource. Hepatol. 59, 318–327.
Snoeck, J., Fellay, J., Bartha, I., Douek, D., Telenti, A., 2011. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirol. 8, 87.
Takamizawa, A., Mori, C., Fuke, I., Manabe, S., Murakami, S., Fujita, J., Onishi, E., Andoh, T., Yoshida, I., Okayama, H., 1991. Structure and organization of the hepatitis C virus genome isolated from human carriers. J. Virol. 65, 1105–1113.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Torres-Puente, M., Cuevas, J.M., Jiménez, N., Bracho, M.A., García-Robles, I., Wróbel, B., Carnicer, F., Del Olmo, J., Ortega, E., Moya, A., González-Candelas, F., 2008. Using evolutionary tools to refine the new hypervariable region 3 within the envelope 2 protein of hepatitis C virus. Infect. Genet. Evol. 8, 74–82.
Tuplin, A., Evans, D.J., Simmonds, P., 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J Gen Virol. 85, 3037–3047.
Washenberger, C.L., Han, J.Q., Kechris, K.J., Jha, B.K., Silverman, R.H., Barton, D.J., 2007. Hepatitis C virus RNA: dinucleotide frequencies and cleavage by RNase L. Virus Res. 130, 85–95.
Zuker, M., 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31, 3406–3415.

**Supplementary material**

Supplementary data to this article can be found online at

http://dx.doi.org/10.1016/j.meegid.2017.01.010

## 2.8- Chapter 8: Effect of RNA substitution models on viroid and RNA virus phylogenetic reconstructions

***This work is in the final stage of preparation.***

**Effect of RNA substitution models on viroid and RNA virus phylogenetic reconstructions**

Juan Ángel Patiño-Galindo[1], Fernando González-Candelas[1] and Oliver G. Pybus[2]

[1]Unidad Mixta Infección y Salud Pública FISABIO-CSISP / Universitat de València.

CIBERESP, Valencia. SPAIN.

[2]Department of Zoology, University of Oxford, Oxford, United Kingdom

**Abstract**

Many viroids and RNA viruses have genomes that exhibit secondary structure, with paired nucleotides forming stems and loops. Such structures violate a key assumption of most methods of phylogenetic reconstruction, that sequence change is independent among sites. Despite this, phylogenetic analyses of these agents rarely utilize evolutionary models that account for RNA secondary structure. Here, we assess the effect of using RNA-specific nucleotide substitution models on the phylogenetic inference of viroids and RNA viruses. We obtained datasets comprising full-genome nucleotide sequences from 6 viroid and 10 single-stranded RNA virus species. For each dataset we inferred consensus RNA secondary structures, then evaluated different DNA and RNA substitution models for phylogenetic reconstruction. We used model selection to choose the best-fitting model and estimated Bayesian phylogenies. Further, for each dataset we generated and compared Robinson-Foulds (RF) statistics in order to test whether the distributions of trees generated under alternative models are different to each other. In all datasets, the best-fitting model was one which considers RNA secondary structure: RNA models that allow a non-zero rate of double substitution (RNA16A, RNA16C) fitted best both in viral and viroid datasets. In 14 of 16 datasets, the use of an RNA-specific model led to significantly longer tree lengths, but only in 2 of 16 it had a significant effect on Robinson-Foulds statistics. In conclusion, the use of an RNA model for phylogenetic inference of viroids and RNA viruses provides a better fit in the reconstructed trees for these organisms and significantly affects branch length estimates.

## Introduction

In order to avoid the problem of unobserved evolutionary changes and thereby accurately estimate genetic distances among taxa, modern molecular systematics relies upon the use of nucleotide (or codon or amino acid) substitution models. These models make assumptions about the process of molecular evolution, for example whether nucleotides vary in frequency, or whether substitution rates vary among nucleotides (Posada & Crandall 2001) or codon positions (Shapiro et al. 2006).

The existence of RNA secondary structures, such as stems (also called hairpins) violates a key assumption made by most methods of phylogenetic reconstruction; that evolutionary changes are independent among sites (Nasrallah et al. 2011). Stems are comprised of nucleotide sequences that form base-pairings with complementary regions from the same strand. Among the 16 possible base-pairings that can potentially occur, only six (the Watson-Crick pairs AU, UA, GC, CG plus GU and UG) are stable enough to form actual base-pairs (the remaining base-pairings are called Mismatches, MM). RNA structures play important roles in RNA viruses and viroids, for example in viral/viroid replication (Hutchins et al. 1986; Damgaard et al. 2004), translation (Pelletier & Sonenberg 1988), and immune evasion (Tellam et al. 2008). Hence, nucleotide changes that disrupt the most stable Watson-Crick pairs are often deleterious and RNA secondary structures can impose strong evolutionary constraints on sequence evolution. In order to maintain RNA structure, change in one base of a pair must matched by a complementary, compensatory change in the other base. One consequence of this evolutionary constraint is that the number of nucleotide changes

estimated from unpaired sites is expected to be higher than that from paired sites (Nasrallah et al. 2011). An association between the presence of base pairing and amino acid conservation has been reported for HIV-1 (Sanjuán & Bordería 2011; Snoeck et al. 2011).

In order to accommodate the among-site correlations imposed by RNA secondary structure, various types of RNA-specific substitution models for phylogenetic inference have been developed. The 6-state (RNA6A-E) models discard all mismatched sites from analysis, whilst the 7-state (RNA7A-G) models pools all mismatched sites into a single state (Tillier & Collins 1998).16-state models (RNA16A-F, I-K) take into account all 16 possible pairs that the four nucleotides could form (Schöniger & von Haeseler 1994; Muse 1995). RNA16 models can be classified in three different types: (i) “all pairs” models (RNA16A, B, I, J and K), in which each of the 16 dinucleotides has its own equilibrium frequency; (ii) “stable sets” models (RNA16D, E and F), in which the equilibrium frequencies of three different types of base pairs, MM, Watson-Crick and wobble (UG, GU), are different; and (iii) “stable pairs” model (RNA16C), considered as an extension of an RNA7 model, in which the 10 possible MM have a single equilibrium frequency (Savill et al. 2001; Allen & Whelan 2014).

Previous studies of ribosomal RNA (rRNA) genes have concluded that RNA-specific models outperform standard nucleotide substitution models when describing the evolution of structured RNA sequences (Savill et al. 2001; Kosakovsky Pond et al. 2007), as assessed by model comparisons using the Akaike Information Criterion (AIC) (Linhart & Zucchini 1986). In agreement with these studies, Allen & Whelan (2014) compared different nucleotide and RNA models for 287 human RNA gene families,

most of them microRNAs and snoRNAs, and concluded that RNA models outperformed nucleotide substitution models in most cases, since the former yielded the lowest corrected AIC (AICc) values.

Conserved RNA secondary structures have been reported to exist in the genomes of linear RNA viruses, such as in species of the *Flaviviridae* family (Thurner et al. 2004; Mauger et al. 2015) and HIV-1 (Watts et al. 2009). Hepatitis Delta Virus (HDV) and viroids, which exist as circular RNA genomes, present exceptionally highly structured genomes, as >70% of sites in their genomes form base-pairs (Wang et al. 1986; Sanjuán et al. 2006). Despite this, phylogenetic reconstructions of RNA viruses (including HDV) and viroids have not been generated using RNA models, thus ignoring the constraints that these structures impose on their genome evolution.

The goal of this study is to investigate whether RNA-specific substitution models outperform standard nucleotide substitution models when applied to different sets of full-genome sequences from RNA viruses and viroids. Further, we measure the degree to which phylogenetic inference is affected, in terms of estimated branch lengths and tree topologies, when an RNA-specific model is used to describe the evolution of paired sites in the genomes of these pathogens.

## Materials and methods

### *Datasets and alignments*

Full-genome nucleotide sequences from 6 viroid species [tomato apical stunt pospiviroid (TASVd), citrus exocortis viroid (CEVd), columnea latent viroid (CLVd),

grapevine yellow speckle viroid (GYSVd), Australian grapevine viroid (AGVd), potato spindle tuber viroid (PSTVd)] and 10 single-stranded RNA virus species [hepatitis delta virus (HDV), Sudan-ebolavirus (SUDV), dengue virus (DENV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), foot and mouth disease virus (FMDV), measles virus (MeV), rabies virus (RV) rubella virus (RuV) and mumps virus (MuV)] were downloaded in April, 2015. Viroid and HDV sequences were downloaded from GenBank whilst viral genomes were obtained from the Virus Pathogen Database and Analysis Resource, VIPRBRC (http://www.viprbrc.org). Only full genome sequences, which included untranslated regions, were considered. Alignments for each species were generated using MAFFT ("align- G-ins- 1" progressive method strategy) (Katoh & Standley 2013b) and positions with a high proportion of gaps were removed with TrimAl (Capella-Gutiérrez et al. 2009). Given that "gappy" positions were rare and represented insertions absent in most taxa, excluding them had no influence on the overall inferred RNA secondary structure.

*RNA-secondary structure inference*

For each species, RNA minimum free-energy consensus secondary structures were predicted using RNAalifold, as implemented in the Vienna Package 2.0 (Lorenz et al. 2011). The folding temperature was set to 25° and 37° C for viroids and viruses, respectively, which, according to Sanjuán et al. (2006), corresponds to the temperatures at which these pathogens replicate. RNA molecules were assumed to be circular for HDV and viroids. Arc diagrams of the obtained structures, which represent base-paired nucleotides along each genome, were plotted with the R4RNA package for R (Lai et al. 2012; R Core Team 2014).

The conservation of the RNA secondary structure within each dataset was tested using RNAz (Gruber et al. 2007) by calculating the Structure Conservation Index (SCI). An SCI = 0 indicates that RNAalifold did not find a consensus structure, while a SCI ≈ 1 reflects a set of perfectly conserved structures (Washietl et al. 2005). Consequently, only those datasets with an overall SCI ≥ 0.70 were retained for further analysis, in order to ensure that the RNA secondary structures under investigation were evolutionary conserved.

*Model selection and phylogenetic analyses*

For each dataset, the best-fitting substitution model for phylogenetic reconstruction was chosen using a Perl script included in the package PHASE-3.0 ("model_selection.pl"; Allen & Whelan 2014). The inputs to this script were (i) the sequence alignment, (ii) the inferred secondary structure and, (iii) an initial neighbor-joining tree, estimated under the Tamura-Nei model using Mega version 5 (Tamura et al. 2011). The Perl script compares two DNA substitution models (HKY and GTR), sixteen different RNA substitution models (seven RNA7 and nine RNA16 models), and the inclusion or exclusion of a gamma distribution model of among-site rate variation. The script identifies the best-fitting model as that with the lowest corrected Akaike's Information Criterion value (Akaike 1974; Burnham & Anderson 2002): AICc= - ln(L) +2k +2k(k+1)/(n-k-1), where K is the number of parameters, L is the likelihood, and n is sample size.

Phylogenetic trees were estimated using the Bayesian Monte Carlo Markov Chain (MCMC) approach implemented in the program mcmcphase that is part of the package PHASE-3.0. This program allows the inference of a phylogenetic tree under a

"mixed model", in which a DNA substitution model is assigned to unpaired positions and an RNA substitution model is assigned to paired positions. Phylogenetic trees were estimated using the best-fitting model, which was always a mixed model, or the DNA-only model. At least two independent MCMC runs, each with >1,000,000 states, were computed and a 10% burn-in was removed from each before analysis.

After combining the output of both MCMC runs, convergence was checked visually by plotting sampled values of the likelihood, posterior and priors. After convergence was confirmed, an extended majority rule consensus phylogenetic tree was obtained for each dataset using the program "mcmcsummarize" from the PHASE package. The phylogeny obtained under the mixed model (which, for all datasets, was found to be the best-fitting model) was then used as a fixed topology to estimate branch lengths, by running mcmcphase with either the DNA or the mixed substitution model.

Next, sites in each sequence alignment were partitioned into two independent datasets that included only paired or unpaired sites. Branch lengths were estimated separately from these two partitions, using the same fixed topology as above. A DNA substitution model was used for the unpaired sites partition, and either the best-fitting DNA substitution model or the RNA substitution model was used for the paired sites partition.

*Comparison of branch lengths and tree topologies*

Tree lengths (the sum of all branch lengths in a phylogeny) were calculated from the consensus trees obtained from the complete alignments. Tree lengths obtained from paired sites (either under a DNA or RNA substitution model) and

unpaired sites (always under a DNA substitution model) were calculated in the same way. Branch lengths estimated under the DNA and mixed substitution models were compared by means of paired Wilcoxon tests.

To assess the effects on tree topology of using or not an RNA specific model, we computed distributions of Robinson-Foulds (RF) distances. The RF distance between two tree topologies is a measure of how different they are (Robinson & Foulds 1981). We computed three different distributions of RF distances: (i) between pairs of tree topologies sampled from the same posterior distribution when an RNA substitution model was included in the analysis, (ii) between pairs of tree topologies sampled from the same posterior distribution, when an RNA substitution model was *not* included in the analysis, and (iii) between a tree from the posterior used in (i) and a tree from the posterior used in (ii). For cases (i) and (ii), trees were sampled without replacement, thus ensuring that a given MCMC state could not be compared with itself. A total of 9000 pairwise tree comparisons were made. All RF distances from a given dataset were normalized according to the number of taxa 2 x number of taxa - 6, for unrooted trees). Distributions (i) and (ii) represent the degree of statistical uncertainty in tree topology arising from inference under a given substitution model, whereas distribution (iii) represents the degree of difference in tree topologies obtained by inference under two different models. Thus, a comparison of distribution (iii) with distributions (i) and (ii) allows us to determine whether the effect on tree topology of using an RNA-substitution model is greater or less than estimation uncertainty alone.

We assessed whether distributions (i) and (ii) were significantly different from distribution (iii) by performing 9000 pairwise comparisons between RF distances

randomly sampled from distributions (i) or (ii) and from distribution (iii). The probability that the two distributions are different is computed as the number of cases in which the value of RF distance from (iii) is larger than the value sampled from (i) or (ii) divided by the total number of comparisons (Abecasis et al. 2009). p-values obtained from the same virus/viroid were then corrected with the false discovery rate method (FDR; Benjamini & Hochberg 1995). The distributions of normalized RF distances and their statistical comparisons were computed using an R script (available on request) that utilizes the phangorn package for R (Schliep 2016).

## Results

### *RNA secondary structure inference*

Structure Conservation Index (SCI) values were calculated with RNAz. Values of SCI ≤ 0.70 were found in only five viral datasets: HCV (SCI=0.40), DENV (0.40), HIV-1 (0.66), RV (0.66) and HDV (0.66) which, correspondingly, were also the virus species with large average pairwise genetic distances (Table 1). For genetically diverse viruses like these, the evolutionary conservation of secondary structure will be higher at the sub-genomic level. Therefore, for DENV, HIV-1, RV and HDV we attempted to infer secondary structures separately for taxonomic units below the species level (e.g. subtypes, genotypes, etc.), with the aim of obtaining a consensus structure comprising paired-sites that are present in >75% of genotypes/subtypes within a species. However, for all cases except HDV, the percentage of paired sites along the genome that were conserved among genotypes was too low (≤12%). For HDV, we found 46% of paired sites along the genome were conserved among the 8 HDV genotypes in the

virus, each with SCI > 0.70 separately. Therefore, for HCV, HIV-1, DENV and RV, we analyzed a less diverse sub-genomic taxonomic unit rather than the whole viral species (subtype 1b for HCV, genotype 4 for DENV, subtype B for HIV-1 and lineage C1 for RV). All these genotype/subtype datasets had SCI > 0.70 and were therefore analyzed further. Arc diagrams representing the RNA minimum free-energy consensus secondary structures that were obtained with RNAalifold for each dataset with SCI > 0.70 are shown in Supplementary Figure S.1. The percentage of nucleotides forming base-pairs of these alignments, which were further analyzed, ranged between 46% (HDV) and 78% (AGVd) (Table 1).

*Model selection and phylogenetic analyses*

For each analyzed dataset, the best-fitting model (i.e. the model with the lowest AICc value) was a mixed model which assigned a DNA substitution model (either GTR or HKY) to unpaired sites and a RNA16 substitution model to paired sites (Table 1).

**Table 1.** Size (number of taxa and sequence length), overall mean genetic distance, Structure Conservation Index (SCI), percentage of base-paired nucleotides and best-fitting evolutionary model of each viroid and virus dataset analyzed

| | n (taxa) | sequence length (nt) | mean pairwise genetic distance (p-distance ± SE) | SCI | % nucleotides forming base-pairs | Best-fitting model |
|---|---|---|---|---|---|---|
| Viroids | | | | | | |
| TASVd | 22 | 374 | 0.036 ± 0.004 | 0.91 | 68% | HKY_Γ+RNA16C_Γ |
| CeVd | 178 | 369 | 0.041 ± 0.005 | 0.92 | 70% | GTR_Γ+RNA16E_Γ |
| CLVd | 14 | 379 | 0.061 ± 0.008 | 0.88 | 68% | GTR_Γ+RNA16A_Γ |
| GYSVd | 24 | 352 | 0.128 ± 0.012 | 0.84 | 65% | GTR_Γ+RNA16C_Γ |
| AGVd | 27 | 368 | 0.020 ± 0.004 | 0.91 | 78% | HKY_Γ+RNA16C |
| PSTVd | 88 | 356 | 0.019 ± 0.003 | 0.97 | 69% | HKY_Γ+RNA16C_Γ |
| Viruses | | | | | | |
| HDV | 121 | 1543 | 0.204 ± 0.005 | 0.66* | 46%# | GTR_Γ+RNA16D_Γ |
| Sudan Ebolavirus | 7 | 18875 | 0.032 ± 0.001 | 0,90 | 66% | GTR_Γ+RNA16A |
| DENV | 23 | 10628 | 0.263 ± 0.003 | 0.40* | NC | NC |
| DENV-4 | 8 | 10628 | 0.088 ± 0.002 | 0.75 | 63% | GTR_Γ+RNA16A_Γ |
| HCV | 42 | 9584 | 0.292 ± 0.002 | 0.40* | NC | NC |
| HCV-1b | 20 | 9584 | 0.087± 0.001 | 0.82 | 67% | GTR_Γ+RNA16A_Γ |
| HIV-1 | 18 | 9681 | 0.126 ± 0.002 | 0.64* | NC | NC |
| HIV-1B | 33 | 9681 | 0.056 ± 0.001 | 0.74 | 59% | GTR_Γ+RNA16D_Γ |
| FMDV | 19 | 8192 | 0.135 ± 0.002 | 0.75 | 61% | GTR_Γ+RNA16D_Γ |
| Measles | 20 | 15893 | 0.042 ± 0.001 | 0.89 | 64% | GTR_Γ+RNA16A_Γ |
| Rubella | 35 | 9758 | 0.060 ± 0.002 | 0.90 | 66% | GTR_Γ+RNA16A_Γ |
| Mumps | 20 | 15355 | 0.045 ± 0.001 | 0.86 | 63% | GTR_Γ+RNA16A_Γ |
| Rabies | 26 | 11923 | 0.111 ± 0.001 | 0.66 | 5%# | GTR_Γ+RNA16A_Γ |
| Rabies C1 | 20 | 11923 | 0.088 ± 0.001 | 0.74 | 64% | GTR_Γ+RNA16A_Γ |

* SCI below 0.70.
# Percentage of nucleotides forming base pairing, after obtaining a consensus structure comprising paired-sites that are present in >75% of genotypes/subtypes within a species.
NC: not computed

Bayesian phylogenies were estimated using mcmcphase, which forms part of the PHASE-3.0 package. To examine the effect of including a RNA substitution model in the analysis, we estimated branch lengths on a fixed topology under two different substitution models: first, using the best-fit model (which, as noted above, was always

a mixed model), and second, using the best-fitting DNA substitution model. Tree lengths (the sum of all branch lengths) obtained under the two abovementioned models were statistically compared using paired Wilcoxon tests. Ratios of the tree lengths obtained the mixed and DNA models (i.e. L(mixed)/L(DNA)) were calculated and statistically compared (see Table 2). While the effect on branch length estimates of using a mixed model was near zero for PSTVd and AGvd (ratios= 0.99 and 1.00, respectively; p-values > 0.05), for the other viral and viroid datasets there was a significant increase in tree length of > 25% (p-values <0.05). The largest effects were observed for TasVd, CLVd and GYSVd, with L(mixed)/L(DNA) ratios of 6.532, 2.795, and 2.675, respectively.

We also compared the estimated tree lengths obtained independently for unpaired and paired sites. The L(paired)/L(unpaired) ratios obtained under the DNA model reported in Table 2 reflect the evolutionary constraints imposed by base-pairing; with the exception of PSTVd, tree lengths estimated from paired sites were > 29% shorter than those estimated from unpaired sites. However, when an RNA model was used for paired sites, the L(paired)/L(unpaired) ratios increased and, in several cases, paired sites under an RNA model yielded remarkably larger tree lengths than unpaired sites (AGVd, GYSVd, HDV, MeV, MuV, DENV-GT4, RV) (Table 2).

**Table 2.** Estimates and comparisons of tree lengths (L) estimated from DNA and mixed models, for all sites, paired sites, and unpaired sites

| | L(DNA model) | L(mixed model) | Ratio (mixed/DNA) | P value log(DNA vs mixed) | L(unpaired sites) | L(paired sites, DNA model) | L(paired sites, RNA model) | Ratio(paired-DNA model/unpaired) | Ratio(paired-RNA model/unpaired) |
|---|---|---|---|---|---|---|---|---|---|
| Viroids | | | | | | | | | |
| TASVd | 0.47 | 3.07 | 6.532 | <0.001 | 3.05 | 0.51 | 1.82 | 0.167 | 0.597 |
| AGVd | 4.91 | 4.93 | 1.004 | 0.341 | 0.55 | 0.31 | 1.13 | 0.563 | 2.055 |
| CeVd | 30.41 | 33.81 | 1.111 | <0.001 | 34.18 | 33.4 | 31.42 | 0.977 | 0.919 |
| CLVd | 0.44 | 1.23 | 2.795 | <0.001 | 1.36 | 0.74 | 1.01 | 0.544 | 0.743 |
| GYSVd | 0.77 | 2.06 | 2.675 | <0.001 | 2.11 | 0.88 | 2.21 | 0.417 | 1.047 |
| PSTVd | 17.15 | 17.04 | 0.994 | 0.457 | 17.16 | 17.25 | 17.16 | 1.005 | 1.000 |
| Viruses | | | | | | | | | |
| HDV | 9.09 | 12.15 | 1.337 | <0.001 | 12.44 | 7.50 | 15.35 | 0.603 | 1.234 |
| Sudan Ebolavirus | 0.07 | 0.10 | 1.429 | <0.001 | 0.10 | 0.05 | 0.13 | 0.500 | 1.300 |
| DENV-4 | 0.51 | 0.61 | 1.196 | <0.001 | 0.66 | 0.45 | 0.89 | 0.682 | 1.348 |
| HCV-1b | 1.16 | 1.74 | 1.500 | <0.001 | 1.76 | 0.87 | 1.84 | 0.494 | 1.045 |
| HIV-1 B | 1.49 | 2.15 | 1.443 | <0.001 | 2.15 | 0.96 | 2.20 | 0.446 | 1.023 |
| FMDV | 2.00 | 2.55 | 1.275 | <0.001 | 2.61 | 1.48 | 2.65 | 0.567 | 1.015 |
| Measles | 0.33 | 0.42 | 1.273 | <0.001 | 0.42 | 0.30 | 0.79 | 0.714 | 1.881 |
| Rubella | 0.71 | 1.06 | 1.493 | <0.001 | 1.06 | 0.60 | 1.33 | 0.566 | 1.255 |
| Mumps | 0.35 | 0.44 | 1.257 | <0.001 | 0.44 | 0.31 | 0.84 | 0.705 | 1.909 |
| Rabies C1 | 0.94 | 1.15 | 1.223 | <0.001 | 1.15 | 0.87 | 2.07 | 0.757 | 1.800 |

*Topology could not be fixed for branch lengths inference due to unresolved bipartitions, and a Wilcoxon rank sum test was performed instead of a paired test.

For each dataset analyzed, three different RF distance distributions were obtained as described above (Figure 1). The randomization tests evidenced that only for HDV and HCV-1b a significantly different distribution of RF distances was obtained when comparing tree states sampled from the same posterior (under a mixed model) than from comparing tree states from the two different posterior distributions, with shorter RF-distances under the mixed model in both cases (HDV: p-value=0.016; HCV: p-value= 0.027). In HCV, a significantly different distribution of shorter RF distances was also obtained when comparing tree states sampled under the DNA model with that obtained comparing both posterior distributions (p- value = 0.048). All the reported significant p-values were still significant after FDR correction. For these viruses, the consensus phylogenetic trees obtained under the mixed model presented more highly-supported nodes (defined by a posterior probability ≥ 0.90) than those obtained under a DNA-only model: 82 (mixed model) vs 68 (DNA model) in HDV, and 14 vs 13 in HCV (Supplementary Figure S.2). In HIV-1 subtype B, the randomization test yielded a p-value = 0.065 and could not be considered significant. However, the use of a mixed model increased notably the number of well-supported clades in the consensus phylogenetic tree inferred, from 10 to 23 (Supplementary Figure S.2). The RF distances obtained when comparing the consensus trees (DNA vs mixed model) of these three viruses were 0.22 for HDV, 0.29 for HCV-1b and 0.40 for HIV-1 subtype B.

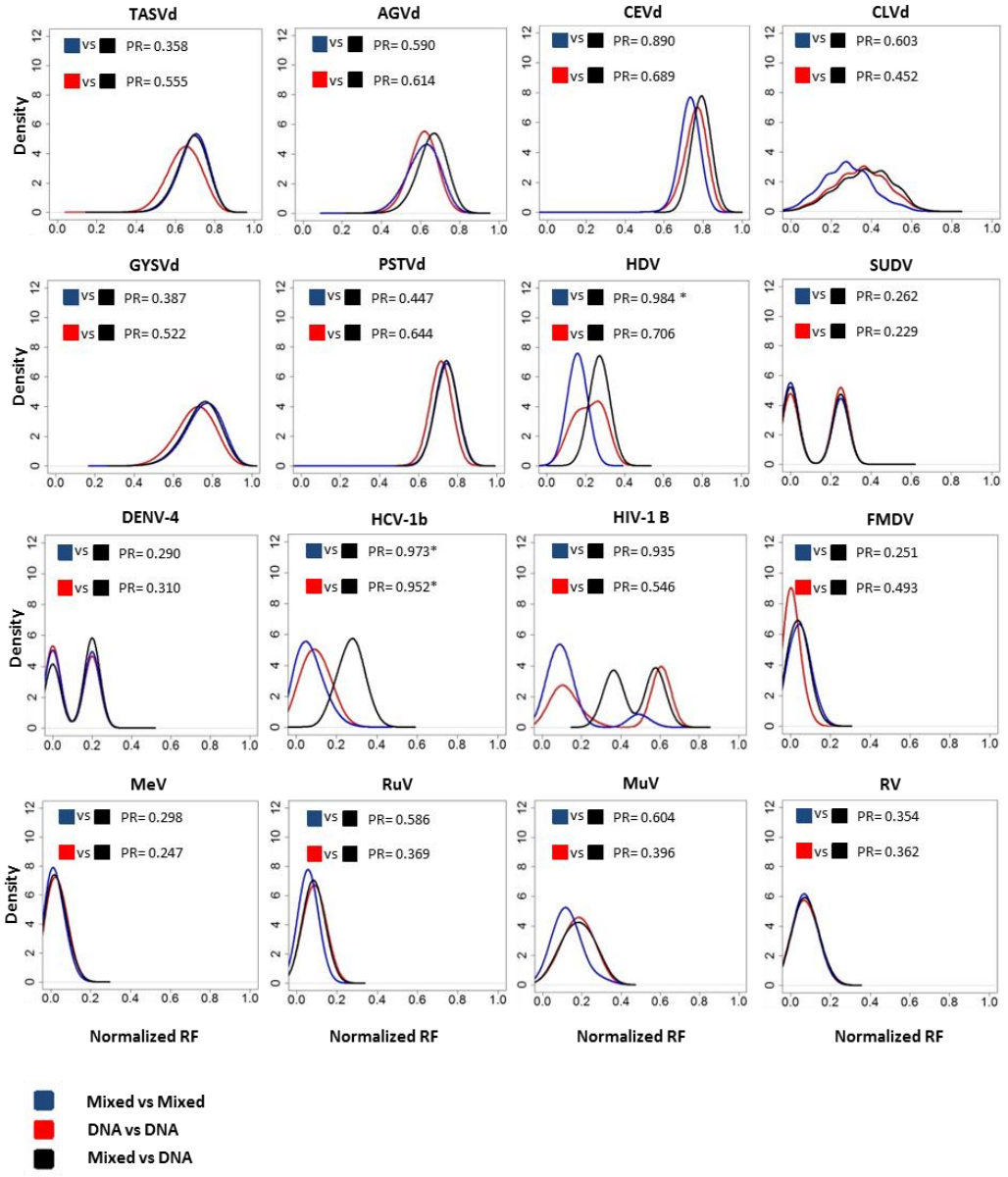



**Figure 1.** Density plots representing, for each dataset, the RF distributions obtained by comparing tree states from the same posterior distribution (either including or excluding the RNA model) or from the two different posterior distributions. The results of the randomization tests are shown as the proportion of cases when an RF value obtained comparing states from the same posterior (blue=under mixed model; red=under DNA model) was lower than the RF value obtained comparing states from the two different posterior distributions (black= mixed vs DNA models). Significant values after FDR correction are labelled with "*".

**Discussion**

In this work we assessed the effect of RNA substitution models on the phylogenetic inference of viroids and RNA viruses based on complete genome sequences. We first investigated whether using an RNA-specific model provides a better fit to the data than the conventional DNA substitution models that are widely used to study viral evolution.

In all datasets the best-fit model was a mixed model that uses a nucleotide model for unpaired sites and an RNA model for paired sites. It is important to note that 16-state RNA substitution models outperformed 7-state RNA models in all instances. The main difference between these families of RNA models is that 7-state models pool all mismatches (pairs of nucleotides that do not form stable base-pairs) in a single state in the Markov chain while 16-state models consider each mismatch as separate state. A special case is RNA16C, in which the 10 different MM have the same transition probabilities, thus being considered an extension of an RNA7 model (Savill et al. 2001; see also the PHASE 3.0 manual at https://github.com/james-monkeyshines/rna-phase-3).

For most of the analyzed viroids, the RNA16C model was chosen as the best-fitting model, whereas for RNA viruses, RNA16A was chosen as the best-fitting model in most cases. RNA16A and C have been previously reported to fit best in different non-coding RNA datasets because, unlike other RNA16 models, they allow a non-zero rate of double substitutions, thus counting complementary changes as a single step (Savill et al. 2001). Allen & Whelan (2014) assessed the best-fitting models for analyzing the evolution of human non-coding RNAs and found that, for the majority of

RNA types, stable pairs models (RNA7A-G and RNA16C) and stable sets models (RNA16D, E, F) fitted the best for such data. They concluded that the former were usually selected in datasets where few changes occurred, while the later were selected when the consensus secondary structure contained higher proportions of paired sites. Our results suggest that models allowing non-zero rates of double substitutions fit best for viroids and viruses.

Bayesian phylogenies were inferred using the best-fitting mixed model and using a DNA substitution model. This allowed us to assess the differences in estimates of branch lengths and trees topologies when an RNA model is included in analysis. In all datasets (except PSTVd and AGVd), the use of a RNA model led to trees with longer branch lengths. Among those datasets where the use of an RNA model led to a significant increase in branch lengths, the increase in total tree length ranged between 11% (CEVd) up to 653% (TASVd). Under a DNA model, the tree lengths estimated from paired sites were always much shorter than those estimated from unpaired sites, and such differences were reduced when the RNA model was used for paired sites. A lower number of substitutions at paired sites, compared to unpaired sites, is expected due to the likely higher evolutionary constraints at paired sites (Nasrallah et al. 2011). However, in some datasets tree lengths estimated from paired sites under a RNA model were considerably larger than those estimated from unpaired sites (especially in AGVd, measles, mumps and rabies virus; Table 2). These results suggest that, in such cases, while the use of DNA models strongly underestimates the number of substitutions, RNA models may lead to an overestimate of the number of substitutions along the inferred tree. It is important to note that PHASE-3.0 gives the estimated branch lengths as expected number of substitutions per nucleotide, even when a RNA

model is included (and not as expected number of substitutions per base-pair), thus allowing the comparison of branch lengths estimated under different models (Allen & Whelan 2014).

Compared to viruses, viroid phylogenies were characterized by presenting larger values of RF distances between sampled from the posterior distributions, regardless of the evolutionary model used. Furthermore, the comparisons of RF distributions show that, with the exception of HCV and HDV, the use of a mixed model to infer viral and viroid phylogenies had no significant effect on the tree topologies. In those significant cases, including an RNA model was associated to an increase in the number of well-supported nodes from the resulting consensus phylogenetic tree.

The RF distributions for SUDV, DENV-4 and HIV-1B were bimodal. In SUDV and DENV-4 this can be explained because there were few sequences, and RFs were either zero or very low, because tree states were either identical or very similar. In the case of HIV-1B, the RF distribution obtained from the posterior distribution sampled under the RNA model was much less bimodal than the others, and could be associated with a substantial reduction in the topologic uncertainty that was obtained under the mixed model: in the consensus tree, the number of well supported nodes increased from 10 (DNA-only) to 23 (mixed model).

One of the limitations that may hamper the use of RNA models for phylogenetic inference is the lack of reliable, and representative RNA structures at the taxonomic unit to be analyzed. In this work, we used consensus RNA structures inferred only by computational approaches as input, although the accuracy of the RNA structures used in the analyses could have been improved by using experimental approaches, such as

RNAse mapping or SHAPE reactivity (Wilkinson et al. 2006). However, to date there are very few secondary structures of complete viral genomes obtained experimentally and they have obtained from single genome sequences (Watts et al. 2009; Mauger et al. 2015), thus ignoring the diversity in RNA secondary structure that is known to exist, even below the species level (Tuplin et al. 2004; Mauger et al. 2015). Because of this lack of representative, experimental RNA secondary structures, we used a computational method implemented in RNAalifold which allowed the inference of the consensus structure of alignments of different, yet related, RNA sequences. This method is known to improve the prediction of secondary structures compared to those obtained only with individual sequences, and allows to obtain a representative structure for the analyzed dataset (Hofacker et al. 2002; Bernhart et al. 2008). Furthermore, we only included in the analyses those datasets corresponding to taxonomic levels showing evolutionarily conserved structures, to ensure that the inferred structures fitted well for each dataset. However, it is important to mention that, *in vivo*, the same primary sequence can fold into alternative structures (Schultes & Bartel 2000). In this way, differences between RNA structures existing *in vivo* and those inferred computationally are expected to exist, and such differences should be larger in the case of viruses with linear RNA genomes than in HDV or viroids, which tend to form simpler, rod-like structures. For this reason, the discussed results should only be interpreted with some caution, as they have been derived in an *in silico* context.

In conclusion, we found that in all viroid and RNA virus datasets analyzed, the selective constraints imposed by RNA secondary structures have significant impacts upon phylogenetic reconstructions. Model selection analyses concluded that, in all

cases, assigning an RNA model to paired sites outperformed the use of a DNA-only model for phylogenetic reconstruction from complete genome sequences. The effect of phylogenetic inference method on branch lengths is significant in most of the datasets analyzed. However, with some exceptions the use of an RNA model does not have a significant effect on the topology inferred. Furthermore, the high uncertainty that characterizes phylogenetic inference of viroid datasets did not decrease when these models were included. To date, viral and viroid phylogenies have always been reconstructed using DNA substitution models. However, in light of our results, we recommend that such analyses should consider the inclusion of RNA models, as they describe better the evolution of paired sites. In addition, it would be of particular importance for phylogeny software that implements molecular clock models, such as BEAST (Drummond & Rambaut 2007), to include the option of using RNA substitution models, as diversification dates and evolutionary rates inferred for RNA viruses under RNA models might be different from estimates obtained without considering secondary structure.

## References


Abecasis AB, Vandamme A-M, Lemey P. 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J. Virol. 83:12917–12924.

Akaike H. 1974. A New Look at the Statistical Model Identification. Autom. Control 19:716–723.

Allen JE, Whelan S. 2014. Assessing the state of substitution models describing noncoding RNA evolution. Genome Biol. Evol. 6:65–75.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57:289–300.

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474.

Burnham K, Anderson D. 2002. Model selection and multi-model inference: a practical information-theoretic approac. Springer Verlag

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Damgaard CK, Andersen ES, Knudsen B, Gorodkin J, Kjems J. 2004. RNA interactions in the 5' region of the HIV-1 genome. J. Mol. Biol. 336:369–379.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Gruber AR, Neuböck R, Hofacker IL, Washietl S. 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Res. 35:W335-8.

Hofacker IL, Fekete M, Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 319:1059–1066.

Hutchins CJ, Rathjen PD, Forster AC, Symons RH. 1986. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. Nucleic Acids Res. 14:3627–3640.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kosakovsky Pond SL, Mannino F V, Gravenor MB, Muse S V, Frost SDW. 2007. Evolutionary model selection with a genetic algorithm: a case study using stem RNA. Mol. Biol. Evol. 24:159–170.

Lai D, Proctor JR, Zhu JYA, Meyer IM. 2012. R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Res. 40:e95.

Linhart H, Zucchini W. 1986. Model selection. John Wiley & Sons, New York

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6:26.

Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, Weeks KM. 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc. Natl. Acad. Sci. U. S. A. 112:3692–3697.

Muse S V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics 139:1429–1439.

Nasrallah CA, Mathews DH, Huelsenbeck JP. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst. Biol. 60:60–73.

Pelletier J, Sonenberg N. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334:320–325.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical computing.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Sanjuán R, Bordería A V. 2011. Interplay between RNA structure and protein evolution in HIV-1. Mol. Biol. Evol. 28:1333–1338.

Sanjuán R, Forment J, Elena SF. 2006. In silico predicted robustness of viroids RNA secondary structures. I. The effect of single mutations. Mol. Biol. Evol. 23:1427–1436.

Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157:399–411.

Schliep K. 2016. Package “phangorn”. Phylogenetic Analysis in R. CRAN. Available at https://cran.r-project.org/web/packages/phangorn/index.html

Schöniger M, von Haeseler A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. 3:240–247.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7–9.

Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. 2011. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural

constraints. Retrovirology 8:87.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Tellam J, Smith C, Rist M, Webb N, Cooper L, Vuocolo T, Connolly G, Tscharke DC, Devoy MP, Khanna R. 2008. Regulation of protein translation through mRNA structure influences MHC class I loading and T cell recognition. Proc. Natl. Acad. Sci. U. S. A. 105:9319–9324.

Thurner C, Witwer C, Hofacker IL, Stadler PF. 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. 85:1113–1124.

Tillier ER, Collins RA. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002.

Tuplin A, Evans DJ, Simmonds P. 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J. Gen. Virol. 85:3037–3047.

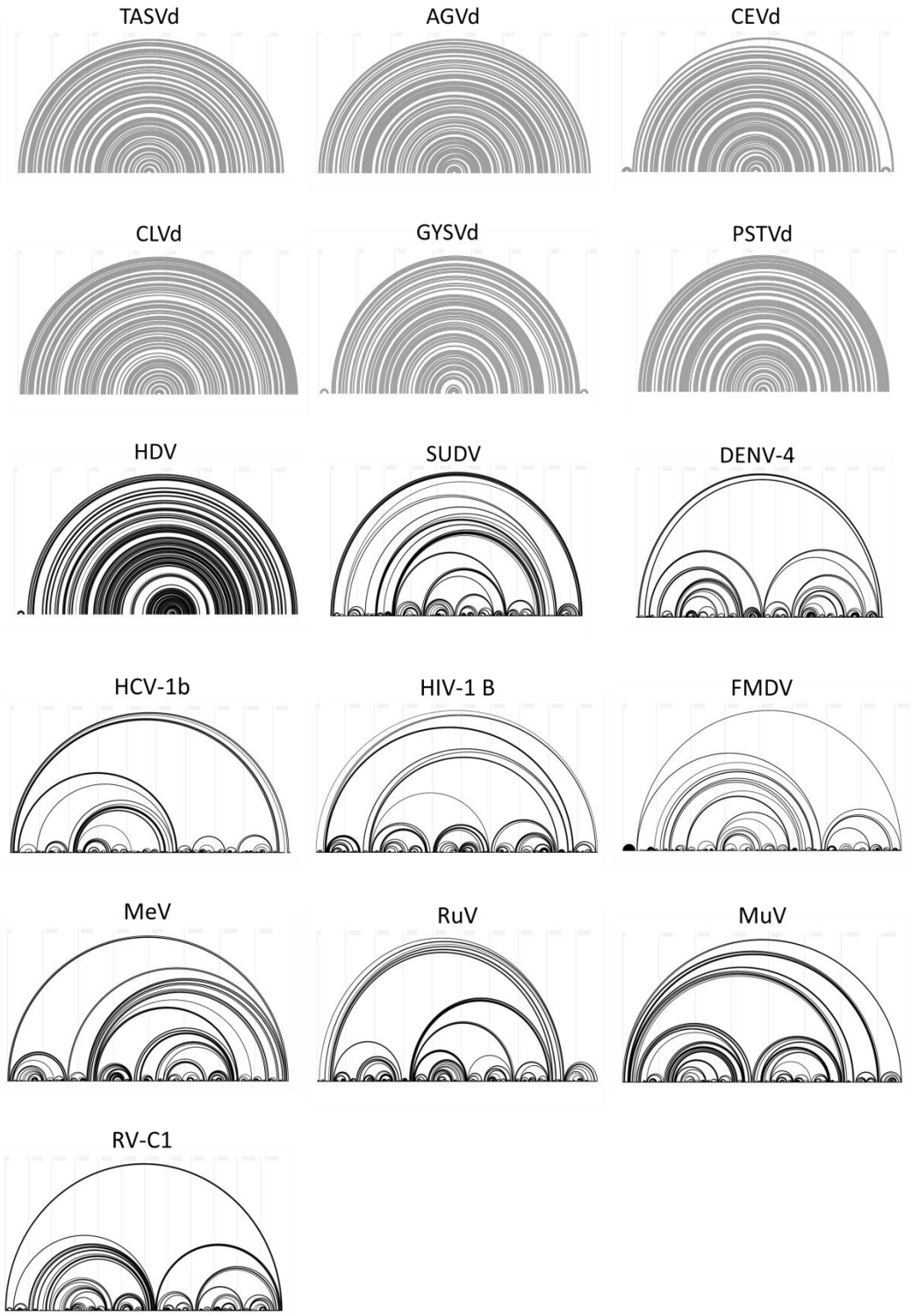
Wang KS, Choo QL, Weiner AJ, Ou JH, Najarian RC, Thayer RM, Mullenbach GT, Denniston KJ, Gerin JL, Houghton M. 1986. Structure, sequence and expression of the hepatitis delta (delta) viral genome. Nature 323:508–514.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U. S. A. 102:2454–2459.

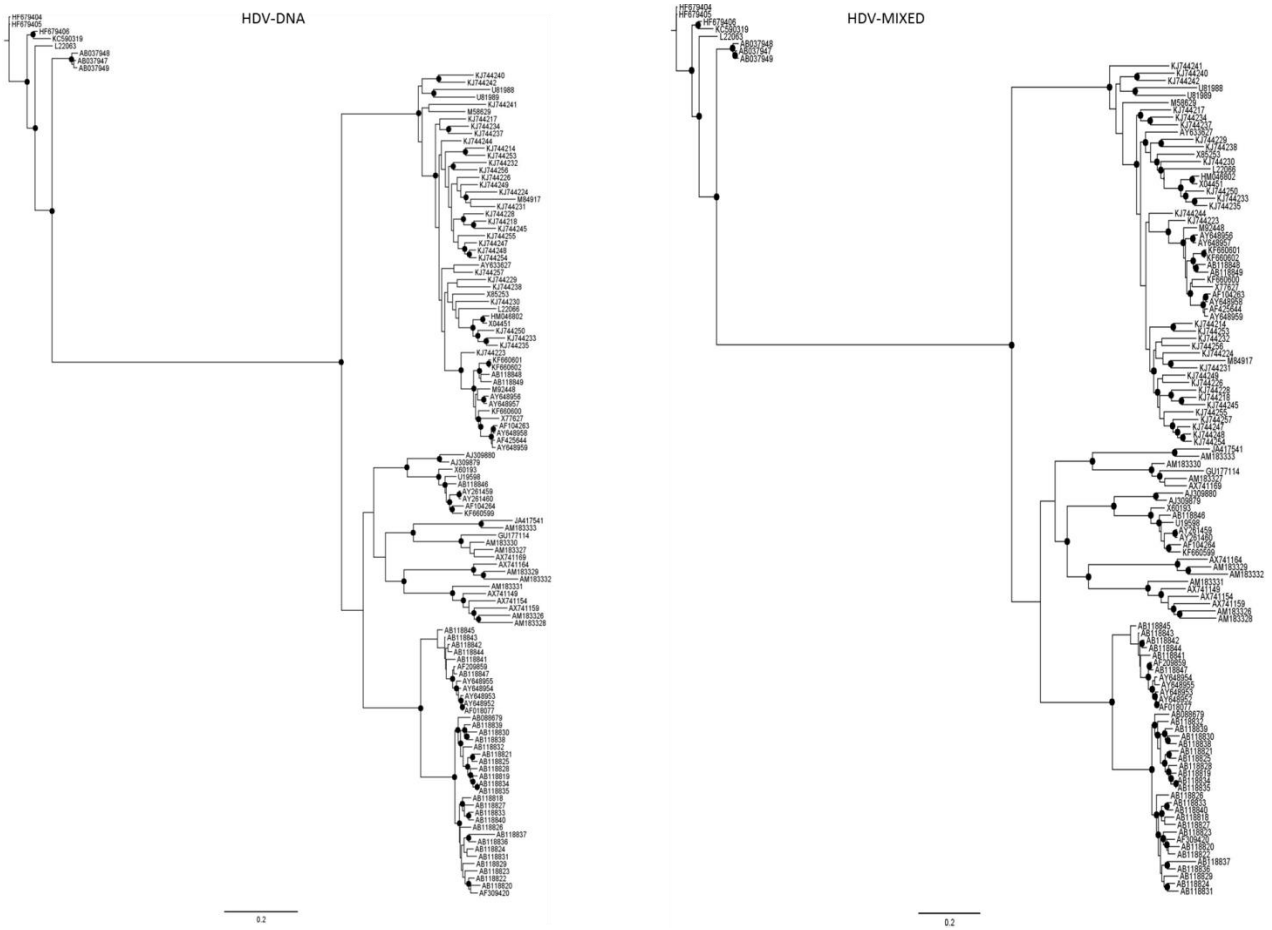
Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460:711–716.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2’-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat. Protoc. 1:1610–1616.

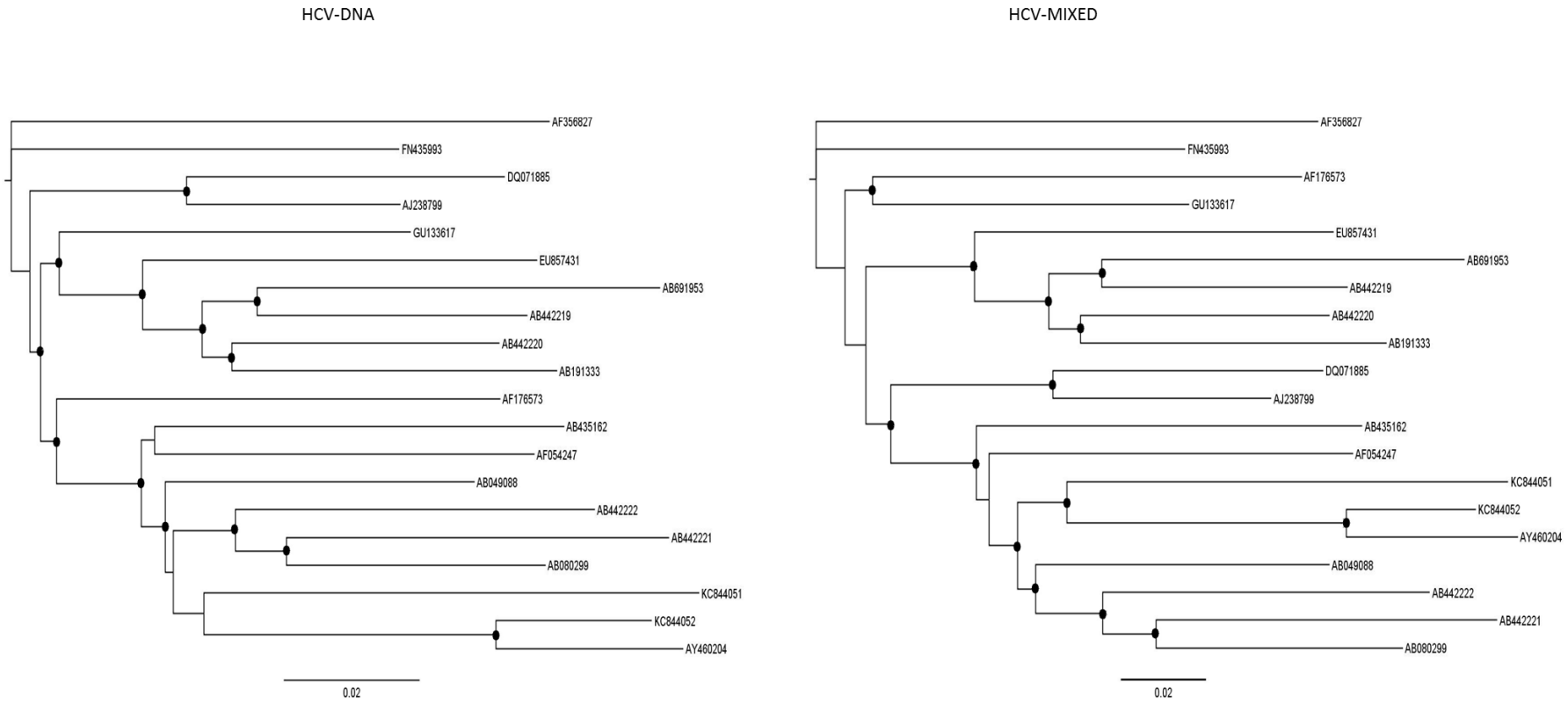
## Supplementary material



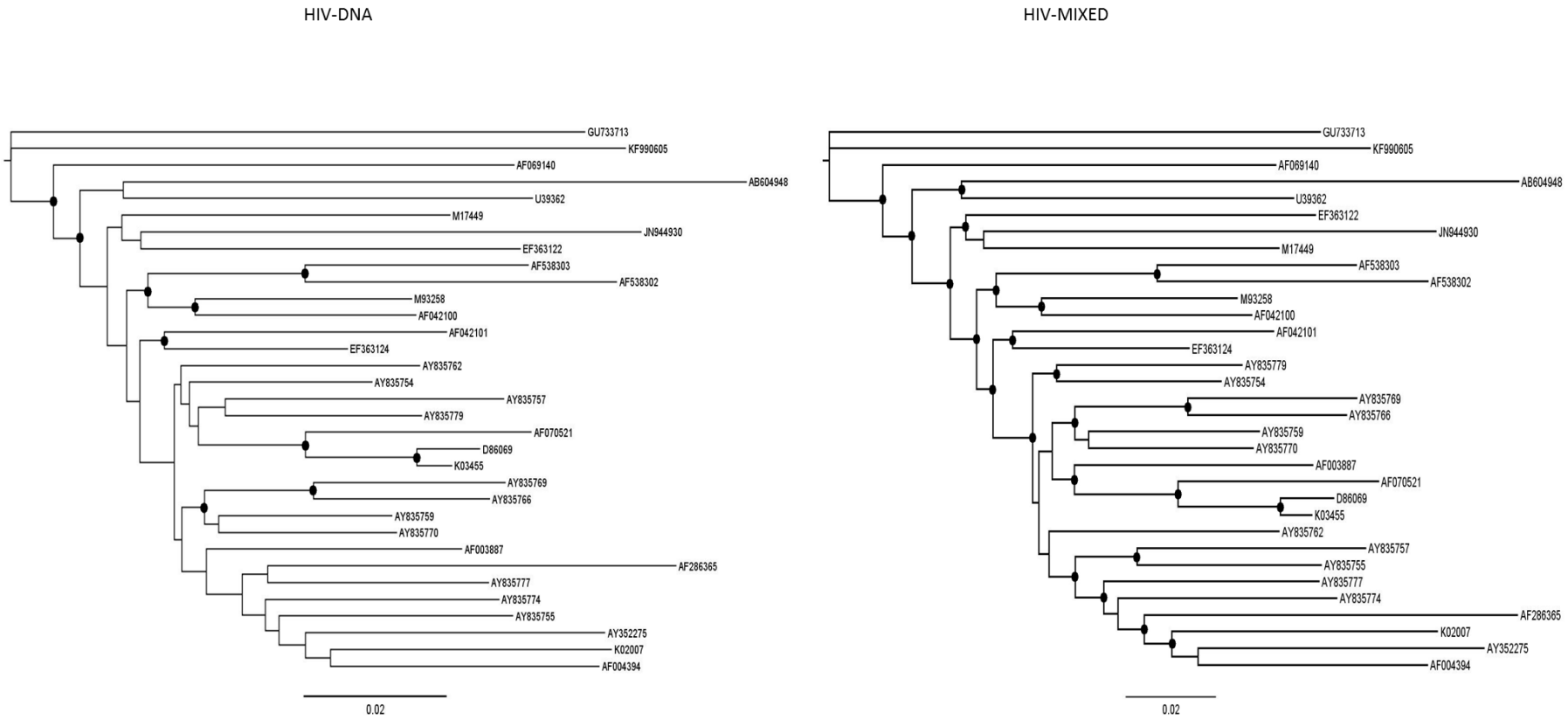

**Supplementary Figure S.1.** Arc plots representing the consensus RNA secondary structure inferred with RNAalifold of each dataset analyzed with PHASE-3.0. Arcs represent base-pairing relationships.




**Supplementary Figure S.2.** Phylogenetic trees of HDV, HCV-1b and HIV-1B obtained with PHASE-3.0 either including or excluding the RNA model. Black circles represent nodes supported by PPs ≥0.90.




**Supplementary Figure S.2 (cont).** Phylogenetic trees of HDV, HCV-1b and HIV-1B obtained with PHASE-3.0 either including or excluding the RNA model. Black circles represent nodes supported by PPs ≥0.90.




**Supplementary Figure S.2 (cont).** Phylogenetic trees of HDV, HCV-1b and HIV-1B obtained with PHASE-3.0 either including or excluding the RNA model. Black circles represent nodes supported by PPs ≥0.90.

# *3.* Discussion and conclusions

### 3.1- Discussion

This thesis was aimed to address several questions about the molecular evolution of RNA viruses. Although the work has mainly focused on the molecular epidemiology of HIV and HCV, it also intended to address other topics: estimating and comparing the genomic evolutionary rates of the main HIV-1 subtypes and CRFs, elucidating the different selective constraints shaping the genomes of HCV-1a and 1b, and assessing the effect of RNA models on the phylogenetic inference of RNA viruses and viroids.

HIV has caused one of the largest pandemics in the history of humankind, causing more than 36 million deaths (Cleves 2008; UNAIDS 2015) worldwide. Although efforts for HIV prevention and treatment have managed to stabilize the number of infections, its incidence is still growing in many countries, including Spain, particularly in risk groups such as MSM (ECDC 2013). In chapters 1, 2, 3 and 4, I performed molecular epidemiology analyses in two different Spanish regions: the Basque Country and the Comunitat Valenciana (CV), for which I had access to a total of 1,727 and 1,806 HIV-1 sequences from different patients respectively. These datasets were analyzed by means of phylogenetic and coalescent analyses for detecting and characterizing local transmission clusters occurring in these regions. Also, univariate and multinomial comparisons were performed to detect epidemiological differences between HIV-1 subtypes and population groups.

I found differences between both Spanish regions in the proportion of patients being included in transmission clusters (Basque Country < 30%; CV = 57%). Such differences could be due to different definitions for transmission cluster, more

stringent in the work from the Basque Country than in that from the CV. Furthermore, the inclusion of chronically infected patients in the analyses from the Basque Country may have hampered the identification of transmission clusters, due to the longer external branches that would represent these patients in the phylogenetic trees. However, Yebra et al. (2013) also found a low proportion of patients belonging to transmission clusters (<20%) following a similar cluster definition to the one used here. These results suggest that, in the different Spanish regions, the contribution of local transmissions to their HIV subepidemics may not be the same.

In line with the epidemic scenario in Western Europe and with the results obtained by Yebra et al. (2013), MSM represented the most vulnerable group to HIV infections, both in the Basque Country and Valencia. They were significantly more prone to form transmission clusters than other risk groups, even in the Basque Country, where they were not the most common transmission risk group. MSM were also the majority group in large transmission clusters, especially in the most recent ones. The increased vulnerability of this group is evidenced by a very large transmission cluster found in the city of Valencia, affecting more than 100 patients between 2010 and 2014. Delgado et al. (2015) also found a very large (n = 100) HIV-1 subtype F transmission cluster affecting MSM from different Spanish localities. Bezemer et al. (2015) analyzed the HIV-1 B epidemic among MSM in the Netherlands until 2013, and found an increasing HIV epidemic in which many of the subepidemics were longstanding.

More than 15% of the patients from the CV were infected with non-B variants, revealing a diverse HIV epidemic. This result is in line with those obtained by González-

Alba et al. (2011) in Madrid. The high migration and tourism rates existing both in Madrid and the CV may explain the complexity of their HIV epidemics, and this appears to be evidenced in the significant association that I found between non-B subtypes and migrants. Interestingly, in this thesis I have also reported the detection of two transmission clusters of CRF19_cpx occurring in the CV among local MSM, as the first evidence of the expansion of this highly pathogenic variant outside Cuba. In this way, although my results suggest that non-B HIV-1 variants are not well established among locals yet, this situation could change soon without intensified HIV prevention campaigns in vulnerable groups, such as migrants and MSM.

It is important to mention that no HIV-1 outbreaks had been reported until 2016 in the CV, being undetected to the local public health officials when using traditional epidemiological analyses. By means of evolutionary analyses I have detected hundreds of local transmission clusters, each of them potentially representing an outbreak. These results provide evidence for the importance of evolutionary analyses in epidemiology, which should be used in combination with traditional epidemiology approaches for the correct detection of outbreaks.

In Chapter 5, I estimated and compared the genomic evolutionary rates of some of the most prevalent HIV-1 subtypes (A1, B, C, D and G) and CRFs (CRF01_AE and CRF02_AG). Most previous estimates of tMRCA and substitution rate for HIV-1 have been inferred from independent genes (most frequently *pol* and *env*; Abecasis et al. 2009; Wertheim et al. 2012), ignoring other genomic regions that may have different mutation rates and/or selective constraints. After obtaining independent and representative genomic datasets of each HIV-1 variant from a public database (LANL),

their tMRCAs and evolutionary rates were estimated using a Bayesian coalescent phylogenetic method. The estimates obtained in this project evidenced intersubtype differences in the genomic evolutionary rates of HIV-1, with subtypes A1, C and CRF01_AE having higher evolutionary rates than subtypes B, D, G and CRF02_AG. This reveals differences among subtypes and CRFs in their selective constraints, mutation rates, generation times and/or epidemic dynamics (Abecasis et al. 2009; Maljkovic Berry et al. 2007). Interestingly, tMRCA estimates obtained from genomes were congruent with previous estimates (Gray et al. 2009; Abecasis et al. 2009; Hemelaar 2012; Yebra et al. 2016) but the 95% HPDs were usually narrower, which evidences that our estimates are more accurate for estimating evolutionary parameters as a consequence of analyzing more informative sequences. The results obtained provide estimates that can be used as prior distributions in future Bayesian coalescent analyses of specific HIV-1-subtypes and genes.

A significant part of this thesis has been dedicated to investigate the molecular evolution of HCV, with special interest in assessing the clinical consequences of the genotypic diversity of HCV (7 genotypes and 67 subtypes) regarding their potential sensitivity to DAAs. In Chapter 6, I assessed whether the 6 genotypes causing the HCV pandemic, and their respective subtypes, presented differences in the frequency of naturally-occurring amino acid variants involved in resistance to DAAs (RAVs). To achieve this goal, I selected 2,901, 2,216 and 1,344 HCV sequences from the *NS3*, *NS5A* and *NS5B* genes, respectively (encoding the proteins that are targets of all approved DAAs) from different public databases: LANL, VIPRBC and EuHCVdb. With these sequences, I analyzed the prevalence of up to 103 RAVs as well as their phylogenetic history and genetic barrier, for each HCV genotype (and subtype). The results obtained

demonstrated that naturally occurring RAVs are common in all HCV genotypes, and that there are differences between genotypes in their susceptibility to different DAAs, either approved or under clinical trials. In addition to an overall low genetic barrier for the selection of RAVs in all HCV genotypes, we also found, by means of phylogenetic analyses, that some of them (such as NS3 Q80K in HCV-1a and HCV-6, and NS5B C316N in HCV-1b) have a high potential to be transmitted between at risk patients. It is important to point out that most DAAs have been developed for HCV genotype 1 (the most prevalent genotype worldwide), so that DAA susceptibility may be lower in patients infected with non-1 HCV genotypes, as appears to be the case of NS5B NNIs. Consequently, the results obtained could help in designing the optimal combinations of antiviral drugs for each HCV genotype.

In chapter 7, I analyzed the distribution of positively selected sites along the genomes of the most prevalent HCV subtypes (1a and 1b). The detection of positively selected sites was performed with MEME, a powerful method based on a ML approach which allows dN/dS to vary among sites and among lineages from a given phylogeny (Murrell et al. 2012). Then, I assessed, using a logistic regression analysis, how these distributions could be affected by the presence of different factors (RNA and protein secondary structure and antibody, CD4 and CD8 T-cell epitopes) along the genomes of these HCV variants. The results obtained evidenced that, although both subtypes have some epidemiological and clinical differences with HCV-1b displaying higher genetic variability than HCV-1a, they share similar selective constraints. Interestingly, although the presence of CD8 T-cell epitopes was associated with positively selected sites, the presence of RNA secondary structure and CD4 T-cell epitopes was associated with

conservation. The results obtained from this study give information about the effect of some of the interactions between HCV and its host on HCV variability.

Finally, in chapter 8 I assessed the effect of using RNA substitution models on the phylogenetic inference of viroids and RNA viruses. For each species, I obtained a dataset of genomic sequences, retrieved from two public databases (GenBank and VIPRBC). Then, I inferred their RNA secondary structure and performed a model comparison test. Finally, I performed different Bayesian phylogenetic analyses either including or excluding the RNA model. The results obtained reveal the importance of using RNA models for phylogenetic inference of RNA viruses and viroids. In all the species analyzed, the best-fitting model was one which takes into account the correlated evolution between both parties of base-pairs. Generally, the use of RNA models led to significantly longer branch length estimates than when only a DNA model is used. Thus, RNA models would correct underestimates of branch lengths made by DNA models. However, with some exceptions the use of an RNA model had no significant effect on tree topology inference. In light of the results obtained, it would be important for phylogeny software used for inferring tMRCAs and evolutionary rates, such as BEAST (Drummond & Rambaut 2007), R8s (Sanderson 2002) or Physher (Fourment & Holmes 2014), to include the option of using RNA substitution models, as diversification dates and evolutionary rates inferred for RNA viruses and viroids under such models might be different from estimates obtained without considering RNA secondary structure.

It is noteworthy that, for the HIV molecular epidemiology analyses performed in this PhD thesis, sequences were obtained by Sanger sequencing. However, along the

development of this PhD thesis, next-generation sequencing (NGS) has gained influence on viral molecular epidemiology. This technique overcomes some of the limitations of the traditional Sanger sequencing method, such as its relatively high cost and low throughput, and is specially promising for the characterization of intra-host variability.

NGS technologies have a very high throughput, generating enormous number of sequences from an epidemic at a very high speed (Cruz-Rivera et al. 2013). Quick et al. recently sequenced Ebola isolates from the recent African epidemic in less than 60 minutes using the MinIon sequencer, thus allowing real time surveillance of the epidemic (Quick et al. 2016) . NGS has also been applied to metagenomic procedures. This has allowed the detection of novel viruses, such as a new arenavirus affecting humans (Palacios et al. 2008).

NGS allows more realistic analyses of the viral quasispecies that are generated within a given patient. This is especially interesting for the detection of rare resistant variants (those with prevalence < 1%, considering the whole intrahost diversity), something very difficult to explore by Sanger sequencing. Detecting these rare variants is very important, because their abundance in the viral population can increase quickly by means of shifts in the viral quasispecies distribution as result of the selective pressures imposed by antiviral treatments. The relevance of low frequency drug resistance variants in treatment failure has already been demonstrated, by means of NGS, to be clinically relevant for HCV and HIV (Nasu et al. 2011; Kyeyune et al. 2016).

NGS can also be useful for the detection of viral outbreaks: the distribution of genetic distances between the viral populations of related cases is expected to be

significantly lower than that obtained from comparing unrelated cases (Escobar-Gutiérrez et al. 2012).

For these reasons, NGS is a recommendable technology for future molecular epidemiology analyses of viruses.

In summary, part of the results obtained in this thesis have a direct application to public health, evidencing that the evolutionary analysis of RNA viruses can provide information regarding their epidemics, such as the detection and characterization of outbreaks and resistant variants, which is usually more difficult to obtain from traditional epidemiologic approaches only. Thus, when possible, both disciplines should be combined into molecular epidemiology approaches for the surveillance of infectious diseases. In addition, this work has also reported other results which can be useful to better understand the mechanics of evolution of RNA viruses, with possible applications in viral phylogenetics or antiviral research.

### 3.2- Conclusions

- The local expansion of HIV-1 transmission clusters, especially among MSM, plays a significant role in shaping the HIV epidemics in the Basque Country and the Comunitat Valenciana. This fact evidences shortcomings in HIV control measures in Spain, at least for specific, vulnerable population groups.
- The high vulnerability of MSM to HIV infections is also reflected in their lower times between transmissions and their significantly higher association with large transmission clusters than other risk groups.

- The results from the molecular epidemiology analyses of HIV also stress the importance of implementing surveillance strategies that use viral sequence information derived from the genotypic analysis of resistance mutations in HIV-infected patients.
- The analysis of nearly complete HIV-1 genomic coding regions leads to more accurate estimations of the viral evolutionary parameters, and has revealed significantly different genomic evolutionary rates between HIV-1 subtypes and CRFs.
- Naturally occurring RAVs are common in all HCV genotytpes, with significant differences between genotypes regarding their susceptibility towards different DAAs. Interestingly, certain RAVs have been efficiently transmitted among individuals, in absence of antiviral treatment.
- Although the most prevalent HCV subtypes (1a and 1b) present some epidemiological, clinical and biological differences, they are subjected to similar selective constrains. Specifically, the presence of secondary structures and CD4 T-cell epitopes is associated with conservation, while the presence of CD8 T-cell epitopes is associated with selection.
- Including RNA evolutionary models for the phylogenetic inference of RNA viruses and viroids outfits their exclusion. Although the use of these models does not have a significant effect on topology inference, they usually lead to estimating longer branches. This could be important when estimating viral tMRCA and evolutionary rates.

# *4.* General bibliography

AASLD/IDSA HCV Guidance Panel. 2015. Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology 62:932–954.

Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, Camacho R, Shapiro B, Rambaut A, Vandamme A-M. 2007. Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J. Virol. 81:8543–8551.

Abecasis AB, Vandamme A-M, Lemey P. 2009. Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution. J. Virol. 83:12917–12924.

Abecasis AB, Wensing AMJ, Paraskevis D, Vercauteren J, Theys K, Van de Vijver DAMC, Albert J, Asjö B, Balotta C, Beshkov D, et al. 2013. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. Retrovirology 10:7.

Aiewsakun P, Katzourakis A. 2015. Time dependency of foamy virus evolutionary rate estimates. BMC Evol. Biol. 15:119.

Aiewsakun P, Katzourakis A. 2016. Time-Dependent Rate Phenomenon in Viruses. J. Virol. 90:7184–7195.

Akaike H. 1974. A New Look at the Statistical Model Identification. Autom. Control 19:716–723.

Alcantara LCJ, Cassol S, Libin P, Deforche K, Pybus OG, Van Ranst M, Galvão-Castro B, Vandamme A-M, de Oliveira T. 2009. A standardized framework for accurate, high-throughput genotyping of recombinant and non-recombinant viral sequences. Nucleic Acids Res. 37:W634-42.

Alizon S, Fraser C. 2013. Within-host and between-host evolutionary rates across the HIV-1 genome. Retrovirology 10:49.

Allen JE, Whelan S. 2014. Assessing the state of substitution models describing noncoding RNA evolution. Genome Biol. Evol. 6:65–75.

Alves R, Queiroz ATL, Pessoa MG, da Silva EF, Mazo DFC, Carrilho FJ, Carvalho-Filho RJ, de Carvalho IMVG. 2013. The presence of resistance mutations to protease and polymerase inhibitors in Hepatitis C virus sequences from the Los Alamos databank. J. Viral Hepat. 20:414–421.

Ambrosioni J, Junier T, Delhumeau C, Calmy A, Hirschel B, Zdobnov E, Kaiser L, Yerly S. 2012. Impact of highly active antiretroviral therapy on the molecular epidemiology of newly diagnosed HIV infections. AIDS 26:2079–2086.

Antoniadou Z-A, Kousiappa I, Skoura L, Pilalas D, Metallidis S, Nicolaidis P, Malisiovas N, Kostrikis LG. 2014. Short communication: molecular epidemiology of HIV type 1 infection in northern Greece (2009-2010): evidence of a transmission cluster of HIV type 1 subtype A1 drug-resistant strains among men who have sex with men. AIDS Res. Hum. Retroviruses 30:225–232.

Asselah T, Boyer N, Saadoun D, Martinot-Peignoux M, Marcellin P. 2016. Direct-acting antivirals for the treatment of hepatitis C virus infection: optimizing current IFN-

free treatment and future perspectives. Liver Int. 36 Suppl 1:47–57.

Baele G, Lemey P, Bedford T, Rambaut A, Suchard MA, Alekseyenko A V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.

Baggaley RF, White RG, Boily M-C. 2010. HIV transmission risk through anal intercourse: systematic review, meta-analysis and implications for HIV prevention. Int. J. Epidemiol. 39:1048–1063.

Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, Marx PA, Hahn BH, Sharp PM. 2003. Hybrid origin of SIV in chimpanzees. Science 300:1713.

Baltimore D. 1971. Expression of animal virus genomes. Bacteriol. Rev. 35:235–241.

Bartels DJ, Zhou Y, Zhang EZ, Marcial M, Byrn RA, Pfeiffer T, Tigges AM, Adiwijaya BS, Lin C, Kwong AD, et al. 2008. Natural prevalence of hepatitis C virus variants with decreased sensitivity to NS3.4A protease inhibitors in treatment-naive subjects. J. Infect. Dis. 198:800–807.

Bartenschlager R, Lohmann V, Penin F. 2013. The molecular and structural basis of advanced antiviral therapy for hepatitis C virus infection. Nat. Rev. Microbiol. 11:482–496.

Bártolo I, Abecasis AB, Borrego P, Barroso H, McCutchan F, Gomes P, Camacho R, Taveira N. 2011. Origin and Epidemiological History of HIV-1 CRF14_BG. PLoS One 6:e24130.

Beebee T, Rowe G. 2008. An introduction to molecular ecology. 2nd ed. Oxford University Press.

Bello G, Afonso JM, Morgado MG. 2012. Phylodynamics of HIV-1 subtype F1 in Angola, Brazil and Romania. Infect. Genet. Evol. 12:1079–1086.

Bello G, Aulicino PC, Ruchansky D, Guimarães ML, Lopez-Galindez C, Casado C, Chiparelli H, Rocco C, Mangano A, Sen L, et al. 2010. Phylodynamics of HIV-1 circulating recombinant forms 12_BF and 38_BF in Argentina and Uruguay. Retrovirology 7:22.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57:289–300.

Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. 2008. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474.

Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, Brookmeyer R. 2012. Global epidemiology of HIV infection in men who have sex with men. Lancet 380:367–377.

Bezemer D, Cori A, Ratmann O, van Sighem A, Hermanides HS, Dutilh BE, Gras L, Rodrigues Faria N, van den Hengel R, Duits AJ, et al. 2015. Dispersion of the HIV-1 Epidemic in Men Who Have Sex with Men in the Netherlands: A Combined Mathematical Model and Phylogenetic Analysis. PLoS Med. 12:e1001898.

Bezemer D, de Wolf F, Boerlijst MC, van Sighem A, Hollingsworth TD, Prins M, Geskus RB, Gras L, Coutinho RA, Fraser C. 2008. A resurgent HIV-1 epidemic among men who

have sex with men in the era of potent antiretroviral therapy. AIDS 22:1071–1077.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet. 22:437–446.

Bracho MA, Sentandreu V, Alastrué I, Belda J, Juan A, Fernández-García E, Santos C, Zafra T, Tasa T, Colomina S, et al. 2014. Emerging trends in CRF02_AG variants transmission among men who have sex with men in Spain. J. Acquir. Immune Defic. Syndr. 65:e130-3.

Bradshaw D, Matthews G, Danta M. 2013. Sexually transmitted hepatitis C infection: the new epidemic in MSM? Curr. Opin. Infect. Dis. 26:66–72.

Bukh J, Miller R, Purcell R. 1995. Genetic Heterogeneity of Hepatitis C Virus: Quasispecies and Genotypes. Semin. Liver Dis. 15:41–63.

Buonaguro L, Tornesello ML, Buonaguro FM. 2007. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. J. Virol. 81:10209–10219.

Burnham K, Anderson D. 2002. Model selection and multi-model inference: a practical information-theoretic approac. Springer Verlag.

Campo DS, Dimitrova Z, Mitchell RJ, Lara J, Khudyakov Y. 2008. Coordinated evolution of the hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 105:9685–9690.

Cannon NA, Donlin MJ, Fan X, Aurora R, Tavis JE, Virahep-C Study Group. 2008. Hepatitis C virus diversity and evolution in the full open-reading frame during antiviral therapy. PLoS One 3:e2123.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

de Carvalho IMVG, Alves R, de Souza PAV-M, da Silva EF, Mazo D, Carrilho FJ, Queiroz ATL, Pessoa MG. 2014. Protease inhibitor resistance mutations in untreated Brazilian patients infected with HCV: novel insights about targeted genotyping approaches. J. Med. Virol. 86:1714–1721.

Casado G, Thomson MM, Sierra M, Nájera R. 2005. Identification of a novel HIV-1 circulating ADG intersubtype recombinant form (CRF19_cpx) in Cuba. J. Acquir. Immune Defic. Syndr. 40:532–537.

Cento V, Mirabelli C, Salpini R, Dimonte S, Artese A, Costa G, Mercurio F, Svicher V, Parrotta L, Bertoli A, et al. 2012. HCV genotypes are differently prone to the development of resistance to linear and macrocyclic protease inhibitors. PLoS One 7:e39652.

Choisy M, Woelk CH, Guégan J-F, Robertson DL. 2004. Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes. J. Virol. 78:1962–1970.

Christin P-A, Besnard G, Edwards EJ, Salamin N. 2012. Effect of genetic convergence on phylogenetic inference. Mol. Phylogenet. Evol. 62:921–927.

Clarke DK, Duarte E a, Elena SF, Moya a, Domingo E, Holland J. 1994. The red queen

reigns in the kingdom of RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 91:4821–4824.

Cleaveland S, Laurenson MK, Taylor LH. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 356:991–999.

Cleves J. 2008. UNAIDS: the first 10 years, 1996-2006. Jt. United Nations Program. HIV/Aids:285.

Combet C, Garnier N, Charavay C, Grando D, Crisan D, Lopez J, Dehne-Garcia A, Geourjon C, Bettler E, Hulo C, et al. 2007. euHCVdb: the European hepatitis C virus database. Nucleic Acids Res. 35:D363–D366.

Costagliola D, Descamps D, Assoumou L, Morand-Joubert L, Marcelin A-G, Brodard V, Delaugerre C, Mackiewicz V, Ruffault A, Izopet J, et al. 2007. Prevalence of HIV-1 drug resistance in treated patients: a French nationwide study. J. Acquir. Immune Defic. Syndr. 46:12–18.

Courgnaud V, Salemi M, Pourrut X, Mpoudi-Ngole E, Abela B, Auzel P, Bibollet-Ruche F, Hahn B, Vandamme A-M, Delaporte E, et al. 2002. Characterization of a novel simian immunodeficiency virus with a vpu gene from greater spot-nosed monkeys (Cercopithecus nictitans) provides new insights into simian/human immunodeficiency virus phylogeny. J. Virol. 76:8298–8309.

Croissant Y. 2015. Package mlogit. CRAN. Available at: https://cran.r-project.org/web/packages/mlogit/index.html

Cruz-Rivera M, Forbi JC, Yamasaki LHT, Vazquez-Chacon CA, Martinez-Guarneros A, Carpio-Pedroza JC, Escobar-Gutiérrez A, Ruiz-Tovar K, Fonseca-Coronado S, Vaughan G. 2013. Molecular epidemiology of viral diseases in the era of next generation sequencing. J. Clin. Virol. 57:378–380.

Cuevas JM, Gonzalez M, Torres-Puente M, Jiménez-Hernández N, Bracho MA, García-Robles I, González-Candelas F, Moya A. 2009. The role of positive selection in hepatitis C virus. Infect. Genet. Evol. 9:860–866.

Cuevas JM, Torres-Puente M, Jiménez-Hernández N, Bracho MA, García-Robles I, Carnicer F, Olmo JD, Ortega E, Moya A, González-Candelas F. 2008. Refined analysis of genetic variability parameters in hepatitis C virus and the ability to predict antiviral treatment response. J. Viral Hepat. 15:578–590.

Cuevas MT, Muñoz-Nieto M, Thomson MM, Delgado E, Iribarren JA, Cilla G, Fernández-García A, Santamaría JM, Lezaun MJ, Jiménez L, et al. 2009. HIV-1 transmission cluster with T215D revertant mutation among newly diagnosed patients from the Basque Country, Spain. J. Acquir. Immune Defic. Syndr. 51:99–103.

Cuypers L, Li G, Neumann-Haefelin C, Piampongsant S, Libin P, Van Laethem K, Vandamme A-M, Theys K. 2016. Mapping the genomic diversity of HCV subtypes 1a and 1b: Implications of structural and immunological constraints for vaccine and drug development. Virus Evol. 2:vew024.

Damgaard CK, Andersen ES, Knudsen B, Gorodkin J, Kjems J. 2004. RNA interactions in the 5’ region of the HIV-1 genome. J. Mol. Biol. 336:369–379.

Delatorre E, Bello G. 2013. Phylodynamics of the HIV-1 epidemic in Cuba. PLoS One

8:e72448.

Delgado E, Cuevas MT, Domínguez F, Vega Y, Cabello M, Fernández-García A, Pérez-Losada M, Castro MÁ, Montero V, Sánchez M, et al. 2015. Phylogeny and Phylogeography of a Recent HIV-1 Subtype F Outbreak among Men Who Have Sex with Men in Spain Deriving from a Cluster with a Wide Geographic Circulation in Western Europe. PLoS One 10:e0143325.

DGSP. 2014. Vigilancia epidemiológica del VIH/SIDA enEspaña: Sistema de información sobre nuevos diagnósticos del VIH y registro nacional de casos de SIDA.

Diez M, Bleda MJ, Varela JR, Ordonana J, Azpiri MA, Vall M, Santos C, Viloria L, de Armas C, Urena JM, et al. 2014. Trends in HIV testing, prevalence among first-time testers, and incidence in most-at-risk populations in Spain: the EPI-VIH Study, 2000 to 2009. Euro Surveill. 19:20971.

Domingo E, Sabo D, Taniguchi T, Weissmann C. 1978. Nucleotide sequence heterogeneity of an RNA phage population. Cell 13:735–744.

Donnelly P, Tavaré S. 1995. Coalescents and genealogical structure under neutrality. Annu. Rev. Genet. 29:401–421.

Drake JW, Holland JJ. 1999. Mutation rates among RNA viruses. Proc. Natl. Acad. Sci. U. S. A. 96:13910–13913.

Drozdetskiy A, Cole C, Procter J, Barton GJ. 2015. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 43:W389-w94.

Drummond A, Pybus OG, Rambaut A. 2003. Inference of viral evolutionary rates from molecular sequences. Adv. Parasitol. 54:331–358.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8:114.

Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol. Biol. Evol. 29:1969–1973.

Duchêne S, Holmes EC, Ho SYW. 2014. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc. Biol. Sci. 281.

Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. 9:267–276.

ECDC. 2013. Thematic report : Men who have sex with men. Monitoring implementation of the Dublin Declaration on Partnership to fight HIV/AIDS in Europe and Central Asia: 2012 progress report.

ECDC/WHO. 2010. HIV/AIDS Surveillance in Europe 2009.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Edwards R a., Rohwer F. 2005. Opinion: Viral metagenomics. Nat. Rev. Microbiol. 3:504–510.

Elena SF, Sanjuán R. 2005. Adaptive value of high mutation rates of RNA viruses: separating causes from consequences. J. Virol. 79:11555–11558.

Escobar-Gutiérrez A, Vazquez-Pichardo M, Cruz-Rivera M, Rivera-Osorio P, Carpio-Pedroza JC, Ruíz-Pacheco JA, Ruiz-Tovar K, Vaughan G. 2012. Identification of hepatitis C virus transmission using a next-generation sequencing approach. J. Clin. Microbiol. 50:1461–1463.

European Association for Study of Liver. 2015. EASL Recommendations on Treatment of Hepatitis C 2015. J. Hepatol. 63:199–236.

EXCOFFIER L, LISCHER HEL. 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10:564–567.

Fares MA, Moya a, Escarmís C, Baranowski E, Domingo E, Barrio E. 2001. Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol. Biol. Evol. 18:10–21.

Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, Tatem AJ, Sousa JD, Arinaminpathy N, Pépin J, et al. 2014. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. Science 346:56–61.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fourment M, Holmes E. 2014. Novel non-parametric models to estimate evolutionary rates and divergence times from heterochronous sequence data. BMC Evol. Biol. 14:163.

Frange P, Meyer L, Deveau C, Tran L, Goujard C, Ghosn J, Girard P-M, Morlat P, Rouzioux C, Chaix M-L. 2012. Recent HIV-1 Infection Contributes to the Viral Diffusion over the French Territory with a Recent Increasing Frequency. PLoS One 7:e31695.

Frank C, Mohamed MK, Strickland GT, Lavanchy D, Arthur RR, Magder LS, El Khoby T, Abdel-Wahab Y, Aly Ohn ES, Anwar W, et al. 2000. The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. Lancet 355:887–891.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Geller R, Domingo-Calap P, Cuevas JM, Rossolillo P, Negroni M, Sanjuán R. 2015. The external domains of the HIV-1 envelope are a mutational cold spot. Nat. Commun. 6:8571.

Geller R, Estada Ú, Peris JB, Andreu I, Bou J-V, Garijo R, Cuevas JM, Sabariegos R, Mas A, Sanjuán R. 2016. Highly heterogeneous mutation rates in the hepatitis C virus genome. Nat. Microbiol. 1:16045.

Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW. 2008. A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc. Natl. Acad. Sci. U. S. A. 105:20362–20367.

Gilbert MTP, Rambaut A, Wlasiuk G, Spira TJ, Pitchenik AE, Worobey M. 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. U. S. A. 104:18566–18570.

González-Alba JM, Holguín A, Garcia R, García-Bujalance S, Alonso R, Suárez A, Delgado R, Cardeñoso L, González R, García-Bermejo I, et al. 2011. Molecular surveillance of HIV-1 in Madrid, Spain: a phylogeographic analysis. J. Virol. 85:10755–10763.

Grabowski MK, Redd AD. 2014. Molecular tools for studying HIV transmission in sexual networks. Curr. Opin. HIV AIDS 9:126–133.

Gray RR, Tatem AJ, Lamers S, Hou W, Laeyendecker O, Serwadda D, Sewankambo N, Gray RH, Wawer M, Quinn TC, et al. 2009. Spatial phylodynamics of HIV-1 epidemic emergence in east Africa. AIDS 23:F9–F17.

Grenfell BT. 2004. Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science 303:327–332.

Gruber AR, Neuböck R, Hofacker IL, Washietl S. 2007. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Res. 35:W335-8.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Hahn BH, Shaw GM, De Cock KM, Sharp PM. 2000. AIDS as a zoonosis: scientific and public health implications. Science 287:607–614.

Hajarizadeh B, Grebely J, Dore GJ. 2013. Epidemiology and natural history of HCV infection. Nat. Rev. Gastroenterol. Hepatol. 10:553–562.

Heeney JL, Dalgleish AG, Weiss RA. 2006. Origins of HIV and the evolution of resistance to AIDS. Science 313:462–466.

Heled J, Drummond AJ. 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.

Hemelaar J. 2012. The origin and diversity of the HIV-1 pandemic. Trends Mol. Med. 18:182–192.

Ho SYW, Duchêne S, Molak M, Shapiro B. 2015. Time-dependent estimates of molecular evolutionary rates: evidence and causes. Mol. Ecol. 24:6007–6012.

Hofacker IL, Fekete M, Stadler PF. 2002. Secondary structure prediction for aligned RNA sequences. J. Mol. Biol. 319:1059–1066.

Holguín A, Álvarez A, Soriano V. 2005. Heterogeneous nature of HIV-1 recombinants spreading in Spain. J. Med. Virol. 75:374–380.

Holmes EC. 2003. Error thresholds and the constraints to RNA virus evolution. Trends Microbiol. 11:543–546.

Holmes EC. 2004. The phylogeography of human viruses. Mol. Ecol. 13:745–756.

Holmes EC. 2008. Evolutionary history and phylogeography of human viruses. Annu. Rev. Microbiol. 62:307–328.

Holmes EC. 2011. What does virus evolution tell us about virus origins? J. Virol. 85:5247–5251.

Holmes EC, Drummond AJ. 2007. The evolutionary genetics of viral emergence. Curr. Top. Microbiol. Immunol. 315:51–66.

Holmes EC, Zhang LQ, Robertson P, Cleland A, Harvey E, Simmonds P, Leigh Brown AJ. 1995. The molecular epidemiology of human immunodeficiency virus type 1 in Edinburgh. J. Infect. Dis. 171:45–53.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682.

Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153–159.

Hué S, Clewley JP, Cane PA, Pillay D. 2004. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. AIDS 18:719–728.

Hué S, Clewley JP, Cane PA, Pillay D. 2005. Investigation of HIV-1 transmission events by phylogenetic methods: requirement for scientific rigour. AIDS 19:449–450.

Hué S, Pillay D, Clewley JP, Pybus OG. 2005. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. Proc. Natl. Acad. Sci. U. S. A. 102:4425–4429.

Huelsenbeck JP, Hillis DM. 1993. Succes of Phylogenetic Methods in the Four-Taxon Case. Syst. Biol. 42:247–264.

Hughes AL, Friedman R. 2005. Variation in the pattern of synonymous and nonsynonymous difference between two fungal genomes. Mol. Biol. Evol. 22:1320–1324.

Hughes AL, Hughes MAK. 2007. More effective purifying selection on RNA viruses than in DNA viruses. Gene 404:117–125.

Hughes GJ, Fearnhill E, Dunn D, Lycett SJ, Rambaut A, Leigh Brown AJ, UK HIV Drug Resistance Collaboration. 2009. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathog. 5:e1000590.

Humphreys I, Fleming V, Fabris P, Parker J, Schulenberg B, Brown A, Demetriou C, Gaudieri S, Pfafferott K, Lucas M, et al. 2009. Full-length characterization of hepatitis C virus subtype 3a reveals novel hypervariable regions under positive selection during acute infection. J. Virol. 83:11456–11466.

Hutchins CJ, Rathjen PD, Forster AC, Symons RH. 1986. Self-cleavage of plus and minus RNA transcripts of avocado sunblotch viroid. Nucleic Acids Res. 14:3627–3640.

Jenkins GM, Rambaut A, Pybus OG, Holmes EC. 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. 54:156–165.

Johnson VA, Calvez V, Gunthard HF, Paredes R, Pillay D, Shafer RW, Wensing AM, Richman DD. Update of the drug resistance mutations in HIV-1: March 2013. Top. Antivir. Med. 21:6–14.

Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. 2005. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. Science 309:1577–1581.

Kaleebu P, French N, Mahe C, Yirrell D, Watera C, Lyagoba F, Nakiyingi J, Rutebemberwa A, Morgan D, Weber J, et al. 2002. Effect of human immunodeficiency virus (HIV) type 1 envelope subtypes A and D on disease progression in a large cohort of HIV-1-positive persons in Uganda. J. Infect. Dis. 185:1244–1250.

Kapoor A, Simmonds P, Gerold G, Qaisar N, Jain K, Henriquez JA, Firth C, Hirschberg DL, Rice CM, Shields S, et al. 2011. Characterization of a canine homolog of hepatitis C virus. Proc. Natl. Acad. Sci. U. S. A. 108:11608–11613.

Kapoor A, Simmonds P, Scheel TKH, Hjelle B, Cullen JM, Burbelo PD, Chauhan L V, Duraisamy R, Sanchez Leon M, Jain K, et al. 2013. Identification of rodent homologs of hepatitis C virus and pegiviruses. MBio 4:e00216-13.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kilbourne ED. 1973. The molecular epidemiology of influenza. J. Infect. Dis. 127:478–487.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press

Kingman JFC. 1982. The coalescent. Stoch. Process. their Appl. 13:235–248.

Koonin E V, Dolja V V. 2012. Expanding networks of RNA virus evolution. BMC Biol. 10:54.

Korber B, Myers G. 1992. Signature pattern analysis: a method for assessing viral sequence relatedness. AIDS Res. Hum. Retroviruses 8:1549–1560.

Kosakovsky Pond SL, Frost SDW. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208–1222.

Kosakovsky Pond SL, Mannino F V, Gravenor MB, Muse S V, Frost SDW. 2007. Evolutionary model selection with a genetic algorithm: a case study using stem RNA. Mol. Biol. Evol. 24:159–170.

Kosakovsky Pond SL, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. In: Nielsen R (ed). Statistical methods in molecular evolution. Springer, New York.

Kourí V, Alemán Y, Pérez L, Pérez J, Fonseca C, Correa C, Aragonés C, Campos J, Álvarez D, Schrooten Y, et al. 2012. High frequency of antiviral drug resistance and non-B subtypes in HIV-1 patients failing antiviral therapy in Cuba. J. Clin. Virol. 55:348–355.

Kouri V, Khouri R, Alemán Y, Abrahantes Y, Vercauteren J, Pineda-Peña A-C, Theys K, Megens S, Moutschen M, Pfeifer N, et al. 2015. CRF19_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba. EBioMedicine 2:244–254.

Kouyos RD, von Wyl V, Yerly S, Böni J, Taffé P, Shah C, Bürgisser P, Klimkait T, Weber R, Hirschel B, et al. 2010. Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. J. Infect. Dis. 201:1488–1497.

Koziel MJ. 2005. Cellular immune responses against hepatitis C virus. Clin. Infect. Dis. 41 Suppl 1:S25-31.

Kuhner MK, Yamato J, Felsenstein J. 1995. Estimating effective population size and mutation rate from sequence data using metropolis-hastings sampling. Genetics 140:1421–1430.

Kuiken C, Yusim K, Boykin L, Richardson R. 2005. The Los Alamos hepatitis C sequence database. Bioinformatics 21:379–384.

Kuiken CL, Foley B, Hahn B, Marx P, McCutchan F, Mellors JW, Mullins S, Wolinski S, Korber B. 1999. The human retroviruses and AIDS 1999 compendium.

Kuiken CL, Foley B, Leitner T, Apetrei C, Mizrachi Y, Mullins JI, Rambaut A, Wolinsky SM, Korber B. 2012. HIV Sequence Compendium 2012. New Mexico: Theoretical Biology and Niophysics Group. Los Alamos national Laboratory

Kyeyune F, Gibson RM, Nankya I, Venner C, Metha S, Akao J, Ndashimye E, Kityo CM, Salata RA, Mugyenyi P, et al. 2016. Low-Frequency Drug Resistance in HIV-Infected Ugandans on Antiretroviral Treatment Is Associated with Regimen Failure. Antimicrob. Agents Chemother. 60:3380–3397.

Lai D, Proctor JR, Zhu JYA, Meyer IM. 2012. R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Res. 40:e95.

Lamonaca V, Missale G, Urbani S, Pilli M, Boni C, Mori C, Sette A, Massari M, Southwood S, Bertoni R, et al. 1999. Conserved hepatitis C virus sequences are highly immunogenic for CD4(+) T cells: implications for vaccine development. Hepatology 30:1088–1098.

Legrand-Abravanel F, Henquell C, Le Guillou-Guillemette H, Balan V, Mirand A, Dubois M, Lunel-Fabiani F, Payan C, Izopet J. 2009. Naturally occurring substitutions conferring resistance to hepatitis C virus polymerase inhibitors in treatment-naive patients infected with genotypes 1-5. Antivir. Ther. 14:723–730.

Leigh Brown AJ, Lycett SJ, Weinert L, Hughes GJ, Fearnhill E, Dunn DT, UK HIV Drug Resistance Collaboration. 2011. Transmission network parameters estimated from HIV sequences for a nationwide epidemic. J. Infect. Dis. 204:1463–1469.

Lepage T, Lawi S, Tupper P, Bryant D. 2006. Continuous and tractable models for the variation of evolutionary rates. Math. Biosci. 199:216–233.

Lewis F, Hughes GJ, Rambaut A, Pozniak A, Leigh Brown AJ. 2008. Episodic sexual transmission of HIV revealed by molecular phylodynamics. PLoS Med. 5:e50.

Li G, Piampongsant S, Faria NR, Voet A, Pineda-Peña A-C, Khouri R, Lemey P, Vandamme A-M, Theys K. 2015. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology 12:18.

Li K, Lemon SM. 2013. Innate immune responses in hepatitis C virus infection. Semin. Immunopathol. 35:53–72.

Lin Y-Y, Liu C, Chien W-H, Wu L-L, Tao Y, Wu D, Lu X, Hsieh C-H, Chen P-J, Wang H-Y, et al. 2015. New insights into the evolutionary rate of hepatitis B virus at different biological scales. J. Virol. 89:3512–3522.

Lindenbach BD, Rice CM. 2005. Unravelling hepatitis C virus replication from genome to function. Nature 436:933–938.

Linhart H, Zucchini W. 1986. Model selection. John Wiley & Sons, New York

Liu TF, Shafer RW. 2006. Web Resources for HIV Type 1 Genotypic-Resistance Test Interpretation. Clin. Infect. Dis. 42:1608–1618.

López-Labrador FX, Moya A, Gonzàlez-Candelas F. 2008. Mapping natural polymorphisms of hepatitis C virus NS3/4A protease and antiviral resistance to inhibitors in worldwide isolates. Antivir. Ther. 13:481–494.

Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6:26.

Lyons S, Kapoor A, Sharp C, Schneider BS, Wolfe ND, Culshaw G, Corcoran B, McGorum BC, Simmonds P. 2012. Nonprimate hepaciviruses in domestic horses, United kingdom. Emerg. Infect. Dis. 18:1976–1982.

Maddison WP, Maddison DR. 1989. Interactive analysis of phylogeny and character evolution using the computer program MacClade. Folia Primatol. (Basel). 53:190–202.

Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SYW, Shapiro B, Pybus OG, Allain J-P, Hatzakis A. 2009. The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. PLoS Med. 6:e1000198.

Di Maio VC, Cento V, Mirabelli C, Artese A, Costa G, Alcaro S, Perno CF, Ceccherini-Silberstein F. 2014. Hepatitis C virus genetic variability and the presence of NS5B resistance-associated mutations as natural polymorphisms in selected genotypes could affect the response to NS5B inhibitors. Antimicrob. Agents Chemother. 58:2781–2797.

Maljkovic Berry I, Ribeiro R, Kothari M, Athreya G, Daniels M, Lee HY, Bruno W, Leitner T. 2007. Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases. J. Virol. 81:10625–10635.

Mangia A, Santoro R, Minerva N, Ricci GL, Carretta V, Persico M, Vinelli F, Scotto G, Bacca D, Annese M, et al. 2005. Peginterferon alfa-2b and ribavirin for 12 vs. 24 weeks in HCV genotype 2 or 3. N. Engl. J. Med. 352:2609–2617.

Manrubia SC, Escarmís C, Domingo E, Lázaro E. 2005. High mutation rates, bottlenecks, and robustness of RNA viral quasispecies. Gene 347:273–282.

Margeridon-Thermet S, Le Pogam S, Li L, Liu TF, Shulman N, Shafer RW, Najera I. 2014. Similar prevalence of low-abundance drug-resistant variants in treatment-naive patients with genotype 1a and 1b hepatitis C virus infections as determined by ultradeep pyrosequencing.Kaushik-Basu N, editor. PLoS One 9:e105569.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. Bioinformatics 16:562–563.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462–2463.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol. 1:vev003.

Martin DP, Posada D, Crandall KA, Williamson C. 2005. A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res. Hum. Retroviruses 21:98–102.

Mauger DM, Golden M, Yamane D, Williford S, Lemon SM, Martin DP, Weeks KM. 2015. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc. Natl. Acad. Sci. U. S. A. 112:3692–3697.

McCloskey RM, Liang RH, Joy JB, Krajden M, Montaner JSG, Harrigan PR, Poon AFY. 2015. Global origin and transmission of hepatitis C virus nonstructural protein 3 Q80K polymorphism. J. Infect. Dis. 211:1288–1295.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654.

McPhee F, Hernandez D, Yu F, Ueland J, Monikowski A, Carifa A, Falk P, Wang C, Fridell R, Eley T, et al. 2013. Resistance analysis of hepatitis C virus genotype 1 prior treatment null responders receiving daclatasvir and asunaprevir. Hepatology 58:902–911.

Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, Barnes E. 2015. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology 61:77–87.

Meyer AG, Spielman SJ, Bedford T, Wilke CO. 2015. Time dependence of evolutionary metrics during the 2009 pandemic influenza virus outbreak. Virus Evol. 1.

Minin VN, Bloomquist EW, Suchard MA. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Mol. Biol. Evol. 25:1459–1471.

Mohd Hanafiah K, Groeger J, Flaxman AD, Wiersma ST. 2013. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. Hepatology 57:1333–1342.

Moya A, Elena SF, Bracho A, Miralles R, Barrio E. 2000. The evolution of RNA viruses: a population genetics view. Proc. Natl. Acad. Sci. USA 97:6967–6973.

Moya A, Holmes EC, González-Candelas F. 2004. The population genetics and evolutionary epidemiology of RNA viruses. Nat. Rev. Microbiol. 2:279–288.

Murphy DG, Willems B, Vincelette J, Bernier L, Côté J, Delage G. 1996. Biological and clinicopathological features associated with hepatitis C virus type 5 infections. J. Hepatol. 24:109–113.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet. 8:e1002764.

Muse S V. 1995. Evolutionary analyses of DNA sequences subject to constraints of secondary structure. Genetics 139:1429–1439.

Nasrallah CA, Mathews DH, Huelsenbeck JP. 2011. Quantifying the impact of dependent evolution among sites in phylogenetic inference. Syst. Biol. 60:60–73.

Nasu A, Marusawa H, Ueda Y, Nishijima N, Takahashi K, Osaki Y, Yamashita Y, Inokuma T, Tamada T, Fujiwara T, et al. 2011. Genetic heterogeneity of hepatitis C virus in association with antiviral therapy determined by ultra-deep sequencing. PLoS One 6:e24907.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. 3:418–426.

Nei M. 2005. Selectionism and neutralism in molecular evolution. Mol. Biol. Evol. 22:2318–2342.

Nickle DC, Rolland M, Jensen MA, Pond SLK, Deng W, Seligman M, Heckerman D, Mullins JI, Jojic N. 2007. Coping with viral diversity in HIV vaccine design. PLoS Comput. Biol. 3:e75.

Niculescu I, Paraschiv S, Paraskevis D, Abagiu A, Batan I, Banica L, Otelea D. 2015. Recent HIV-1 Outbreak Among Intravenous Drug Users in Romania: Evidence for Cocirculation of CRF14_BG and Subtype F1 Strains. AIDS Res. Hum. Retroviruses 31:488–495.

Okamoto H, Kojima M, Okada S, Yoshizawa H, Iizuka H, Tanaka T, Muchmore EE, Peterson DA, Ito Y, Mishiro S. 1992. Genetic drift of hepatitis C virus during an 8.2-year infection in a chimpanzee: variability and stability. Virology 190:894–899.

de Oliveira T, Deforche K, Cassol S, Salminen M, Paraskevis D, Seebregts C, Snoeck J, van Rensburg EJ, Wensing AMJ, van de Vijver DA, et al. 2005. An automated genotyping system for analysis of HIV-1 and other microbial sequences. Bioinformatics 21:3797–3800.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. Virology 265:218–225.

Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan P-L, Hui J, Marshall J, et al. 2008. A new arenavirus in a cluster of fatal transplant-associated diseases. N. Engl. J. Med. 358:991–998.

Pang PS, Planet PJ, Glenn JS. 2009. The evolution of the major hepatitis C genotypes correlates with clinical response to interferon therapy. PLoS One 4:e6579.

Paolucci S, Fiorina L, Mariani B, Gulminetti R, Novati S, Barbarini G, Bruno R, Baldanti F. 2013. Naturally occurring resistance mutations to inhibitors of HCV NS5A region and NS5B polymerase in DAA treatment-naïve patients. Virol. J. 10:355.

Paolucci S, Fiorina L, Piralla A, Gulminetti R, Novati S, Barbarini G, Sacchi P, Gatti M, Dossena L, Baldanti F. 2012. Naturally occurring mutations to HCV protease inhibitors in treatment-naïve patients. Virol. J. 9:245.

Paraskevis D, Lemey P, Salemi M, Suchard M, Van De Peer Y, Vandamme A-M. 2003. Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using

bayesian inference: implications for the origin of HIV-1. Mol. Biol. Evol. 20:1986–1996.

Paraskevis D, Wensing AMJ, Vercauteren J, Vijver DA, Albert J, Asjo B, . on behalf of the SP. 2006. Prevalence of HIV-1 subtypes among newly HIV-1 diagnosed individuals during 2002–2003 in Europe: Evidence for a continuous introduction of non-B subtypes. In: 1st International Workshop on HIV Transmission; Toronto, Canada 200. p. p.31. Abstract N$^{o}$ 34.

Patiño Galindo JA, Torres-Puente M, Gimeno C, Ortega E, Navarro D, Galindo MJ, Navarro L, Navarro V, Juan A, Belda J, Bracho MA, González-Candelas F, et al. 2015. Expansion of the CRF19_cpx Variant in Spain. J. Clin. Virol. 69:146–149.

Patiño-Galindo JÁ, Salvatierra K, González-Candelas F, López-Labrador FX. 2016. Comprehensive Screening for Naturally Occurring Hepatitis C Virus Resistance to Direct-Acting Antivirals in the NS3, NS5A, and NS5B Genes in Worldwide Isolates of Viral Genotypes 1 to 6. Antimicrob. Agents Chemother. 60:2402–2416.

Patiño-Galindo JA, Thomson MM, Pérez-Álvarez L, Delgado E, Cuevas MT, Fernández-García A, Nájera R, Iribarren JA, Cilla G, López-Soria L, et al. 2016. Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain. Infect. Genet. Evol. 40:91–97.

Pelletier J, Sonenberg N. 1988. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature 334:320–325.

Pellicelli AM, Romano M, Stroffolini T, Mazzoni E, Mecenate F, Monarca R, Picardi A, Bonaventura ME, Mastropietro C, Vignally P, et al. 2012. HCV genotype 1a shows a better virological response to antiviral therapy than HCV genotype 1b. BMC Gastroenterol. 12:162.

Pennings PS, Kryazhimskiy S, Wakeley J. 2014. Loss and recovery of genetic diversity in adapting populations of HIV. PLoS Genet. 10:e1004000.

Pérez L, Kourí V, Alemán Y, Abrahantes Y, Correa C, Aragonés C, Martínez O, Pérez J, Fonseca C, Campos J, et al. 2013. Antiretroviral drug resistance in HIV-1 therapy-naive patients in Cuba. Infect. Genet. Evol. 16:144–150.

Pérez L, Thomson MM, Bleda MJ, Aragonés C, González Z, Pérez J, Sierra M, Casado G, Delgado E, Nájera R. 2006. HIV Type 1 molecular epidemiology in cuba: high genetic diversity, frequent mosaicism, and recent expansion of BG intersubtype recombinant forms. AIDS Res. Hum. Retroviruses 22:724–733.

Pérez-Álvarez L, Delgado E, Vega Y, Montero V, Cuevas T, Fernández-García A, García-Riart B, Pérez-Castro S, Rodríguez-Real R, López-Álvarez MJ, et al. 2014. Predominance of CXCR4 tropism in HIV-1 CRF14_BG strains from newly diagnosed infections. J. Antimicrob. Chemother. 69:246–253.

Peters PJ, Pontones P, Hoover KW, Patel MR, Galang RR, Shields J, Blosser SJ, Spiller MW, Combs B, Switzer WM, et al. 2016. HIV Infection Linked to Injection Use of Oxymorphone in Indiana, 2014–2015. N. Engl. J. Med. 375:229–239.

Phillips AN, Cambiano V, Nakagawa F, Brown AE, Lampe F, Rodger A, Miners A, Elford J, Hart G, Johnson AM, et al. 2013. Increased HIV Incidence in Men Who Have Sex

with Men Despite High Levels of ART-Induced Viral Suppression: Analysis of an Extensively Documented Epidemic. PLoS One 8:e55312.

Pickett BE, Greer DS, Zhang Y, Stewart L, Zhou L, Sun G, Gu Z, Kumar S, Zaremba S, Larsen CN, et al. 2012. Virus pathogen database and analysis resource (ViPR): a comprehensive bioinformatics database and analysis resource for the coronavirus research community. Viruses 4:3209–3226.

Pineda-Peña A-C, Faria NR, Imbrechts S, Libin P, Abecasis AB, Deforche K, Gómez-López A, Camacho RJ, de Oliveira T, Vandamme A-M. 2013. Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect. Genet. Evol. 19:337–348.

Plantier J-C, Leoz M, Dickerson JE, De Oliveira F, Cordonnier F, Lemée V, Damond F, Robertson DL, Simon F. 2009. A new human immunodeficiency virus derived from gorillas. Nat. Med. 15:871–872.

Pond SLK, Frost SDW. 2005. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol. Biol. Evol. 22:478–485.

Pond SLK, Frost SDW, Muse S V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Posada D. 2002. Evaluation of methods for detecting recombination from DNA sequences: empirical data. Mol. Biol. Evol. 19:708–717.

Posada D. 2008. jModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25:1253–1256.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. Syst. Biol. 50:580–601.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. 98:13757–13762.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

Pybus OG, Barnes E, Taggart R, Lemey P, Markov P V, Rasachak B, Syhavong B, Phetsouvanah R, Sheridan I, Humphreys IS, et al. 2009. Genetic history of hepatitis C virus in East Asia. J. Virol. 83:1071–1082.

Pybus OG, Cochrane A, Holmes EC, Simmonds P. 2005. The hepatitis C virus epidemic among injecting drug users. Infect. Genet. Evol. 5:131–139.

Pybus OG, Rambaut A, Harvey PH. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. Genetics 155:1429–1437.

Pybus OG, Thézé J. 2016. Hepacivirus cross-species transmission and the origins of the hepatitis C virus. Curr. Opin. Virol. 16:1–7.

Quan P-L, Firth C, Conte JM, Williams SH, Zambrana-Torrelio CM, Anthony SJ, Ellison JA, Gilbert AT, Kuzmin I V, Niezgoda M, et al. 2013. Bats are a major natural reservoir for hepaciviruses and pegiviruses. Proc. Natl. Acad. Sci. U. S. A. 110:8194–8199.

Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R,

Dudas G, Mikhail A, et al. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232.

R Core Team. 2011. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical computing.

R Core Team. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical computing.

Rambaut A, Lam TT, Max Carvalho L, Pybus OG. 2016. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2:vew007.

Raney JL, Delongchamp RR, Valentine CR. 2004. Spontaneous mutant frequency and mutation spectrum for gene A of PHIX174 grown in E. coli. Environ. Mol. Mutagen. 44:119–127.

Rannala B, Yang ZH. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Ripley B, Hornik K, Gebhardt A, Firth D. 2016. Package "MASS." CRAN. Available at https://cran.r-project.org/web/packages/MASS/index.html

Roberts JD, Bebenek K, Kunkel TA. 1988. The accuracy of reverse transcriptase from HIV-1. Science 242:1171–1173.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Rosenberg NA, Nordborg M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3:380–390.

Rousseeuw PJ, Ruts I, Tukey JW. 1999. The Bagplot: A Bivariate Boxplot. Am. Stat. 53:382–387.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko J V., Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832–837.

Salvatierra K. 2014. Resistencia a nuevos antivirales de acción directa en aislados clínicos del virus de la hepatitis C. Ph.D. thesis. University of València, Spain.

Sanderson MJ. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Sanjuán R. 2012. From molecular genetics to phylodynamics: evolutionary relevance of mutation rates across viruses. PLoS Pathog. 8:e1002685.

Sanjuán R, Bordería A V. 2011. Interplay between RNA structure and protein evolution in HIV-1. Mol. Biol. Evol. 28:1333–1338.

Sanjuán R, Forment J, Elena SF. 2006. In silico predicted robustness of viroids RNA secondary structures. I. The effect of single mutations. Mol. Biol. Evol. 23:1427–1436.

Sanjuán R, Nebot MR, Peris JB, Alcamí J. 2013. Immune activation promotes evolutionary conservation of T-cell epitopes in HIV-1. PLoS Biol. 11:e1001523.

Sarobe P, Huarte E, Lasarte JJ, López-Díaz de Cerio A, García N, Borrás-Cuesta F, Prieto J. 2001. Characterization of an immunologically conserved epitope from hepatitis C virus E2 glycoprotein recognized by HLA-A2 restricted cytotoxic T lymphocytes. J. Hepatol. 34:321–329.

Sarrazin C, Zeuzem S. 2010. Resistance to direct antiviral agents in patients with hepatitis C virus infection. Gastroenterology 138:447–462.

Savill NJ, Hoyle DC, Higgs PG. 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. Genetics 157:399–411.

Scheel TKH, Gottwein JM, Mikkelsen LS, Jensen TB, Bukh J. 2011. Recombinant HCV variants with NS5A from genotypes 1-7 have different sensitivities to an NS5A inhibitor but not interferon-α. Gastroenterology 140:1032–1042.

Schliep K. 2016. Package "phangorn". Phylogenetic Analysis in R. CRAN. Available at https://cran.r-project.org/web/packages/phangorn/index.html

Schöniger M, von Haeseler A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. Mol. Phylogenet. Evol. 3:240–247.

Schultes EA, Bartel DP. 2000. One sequence, two ribozymes: implications for the emergence of new ribozyme folds. Science 289:448–452.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7–9.

Shepard CW, Finelli L, Alter MJ. 2005. Global epidemiology of hepatitis C virus infection. Lancet. Infect. Dis. 5:558–567.

Sheridan I, Pybus OG, Holmes EC, Klenerman P. 2004. High-resolution phylogenetic analysis of hepatitis C virus adaptation and its relationship to disease progression. J. Virol. 78:3447–3454.

Simmonds P. 2004. Genetic diversity and evolution of hepatitis C virus--15 years on. J. Gen. Virol. 85:3173–3188.

Simmonds P. 2013. The origin of hepatitis C virus. Curr. Top. Microbiol. Immunol. 369:1–15.

Simmonds P, Holmes EC, Cha TA, Chan SW, McOmish F, Irvine B, Beall E, Yap PL, Kolberg J, Urdea MS. 1993. Classification of hepatitis C virus into six major genotypes and a series of subtypes by phylogenetic analysis of the NS-5 region. J. Gen. Virol. [Internet]:2391–2399.

Simon-Loriere E, Holmes EC, Pagán I. 2013. The effect of gene overlapping on the rate of RNA virus evolution. Mol. Biol. Evol. 30:1916–1928.

Smith DB, Bukh J, Kuiken C, Muerhoff AS, Rice CM, Stapleton JT, Simmonds P. 2014. Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. Hepatology 59:318–327.

Smith DB, Pathirana S, Davidson F, Lawlor E, Power J, Yap PL, Simmonds P. 1997. The origin of hepatitis C virus genotypes. J. Gen. Virol. 78 ( Pt 2):321–328.

Smith JM. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. 34:126–129.

Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. 2011. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology 8:87.

Svarovskaia E, Dvory-Sobol H, Gontcharova V, Chiu S, Hebner CM, Hyland R, Kowdley K, Lawitz E, Gane E, Symonds WT, et al. 2012. Comprehensive resistance testing in patients who relapsed after treatment with sofosbuvir (GS-7977)-containing regimens in phase 2 studie. Hepatology 56 suppl 1:551A.

Svicher V, Cento V, Salpini R, Mercurio F, Fraune M, Beggel B, Han Y, Gori C, Wittkop L, Bertoli A, et al. 2011. Role of hepatitis B virus genetic barrier in drug-resistance and immune-escape development. Dig. Liver Dis. 43:975–983.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Takamizawa A, Mori C, Fuke I, Manabe S, Murakami S, Fujita J, Onishi E, Andoh T, Yoshida I, Okayama H. 1991. Structure and organization of the hepatitis C virus genome isolated from human carriers. J. Virol. 65:1105–1113.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Tellam J, Smith C, Rist M, Webb N, Cooper L, Vuocolo T, Connolly G, Tscharke DC, Devoy MP, Khanna R. 2008. Regulation of protein translation through mRNA structure influences MHC class I loading and T cell recognition. Proc. Natl. Acad. Sci. U. S. A. 105:9319–9324.

Temin HM. 1993. Retrovirus variation and reverse transcription: abnormal strand transfers result in retrovirus genetic variation. Proc. Natl. Acad. Sci. U. S. A. 90:6900–6903.

Thibeault D, Bousquet C, Gingras R, Lagacé L, Maurice R, White PW, Lamarre D. 2004. Sensitivity of NS3 serine proteases from hepatitis C virus genotypes 2 and 3 to the inhibitor BILN 2061. J. Virol. 78:7352–7359.

Thomson MM, Fernández-García A, Delgado E, Vega Y, Díez-Fuertes F, Sánchez-Martínez M, Pinilla M, Castro MÁ, Mariño A, Ordóñez P, et al. 2012. Rapid expansion of a HIV-1 subtype F cluster of recent origin among men who have sex with men in Galicia, Spain. J. Acquir. Immune Defic. Syndr. 59:e49-51.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Thurner C, Witwer C, Hofacker IL, Stadler PF. 2004. Conserved RNA secondary structures in Flaviviridae genomes. J. Gen. Virol. 85:1113–1124.

Tillier ER, Collins RA. 1998. High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. Genetics 148:1993–2002.

Tong X, Li L, Haines K, Najera I. 2014. Identification of the NS5B S282T resistant variant and two novel amino acid substitutions that affect replication capacity in hepatitis C virus-infected patients treated with mericitabine and danoprevir. Antimicrob. Agents Chemother. 58:3105–3114.

Torres-Puente M, Cuevas JM, Jiménez-Hernández N, Bracho MA, García-Robles I, Wrobel B, Carnicer F, del Olmo J, Ortega E, Moya A, et al. 2008. Using evolutionary tools to refine the new hypervariable region 3 within the envelope 2 protein of hepatitis C virus. Infect. Genet. Evol. 8:74–82.

Trifonov V, Rabadan R. 2010. Frequency analysis techniques for identification of viral genetic data. MBio 1(3):e00156-10.

Tuplin A, Evans DJ, Simmonds P. 2004. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequences of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. J. Gen. Virol. 85:3037–3047.

UNAIDS. 1998. Report on the global HIV/AIDS epidemic June 1998. Geneva.

UNAIDS. 2004. 2004 report on the global HIV/AIDS epidemic: 4th global report. Geneva.

UNAIDS. 2013. GLOBAL REPORT: UNAIDS report on the global AIDS epidemic 2013. Geneva

UNAIDS. 2015. Fact sheet 2015 | UNAIDS. Geneva.

UNAIDS. 2016. Fact sheet – Latest global and regional statistics on the status of the AIDS epidemic. Geneva.

UNAIDS (Joint United Nations Programme on HIV/AIDS). 2002. Report on the global HIV/AIDS epidemic. :1–226. Geneva.

Vallari A, Holzmayer V, Harris B, Yamaguchi J, Ngansop C, Makamche F, Mbanya D, Kaptué L, Ndembi N, Gürtler L, et al. 2011. Confirmation of putative HIV-1 group P in Cameroon. J. Virol. 85:1403–1407.

Vega Y, Delgado E, Fernández-García A, Cuevas MT, Thomson MM, Montero V, Sánchez M, Sánchez AM, Pérez-Álvarez L, Spanish Group for the Study of New HIV-1 Diagnoses in Galicia and Basque Country. 2015. Epidemiological Surveillance of HIV-1 Transmitted Drug Resistance in Spain in 2004-2012: Relevance of Transmission Clusters in the Propagation of Resistance Mutations. PLoS One 10:e0125699.

van de Vijver DA, Wensing AMJ, Angarano G, Asjö B, Balotta C, Boeri E, Camacho R, Chaix M-L, Costagliola D, De Luca A, et al. 2006. The Calculated Genetic Barrier for Antiretroviral Drug Resistance Substitutions Is Largely Similar for Different HIV-1 Subtypes. J. Acquir. Immune Defic. Syndr. 41:352–360.

Wang KS, Choo QL, Weiner AJ, Ou JH, Najarian RC, Thayer RM, Mullenbach GT, Denniston KJ, Gerin JL, Houghton M. 1986. Structure, sequence and expression of the hepatitis delta (delta) viral genome. Nature 323:508–514.

Washenberger CL, Han J-Q, Kechris KJ, Jha BK, Silverman RH, Barton DJ. 2007. Hepatitis C virus RNA: dinucleotide frequencies and cleavage by RNase L. Virus Res. 130:85–95.

Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. Proc. Natl. Acad. Sci. U. S. A. 102:2454–2459.

Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW, Swanstrom R, Burch CL, Weeks KM. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome.

Nature 460:711–716.

Wertheim JO, Fourment M, Kosakovsky Pond SL. 2012. Inconsistencies in estimating the age of HIV-1 subtypes due to heterotachy. Mol. Biol. Evol. 29:451–456.

Wertheim JO, Leigh Brown AJ, Hepler NL, Mehta SR, Richman DD, Smith DM, Kosakovsky Pond SL. 2014. The global transmission network of HIV-1. J. Infect. Dis. 209:304–313.

Whelan S, Liò P, Goldman N. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 17:262–272.

Wilkinson E, Engelbrecht S, de Oliveira T. 2015. History and origin of the HIV-1 subtype C epidemic in South Africa and the greater southern African region. Sci. Rep. 5:16897.

Wilkinson KA, Merino EJ, Weeks KM. 2006. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nat. Protoc. 1:1610–1616.

Wyles DL, Gutierrez JA. 2014. Importance of HCV genotype 1 subtypes for drug resistance and response to therapy. J. Viral Hepat. 21:229–240.

Xu Z, Choi J, Yen TS, Lu W, Strohecker A, Govindarajan S, Chien D, Selby MJ, Ou J. 2001. Synthesis of a novel hepatitis C virus protein by ribosomal frameshift. EMBO J. 20:3840–3848.

Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409–418.

Yang Z, Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Mol. Biol. Evol. 14:717–724.

Yebra G, Delgado R, Pulido F, Rubio R, Galán JC, Moreno S, Holguín Á. 2014. Different trends of transmitted HIV-1 drug resistance in Madrid, Spain, among risk groups in the last decade. Arch. Virol. 159:1079–1087.

Yebra G, Holguín A, Pillay D, Hué S. 2013. Phylogenetic and demographic characterization of HIV-1 transmission in Madrid, Spain. Infect. Genet. Evol. 14:232–239.

Yebra G, Kalish ML, Leigh Brown AJ. 2016. Reconstructing the HIV-1 CRF02_AG and CRF06_cpx epidemics in Burkina Faso and West Africa using early samples. Infect. Genet. Evol.

Yebra G, de Mulder M, Martín L, Rodríguez C, Labarga P, Viciana I, Berenguer J, Alemán MR, Pineda JA, García F, et al. 2012. Most HIV type 1 non-B infections in the Spanish cohort of antiretroviral treatment-naïve HIV-infected patients (CoRIS) are due to recombinant viruses. J. Clin. Microbiol. 50:407–413.

Yoder AD, Yang Z. 2000. Estimation of primate speciation dates using local molecular clocks. Mol. Biol. Evol. 17:1081–1090.

Zehender G, Ebranati E, Lai A, Santoro MM, Alteri C, Giuliani M, Palamara G, Perno CF, Galli M, Lo Presti A, et al. 2010. Population dynamics of HIV-1 subtype B in a

cohort of men-having-sex-with-men in Rome, Italy. J. Acquir. Immune Defic. Syndr. 55:156–160.

Zein NN. 2000. Clinical significance of hepatitis C virus genotypes. Clin. Microbiol. Rev. 13:223–235.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zuckerkandl E, Pauling L. 1962. Horizons in Biochemistriy. In: Kasha M, Pullman B, editors. Horizons in Biochemistry. New York: Academic Press. p. 189–225.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 31:3406–3415.