


# SCIENTIFIC REPORTS

OPEN

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

Juan Ángel Patiño-Galindo<sup>1,2</sup>, Manoli Torres-Puente<sup>1</sup>, María Alma Bracho<sup>1,2</sup>, Ignacio Alastrué<sup>3</sup>, Amparo Juan<sup>3</sup>, David Navarro<sup>4,5</sup>, María José Galindo<sup>4</sup>, Dolores Ocete<sup>6</sup>, Enrique Ortega<sup>6</sup>, Concepción Gimeno<sup>5,6</sup>, Josefina Belda<sup>7</sup>, Victoria Domínguez<sup>8</sup>, Rosario Moreno<sup>8</sup> & Fernando González-Candelas<sup>1,2</sup> 

HIV infections are still a very serious concern for public health worldwide. We have applied molecular evolution methods to study the HIV-1 epidemics in the Comunidad Valenciana (CV, Spain) from a public health surveillance perspective. For this, we analysed 1804 HIV-1 sequences comprising protease and reverse transcriptase (PR/RT) coding regions, sampled between 2004 and 2014. These sequences were subtyped and subjected to phylogenetic analyses in order to detect transmission clusters. In addition, univariate and multinomial comparisons were performed to detect epidemiological differences between HIV-1 subtypes, and risk groups. The HIV epidemic in the CV is dominated by subtype B infections among local men who have sex with men (MSM). 270 transmission clusters were identified (>57% of the dataset), 12 of which included  $\geq 10$  patients; 11 of subtype B (9 affecting MSMs) and one (n = 21) of CRF14, affecting predominately intravenous drug users (IDUs). Dated phylogenies revealed these large clusters to have originated from the mid-80s to the early 00s. Subtype B is more likely to form transmission clusters than non-B variants and MSMs to cluster than other risk groups. Multinomial analyses revealed an association between non-B variants, which are not established in the local population yet, and different foreign groups.

Of the four phylogenetic groups which comprise HIV-1 (M, N, O, P), group M is the causal agent of the AIDS pandemic<sup>1,2</sup>. The latest UNAIDS/WHO report<sup>3</sup> estimated in almost 37 million the number of persons infected with HIV globally, with approximately 2.1 million new infections in 2015.

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<sup>4</sup>. 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<sup>5,6</sup>. The increasing incidence among MSM is remarkably high in the age range 20–35 years<sup>6</sup>. HIV infections in Spain also affect the foreign population disproportionately: in 2012, 35% of the new diagnosis corresponded to immigrants or persons of foreign origin<sup>6</sup>.

There exist nine subtypes (denoted as A, B, C, D, F, G, H, J and K) and at least 61 circulating recombinant forms (CRFs) within HIV-1 group M<sup>7</sup>. 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<sup>8,9</sup>. Genetic and antigenic differences among HIV-1 subtypes and CRFs are also a challenge for the development of an effective HIV-1 vaccine<sup>10</sup>.

<sup>1</sup>Unidad Mixta Infección y Salud Pública FISABIO-CSISP/Universidad de Valencia-I2SysBio, Valencia, 46180, Spain.

<sup>2</sup>CIBER of Epidemiology and Public Health, Instituto de Salud Carlos III, Madrid, 28029, Spain. <sup>3</sup>Unidad Prevención del SIDA y otras ITS, Valencia, 46017, Spain. <sup>4</sup>Hospital Clínico Universitario, Valencia, 46010, Spain. <sup>5</sup>Dpto. Microbiología, Universidad de Valencia, 46080, Valencia, Spain. <sup>6</sup>Consorcio Hospital General Universitario, Valencia, 46014, Spain.

<sup>7</sup>Unidad Prevención del SIDA y otras ITS, Alicante, 03010, Spain. <sup>8</sup>Hospital General Universitario, Castello, 12004, Spain. Correspondence and requests for materials should be addressed to F.G.-C. (email: [fernando.gonzalez@uv.es](mailto:fernando.gonzalez@uv.es))

	B (n = 1512)	A1 (n = 34)	F1 (n = 34)	G (n = 20)	CRF02_AG (n = 66)	CRF14_BG (n = 28)	Others (n = 110)*	Total (n = 1804)
<b>Gender</b>								
Male	818	10	13	5	24	11	49	930
Female	115	6	6	8	25	6	24	190
UNK	579	18	15	7	17	11	37	684
<b>Geographical origin</b>								
Spain	579	6	5	3	11	8	25	637
W. Europe and N. America	27	0	0	0	1	0	1	29
Eastern Europe	17	6	5	0	0	5	6	39
Africa and M. East	13	3	3	10	26	0	12	67
Latin America	141	1	3	0	7	3	19	174
Others	1	0	0	0	0	0	2	3
UNK	734	18	18	7	21	12	45	855
<b>Risk group</b>								
HT	124	8	6	9	29	2	26	204
MSM	586	3	6	0	11	1	30	637
IDU	85	3	0	2	3	12	5	110
Other	2	0	0	0	0	0	0	2
UNK	715	20	22	9	23	13	49	851
<b>Clustering</b>								
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)	279	0	0	0	0	21	0	300
<b>Age: mean (min–max)</b>	35.25 (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)	34.82 (0–76)

**Table 1.** Distribution of HIV cases in the dataset (n = 1804) classified by viral subtype, gender, nationality, risk group, age and clustering status. \*The “Others” subset includes 10 subtype C, 1 subtype D, 13 CRF19\_cpx, 12 CRF12\_BF, 9 CRF06\_cpx, 7 CRF47\_BF and 58 other (mostly unassigned) recombinant sequences.

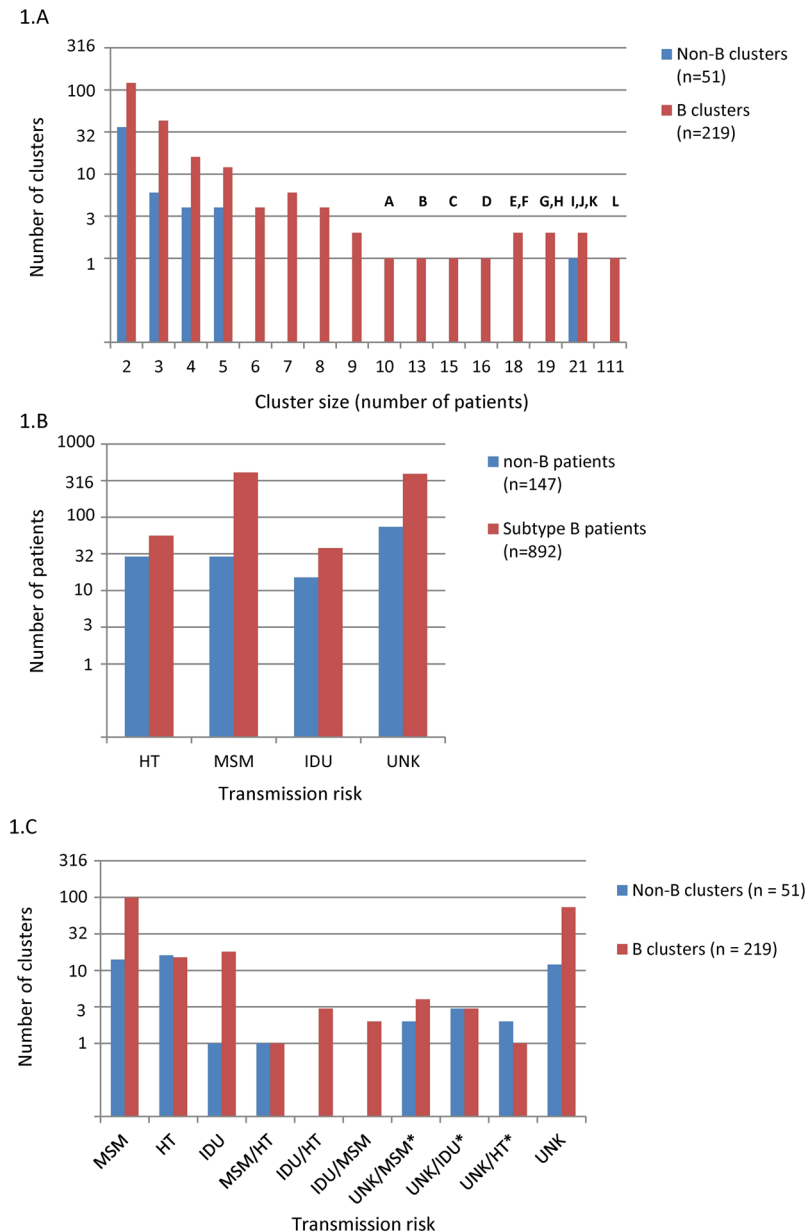
Worldwide, the most prevalent HIV-1 subtype is C, accounting for around 50% of all cases. However, the HIV epidemic in Europe, particularly among MSM, is mainly driven by subtype B<sup>11</sup>, with frequent reported transmission clusters affecting this group<sup>12–20</sup>. However, there is evidence for an increased introduction of non-B subtypes<sup>11</sup>. For instance, in a sample of 206 patients from Spain, Abecasis and colleagues<sup>11</sup> found that CRF02\_AG was the second most prevalent HIV-1 variant after subtype B (prevalence <2%).

The Comunidad 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. Here, we have analysed more than 1,800 HIV-1 *pol* sequences obtained between 2004 and 2014 from different patients in the CV. We have used molecular phylogenetic analyses to complement HIV surveillance tasks of the Public Health Directorate of the CV Regional Government. Specifically, we have used this information to infer the distribution of HIV-1 subtypes, to analyse the introductions (and further local expansion) of this virus in the CV, to identify the emergence of new viral variants in this region, and also to analyse which risk 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.

## Results

Of the 1804 sequences analysed, 1512 (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 = 3.7%), 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, 83.04% (930/1120) were male vs 16.96% (190/1120) female; 67.12% (637/949) were native from Spain (determined by place of birth) vs 32.87% (312/949) immigrants or of foreign origin; 66.84% (637/953) were MSM, 21.41% (204/953) heterosexual (HT) and 11.54% (110/953) IDUs. One patient was infected vertically and other one was haemophilic. The mean age was 34.82 years (range 0 to 76) (Table 1). MSMs were more likely to be Spaniards than non-MSM (Fisher’s Exact test, FET: p-value =  $5.5 \times 10^{-5}$ , odds-ratio, OR = 1.84 (1.35–2.48)).

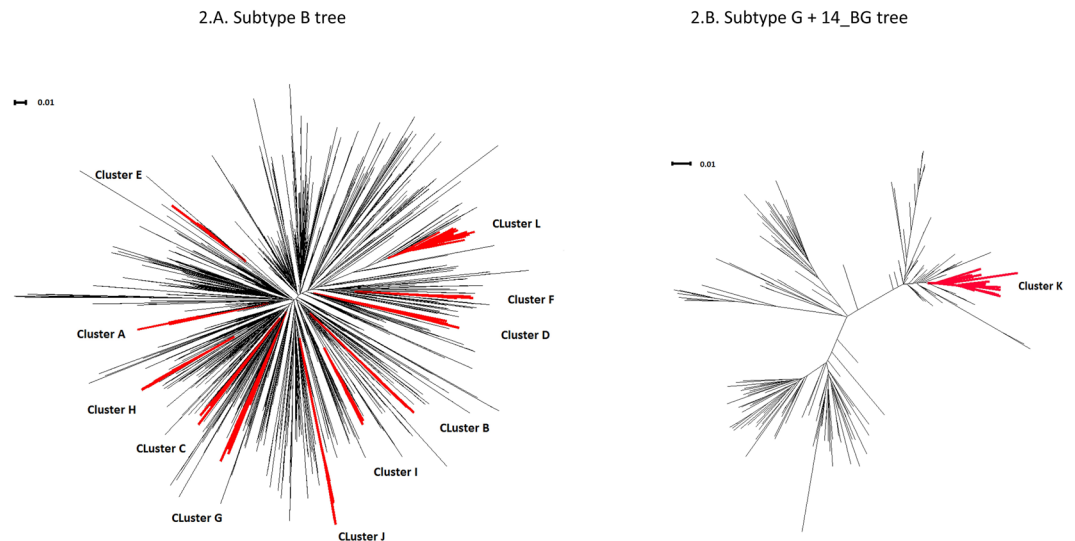
Phylogenetic analyses revealed the existence of 270 transmission clusters, with sizes ranging from 2 to 111 patients (Fig. 1A). In total, 1039 patients from the dataset were included in a transmission cluster (57.5%), with 302 patients (16.7%) included in large clusters of 10 or more patients (Table 1; Fig. 1A). Among the 892 patients clustering in transmission groups of subtype B, 407 (45.6%) were MSM, 56 (6.3%) HTs, 38 (4.2%) IDUs and 391 (43.8%) of unknown transmission route. On the other hand, of the 147 patients clustering in non-B clusters 29



**Figure 1.** (A) Distribution of sizes (log<sub>10</sub> scale) of the 270 transmission clusters found in the CV (2004–2014) through phylogenetic analysis. Block letters indicate the 12 clusters that were analysed 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.

(19.7%) were HTs, 29 (19.7%) MSM, 15 (10.2%) IDUs and (50.3%) 74 of unknown transmission route (Fig. 1B). Among the transmission clusters that could be classified by the most frequent risk of transmission ( $\geq 50\%$ ) for the CV patients included in them, the most prevalent transmission clusters of subtype B were those classified as MSM (n = 99/219, 45.2%), followed by IDUs (n = 18/219, 8.2%) and HTs (n = 15/219, 6.8%). However, for non-B transmission clusters, those of HTs were the most frequent ones (n = 16/51, 31.4%), followed by MSM (n = 14/51, 27.5%) and IDUs (n = 1/51, 2.0%). 87 of the 219 (39.7%) subtype B clusters and 20 of 51 (39.2%) non-B clusters could not be classified based on a majority risk of transmission (Fig. 1C). Subtype B sequences were more likely to be part of a transmission cluster than non-B sequences (FET: p-value = 0.007, OR = 1.42, 95% CI = 1.09–1.84). Also, MSM were more likely to be part of a transmission cluster than other risk groups (FET: p-value =  $1.68 \times 10^{-4}$ , OR = 1.56, 95% CI = 1.23–1.99).

Twelve large transmission clusters, which included at least 10 patients from the CV (Fig. 2), were detected: 11 were from subtype B and included a total of 281 patients (Fig. 2A). One cluster of 21 patients corresponded to CRF14\_BG (Cluster K, Fig. 2B). 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). Nine



**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).

of the large 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 (Supplementary Figs 1 to 12) revealed that they originated between 1984 (cluster G, unclassified transmission route) and 2005 (cluster B, MSM). Only transmission clusters that included MSM as the main transmission risk were originated since 2000 (Table 2; Supplementary Figs 1 to 12). Bagplots (bivariate representations of boxplots<sup>21</sup>) 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 Fig. 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).

Despite most patients infected with subtype B were Spanish natives (579 Spaniards vs 199 foreigners), a majority of those infected with non-B variants were foreigners (58 vs 113; FET:  $p$ -value  $< 2.2 \times 10^{-16}$ , OR = 5.66, 95% CI = 3.91–8.23). This difference was also observed when considering only patients belonging to a transmission cluster (subtype B: 378 vs 114; non-B: 31 vs 47; FET:  $p$ -value =  $1.8 \times 10^{-10}$ ; OR = 5.01, 95% CI = 2.96–8.57) (Supplementary Table S1).

A total of 88 subtype B, and 9 non-B transmission clusters did not include patients of foreign origin. These subtype B clusters included 178 Spanish patients and 122 of unknown origin, and the non-B clusters included 13 Spanish patients and 14 of unknown origin. A total of 49 subtype B, and 7 non-B transmission clusters were formed by both Spanish and non-Spanish patients. Altogether, these mixed subtype B clusters included 200 Spanish, 87 non-Spanish (mostly from Latin America) and 113 of unknown origin, and the mixed non-B clusters included 18 Spanish, 10 non-Spanish (mostly from Latin America or Eastern Europe) and 13 of unknown origin. All large transmission clusters but one (cluster G, for which there was not information available in 17 of its 19 patients) presented this characteristic. Finally, a total of 16 subtype B and 25 non-B transmission clusters included only patients of foreign origin. These subtype B clusters included 27 foreign patients (mostly from Latin America) and 16 of unknown origin, and the non-B clusters included 37 foreign patients (mostly from Africa/Middle East and Latin America) and 20 of unknown origin (Supplementary Table S1).

In addition to the patients' origin, other variables considered for the multinomial analysis presented differences between subtypes. So, subtype B sequences were disproportionately more present in males compared to females (FET:  $p$ -value  $< 2.2 \times 10^{-16}$ , OR = 4.75, 95% CI = 3.29–6.86), in MSM with respect to HTs and IDUs ( $p$ -value  $< 2.2 \times 10^{-16}$ , OR = 7.4, 95% CI = 4.87–11.31 and  $p$ -value =  $2 \times 10^{-5}$ , OR = 3.37, 95% CI = 1.90–5.89, respectively), and in IDUs compared to HTs ( $p$ -value = 0.0038, OR = 2.19, 95% CI = 1.26–3.88). There were also differences between subtypes with respect to age distribution (Kruskal-Wallis test:  $P = 0.0013$ ,  $\text{Chi}^2 = 21.8$ ,  $\text{df} = 6$ ), with CRF02\_AG patients being significantly younger than B patients (Games-Howell post-hoc test result for this comparison:  $p = 0.001$ ,  $t = 4.37$ ,  $\text{df} = 58$ ). 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 906 patients for whom there was information for all the variables considered: sex, risk group, country of origin, clustering status and age. We considered 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  $R^2 = 0.32$ ,  $\text{Chi}^2 = 421.34$ ,  $p$ -value  $< 2.2 \times 10^{-16}$ ). We checked that this subset was representative of the global set ( $n = 1804$ ) by means of FETs. The  $p$ -values obtained for all the analysed variants were  $> 0.05$ , including the distribution of patients from the different cluster-size categories (no cluster, small cluster, medium cluster and large cluster) that would result if only sequences included in the multinomial analysis had been taken into account for cluster detection, (FET:  $P > 0.10$ ; Supplementary Table S2). This dataset was also genetically representative of the whole dataset (Supplementary Fig. S13). The

Transmission 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)

**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). #Time measured in years. \*Substitutions per site and year.

multivariate model shows that the chances 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 = 0.029, OR = 36.2, 95% CI = 1.45–904.0), in IDUs (p-value =  $9.44 \times 10^{-5}$ , OR = 257, 95% CI = 15.8–4160) and in those patients forming large transmission clusters (p-value =  $1.6 \times 10^{-4}$ , OR = 36.9, 95% CI = 5.67–240) (Table 3).

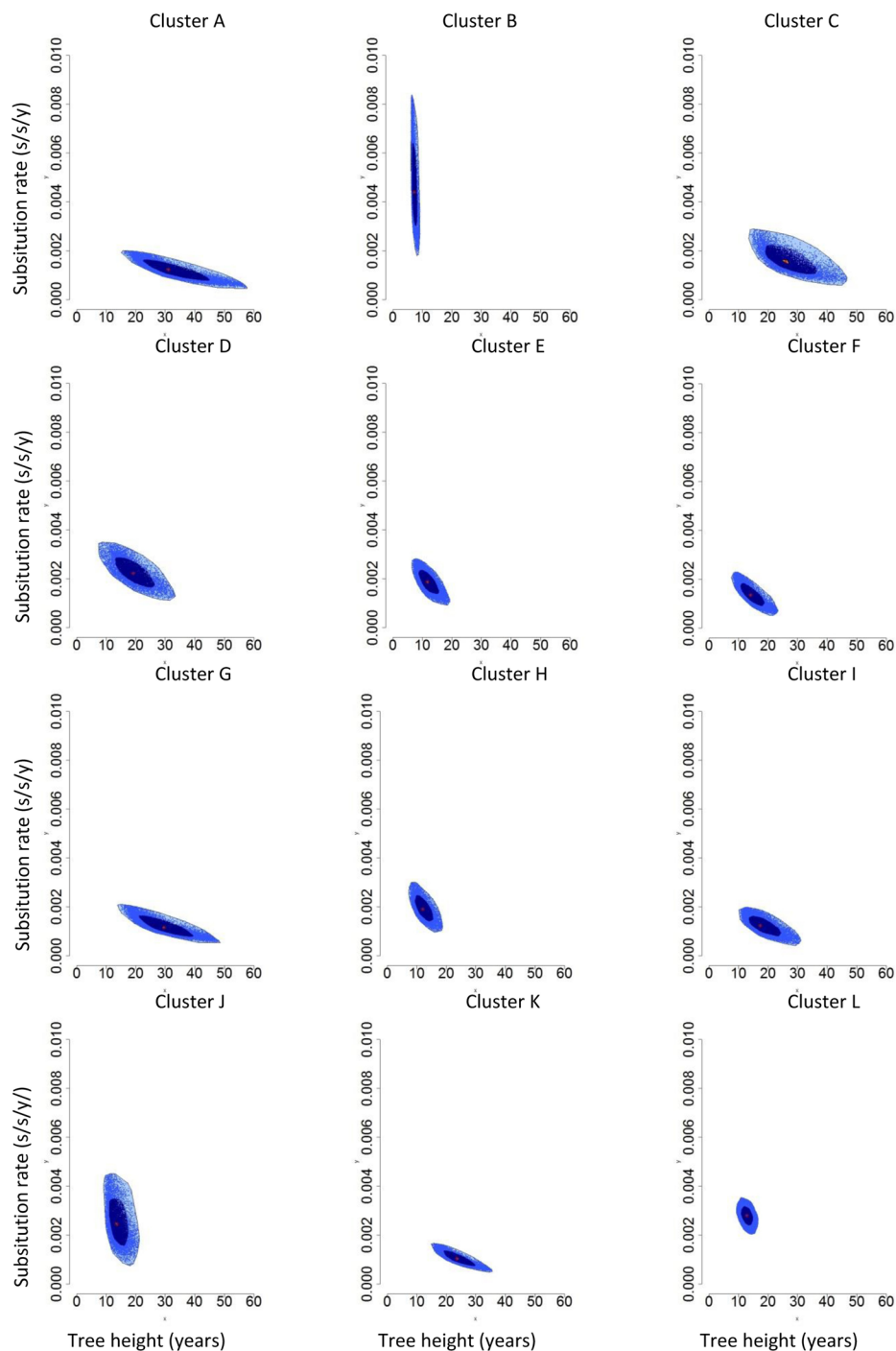
Pairwise genetic distances (GD) between individuals from a given transmission cluster ranged between 0.00 substitutions/site (s/s), in clusters of sizes of 2, 3, 5, 7, 8, 15, 16, 18 and 111 individuals, and 0.073 s/s, in a cluster of 6 individuals. For most pairwise comparisons, GD was lower than 0.045 s/s, regardless the size of the transmission cluster (Fig. 4; Supplementary Table S3).

## Discussion

We have studied the HIV epidemic in the Comunidad Valenciana by analysing, with molecular and evolutionary tools, 1804 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 = 3.7%), in agreement with previous estimates for Spain<sup>11,22</sup>.

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 be included in a transmission cluster, as well as for Spanish natives, than other risk groups. Considering the 12 largest clusters (size  $\geq 10$  patients), this transmission risk was the most frequent in 9 of them. Excluding transmission cluster A (MSM, tMRCA = 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. Previous analyses in the Spanish regions of Madrid<sup>23</sup> and the Basque Country<sup>24</sup> detected lower proportions of clustering patients (18 and 27% of their analysed sequences, respectively). These results suggest that the importance of local transmissions may not be the same in different Spanish regions. In addition, they also found evidences for an increased vulnerability to HIV of the Spanish MSM community in recent years. A more detailed analysis of cluster L is provided elsewhere<sup>25</sup>.

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 analysed 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<sup>26–28</sup>. 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 analysed time-span non-B HIV variants were not well established among Spanish locals in this region. This is exemplified by the fact that almost half of the non-B transmission clusters did not include any Spanish patient. 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. Previous molecular epidemiology studies of HIV-1 in Western Europe also reported non-B variants to be associated with the migrant population<sup>11</sup>, mainly affecting HTs. In the Spanish region of Madrid, where non-B variants have been found with a prevalence of up to 37%<sup>23</sup>, more than 70% of non-B sequences were found to belong to non-Spanish patients<sup>29</sup>. Other works in different Western European countries have reported differences in the prevalence of



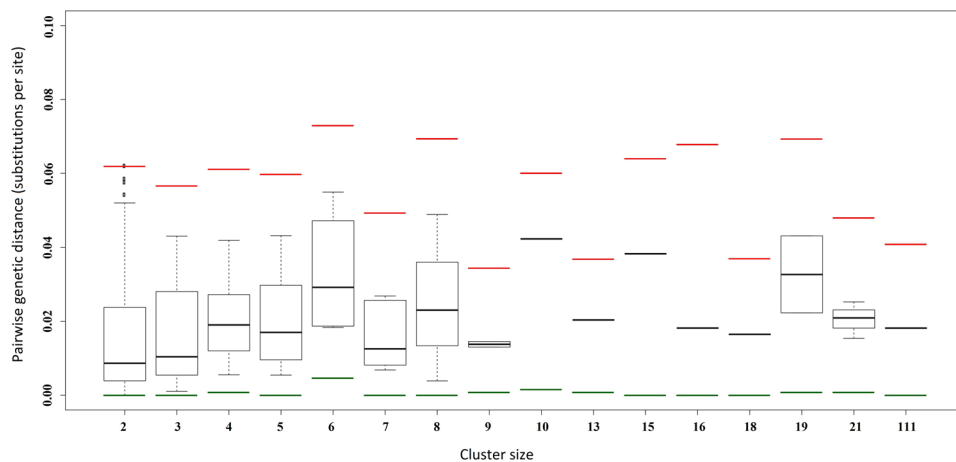
**Correlation between median tree heights and median rates of clusters A to L:  $P=0.012$ ,  $R=-0.70$**

**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).

non-B variants in the migrant population. Similarly to our results, in countries such as Switzerland (estimated prevalence of non-B HIV sequences = 21%<sup>30</sup>) or the Belgian-Luxemburg population (prevalence >45%<sup>11, 31</sup>) nearly 70% of non-B variants have been associated to patients of migrant origin<sup>30, 31</sup>. These results contrast with those from Portugal, the Western European country where subtype B displays the lowest prevalence<sup>11</sup>. An study in the Portuguese region of Minho (close to the border with Spain) found non-B variants to have a prevalence of 73%, and only 10% of these sequences belonged to non-Portuguese patients<sup>32</sup>, thus evidencing that non-B infections are already established among locals.

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	1.674	0.094265	\$	11.0	0.664	181
02_AG:RISK(HT)	1.07	0.579	1.850	0.064773	\$	2.91	0.937	9.07
14_BG:RISK(UDI)	5.55	1.42	3.905	9.44E-05	***	257.0	15.8	4160
14_BG:Eastern Europe	3.29	1.09	3.023	0.002502	**	26.9	3.18	227
F:Eastern Europe	3.51	1.08	3.250	0.001155	**	33.5	4.03	279
OTHERS: Eastern Europe	2.01	0.583	3.448	0.000566	***	7.47	2.38	23.5
A1:Eastern Europe	2.78	0.795	3.493	0.000478	***	16.1	3.38	76.3
F1:Africa and Middle East	3.50	1.05	3.353	0.000799	***	33.3	4.29	258
G:Africa and Middle East	3.78	1.17	3.236	0.001218	**	43.7	4.43	431
OTHERS: Africa and Middle East	2.77	0.554	5.009	5.48E-07	***	16.0	5.41	47.4
A1:Africa and Middle East	1.92	0.993	1.932	0.053351	\$	6.81	0.973	47.7
02_AG:Africa and Middle East	3.76	0.580	6.478	9.28E-11	***	42.9	13.8	134
14_BG:Latin America	2.71	1.38	1.955	0.050547	\$	15.0	0.994	225
OTHERS:Latin America	1.19	0.352	3.391	0.000696	**	3.30	1.66	6.58
14_BG:ClusterLARGE	3.61	0.955	3.776	0.000159	***	36.9	5.67	240

**Table 3.** Results of the multinomial analysis (only significant associations between each HIV variant and the categories compared at each variable are shown). \*\*\* $P < 0.001$ ; \*\* $0.001 < P < 0.01$ ; \* $0.01 < P < 0.05$ ; \$ $0.05 < P < 0.1$ .



**Figure 4.** Boxplots representing mean pairwise genetic distances (s/s) between individuals within transmission clusters of different sizes, as estimated with the TN93 + GAMMA (4 CAT) model. Green and red bars represent the minimum and maximum pairwise distances, respectively, between a pair of individuals from a given cluster size.

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<sup>33,34</sup>. Previous phylogeographic analyses have suggested that CRF14\_BG originated in the Iberian Peninsula<sup>33</sup>, and its prevalence is increasing in some Eastern European countries, such as Romania, boosted by migration between these countries and Spain<sup>35</sup>. Similarly to Romania, this CRF has been found to be common among IDUs in Greece, where a large transmission network of CRF14\_BG of recent origin (tMRCA = 2009) has been reported to affect more than 70 people, and which origin has been linked to strains from South-western Europe<sup>36</sup>.

We also detected two other transmission clusters of smaller size ( $n = 5$  and  $n = 4$ ) affecting Spanish-native MSMs of another highly pathogenic variant (CRF19\_cpx), which we previously reported as the first evidence of expansion of this variant outside Cuba<sup>37</sup>. 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 MSMs.

The significant association between non-B variants and immigrants reported in this work and in previous molecular epidemiology analyses of HIV-1 in Madrid, Spain<sup>23,29</sup>, as well as the aforementioned examples of emergence of CRFs associated with migratory waves, suggest that high migration and tourism rates in some Spanish regions may contribute to the high genetic diversity of their HIV-1 epidemics. Thus, although our results suggest that non-B HIV groups are not well established in the local population, with only a few of them including patients of both Spanish and non-Spanish origin, this situation could change soon without intensified HIV prevention campaigns focused on vulnerable groups, including migrants, IDUs and MSM.

Although the large number of sequences for which no epidemiological information was available could be a potential limitation of our study, the results reported are representative of the sampled population and represent one of the most comprehensive analysis of the HIV pandemics in Spain<sup>22,23,29,38</sup>.

We have found a significant, negative correlation 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<sup>39–41</sup>. 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 bottlenecks that occur at viral transmission, which might originate a transient accumulation of deleterious mutations. Consequently, these results suggest that TDRP is an important factor to consider in molecular epidemiology, even when datasets are obtained from the same population and at short timescales (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, in order to prevent spurious clustering by convergent evolution. In our analyses, we did not remove those positions because their impact on the detection of transmission clusters has been demonstrated to be irrelevant<sup>42</sup>. Furthermore, a great majority of our sequences were known to derive from treatment-naïve patients (only 66 patients were known to have been treated).

Additionally, although the identification or detection of transmission clusters is commonly based only on phylogenetic cluster support, it is also often based on criteria that combine phylogenetic support, usually high bootstrap values, with genetic distance (GD) thresholds<sup>43</sup>. In this work, we did not consider any GD threshold value as criterion to define transmission clusters. This is because the rationale for genetic distance cut-offs is rarely provided, with very different thresholds being reported in the literature<sup>44</sup>, and larger GD values are expected for older transmission events. In this way, choosing distance-based thresholds can lead to underestimating transmission clusters, especially those that span long time periods<sup>43</sup>, as it is the case of our dataset. Nevertheless, a great majority of GDs estimated from the pairwise comparisons performed here were lower than 0.045/s. Thus, almost all the transmission clusters reported here would be characterized by both high phylogenetic support and low GD.

In conclusion, our results provide evidence that the HIV-1 epidemic in the CV is dominated by subtype B, especially among local MSMs. Although there was 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, urges to increase efforts on HIV testing and prevention campaigns to prevent their further expansion.

## Materials and Methods

**Dataset.** A total of 1804 PR/RT sequences were obtained from newly HIV diagnosed people at six different hospitals and two HIV counselling and testing centres (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 and subtyped as detailed in Patiño-Galindo *et al.*<sup>25</sup>.

**Detection of local transmission clusters.** Detection of transmission clusters, was performed following the same two-step procedure detailed in Patiño-Galindo *et al.*<sup>25</sup> Briefly, in the first step, we obtained independent phylogenetic trees for each HIV-1 subtype and CRF detected, which included reference sequences retrieved from the Los Alamos HIV database spanning the analysed PR/RT region ( $n = 1787$  subtype B, 2097 A1, 720 F1, 854 G + CRF14\_BG and 2157 CRF02\_AG, as well as 5720 sequences from all the HIV-1 subtypes/CRFs that were rare in the CV dataset). The initial trees were reconstructed with FastTree 2.1 in order to detect potential transmission clusters, defined as clades with SH-like support  $\geq 0.70$ , and containing  $\geq 90\%$  sequences from the CV dataset<sup>17</sup>.

In the second step, these potential clusters were confirmed with ML phylogenies obtained with PhyML 3.0<sup>45</sup> in which the following additional reference sequences were included: (i) the 10 sequences with the highest similarity to each CV sequence from the potential clusters, as retrieved with a BLASTN search at the NCBI server (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>); (ii) 854 subtype B, 77 A1, 74 F1, 173 G + CRF14\_BG, 101 CRF02\_AG as well as 482 reference sequences from the other HIV-1 subtypes/CRFs. Only those potential clusters which contained more than 90% of sequences from the CV and grouped in the ML tree with aLRT support  $\geq 0.98$  were considered as confirmed. Clusters were then classified depending on the major risk of transmission ( $\geq 50\%$ ) for the corresponding patients from the CV.

**Dated phylogenies.** Dated phylogenies of large transmission clusters (those containing at least 10 patients from the CV<sup>46</sup>) were obtained with BEAST v1.8.1<sup>47</sup>, similarly as in Patiño-Galindo *et al.*<sup>25</sup>. 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)<sup>17</sup> was placed on the ucl.d.mean parameter.

**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 906 of the 1804 sequences were included in this analysis. Seven groups of different HIV-1 subtypes and CRFs were used: A1 (n = 14), B (n = 752), 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 S4). All the statistical analyses were performed using R<sup>48</sup>. The *mlogit* R package<sup>49</sup> was used for the multinomial analysis.

**Estimation of pairwise genetic distances.** Pairwise genetic distances (GD) between individuals within transmission clusters were estimated, as substitutions per site (s/s), with the TN93 +  $\Gamma$ (4CAT) model, using *Perl* and *R* scripts (available upon request) dependent on the *ape* R package<sup>50</sup>.

**Sequence information.** GenBank accession numbers for the CV sequences used in this study are HF567872–HF567912 and MF403205–MF404967.

## References

- Hahn, B. H., Shaw, G. M., De Cock, K. M. & Sharp, P. M. AIDS as a zoonosis: scientific and public health implications. *Science* **287**, 607–14 (2000).
- Plantier, J.-C. *et al.* A new human immunodeficiency virus derived from gorillas. *Nat. Med.* **15**, 871–2 (2009).
- UNAIDS. Fact sheet – Latest global and regional statistics on the status of the AIDS epidemic. (2016).
- ECDC. 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. (2013).
- UNAIDS (Joint United Nations Programme on HIV/AIDS). Report on the global HIV/AIDS epidemic. 1–226 (2002).
- DGSP. Vigilancia epidemiológica del VIH/SIDA en España: Sistema de información sobre nuevos diagnósticos del VIH y registro nacional de casos de SIDA. (2014).
- Kuiken, C. L. *et al.* *HIV Sequence Compendium 2012* (2012).
- Kaleebu, P. *et al.* 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**, 1244–50 (2002).
- Kouri, V. *et al.* CRF19\_cpx is an Evolutionary fit HIV-1 Variant Strongly Associated With Rapid Progression to AIDS in Cuba. *EBioMedicine* **2**, 244–54 (2015).
- Nickle, D. C. *et al.* Coping with viral diversity in HIV vaccine design. *PLoS Comput. Biol.* **3**, e75 (2007).
- Abecasis, A. B. *et al.* HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology* **10**, 7 (2013).
- Kouyos, R. D. *et al.* Molecular epidemiology reveals long-term changes in HIV type 1 subtype B transmission in Switzerland. *J. Infect. Dis.* **201**, 1488–97 (2010).
- Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A. & Leigh Brown, A. J. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* **5**, e50 (2008).
- Leigh Brown, A. J. *et al.* Transmission network parameters estimated from HIV sequences for a nationwide epidemic. *J. Infect. Dis.* **204**, 1463–9 (2011).
- Cuevas, M. T. *et al.* 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 (2009).
- Zehender, G. *et al.* 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**, 156–60 (2010).
- Hué, S., Pillay, D., Clewley, J. P. & Pybus, O. G. Genetic analysis reveals the complex structure of HIV-1 transmission within defined risk groups. *Proc. Natl. Acad. Sci. USA* **102**, 4425–4429 (2005).
- Bezemer, D. *et al.* 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 (2015).
- Magiorkinis, G. *et al.* The global spread of HIV-1 subtype B epidemic. *Infect. Genet. Evol.* **46**, 169–179 (2016).
- Beloukas, A. *et al.* Molecular epidemiology of HIV-1 infection in Europe: An overview. *Infect. Genet. Evol.* **46**, 180–189 (2016).
- Rousseeuw, P. J., Ruts, I. & Tukey, J. W. The Bagplot: A Bivariate Boxplot. *Am. Stat.* **53**, 382–387 (1999).
- Yebra, G. *et al.* 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**, 407–13 (2012).
- Yebra, G., Holguín, A., Pillay, D. & Hué, S. Phylogenetic and demographic characterization of HIV-1 transmission in Madrid, Spain. *Infect. Genet. Evol.* **14**, 232–9 (2013).
- Patiño-Galindo, J. A. *et al.* Transmission dynamics of HIV-1 subtype B in the Basque Country, Spain. *Infect. Genet. Evol.* **40**, 91–7 (2016).
- Patiño-Galindo, J. Á. *et al.* Identification of a large, fast-expanding HIV-1 subtype B transmission cluster among MSM in Valencia, Spain. *PLoS One* **12**, e0171062 (2017).
- Hemelaar, J. The origin and diversity of the HIV-1 pandemic. *Trends Mol. Med.* **18**, 182–92 (2012).
- Buonaguro, L., Tornesello, M. L. & Buonaguro, F. M. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *J. Virol.* **81**, 10209–19 (2007).
- Bello, G., Afonso, J. M. & Morgado, M. G. Phylodynamics of HIV-1 subtype F1 in Angola, Brazil and Romania. *Infect. Genet. Evol.* **12**, 1079–86 (2012).
- González-Alba, J. M. *et al.* Molecular surveillance of HIV-1 in Madrid, Spain: a phylogeographic analysis. *J. Virol.* **85**, 10755–63 (2011).
- von Wyl, V. *et al.* The role of migration and domestic transmission in the spread of HIV-1 non-B subtypes in Switzerland. *J. Infect. Dis.* **204**, 1095–103 (2011).
- Dauwe, K. *et al.* Characteristics and spread to the native population of HIV-1 non-B subtypes in two European countries with high migration rate. *BMC Infect. Dis.* **15**, 524 (2015).
- Carvalho, A. *et al.* Analysis of a local HIV-1 epidemic in Portugal highlights established transmission of non-B and non-G subtypes. *J. Clin. Microbiol.* **53**, 1506–14 (2015).
- Bártolo, I. *et al.* Origin and Epidemiological History of HIV-1 CRF14\_BG. *PLoS One* **6**, e24130 (2011).
- Pérez-Álvarez, L. *et al.* Predominance of CXCR4 tropism in HIV-1 CRF14\_BG strains from newly diagnosed infections. *J. Antimicrob. Chemother.* **69**, 246–53 (2014).

35. Niculescu, I. *et al.* 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**, 488–95 (2015).
36. Paraskevis, D. *et al.* Economic recession and emergence of an HIV-1 outbreak among drug injectors in Athens metropolitan area: a longitudinal study. *PLoS One* **8**, e78941 (2013).
37. Patiño Galindo, J. A. *et al.* Expansion of the CRF19\_cpx Variant in Spain. *J. Clin. Virol.* **69**, 146–149 (2015).
38. Yebra, G. *et al.* Different trends of transmitted HIV-1 drug resistance in Madrid, Spain, among risk groups in the last decade. *Arch. Virol.* **159**, 1079–87 (2014).
39. Ho, S. Y. W. & Shapiro, B. Skyline-plot methods for estimating demographic history from nucleotide sequences. *Mol. Ecol. Resour.* **11**, 423–434 (2011).
40. Aiwusakun, P. & Katzourakis, A. Time dependency of foamy virus evolutionary rate estimates. *BMC Evol. Biol.* **15**, 119 (2015).
41. Aiwusakun, P. & Katzourakis, A. Time-Dependent Rate Phenomenon in Viruses. *J. Virol.* **90**, 7184–95 (2016).
42. Hué, S., Clewley, J. P., Cane, P. A. & Pillay, D. HIV-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS* **18**, 719–28 (2004).
43. Hassan, A. S., Pybus, O. G., Sanders, E. J., Albert, J. & Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data. *AIDS* **31**, 1211–1222 (2017).
44. Grabowski, M. K. & Redd, A. D. Molecular tools for studying HIV transmission in sexual networks. *Curr. Opin. HIV AIDS* **9**, 126–33 (2014).
45. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–21 (2010).
46. Lunar, M. M. *et al.* Bridging epidemiology with population genetics in a low incidence MSM-driven HIV-1 subtype B epidemic in Central Europe. *BMC Infect. Dis.* **15**, 65 (2015).
47. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 214 (2007).
48. R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical computing, 2014).
49. Croissant, Y. *Package mlogit.* (2015).
50. Paradis, E. *et al.* Package 'ape' (2017).

## Acknowledgements

This work was partly supported by project BFU2014-58656-R from Ministerio de Economía y Competitividad (Spanish Government) and project PROMETEO/2016/122 from Generalitat Valenciana to F.G.C., project CP07/00078 from Ministerio de Sanidad y Consumo (Spanish Government) to M.A.B., and Gilead Science S.L. J.A.P.G. was recipient of a fellowship (FPU-AP2010-0561) from Ministerio de Educación, Cultura y Deporte, Spanish Government. The funders played no role on the design, analysis and conclusions from this work. We thank the CRIVIH for their support of this investigation. All the authors declare to have no conflict of interests.

## Author Contributions

J.A.P.G., M.T.P., M.A.B. and F.G.C. designed the study. M.T.P. and M.A.B. performed the experiments. J.A.P.G. and F.G.C. analysed the data. M.T.P., M.A.B., I.A., A.J., D.N., M.J.G., D.O., E.O., C.G., J.B., V.D., R.M. contributed data. J.A.P.G. and F.G.C. wrote the paper. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-10286-1

**Competing Interests:** The authors declare that they have no competing interests.

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



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

© The Author(s) 2017