# Rapid CD4 cell loss is caused by specific CRF01_AE cluster with V3 signatures favoring CXCR4 usage

## Short title: CRF01_AE cluster with high pathogenicity

Hongshuo Song[a], Weidong Ou[a], Yi Feng[a], Junli Zhang[a], Fan Li[a], Jing Hu[a], Hong Peng[a], Hui Xing[a], Liying Ma[a], Qiuxiang Tan[b], Beili Wu[b], and Yiming Shao[a,c,1]

[a] State Key Laboratory for Infectious Disease Prevention and Control, National Center for AIDS/STD Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China; [b] Chinese Academy of Sciences, Shanghai 201203, China; [c] Center of Infectious Diseases, Peking University, Beijing 100191, China.

[1]Correspondence to:
Dr. Yiming Shao, M.D., Ph.D.
National Center for AIDS/STD Control and Prevention
Chinese Center for Disease Control and Prevention
Beijing 102206, China
Phone: 86-10-82805014
Email: yshao@bjmu.edu.cn

**Key words:** CRF01_AE; Coreceptor tropism; CXCR4; HIV-1 pathogenesis

## Abstract

HIV-1 evolved into various genetic subtypes and circulating recombinant forms (CRFs) in the global epidemic, with the same subtype or CRF usually having similar phenotype. Being one of the world's major CRFs, CRF01_AE infection was reported to associate with higher prevalence of CXCR4 (X4) viruses and faster CD4 decline. However, the underlying mechanisms remain unclear. We identified eight phylogenetic clusters of CRF01_AE in China and hypothesized that they may have different phenotypes. In the national HIV molecular epidemiology survey, we discovered that people infected by CRF01_AE cluster 4 had significantly lower CD4 count (391 vs. 470, $p < 0.0001$) and higher prevalence of predicted X4-using viruses (17.1% vs. 4.4%, $p < 0.0001$) compared to those infected by cluster 5. In a MSM cohort, X4-using viruses were only isolated from sero-convertors infected by cluster 4, which associated with rapid CD4 loss within the first year of infection (141 vs. 440, $p = 0.01$). Using co-receptor binding model, we identified unique V3 signatures in cluster 4 that favor CXCR4 usage. We demonstrate for the first time that HIV-1 phenotype and pathogenicity can be determined at the phylogenetic cluster level in a single subtype. Since its initial spread to human from chimpanzee in 1930s, HIV-1 remains undergoing rapid evolution in larger and more diverse population. The divergent phenotype evolution of two major CRF01_AE clusters highlights the importance in monitoring the genetic evolution and phenotypic shift of HIV-1 to provide early warning for the appearance of more pathogenic strains such as CRF01_AE cluster 4.

## Significance Statement

Past studies on HIV-1 evolution were mainly at the genetic level. This study provides well-matched genotype and phenotype data and demonstrates disparate pathogenicity of two major CRF01_AE clusters. While both CRF01_AE cluster 4 and cluster 5 are mainly spread through the MSM route, cluster 4 but not cluster 5 causes fast CD4 loss, which is associated with the higher prevalence CXCR4 viruses in cluster 4. The higher CXCR4 use tendency in cluster 4 is derived from its unique V3 loop favoring CXCR4 binding. This study for the first time demonstrates disparate HIV-1 phenotype between different phylogenetic clusters. It is important to monitor HIV-1 evolution at both the genotype and phenotype level to identify and control more pathogenic HIV-1 strains.

## Introduction

During global transmission, HIV-1 evolved into various subtypes and their hybrids, the so called circulating recombinant forms (CRFs) (1). CRF01_AE, one of the major CRFs spreads mainly in Southeast Asia and China (1, 2). Early studies in Thailand reported faster CD4 loss and shorter survival of CRF01_AE infected people compared to infections in western countries where subtype B predominate (3-6). In recent years, faster disease progression of CRF01_AE was also reported in China (7-9). Although past studies reported high prevalence of X4 viruses and fast disease progression of CRF01_AE infections (7, 8, 10, 11), the data were mainly based on samples without known infection time and genotypic prediction without phenotypic confirmation. It is still unclear how early X4 viruses can emerge during natural CRF01_AE infection and whether the genotypic prediction is reliable.

The epidemic of CRF01_AE in China was initiated by multiple phylogenetic clusters introduced from Thailand in 1990s (2). We formerly identified eight CRF01_AE clusters, which have different geographic distributions and epidemic patterns (2, 12). Therefore, we hypothesized that they may have different phenotypes. We used the national HIV molecular epidemiology survey (NHMES) dataset for screening and a men-who-have-sex-with-men (MSM) sero-incidence cohort for in-depth study to determine the time of X4 virus emergence, and phenotypically confirmed viral tropism using matched viral isolates. We observed significantly lower CD4 counts in CRF01_AE cluster 4 compared to cluster 5 in the NHMES dataset. Focusing on the MSM cohort, we demonstrated higher prevalence of X4 phenotype in cluster 4 among people within the first year of

101 infection. We further determined the genetic and structural basis favoring X4 co-
102 receptor usage in cluster 4. This is the first demonstration that different clusters from the
103 same HIV-1 subtype can cause disparate rate of disease progression and bridged the
104 missing link between CRF01_AE genetic makeup, phenotype and clinical outcomes.
105
106
107 **Results**
108 **Lower CD4 T cell count and higher prevalence of X4-using virus in CRF01_AE**
109 **cluster 4**
110 We compared the CD4 T cell counts of 1118 CRF01_AE, 633 CRF07_BC and 123
111 subtype B newly diagnosed HIV-1 positive participants in the China's NHMES study and
112 found no differences (Fig. 1A). When analyzing the two major CRF01_AE clusters
113 which contribute greatly to China's MSM epidemic, however, the CD4 count in cluster 4
114 was significantly lower than in cluster 5 ($p < 0.0001$) (Fig. 1B). Since CD4 counts
115 decline with time during natural infection, we distinguished cases of recent infection
116 from long-term infection by HIV-1 Limiting Antigen-Avidity assay. Significantly lower
117 CD4 count in cluster 4 was found in both recent and long-term infection groups ($p =$
118 0.0004 and $p < 0.0001$, respectively) (Fig.1C-D). The proportion of people with CD4
119 below 200 was higher in cluster 4 than in cluster 5 (9.2 % vs.1.8 %, $p = 0.0001$) (Fig.
120 1B), while the rates of recent infection were nearly the same between the two clusters
121 (31.3% vs. 31.4%). The results clearly showed that the CRF01_AE cluster 4 but not
122 cluster 5 leads to rapid CD4 cell loss after infection.

Because CXCR4 tropism is associated with lower CD4 cell count (13-17), we investigated the prevalence level of X4 viruses in clusters 4 and 5 using genotypic prediction. Geno2pheno prediction (FPR cutoff = 2%) showed higher prevalence of X4 genotype in cluster 4 than in cluster 5 (17.1% vs. 4.4%, $p < 0.0001$) (Fig. 1E). In cluster 4, the X4 genotype were "concentrated" among patients with low CD4 counts (Fig. 1E). This result indicates a strong association between higher X4 prevalence and lower CD4 count in CRF01_AE cluster 4 infection.

**High prevalence of X4-using phenotype in CRF01_AE cluster 4 among recently infected individuals**

To further characterize the mechanism of fast CD4 cell loss in CRF01_AE cluster 4, we focused on a sero-incidence cohort with around two thousands of MSMs from Beijing Chaoyang District (the CYM cohort). For the 135 MSM sero-convertors, we first performed Illumina deep sequencing on all 78 participants with archived first year blood samples and successfully sequenced 71 of them, with an average of 65,000 sequencing reads per participant (SI Appendix, Figs. S1-S2, Tables S1-S2). Analysis of Geno2pheno FPR distribution among the plasma viral quasispecies in each participant again found higher frequency of X4-using variants in people infected by CRF01_AE cluster 4 than by cluster 5, CRF07_BC and subtype B (Fig. 2 and SI Appendix, Fig. S3 and Table S2). Together, deep sequencing on samples collected within the first year of infection confirmed the observation in large cross-sectional NHMES study.

To confirm the genetic prediction phenotypically, we isolated viruses using cryopreserved PBMC from the deep sequenced CRF01_AE participants and conducted

coreceptor tropism assays using GHOST cell lines. A total of 24 viruses were successfully isolated. Five showed X4-using phenotype and the remaining 19 were R5-only phenotype (Fig. 3A-C). All of the five isolates with X4-using phenotype, including four dual-tropic and one exclusively X4-tropic belonged to cluster 4 (Fig. 3A and SI Appendix, Fig. S4) (The method to determine coreceptor tropism was described in Materials and Methods and Fig. S4). This indicates a much higher prevalence of X4-using phenotype in cluster 4 than in cluster 5 (31.3% vs. 0%), consistent with the result of genotypic prediction.

**Individuals harboring X4 viruses had significantly lower CD4 T cell count**

Previous studies demonstrated the association between X4-using phenotype and lower CD4 count, mainly for subtype B HIV-1 (13-17). However, a recent study based solely on genotypic prediction failed to find such an association in CRF01_AE infected people in China (though there is a trend that people with CD4<50 tend to have Geno2pheno FPR<5) (8). Because genotypic prediction could overestimate the actual X4 prevalence, we therefore compared the CD4 T cell count base on virus phenotype (Fig. 3D). Among the 24 phenotype-confirmed participants, those with X4-using phenotype had significantly lower CD4 counts compared to those with R5-only phenotype in cluster 4 (141 vs. 440, $p = 0.003$) and in cluster 5 (141 vs. 441, $p = 0.01$ ) (Fig. 3A and 3D). With well-matched phenotype data, we confirmed that X4-using phenotype is associated with significantly lower CD4 count in CRF01_AE. The higher prevalence of X4 phenotype in cluster 4 during the first year of infection and its association with lower CD4 count

explained the reason why cluster 4 had lower CD4 count compared to cluster 5 starting from early infection stage (Fig. 1C-D).

**Genetic and structural determinants for higher CXCR4 usage propensity**

To explore the mechanism for higher X4-using propensity exhibited by cluster 4 viruses, we first compared the V3 loop sequences between clusters 4 and 5. We found that cluster 4 viruses have two highly conserved basic amino acids at positions 13 and 32 in its V3 loop (R13 and K32, HXB2 numbering R308 and K327), which were present in only about 8% of the cluster 5 viruses (Fig. 4A-B). These two conserved amino acids confer cluster 4 viruses a higher positively charged V3 loop, upon which fewer mutations may be required to switch from the R5 to X4 phenotype. Structure analysis using V3-docking models (18) suggested that residue R13 in cluster 4 potentially forms salt bridges with D262 and E277 and a hydrogen bond with the side chain of H281 in the CXCR4 coreceptor, while the corresponding residue in cluster 5 may form hydrogen bonds with K22 and D276 in the CCR5 coreceptor (Fig. 4C). Because the ligand binding pocket of CXCR4 is more negatively charged than that in CCR5, the positively charged residue R13 in cluster 4 viruses is more favored. The residue K32 in cluster 4 may form salt bridges with CXCR4's N terminus, which contains more acidic residues than CCR5's N terminus (Fig. 4C). Therefore, the highly conserved residues R13 and K32 in cluster 4 may be key determinants for the higher tendency of X4 usage. Interestingly, the V3 region of cluster 4 has less variations compared to the CRF01_AE ancestor sequences from Thailand, notably the preservation of residue K32, while cluster 5 is more divergent (Fig. 4A and SI Appendix, Fig. S5).

Despite the fact that R13 and K32 were present in vast majority of cluster 4 viruses, only about 30% of individuals in cluster 4 were phenotypically confirmed to harbor X4 viruses. This indicates the existence of additional determinants to shift to the X4 phenotype. To further identify key amino acids governing the phenotype switch, we analyzed the genetic composition of PBMC isolates at the single-genome level (Fig. 5 and SI Appendix, Fig. S6 and Table S3). In order to "sieve out" the X4-using variant(s) from the entire viral population (that is, the exact sequence(s) accounting for the X4-using phenotype), we further sequenced the viruses released from the GHOST.X4 cell culture by SGA for the five X4-using samples (SI Appendix, Table S3). Comparing the genetic composition between the PBMC viral isolates and the concurrent plasma viral population showed two different patterns: in majority of R5 subjects (15 of 19), the V3 lineages in the PBMC isolate and in plasma were in proportion, that is, the predominant lineage in the PBMC isolate was also the predominant one in plasma (Fig. 5A and SI Appendix, Table S3). In contrast, in four of the five X4 isolates (with the exception of CYM194), the predominant, phenotypically confirmed X4 lineage in the PBMC isolate existed as minor variant in plasma (Fig. 5B and SI Appendix, Table S3). All V3 lineages detected in the viral isolates from PBMC were present in plasma, and more V3 lineages were detected in plasma than in the primary isolates (Fig. 5 and SI Appendix, Table S3).

Several V3 alterations were found in the phenotypically confirmed X4 sequences. First, all X4-using sequences lost the N-linked glycan site at the beginning of the V3 loop (V3 positions 6-8, HXB2 numbering 301-303), mostly by T to I substitution at position 8 (Fig. 6A). Second, while residue N was invariably found in all R5 sequences at position 7, all

but one X4-using sequences had residue K at this position (Fig. 6A). Third, either E or D were found at position 25 in R5 sequences, however, non-E/D substitutions (S/A/G) were present in majority of X4-using sequences (Fig. 6A). Interestingly, none of the X4 sequences have positively charged amino acid R or K at V3 position 11 or 25, which are important for X4 usage in other HIV-1 subtypes (19-21). This implies different evolutionary pathway of coreceptor switching in CRF01_AE HIV-1. In genotypic prediction, all X4-using sequences had Geno2pheno FPR values below 2%, and V3 net charge no less than 5. In contrast, all R5 sequences had FPR values higher than 2%, and V3 net charge no more than 5. As expected, cluster 4 has an overall higher V3 net charge than cluster 5 (Fig. 6A). Notably, five sequences in cluster 4 with FPR below 5% were in fact R5-only phenotype (Fig. 6A). Therefore, using FPR 5% as the cutoff may significantly overestimate the prevalence of X4 phenotype in CRF01_AE.

Using the CCR5-V3 and CXCR4-V3 complex models (22, 23) , we also investigated the role of V3 positions 7, 8 and 25 in viral tropism from a structural perspective. In the model of CCR5-V3 complex, residue T8 in R5 V3 loop is surrounded by hydrophilic amino acids, suggesting that residue T8 is more favored by hydrophilic environment. However, in the model of CXCR4-V3 complex, residue I8 in the X4 V3 loop is surrounded by hydrophobic amino acids (Fig. 6B). This could explain why all X4 viruses have T to I/M substitutions at position 8 because the residue T8 may not well fit the hydrophobic environment within CXCR4's ligand binding pocket. In the CXCR4-V3 complex, V3 position 7 is surrounded by negatively charged residues, which favor the interaction with positively charged residues K7 in X4 sequences (Fig. 6B). Compared to

237 the corresponding region in the ligand binding pocket of CXCR4, the ligand binding
238 pocket in CCR5 around V3 position 25 contains more positively charged residues,
239 which are favored for interacting with the negatively charged residues D/E25. This
240 explained why all R5 viruses have D/E at V3 position 25. However, D/E may be less
241 favored in the less positively charged environment in the ligand binding pocket of
242 CXCR4 (Fig. 6B). Therefore, non-D/E substitutions as observed in X4-using sequences
243 would be required for efficient X4 binding (Fig. 6B). Taken together, genetic analysis in
244 combination with structural modeling showed that specific V3 substitutions at position 7,
245 8 and 25 may be required to achieve X4-using phenotype in the context of CRF01_AE
246 cluster 4 envelope.
247
248 **Discussion**
249 Since the initial introduction to human in the early 20th century, HIV-1 evolved
250 genetically and biologically with faster pace than other viruses, due to both the high
251 error-prone nature of its reverse transcriptase and the unusual transmission routes,
252 such as drug injections, heterosexual transmission and MSM activities. The Chinese
253 HIV-1 epidemic with multiple subtypes and their genetic clusters circulating at the same
254 time provides a unique opportunity to monitor virus evolution at both the genotype and
255 phenotype levels.
256
257 Past studies on HIV-1 evolution were mainly focused on virus genotype not phenotype,
258 because the later takes longer observation time and requires large well-matched
259 samples. In this study, we focused on various genetic clusters of CRF01_AE HIV-1 in

China. In the large cross sectional data from the NHMES, we discovered significant difference in CD4 count between people infected by CRF01_AE cluster 4 and 5, at both early and later stage of infection. The lower CD4 count in cluster 4 is directly associated with the higher prevalence of X4 virus based on genotypic prediction. This genotype-based observation in large population was further confirmed by well-matched genotyping and phenotyping data from a MSM sero-incidence cohort. We observed that among sero-convertors, those harboring X4-using viruses had rapid CD4 loss.

It is usually considered that X4 variants emerge during late stage of infection, with the development of immunodeficiency of the host (24). In subtype B HIV-1, around 50% of patients underwent coreceptor switch, usually after 5 years of infection, which correlated with rapid CD4 decline and faster progression to AIDS (16, 25-30). The exact time of coreceptor switch in different HIV-1 subtypes are not well understood. A recent study did not detect X4 variants in subtype B infected people who were within 2 years of infection (31). We demonstrated here that in certain HIV-1 subtypes or clusters, such as CRF01_AE cluster 4, coreceptor switch can occur much earlier than previously thought. This unusually fast speed of coreceptor switch is associated with the rapid CD4 loss in CRF01_AE cluster 4 compared to cluster 5 as well as other HIV-1 subtypes. Supported by genetic and structural analysis, the unique V3 signatures in cluster 4 (R13 and K32), which confer higher V3 net charge may be the major intrinsic determinant for such a high coreceptor switch tendency. Further efforts are required to understand whether the rapid emergence of X4 virus in vivo is essentially a random event due to accumulation of mutations, or also driven by host factor(s) like immune pressure. In addition,

envelope positions outside of the V3 loop like V1V2 could also play a part in this process. A better understanding of the driving force and evolutionary pathway for coreceptor switch in CRF01_AE cluster 4 may lead to strategies to block the early emergence of X4 virus during infection.

The genetic features of the X4 variants in CRF01_AE cluster 4 are also different from previously found in other HIV-1 subtypes. In particular, none of the X4 sequences in CRF01_AE cluster 4 have positively charged amino acid (R or K) at V3 positions 11 or 25, which are key amino acids for X4 usage observed in other subtypes (19-21). Instead, all of them lost the V3 glycan (the N301 glycan), and nearly all have residue K at V3 position 7. This highlights different evolutionary pathways for coreceptor switching in different HIV-1 subtypes. N301 glycan has previously been shown to be functionally critical for both coreceptor utilization and virus replication in subtype B (32, 33). With compensatory mutations or a high V3 net charge, loss of the N301 glycan leads to switch from R5 to X4 phenotype (32, 33). However, in the absence of compensatory mutations, loss of this glycan can abolish virus replication (32). Possibility due to this high fitness constraint, N301 glycan is highly conserved in naturally occurring subtype B sequences. A recent study found that in the Los Alamos HIV database, N301 glycan is present in as high as 99% of subtype B sequences with R5 phenotype and more than 80% of sequences with X4 phenotype. Differently, in CRF01_AE, 94% of R5 sequences and 39% of X4 sequences have the N301 glycan site (34) .This again indicates that loss of the N301 glycan is an important pathway for coreceptor switch in CRF01_AE HIV-1, but is not a primary route among subtype B viruses. It will be interesting to study in the

306 future whether N301 glycan has a lower fitness barrier in the context of CRF01_AE
307 cluster 4 envelope than in CRF01_AE cluster 5 and other HIV-1 subtypes.
308
309 Another interesting finding is the outgrowth of highly replication competent X4 variants
310 in primary viral isolates from CRF01_AE cluster 4. The low frequency of those X4
311 variants in plasma may not be explained by their low replication fitness, as they rapidly
312 outcompete other lineages in the in vitro setting. Instead, it is more likely due to the
313 compartmentalization of the R5 and X4 viruses in different cell subsets or tissues in vivo.
314 Due to the differential expressions of CCR5 and CXCR4 coreceptors in memory and
315 naïve CD4 T cell subsets (35, 36), R5 and X4 viruses are considered to preferentially
316 replicate in the memory and naïve CD4 subsets, respectively (37-41). It has been
317 shown that naïve CD4 T cells produce viruses at a lower propagation rate than memory
318 T cells (42, 43), possibility due to the relatively low division rate (44, 45). Therefore, in
319 vivo, those minor X4 lineages might compartmentalize in cell subsets or tissues that
320 shut the viruses less efficiently into the blood. Regardless of the mechanism, this
321 observation has its clinical implication: because conventional sequencing method may
322 not be sensitive enough to capture those minor X4 variants in plasma, deep sequencing
323 or phenotypic assay would be required to determine the existence of X4 variants in vivo,
324 especially when using treatment regimens including the CCR5 inhibitor.
325
326 In summary, we for the first time demonstrated that various phylogenetic clusters of the
327 same HIV-1 subtype can have disparate pathogenicity and cause different disease
328 outcomes, which filled the missing link between HIV-1 phylogenetic cluster and viral

329 phenotype. At the phenotype level, CRF01_AE cluster 4 evolved enhanced X4 tropism
330 and viral pathogenesis, while cluster 5 became more attenuated due to a decreased
331 potential of using CXCR4. Whether the process of "phenotype divergence" occurred as
332 a random founder event from the initial seeding clusters, or due to adaptation to
333 different hosts or transmission routs remains to be studied. Our study emphasizes the
334 importance of monitoring HIV-1 genetic drift and phenotype shift at the phylogenetic
335 cluster level in order to timely control the spread of more pathogenic viruses like
336 CRF01_AE cluster 4.
337
338
339 **Materials and Methods**
340 **Study participants**
341 The study participants were from the national HIV molecular epidemiology survey
342 (NHMES) and the Beijing Chaoyang District MSM cohort (the CYM cohort). Written
343 informed consent was obtained from all study participants. See SI Appendix,
344 Supplementary Text for details.
345
346 **Enzyme Immunoassay (EIA)**
347 In order to distinguish recent HIV-1 infections from long-term HIV-1 infections, the
348 Enzyme Immunoassay (EIA) was performed using the Maxim HIV-1 Limiting Antigen-
349 Avidity (LAg-Avidity) EIA kit (Maxim Biomedical). The experiment and data analysis
350 were performed according to manufacturer's instructions (Maxim Biomedical).
351

352 **Viral RNA extraction and cDNA synthesis**

353 Viral RNA was extracted from 200 µl of plasma sample using the QIAamp Viral RNA
354 Mini Kit (Qiagen). RNA was eluted into 50 µl of RNase free water. A total of 17 µl viral
355 RNA was used for cDNA synthesis using the SuperScript III reverse transcriptase
356 (Invitrogen) with the Oligo (dT) primer. The cDNA was immediately used for PCR
357 amplification.

358

359 **Library preparation for sequencing on Illumina MiSeq**

360 The Illumina MiSeq library was prepared using a nested PCR approach. The 8 nt
361 Illumina index and adaptors (P5 and P7) were added to both ends of the second round
362 PCR primers (SI Appendix, Fig. S2 and Table S4). The second round PCR products
363 were gel-purified to remove unspecific bands and primer dimers, and quantified by
364 qPCR using the KAPA SYBR FAST qPCR kit according to manufacturer's instructions
365 (KAPA Biosystems). See SI Appendix, Supplementary Text for details.

366

367 **Next generation sequencing and data analysis**

368 The pooled DNA library was sequenced on an Illumina MiSeq using the MiSeq Reagent
369 Kit v2 (500 cycles, Illumina) as previously described (46). Each pair of fastq reads in
370 files "read 1" and "read 2" were merged by the FLASH software (47). The merged fastq
371 files were then filtered based on data quality on the Galaxy server (48) using the
372 following parameter: no more than 10 bases with Q score lower than 30 in each read.
373 The filtered clean reads were then converted into fasta format. In each individual,
374 identical reads were collapsed into haplotypes after the primer regions were trimmed.

375 The frequency of each haplotype among the total clean reads was calculated. The most
376 frequent haplotype in each individual was used to infer the phylogenetic relationship
377 among all deep sequenced individuals.
378
379 **Genotypic prediction of co-receptor usage**
380 Genotypic prediction of co-receptor usage was performed using the Geno2pheno clonal
381 model (https://coreceptor.geno2pheno.org) (49). For the deep sequencing data, the
382 FPR (false positive rate, the probability of classifying an R5-virus falsely as X4) was
383 obtained for each V3 haplotype. To avoid the impact of potential PCR or sequencing
384 error on data analysis, only V3 haplotypes appeared three times or more in each
385 sample were used for analysis, while V3 singletons and those appeared only twice
386 where discarded. The frequency distribution of FPR value in each sample was obtained
387 by calculating the frequency of each V3 haplotype among the total reads analyzed.
388
389 **Primary virus isolation from PBMC**
390 To obtain primary virus isolates, cryopreserved PBMC from HIV-1 infected patients
391 were co-cultivated with stimulated normal PBMC from healthy donors. In brief, fresh
392 PBMC from healthy donors were stimulated for 3 days in RPMI1640 containing 10%
393 fetal bovine serum (FBS), interleukin 2 (IL-2) (32 U/ml; PeproTech), soluble anti-CD3
394 (0.2 μg/ml; eBioscience) and soluble anti-CD28 (0.2 μg/ml; eBioscience) as described
395 previously(50). After stimulation, cells were washed twice with RPMI1640 to remove the
396 residue simulating antibodies. A total of $10^7$ stimulated PBMC from healthy donor were
397 then mixed with $10^7$ of PBMC from an infected patient. The cell mixtures were then

398 depleted for CD8 T cells using the EasySep Human CD8 Positive Selection Kit
399 (Stemcell Technologies). The CD8-depleted cell mixture was cultured in a T25 flask with
400 RPMI1640 containing 10% fetal bovine serum (FBS) and 32 U/ml IL-2 (PeproTech) for
401 up to 4 weeks. Every 3 days, half volume of the culture supernatant was replaced with
402 fresh medium. Every 7 days, half of the entire culture (including cells) was removed,
403 and $5\times10^6$ of stimulated, CD8 T cell depleted PBMC from healthy donors were added.
404 The p24 concentration in the culture supernatant was measured every week. Majority of
405 cultures achieved peak p24 production around week 3. Cultures with p24 concentration
406 less than 2 ng/ml at week 4 were considered to be failed.
407
408 **Coreceptor tropism determination**
409 Coreceptor tropism of the primary viral isolates was determined using the
410 GHOST(3).CCR5 and GHOST(3).CXCR4 cell lines (51). Both GFP expression in the
411 Ghost cell lines and viral p24 production were used to determine the coreceptor usage.
412 See SI Appendix, Supplementary Text for details.
413
414 **Single genome amplification**
415 Single genome amplification (SGA) was performed as previously described (52). The
416 sequences were aligned using GeneCutter
417 (https://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html), followed by
418 the manual adjustment to obtain the optimal alignment. See SI Appendix,
419 Supplementary Text for details.
420

## Statistical analysis

All statistical analysis was performed using the Prism 7 (GraphPad Software). Statistical differences were determined using two-tailed Mann-Whitney test or Fisher's exact test as indicated in the figure legends. The exact *p* values were provided in the figures.

## Data availability

Newly generated nucleic acid sequences in the current study were deposited in GenBank with accession numbers MH672692-MH673032.

## Acknowledgements

We thank Dr. Cecilia Cheng-Mayer for comments on the manuscript. This work was supported by China National Major Project for Infectious Diseases Control and Prevention, and China Key Project of the State Key Laboratory of Infectious Diseases Control and Prevention.

## Author contributions

YS conceived and designed the study. HS, WO, YF, JZ, FL, JH and HP performed experiments. QT and BW designed and performed the structural analysis. HS, WO, YF, HX, LM, QT, BW and YS analyzed the data. HS, BW and YS wrote and edited the manuscript.

## Competing interests

The authors declare no competing interests.

444

445

446

447

448

449 **References**：


450 1. Taylor BS, Sobieszczyk ME, McCutchan FE, & Hammer SM (2008) The
451 challenge of HIV-1 subtype diversity. *N Engl J Med* 358(15):1590-1602.
452 2. Feng Y*, et al.* (2013) The rapidly expanding CRF01_AE epidemic in China is
453 driven by multiple lineages of HIV-1 viruses introduced in the 1990s. *AIDS*
454 27(11):1793-1802.
455 3. Kilmarx PH*, et al.* (2000) Disease progression and survival with human
456 immunodeficiency virus type 1 subtype E infection among female sex workers in
457 Thailand. *J Infect Dis* 181(5):1598-1606.
458 4. Costello C*, et al.* (2005) HIV-1 subtype E progression among northern Thai
459 couples: traditional and non-traditional predictors of survival. *Int J Epidemiol*
460 34(3):577-584.
461 5. Nelson KE, Costello C, Suriyanon V, Sennun S, & Duerr A (2007) Survival of
462 blood donors and their spouses with HIV-1 subtype E (CRF01 A_E) infection in
463 northern Thailand, 1992-2007. *AIDS* 21 Suppl 6:S47-54.
464 6. Rangsin R*, et al.* (2004) The natural history of HIV-1 infection in young Thai men
465 after seroconversion. *J Acquir Immune Defic Syndr* 36(1):622-629.
466 7. Li X*, et al.* (2014) Evidence that HIV-1 CRF01_AE is associated with low CD4+T
467 cell count and CXCR4 co-receptor usage in recently infected young men who
468 have sex with men (MSM) in Shanghai, China. *PLoS One* 9(2):e89462.
469 8. Li Y*, et al.* (2014) CRF01_AE subtype is associated with X4 tropism and fast HIV
470 progression in Chinese patients infected through sexual transmission. *AIDS*
471 28(4):521-530.
472 9. Chu M*, et al.* (2017) HIV-1 CRF01_AE strain is associated with faster HIV/AIDS
473 progression in Jiangsu Province, China. *Sci Rep* 7(1):1570.
474 10. To SW*, et al.* (2013) Determination of the high prevalence of Dual/Mixed- or X4-
475 tropism among HIV type 1 CRF01_AE in Hong Kong by genotyping and
476 phenotyping methods. *AIDS Res Hum Retroviruses* 29(8):1123-1128.

11. Ng KY, *et al.* (2013) High prevalence of CXCR4 usage among treatment-naive CRF01_AE and CRF51_01B-infected HIV-1 subjects in Singapore. *BMC Infect Dis* 13:90.
12. Li X, *et al.* (2017) Tracing the epidemic history of HIV-1 CRF01_AE clusters using near-complete genome sequences. *Sci Rep* 7(1):4024.
13. Melby T, *et al.* (2006) HIV-1 coreceptor use in triple-class treatment-experienced patients: baseline prevalence, correlates, and relationship to enfuvirtide response. *J Infect Dis* 194(2):238-246.
14. Moyle GJ, *et al.* (2005) Epidemiology and predictive factors for chemokine receptor use in HIV-1 infection. *J Infect Dis* 191(6):866-872.
15. Wilkin TJ, *et al.* (2007) HIV type 1 chemokine coreceptor use among antiretroviral-experienced patients screened for a clinical trial of a CCR5 inhibitor: AIDS Clinical Trial Group A5211. *Clin Infect Dis* 44(4):591-595.
16. Koot M, *et al.* (1993) Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann Intern Med* 118(9):681-688.
17. Brumme ZL, *et al.* (2005) Molecular and clinical epidemiology of CXCR4-using HIV-1 in a large population of antiretroviral-naive individuals. *J Infect Dis* 192(3):466-474.
18. Tan Q, *et al.* (2013) Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. *Science* 341(6152):1387-1390.
19. Huang W, *et al.* (2011) Mutational pathways and genetic barriers to CXCR4-mediated entry by human immunodeficiency virus type 1. *Virology* 409(2):308-318.
20. Berger EA, Murphy PM, & Farber JM (1999) Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annu Rev Immunol* 17:657-700.
21. Ping LH, *et al.* (1999) Characterization of V3 sequence heterogeneity in subtype C human immunodeficiency virus type 1 isolates from Malawi: underrepresentation of X4 variants. *J Virol* 73(8):6271-6281.
22. Tamamis P & Floudas CA (2014) Molecular recognition of CCR5 by an HIV-1 gp120 V3 loop. *PLoS One* 9(4):e95767.
23. Tamamis P & Floudas CA (2013) Molecular recognition of CXCR4 by a dual tropic HIV-1 gp120 V3 loop. *Biophys J* 105(6):1502-1514.
24. Swanstrom R & Coffin J (2012) HIV-1 pathogenesis: the virus. *Cold Spring Harb Perspect Med* 2(12):a007443.
25. Schuitemaker H, *et al.* (1992) Biological phenotype of human immunodeficiency virus type 1 clones at different stages of infection: progression of disease is associated with a shift from monocytotropic to T-cell-tropic virus population. *J Virol* 66(3):1354-1360.

26. Connor RI, Sheridan KE, Ceradini D, Choe S, & Landau NR (1997) Change in coreceptor use correlates with disease progression in HIV-1--infected individuals. *J Exp Med* 185(4):621-628.
27. Koot M*, et al.* (1999) Conversion rate towards a syncytium-inducing (SI) phenotype during different stages of human immunodeficiency virus type 1 infection and prognostic value of SI phenotype for survival after AIDS diagnosis. *J Infect Dis* 179(1):254-258.
28. Verhofstede C, Nijhuis M, & Vandekerckhove L (2012) Correlation of coreceptor usage and disease progression. *Curr Opin HIV AIDS* 7(5):432-439.
29. Moore JP, Kitchen SG, Pugach P, & Zack JA (2004) The CCR5 and CXCR4 coreceptors--central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS Res Hum Retroviruses* 20(1):111-126.
30. Richman DD & Bozzette SA (1994) The impact of the syncytium-inducing phenotype of human immunodeficiency virus on disease progression. *J Infect Dis* 169(5):968-974.
31. Zhou S, Bednar MM, Sturdevant CB, Hauser BM, & Swanstrom R (2016) Deep Sequencing of the HIV-1 env Gene Reveals Discrete X4 Lineages and Linkage Disequilibrium between X4 and R5 Viruses in the V1/V2 and V3 Variable Regions. *J Virol* 90(16):7142-7158.
32. Ogert RA*, et al.* (2001) N-linked glycosylation sites adjacent to and within the V1/V2 and the V3 loops of dualtropic human immunodeficiency virus type 1 isolate DH12 gp120 affect coreceptor usage and cellular tropism. *J Virol* 75(13):5998-6006.
33. Pollakis G*, et al.* (2001) N-linked glycosylation of the HIV type-1 gp120 envelope glycoprotein as a major determinant of CCR5 and CXCR4 coreceptor utilization. *J Biol Chem* 276(16):13433-13441.
34. Joshi A*, et al.* (2017) HIV-1 subtype CRF01_AE and B differ in utilization of low levels of CCR5, Maraviroc susceptibility and potential N-glycosylation sites. *Virology* 512:222-233.
35. Bleul CC, Wu L, Hoxie JA, Springer TA, & Mackay CR (1997) The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proc Natl Acad Sci U S A* 94(5):1925-1930.
36. Lee B, Sharron M, Montaner LJ, Weissman D, & Doms RW (1999) Quantification of CD4, CCR5, and CXCR4 levels on lymphocyte subsets, dendritic cells, and differentially conditioned monocyte-derived macrophages. *Proc Natl Acad Sci U S A* 96(9):5215-5220.
37. Blaak H*, et al.* (2000) In vivo HIV-1 infection of CD45RA(+)CD4(+) T cells is established primarily by syncytium-inducing variants and correlates with the rate of CD4(+) T cell decline. *Proc Natl Acad Sci U S A* 97(3):1269-1274.

38. van Rij RP, *et al.* (2000) Differential coreceptor expression allows for independent evolution of non-syncytium-inducing and syncytium-inducing HIV-1. *J Clin Invest* 106(12):1569.
39. Nishimura Y, *et al.* (2005) Resting naive CD4+ T cells are massively infected and eliminated by X4-tropic simian-human immunodeficiency viruses in macaques. *Proc Natl Acad Sci U S A* 102(22):8000-8005.
40. Ribeiro RM, Hazenberg MD, Perelson AS, & Davenport MP (2006) Naive and memory cell turnover as drivers of CCR5-to-CXCR4 tropism switch in human immunodeficiency virus type 1: implications for therapy. *J Virol* 80(2):802-809.
41. Council OD & Joseph SB (2018) Evolution of Host Target Cell Specificity During HIV-1 Infection. *Curr HIV Res* 16(1):13-20.
42. Eckstein DA, *et al.* (2001) HIV-1 actively replicates in naive CD4(+) T cells residing within human lymphoid tissues. *Immunity* 15(4):671-682.
43. Zhang Z, *et al.* (1999) Sexual transmission and propagation of SIV and HIV in resting and activated CD4+ T cells. *Science* 286(5443):1353-1357.
44. McLean AR & Michie CA (1995) In vivo estimates of division and death rates of human T lymphocytes. *Proc Natl Acad Sci U S A* 92(9):3707-3711.
45. McCune JM, *et al.* (2000) Factors influencing T-cell turnover in HIV-1-seropositive patients. *J Clin Invest* 105(5):R1-8.
46. Williams WB, *et al.* (2015) HIV-1 VACCINES. Diversion of HIV-1 vaccine-induced immunity by gp41-microbiota cross-reactive antibodies. *Science* 349(6249):aab1253.
47. Magoc T & Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957-2963.
48. Blankenberg D, *et al.* (2010) Manipulation of FASTQ data with Galaxy. *Bioinformatics* 26(14):1783-1785.
49. Lengauer T, Sander O, Sierra S, Thielen A, & Kaiser R (2007) Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol* 25(12):1407-1410.
50. Song H, *et al.* (2012) Impact of immune escape mutations on HIV-1 fitness in the context of the cognate transmitted/founder genome. *Retrovirology* 9:89.
51. Vodros D, *et al.* (2001) Quantitative evaluation of HIV-1 coreceptor use in the GHOST3 cell assay. *Virology* 291(1):1-11.
52. Keele BF, *et al.* (2008) Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A* 105(21):7552-7557.

## Figure legends

**Figure 1.** Comparison of CD4 T cell count and prevalence of X4 virus in different HIV-1 subtypes and CRF01_AE clusters. (A) Comparison of CD4 T cell count between individuals infected by CRF01_AE (n=1118), CRF07_BC (n=633) and subtype B (n=123) from the national HIV molecular epidemiology survey. (B-D) Significantly lower CD4 T cell count among individuals infected by CRF01_AE cluster 4 (n=308) than those infected by cluster 5 (n=273) regardless of the stage of infection (B), in the recent infection group (C), and in long-term infection group (D). The small figure in panel B shows the percentage of individuals with CD4 below 200. In each panel, the vertical line, box and whisker represents the median, upper and lower quantiles, and the 5-95 percentile, respectively. The statistical difference in CD4 count between different groups was calculated using two-tailed Mann-Whitney U test. The percentage of subjects with CD4 below 200 was compared using two-tailed Fisher's exact test. (E) Prevalence of predicted X4 viruses in CRF01_AE cluster 4 and cluster 5, and in the lower (<200) and higher (>200) CD4 groups in CRF01_AE cluster 4. The statistical difference was determined using two-tailed Fisher's exact test.

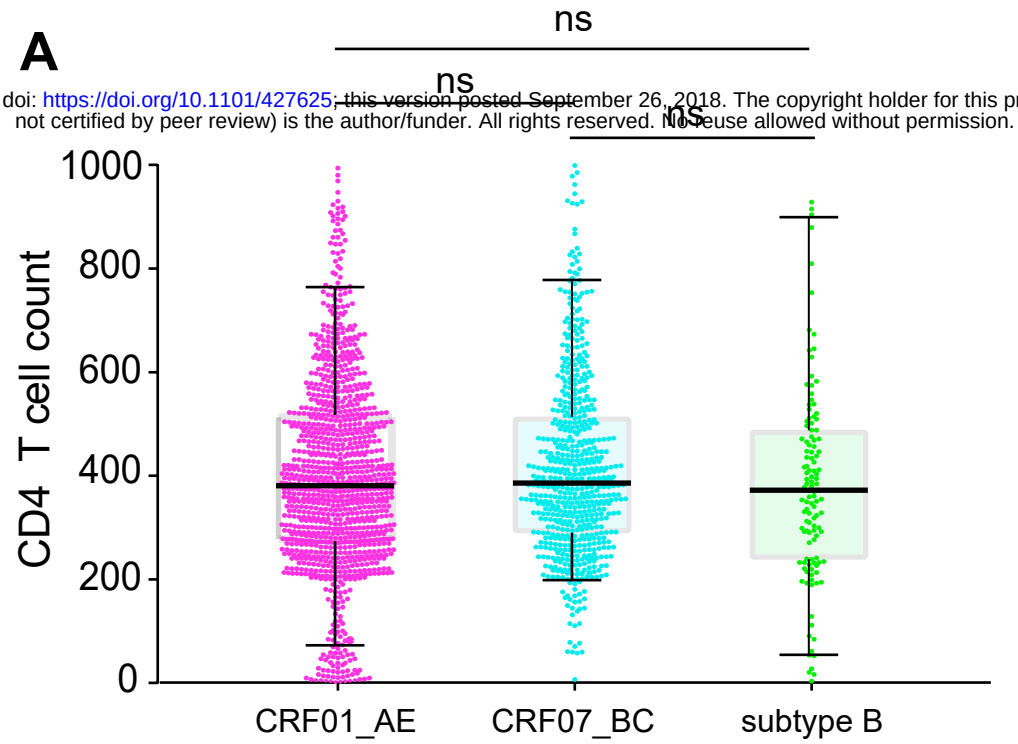
618 **Figure 2.** Higher frequency of predicted X4-using variants in CRF01_AE cluster 4
619 identified by deep sequencing. (A) Phylogenetic relationship of 60 deep sequenced
620 individuals from the CYM cohort. In each individual, the most frequent haplotype among
621 the deep sequencing reads was used for phylogenetic inference. The Neighbor-joining
622 (NJ) tree was constructed using the Kimura 2-parameter evolutionary model with 1000
623 bootstrap replications. In the tree, the branches for CRF01_AE cluster 4 (n=22), cluster
624 5 (n=11), CRF07_BC (n=19) and subtype B (n=8) were color coded. (B) Heatmap
625 showing the frequency distribution of Geno2pheno FPR value among the deep
626 sequencing reads in each individual. The samples in the phylogenetic tree and in the
627 heatmap were matched.
628
629 **Figure 3.** Phenotypic characterization of primary CRF01_AE viral isolates and the
630 association between coreceptor tropism and CD4 count. (A) Coreceptor usage
631 phenotype of 24 primary viral isolates from CRF01_AE cluster 4 (n=16) and cluster 5
632 (n=8) and the corresponding CD4 counts in each individual. (B) Determination of
633 coreceptor tropism using GHOST.CCR5 and GHOST.CXCR4 cell lines. The GFP
634 expression induced by one representative R5-only isolate and one X4-using isolate
635 were shown. (C) Syncytium formation at day 12 post PBMC co-cultivation in the culture
636 of CYM248, an isolate using CXCR4 exclusively. (D) Significantly lower CD4 count in
637 individuals harboring X4-using viruses in CRF01_AE cluster 4. The black vertical line
638 represents the median CD4 count. The statistical difference was calculated using two-
639 tailed Mann-Whitney U test.
640

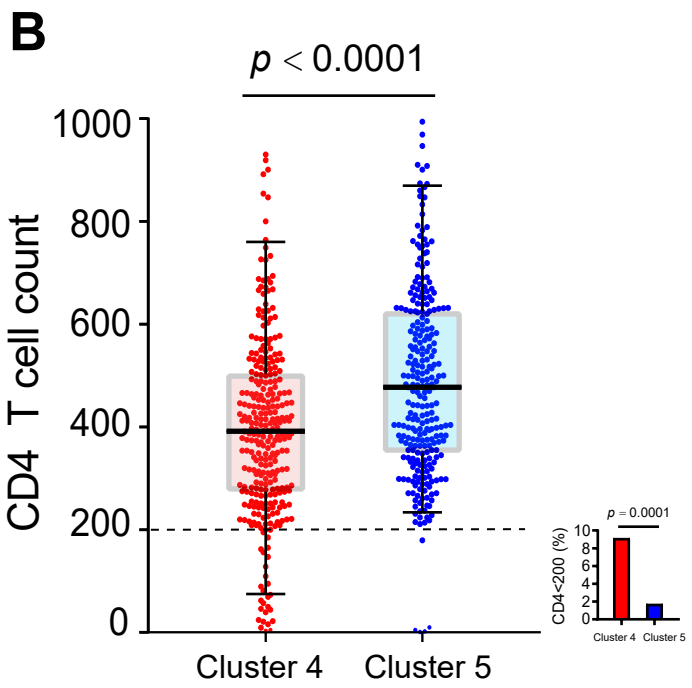
641 **Figure 4.** Genetic determinants and structural basis of the higher X4-using tendency in
642 CRF01_AE cluster 4. (A) The frequency of each V3 amino acid was determined with a
643 total of 385 available sequences from CRF01_AE cluster 4, 328 available sequences
644 from CRF01_AE cluster 5, and 34 sequences from Thailand downloaded from the Los
645 Alamos HIV sequence database (before the year 2000). The plots were generated
646 using the WebLogo tool (https://weblogo.berkeley.edu/). (B) Pie charts showing the
647 prevalence of positively charged amino acid K13 and R32 in CRF01_AE cluster 4 and
648 cluster 5. (C) Structural analysis for V3 position 13 and 32 in binding of the CCR5 and
649 CXCR4 coreceptors using the V3-docking model.
650
651 **Figure 5.** Genetic composition of the PBMC viral isolates. (A-B) SGA-derived gp160
652 sequences from the R5 isolate CYM179 (A) and the dual tropic isolate CYM176 (B)
653 were shown using highlighter plot. In each sample, one sequence with the predominant
654 V3 lineage was used as the master sequence. The synonymous and non-synonymous
655 substitutions compared to the master sequence were shown in green and red,
656 respectively. The corresponding V3 alignment was shown on the right, and different V3
657 lineages were color-coded. The circle plots show the proportion of each V3 lineage in
658 the PBMC viral isolate as detected by SGA and in the plasma as detected by deep
659 sequencing. In CYM176, the red arrows indicate the phenotypically confirmed X4
660 lineage.
661
662 **Figure 6.** Genetic characteristics of phenotypically confirmed X4 sequences and
663 structural modeling for coreceptor binding. (A) V3 amino acids alignment of SGA-

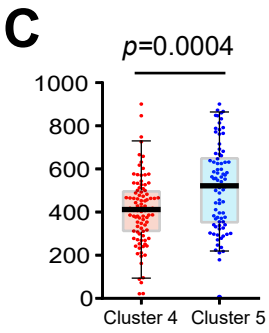
664 derived sequences from the phenotype-confirmed primary viral isolates. Sequences
665 shown in red were phenotypically confirmed X4-using sequences (that is, sequences
666 sieved out from the GHOST.CXCR4 culture by SGA). Sequences shown in blue were
667 the predominant V3 forms in the R5 isolates. Key V3 positions associated with X4-using
668 phenotype were shaded in light blue. (B) Structural modeling for V3 positions 7, 8 and
669 25 in binding of coreceptors CCR5 and CXCR4 using the V3-docking model.

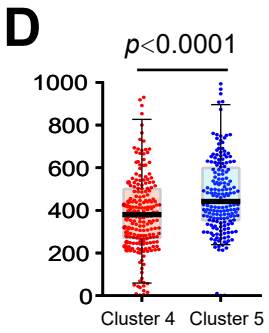
# Figure 1

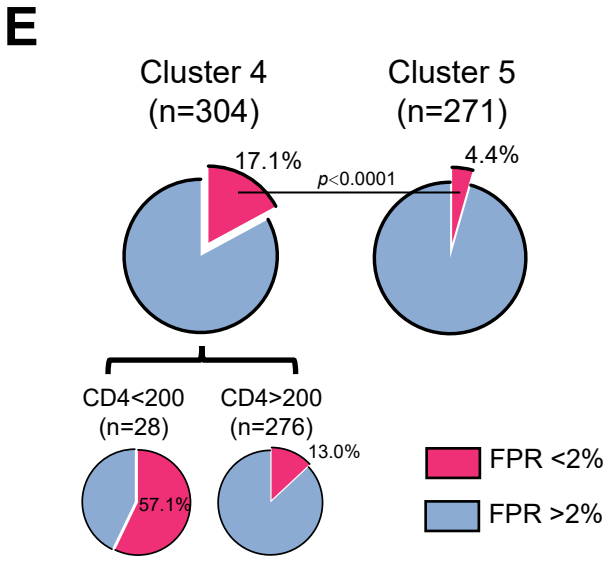




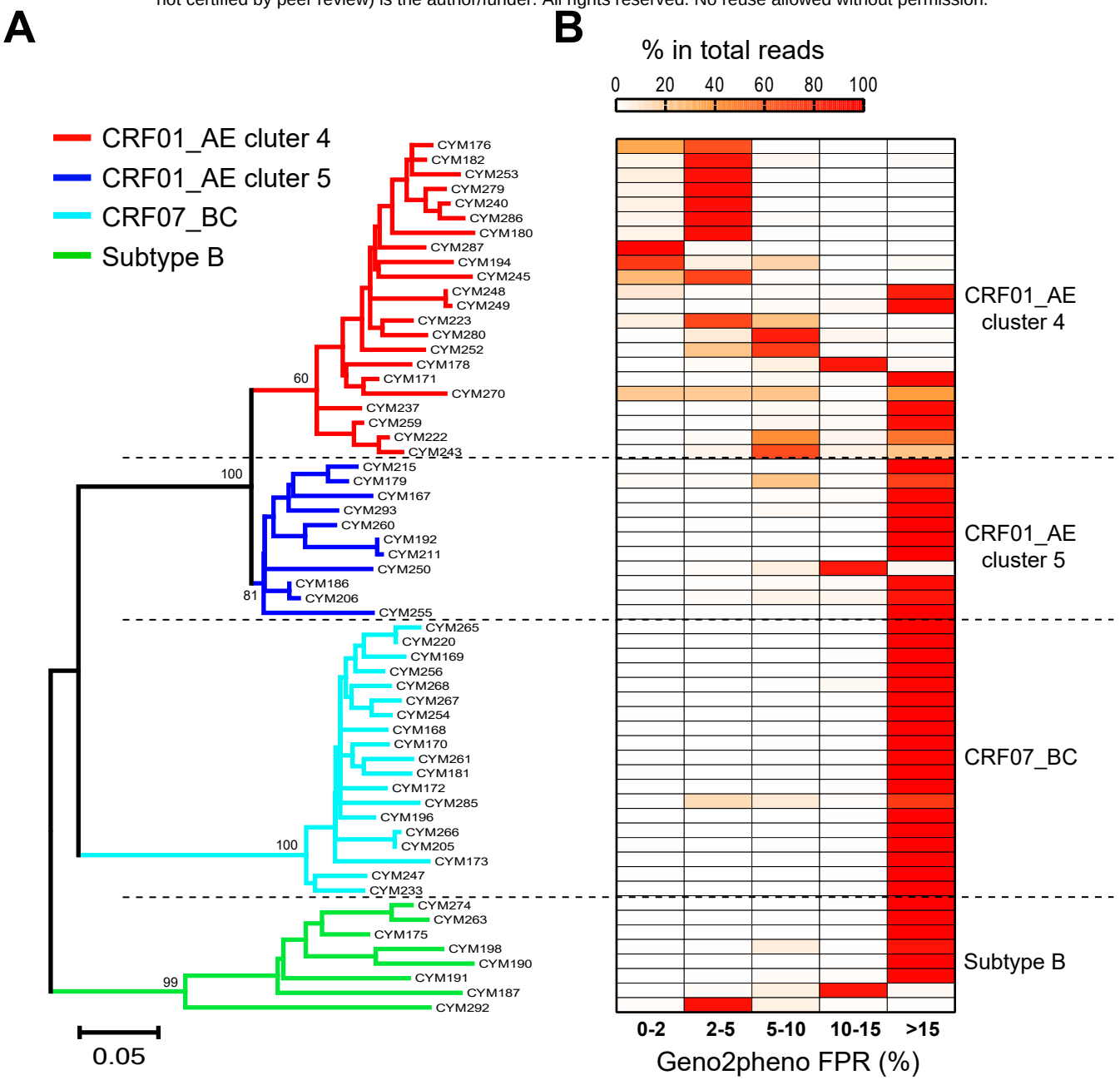



# Figure 2





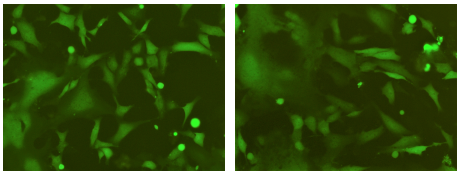
# Figure 3



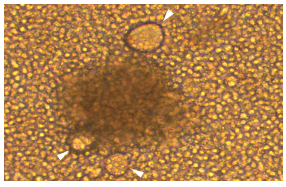
**A**

| Subjects | Phenotype | CD4 count | CD4 average | |
|---|---|---|---|---|
| CYM248 | X4 | 10 | | |
| CYM245 | R5/X4 dual | 136 | | |
| CYM176 | R5/X4 dual | 178 | 141 | **Cluster 4 X4-using** |
| CYM270 | R5/X4 dual | 18 | | |
| CYM194 | R5/X4 dual | 364 | | |
| CYM240 | R5 | 384 | | |
| CYM286 | R5 | 593 | | |
| CYM279 | R5 | 402 | | |
| CYM182 | R5 | 425 | | |
| CYM223 | R5 | 758 | | |
| CYM249 | R5 | 409 | 440 | **Cluster 4 R5-only** |
| CYM252 | R5 | 467 | | |
| CYM178 | R5 | 305 | | |
| CYM280 | R5 | 347 | | |
| CYM171 | R5 | 201 | | |
| CYM243 | R5 | 548 | | |
| CYM250 | R5 | 716 | | |
| CYM293 | R5 | 615 | | |
| CYM192 | R5 | 348 | | |
| CYM167 | R5 | 527 | | |
| CYM215 | R5 | 416 | 441 | **Cluster 5 R5-only** |
| CYM179 | R5 | 197 | | |
| CYM255 | R5 | 325 | | |
| CYM186 | R5 | 381 | | |

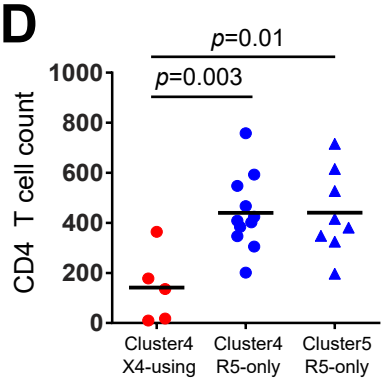
**B**




**C**



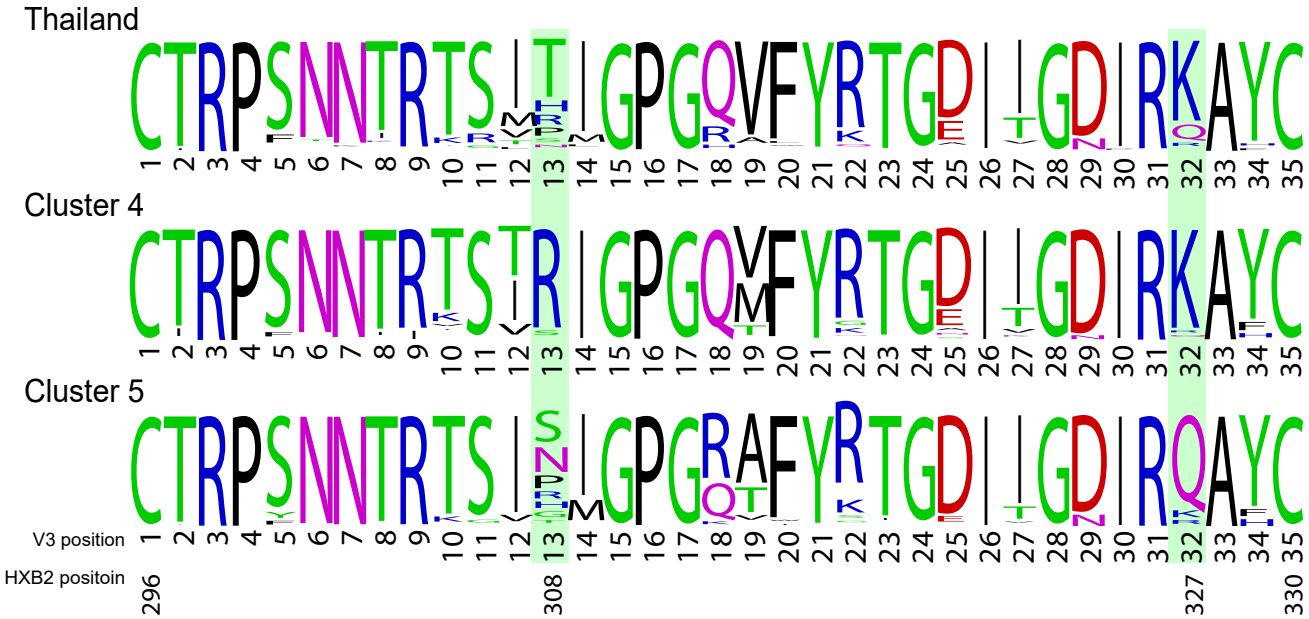

**D**



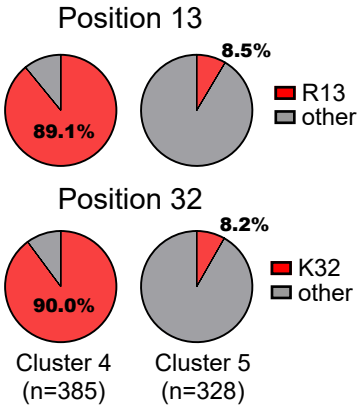
# Figure 4



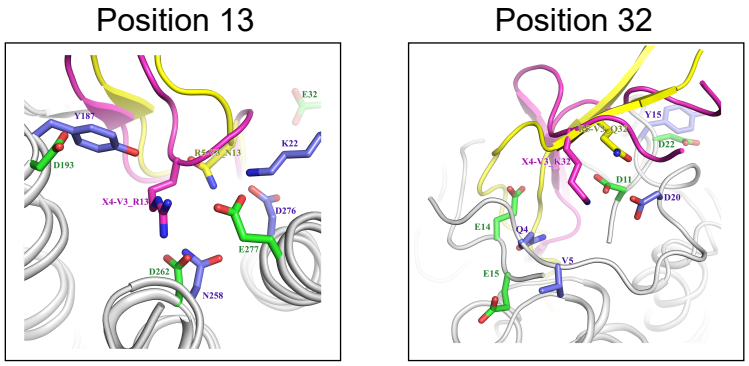
**A**




**B**

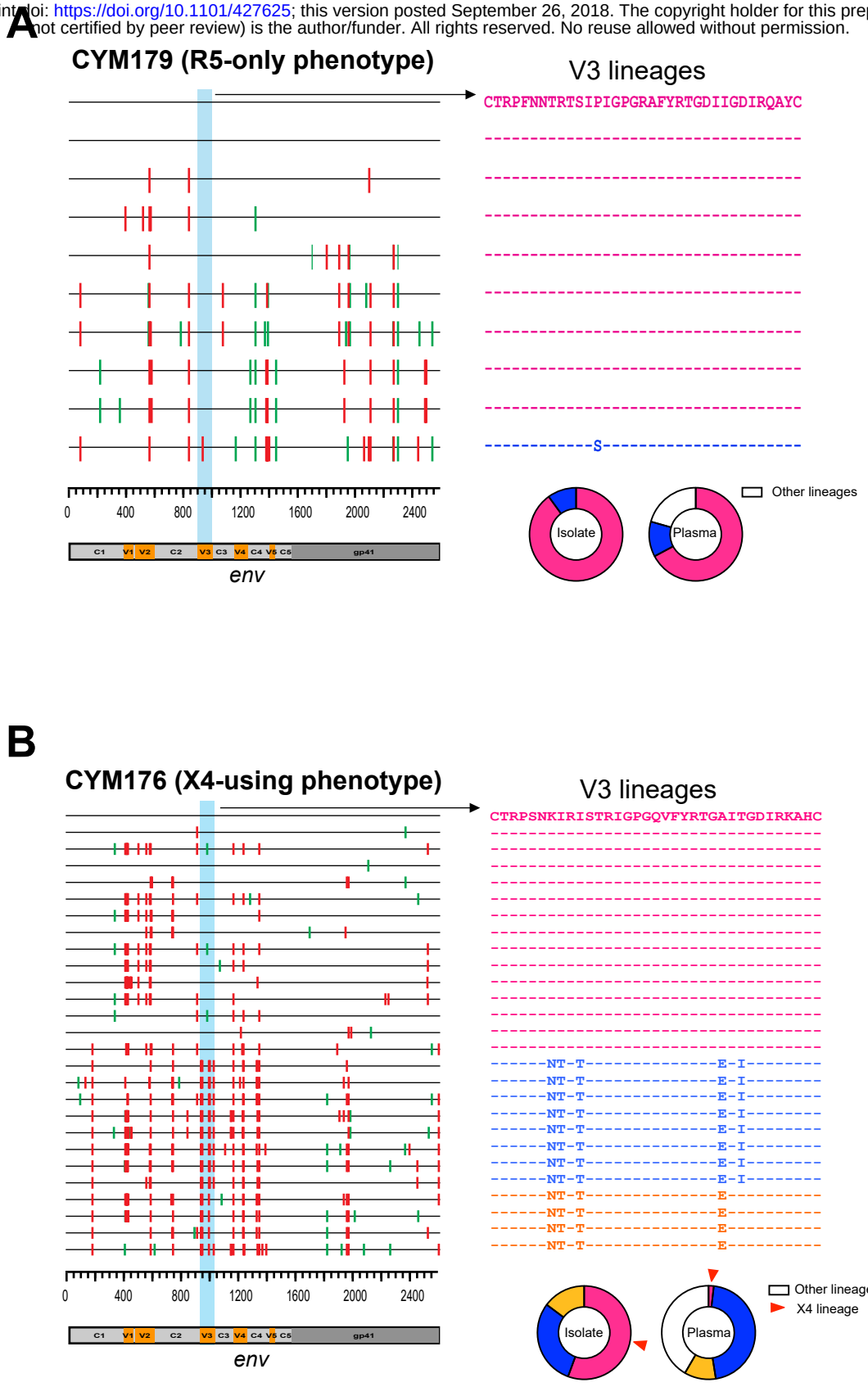



**C**



# Figure 5

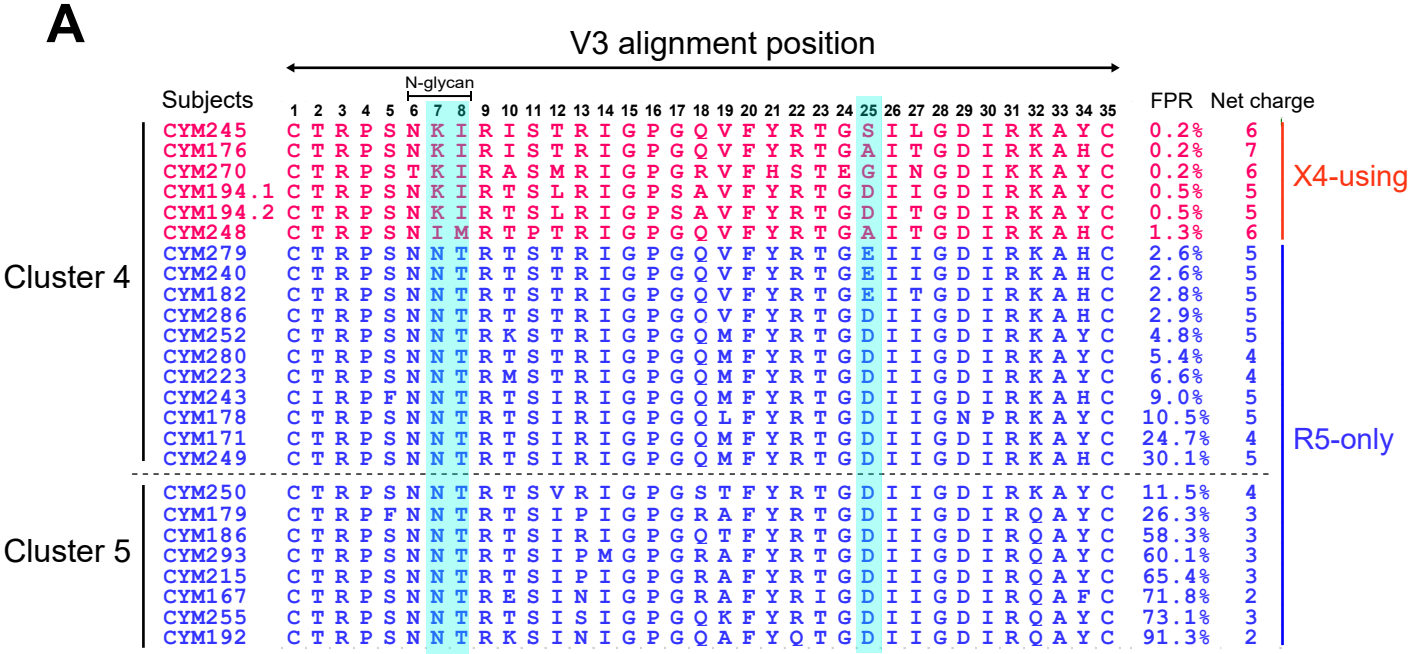




# Figure 6



**A**




**B**