Major article

A systematic molecular epidemiology screen reveals numerous HIV-1 superinfections in the Swiss HIV Cohort Study

Running title: HIV-1 Superinfection screen in the SHCS

Sandra E. Chaudron[1,2], Christine Leemann[1,2], Katharina Kusejko[1,2], Huyen Nguyen[1,2], Nadine Tschumi[1,3], Alex Marzel[1,4], Michael Huber[2], Jürg Böni[2], Matthieu Perreau[5], Thomas Klimkait[6], Sabine Yerly[7], Alban Ramette[8], Hans H. Hirsch[9,10], Andri Rauch[11], Alexandra Calmy[7,12], Pietro Vernazza[13], Enos Bernasconi[14], Matthias Cavassini[15], Karin J. Metzner[1,2,¶], Roger D. Kouyos[1,2,¶], Huldrych F. Günthard[1,2,¶], and the Swiss HIV Cohort Study[^]

1. Department of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Zurich, Switzerland.
2. Institute of Medical Virology, University of Zurich, Zurich, Switzerland.
3. Swiss Tropical and Public Health Institute, Basel, Switzerland.
4. Schulthess Klinik, Zurich, Switzerland.
5. Service of Immunology and Allergy, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland.
6. Department of Biomedicine, University of Basel, Basel, Switzerland.
7. Laboratory of Virology, Geneva University Hospitals, Geneva, Switzerland.
8. Institute for Infectious Diseases, University of Bern, Bern, Switzerland.
9. Division of Infectious Diseases and Hospital Epidemiology, University Hospital Basel, Basel, Switzerland.
10. Clinical Virology, Laboratory Medicine, University Hospital Basel, Basel, Switzerland.
11. Department of Infectious Diseases, Bern University Hospital, University of Bern, Bern, Switzerland.
12. Division of Infectious Diseases and Faculty of medicine, University of Geneva, Geneva, Switzerland.
13. Clinic for Infectiology and Hospital Hygiene, Cantonal Hospital St Gallen, St Gallen, Switzerland.
14. Division of Infectious Diseases, Regional Hospital Lugano, Lugano, Switzerland.
15. Service for Infectious Diseases, Lausanne University Hospital, Lausanne, Switzerland.

[¶]**These authors contributed equally to this work:** Karin J. Metzner, Roger D. Kouyos and Huldrych F. Günthard

[^]**The members of the Swiss HIV Cohort Study are listed in the Acknowledgments**

Abstract

Background: Studying HIV-1 superinfection is important to understand virus transmission, disease progression, and vaccine design. But detection remains challenging, with low sampling frequencies, and insufficient longitudinal samples.

Methods: Using the Swiss HIV Cohort Study(SHCS), we developed a molecular epidemiology screening for superinfections. A phylogeny built from 22,243 HIV-1 partial-polymerase sequences was used to identify potential superinfections among 4,575 SHCS participants with longitudinal sequences. A subset of potential superinfections was tested by near-full-length viral genome sequencing (NFVGS) of bio-banked plasma samples.

Results: Based on phylogenetic and distance criteria, 325 potential HIV-1 superinfections were identified and categorised by their likelihood of being detected as superinfection due to sample misidentification. NFVGS was performed for 128 potential superinfections: Of these, fifty-two were confirmed by NFVGS, 15 were not confirmed, and for 61 sampling did not allow for confirming or rejecting superinfection because the sequenced samples did not include the relevant time points causing the superinfection signal in the original screen. Thus, NFVGS could support 52/67 adequately sampled potential superinfections.

Conclusions: This cohort-based molecular approach identified, to our knowledge, the largest population of confirmed superinfections, showing that, while rare with a prevalence of 1-7%, superinfections are not negligible events.

Key words: HIV-1 Superinfection, Molecular Epidemiology Screening, Phylogenetics.

Introduction

The human immunodeficiency virus type 1(HIV-1) remains a global health challenge with 1.7 million new infections in 2019 [1] despite available combination antiretroviral therapy(ART) which upon success [2], can reduce HIV transmission to almost zero [3]. In 2019, Switzerland had ~16,600 PLWH, with 425 diagnosed in 2018 [4]. Although HIV-1 is typically transmitted from infected to non-infected individuals, HIV-1 superinfection can also occur, i.e. individuals with an already established HIV-1 infection acquiring another HIV-1 strain [5,6]. Since the first reported cases of HIV-1 superinfection in 2002, several others were identified by a range of different approaches [7-9]. These superinfections were primarily identified in association with an unexpected increase in VL or after treatment failure. Typically, cases are molecularly confirmed either by strain-specific polymerase chain reactions(PCR) [6], heteroduplex mobility assays [10], and genetic screening assays [11] to identify different HIV-1 subtypes, and calculate viral sequence ambiguity scores [12] or by reconstruction of sequence-based viral phylogenies obtained from longitudinally sampled sequences [13].

However, there are still unknowns such as, the factors contributing to its acquisition and its incidence in the population [14,15]. The immunological responses associated with HIV-1 superinfection also lack understanding [16,17]. Finally, HIV-1 superinfections leading to rapid or, preventing disease progression is debated [18,19], These uncertainties remain because a systematic assessment of HIV-1 superinfection is challenging. Firstly, HIV-1 superinfection is difficult to distinguish from coinfection due to intra-subtype viral similarity. Accordingly, the most reported HIV-1 superinfections involve distinct HIV-1 subtypes [20].

Secondly, the event is deemed rare, and can be transient, thus missed, if the sampling is inappropriate and the superinfecting strain does not outcompete the established strain [21], making it challenging to assess the within-host viral population dynamics [22]. Finally, sampling size, frequencies and timing are critical to detect HIV-1 superinfections. A large screen of a prospective seroincident cohort in Mombasa revealed 21 HIV-1 superinfection cases [23]. Another large retrospective screening on 4,425 individuals could only confirm 2 of the 14 potential cases re-sequenced [24]. Overall, each studies on HIV-1 superinfection identified at most a dozen cases. In general, longitudinal samples, population sequences or next generation sequences (NGS) linked to dense qualitative epidemiological data from HIV-infected individuals are often not available for systematic screens of large populations and identification of HIV-1 superinfection in large numbers.

We thus established a molecular epidemiology approach to systematically screen for HIV-1 superinfection in cohort studies. We utilised the Swiss HIV Cohort Study(SHCS), a well-characterised cohort of over 20,000 HIV-1 infected individuals, with good representativeness of the Swiss HIV-1 epidemic [25]. This method to identify HIV-1 superinfection cases, was developed with the viral phylogeny of longitudinally sampled HIV-1 *pol* sequences from genotypic HIV-1 drug resistance testing. The process was then validated with HIV-1 near-full-length viral genome sequencing(NFVGS) from longitudinal samples of the potential cases identified within the SHCS.

Methods

The Swiss HIV Cohort Study

The SHCS is a Swiss prospective multi-centre, longitudinal study established in 1988 [25], with 20,845 PLWH enrolled by end 2019. It covers ≥53% of the cumulative number

of HIV-infected individuals, ~75% of all HIV-positive individuals on ART and 72% of individuals diagnosed with acquired immune deficiency syndrome(AIDS). All participants provided written informed consent and the SHCS was approved by the participating institutions' ethics committees. At Semi-annual follow-ups, socio-demographic, behavioural, clinical, laboratory data are obtained, and biological samples are stored in the SHCS biobank. Since 2002, routine HIV genotypic resistance tests(GRT) are performed on baseline plasma samples and for treatment failure. Also, >11,000 GRT were done retrospectively from biobanked plasma samples obtained before 2002 [26]. Overall, ~60% of enrolled individuals have ≥1 HIV-1 partial polymerase(*pol*) gene consensus sequence in the SHCS drug resistance database(DRDB). The DRDB HIV-1 partial *pol* sequences contain the protease(PR: nt. 2,253-2,550), and the reverse transcriptase(RT: nt. 2,550-3,870, at minimum codons 28-225). At the start of this study in 2017, the database contained demographic information on 20,089 individuals.

Data availability **(**See Appendix S1).

Phylogeny reconstruction

We built a phylogeny using an in-house pipeline. Briefly, for the initial screen, all SHCS sequences quality checked i.e. filtered for length (PR ≥250bp, RT ≥500bp) and duplicates. They were aligned (Appendix S2) to HIV-1 HXB2 *pol* gene (Genbank accession number: K03455.1, nt. 2,253-3,870) and known drug resistance mutations from the Stanford HIV DRDB and the International Antiviral Society-USA were removed from the alignment. The sequences were trimmed and the phylogeny was reconstructed with two different tools (Appendix S2). For the validation analyses we used the same process on different genomic area of interest in the near-full-length HIV-1 consensuses.

Identification of potential HIV-1 superinfections

We used two criteria on SHCS participants with ≥2 longitudinal (i.e. different time points) HIV-1 partial *pol* sequences. First, the within-individual maximal patristic distance, obtained by calculating the pairwise patristic distance from the individual's longitudinal sequences (Appendix S2 for R functions). Second, the cluster size, i.e., the number of sequences in the smallest subtree containing all the longitudinal sequences from an individual's most recent common ancestor(MRCA). For superinfection, this cluster in non-monophyletic containing all sequences of the focal patients as well as other SHCS sequences. Thus, to identify potential HIV-1 superinfection, we chose the thresholds of ≥0.05 and ≥20 for within-individual maximum patristic genetic distance and per individual smallest cluster size. Respectively, we tested the sensitivity of these combined thresholds by varying them from ≥0.01 to ≥0.1 (patristic distance) and from ≥5 to ≥250 (smallest cluster size).

We define the estimated time of HIV-1 superinfection for each individual, as the time point with the highest maximal patristic distance. This time point is the most distant to the other time points, thus provides the strongest evidence for HIV-1 superinfection.

Categorisation of the potential HIV-1 superinfections

To assess the likelihood of HIV-1 superinfection regarding sample misidentification, we categorised the topology of the smallest subtree of SHCS participants with ≥2 longitudinal sequences (See Results, Appendix S2 for R functions used). In individuals' smallest tree, each time point was alternatively removed from the phylogeny, a new MRCA and smallest cluster size from it was calculated for the remaining time points. Category 1 cases only have two longitudinal HIV-1 partial *pol* sequences, distant in the phylogeny, and so no new MRCA was found by dropping one or the other sequence.

Category 2 cases have >2 longitudinal sequences. One of the new smallest cluster sizes was smaller than the total number of longitudinal sequences in the phylogeny, for the focal individual. Meaning, all time points except one, cluster together in the phylogeny. Category 3 cases also have >2 longitudinal sequences but a different tree topology. Every new smallest cluster size remained larger than the total number of an individual's longitudinal sequences used in the phylogeny. Meaning, several sequences from the same individual cluster together away from the others in the phylogeny.

Retrospective sequencing of near-full-length HIV-1 genomes

To validate potential HIV-1 superinfections, we next-generation sequenced(NGS) near-full-length HIV-1 genome. Given the limitations of NGS at low virus loads(VL), we only used plasma samples with VL ≥5,000 copies/ml. HIV-1 RNA isolation, cDNA synthesis and PCR amplification were performed from individuals' longitudinal plasma samples, with four overlapping fragments across HIV-1 genome amplified, combined, and NGS with Illumina Mi-Seq (detailed in Appendix S3).

Bioinformatic analysis

We analysed the NGS reads for each time point of a focal individual using an in-house bioinformatic pipeline. Briefly, the NGS reads were trimmed, mapped to HIV-1 HXB2 and the near-full-length viral consensus was reconstructed (Appendix S2). The coverage along the genome, was assessed. The read mapping was repeated using as reference, the new sample' viral consensus, until no further improvement in the coverage. The consensus before the last mapping was used to build the final viral consensus. The consensuses with 2,500 HIV-1 full-length background sequences randomly selected from the Los Alamos HIV Databases, matching the viral subtypes prevalence in the

SHCS (Appendix S4), were used to validate superinfections with phylogeny and our selection criteria as described above.

The regions analysed were HIV-1 full-length, *pol*, *gag* and *env* (Appendix S2). We excluded samples if the amplicon spanning the genomic area of interest failed, and individuals if too many samples failed leading to only <2 sequences available for the analysis.

Statistical analysis

We characterised the confirmed and not confirmed HIV-1 superinfection, the remaining superinfection cases, and a control group. The control group represents 4,250 of 4,575 SHCS individual with ≥2 longitudinal HIV-1 *pol* sequences, not meeting the selection criteria for a potential HIV-1 superinfection. The groups were compared on gender (male or female), ethnicity (white, black or other ethnicity (i.e. hispano-american, Asian and unknown)), risk behaviour such as likely source of infection: men who have sex with men(MSM), heterosexual contacts(HET), and Intravenous drug use(IVD). Having had a positive test for other coinfections such Cytomegalovirus(CMV), Syphilis (caused by Treponema Pallidum) and Hepatitis C(HCV) were also considered. We performed uni- and multivariable logistic regressions considering the confirmed superinfections against the control group.

Average pairwise diversity calculation

The average pairwise diversity(APD) [27,28] was calculated over the third-codon positions of HIV-1 *pol*(PR-RT), based on the viral consensus sequence (Appendix S5). A high APD score reflecting a high within-host diversity may potentially be a useful marker for superinfection. We used the APD score of 0.0336 for high diversity at a given time point to confirm superinfection.

Results

The study population for HIV-1 superinfection in the SHCS

To study HIV-1 superinfection, we started our workflow with all genotypic resistance testing(GRT) HIV-1 partial *pol* sequences in the SHCS DRDB (Figure 1). We then reconstructed the phylogeny of 22,243 sequences linked to 12,397 cohort participants. We restricted the workflow to 4,575 individuals in the phylogeny having ≥2 longitudinal HIV-1 partial *pol* sequences, requirement to screen for HIV-1 superinfection.

Identification of potential HIV-1 superinfections in the SHCS

To screen for HIV-1 superinfection, we used: i) the within-individual maximum patristic distance and ii) the per individual smallest non-monophyletic cluster size. The first criterion reflects the requirement, for sufficiently variable and genetically distant within-individual sequences [29]. The second criterion distinguishes superinfection from transmission chains initiated by the focal individual [24,30,31], and usually, a value ≤0.045 is used to identify transmission clusters in phylogenies [32]. We then varied the maximum patristic distance from ≥0.01 to ≥0.10 and the smallest cluster size from ≥5 to ≥250 sequences, to identify HIV-1 superinfections. The number of cases varied considerably within the smaller thresholds, but for the higher ranges (0.05-0.10 and 20-250) it was less dependent on the thresholds (Figure 2.A). We selected a within-individual maximum patristic distance ≥0.05 and a smallest cluster size ≥20 sequences to identify HIV-1 superinfection (Figure 2.A, 2.B). We identified 325 potential HIV-1 superinfections in the SHCS (Figure 2.C; results comparable using RAxML Table S1).

Categorisation of the potential HIV-1 superinfections in the SHCS

HIV-1 superinfection is described as a much less frequent event compared to initial infection [23]. A previous study showed that ~86% of the analysed cases were linked to misidentified samples or sequences [24]. We thus categorised the potential HIV-1 superinfections to address the evidence for superinfection against potential specimen misidentification (Figure 3). The potential cases were classified into one of 3 categories. The likelihood of being superinfected increases from category 1 to 3, while respectively the likelihood of specimen misidentification decreases. We classified 29 individuals in category 3, 161 in category 2, and 135 in category 1.

Confirming potential HIV-1 superinfection

To validate our approach, we did HIV-1 NFVGS for 128 cases (Figure 4) with available longitudinal plasma samples in the SHCS biobank, around the estimated assumed time of superinfection, and reconstructed the partial *pol* phylogeny. Using only our selection criteria for HIV-1 superinfection, we confirmed 41 (32%, values summarised in Table S2 and Figure S1) superinfections in the SHCS. The *confirmed superinfections* were 10/15(66.7%) category 3, 19/69(27.6%) category 2, and 12/44(27.2%) category 1. The varying confirmation rate, correlating with the hypothesised superinfection categorisation (Figure 3; results consistent using RAxML for phylogeny Table S3).

For the 87 unconfirmed superinfections, a potential reason could be mismatches between the time points sequenced for validation and the ones in the initial screening. To investigate this hypothesis, we looked at whether the overlapping time points between the validation and initial screen analyses, alone would have allowed to identify the superinfections in the initial screen. The *Lack of Evidence* subset*,* are 61 cases not confirmed cases for which we either had only one or no time point matching between the

two analyses, or ≥2 matching time points not sufficient to identify the superinfection in our initial analysis. In these cases, we did not confirm superinfection likely because of less informative time points used for validation, taken outside the critical window of superinfection. Thus, the validation for these cases, does not allow the confirmation of superinfection but also does not contradict the initial screen (and hence it does not provide evidence against superinfection).

The *Discrepant Phylogenetic Pattern* subset, are 26 not confirmed cases, with sampling times sufficiently concordant with the initial screen. i.e., we have multiple sequences matching between the validation and the initial screen analyses. These matching time points were informative enough to initially identify HIV-1 superinfection, hence the discrepancy between the two analyses. Since our method uses consensus sequences to confirm HIV-1 superinfections, we considered a measure of diversity as complementary screening method. The APD was shown [27,28] to be a time measure of diversity within HIV-1 infected individuals, calculated over the HIV-1 partial *pol* gene. High APD scores (≥0.0336 for *pol*(PR-RT) [28]) may be a marker for superinfection, and in our study, 43 cases had a high APD (Figures 4, S2) for at least one time point. Eleven (6 category 2, 5 category 1) belonged to the 26 *Discrepant Phylogenetic Pattern*. Such a high viral diversity in some of these cases still suggests HIV-1 superinfections, thereby potentially resolving the apparent discrepancy. For the confirmed cases, with no evidence of superinfection by APD, we looked at the fraction of ambiguous nucleotides [33], and it supports the conclusion of superinfection (Table S4). Finally, we investigated the remaining 15 cases (6 category 1, 9 category 2), not confirmed with ≥2 longitudinal sequences matching between both analyses, and lower APD. Fourteen were intra-subtype B superinfections, a limitation for detection of superinfections with phylogenetic

approaches, especially if the sampled viruses are close genetically. One was an inter-subtype B and 01_AE superinfection, belonged to the category 1. We NFVGS the 2 available samples, and both identified as subtype 01_AE. This indicates that it is not an inter-subtype HIV-1 superinfection, and that sample misidentification could be the cause of the suspected superinfection.

In total, from the 128 cases which were HIV-1 NFVGS, we confirmed 52 HIV-1 superinfections (77.6% of 67 cases sufficiently concordant with the initial screen).

Basic epidemiology of the HIV-1 superinfections in the SHCS

The confirmed or hypothesised superinfections were similar in demography and basic epidemiology (Tables 1, 2). Of 52 confirmed, 71% are males, 29% females ($OR_{multivariable}$(95% CI): 1.192(0.562-2.518)), similar for the other groups compared. There are more MSM 42%, than HET 29% (0.608(0.244-1.428)) and IVD 29% (0.695(0.243-2.048)). The SHCS is predominantly of white ethnicity (~70% in 2019), also reflected in the confirmed superinfections, with 86% white vs. 8% black (0.622(0.149-1.737) and 6% other ethnicities. For the confirmed superinfections, 23% were already assigned ≥2 subtypes in our database vs. only ~6% in the control group. The ones in the control group are due to recombinant, which clustered with the main subtype in the phylogeny, hence why they were not superinfections. Having ≥2 subtypes is significant in the uni- and multivariable analysis (5.094(2.528-9.546), and 5.409(2.667-10.225)), confirming the hypothesised HIV-1 superinfections. Finally, 87% of confirmed cases were seropositive for CMV(1.201(0.547-3.029)), 40% for HCV(1.763(0.737-3.853)) and 11% for Syphilis(0.788(0.359-1.623)). We see no significant effect of having or having had these coinfections with being HIV-1 superinfected. For the confirmed superinfections, the viral load and treatment history patterns support that superinfection

occurs during treatment failure where an increase of the viral load is often noticeable (Table S5, Figure S3). Finally, the time elapsed analysis between the GRT time points (Figure S4) is in line with the assumption that a smaller sampling frequency allows the detection of HIV-1 superinfections more efficiently.

Discussion

In this work, we developed a molecular epidemiology-based approach for systematic screening of HIV-1 superinfection, using the dense pool of historic samples of the SHCS. We identified 325 potential HIV-1 superinfections, and assessed 128 likely cases by retrospective HIV-1 NFVGS from longitudinal samples in our biobank. We validated our approach, unambiguously confirmed 52 cases (77.6% of 67 sufficiently concordant with the initial screen).

A similar approach found ~86% validation cases involving sample misidentification [24]. Other studies retrospectively analysed cases identified through patients' abnormal laboratory values (e.g. increased VL, decreased CD4 counts, changed resistance patterns) [8,9,20], or using proviral DNA [21,34,35], which was not yet used to screen for superinfection. Also, available tools to investigate dual infections [30] are not yet tailored to identify superinfection, nor to work with individuals' longitudinal sequences. So, although considerable work on superinfection was done, larger systematic population-based longitudinal screens are missing. And with changing guidelines for treatment as prevention independent of clinical or laboratory markers [36,37], the suitable window to systematically study HIV-1 superinfection became very short or borderline impossible. Thus, our study demonstrates the feasibility of a systematic molecular epidemiology-based approach, applicable to cohort studies to screen for HIV-1 superinfection. The systematic aspect of our approach is underlined by our initial screen including 4,575

of12,397 patients from the SHCS DRDB; i.e. we screened for superinfections in ≥30% of patients with available viral sequences.

With our approach and the SHCS resources, we reliably identified and confirmed, to our knowledge, more cases than other studies could [23,38,39]. Notably, we use the combination of two robust criteria, which were often used separately but rarely in combination to characterise superinfections. The confirmation rate per category supports our hypothesised stratification on the likelihood of sample misclassification (66.7-100% in category 3, 27.6% in category 2 and 27.2% in category 1). Considering the similar results between the categories 1 and 2, they could be treated as one category in phylogenetic-based approaches to identify HIV-1 superinfection. Overall, this demonstrates the complexity of using phylogenetics and the appropriate phylogeny reconstruction tool in such analysis [40].

We also demonstrated that the sampling window and frequency is key to systematically screen for HIV-1 superinfection. We could not confirm or eliminate 61 cases for which the time points in common between the initial screen and validation analyses were insufficiently informative. For 26 other cases with ≥2 time points matching between both analyses, 56.7% could not be confirmed highlighting another limitation of such approach regarding the similarity of the virus strains involved in the superinfection. They were intra-subtype B cases that our validation phylogeny could not disentangle, most likely due to sampling time or to the size and diversity of the phylogenetic tree used for our validation compared to the original screening (3,009 vs. 22,243 sequences). These superinfections remain challenging for a systematic phylogenetic-based screen and should involve detailed recombination analysis or haplotype reconstruction, more sensitive to these phenotypes. We thus used the APD as an additional measure of intra-

patient diversity, and validated 52 cases as true superinfections (77.6% of 67 with sufficiently matching and cross-comparable near-full-length and GRT HIV-1 sequences). Overall, our phylogenetic approach like most others, shows limitations and might not detect superinfections involving viral strains from the same subtype, or a strain that does not replace, or only partially outcompetes (at a lower frequency) the original strain. We also acknowledge that our screen may fail to identify superinfections in case of recombination or if the viral strains before and after superinfections stem from epidemiological settings which are only poorly represented in the analysed dataset. Finally, our initial screen was based on partial *pol* sequences, thus we cannot exclude that some cases not selected by this screening method would show signs of superinfection in other genes.

For the potential and confirmed superinfections, we find more MSM than HET and IVDs, and no significant association between confirmed superinfections and these risk factors. Despite the small sample size, this may suggest that superinfection happens independently of the infection route. This also suggest that for MSM and IVDs, HIV-1 superinfections occur in rather similar networks with closely related viral strains, which are challenging to detect using phylogenetic-based approaches and thus to estimate the true prevalence.

To our knowledge, this is the most extensive screen for superinfections in a cohort study with the largest number of confirmed cases. It confirms that superinfections are rare but not negligible events, with an estimated prevalence of 1 to 7%, most likely an underestimation since detection is challenging. Nevertheless, this work paves the way for follow-up studies to benefit from the sample size and the NGS data generated to molecularly characterise HIV-1 superinfection and investigate other risk factors

associated. Better molecular characterisation and risk factors understanding could provide further insights into HIV-transmission and pathogenesis, benefit HIV vaccine research, and enable preventive measures to raise awareness on HIV-1 superinfection in the community. Overall, this work sets the groundwork to use any detailed cohort database like the SHCS, to systematically study HIV-1 superinfection and its biological mechanisms.

Acknowledgments

We thank the patients who participate in the SHCS; the physicians and study nurses for excellent patient care; A. Scherrer, A. Traytel, S. Wild, and K. Kusejko from the SHCS Data Centre for data management; and D. Perraudin and M. Amstad for administrative assistance.

Members of the Swiss HIV Cohort Study: K. Aebi-Popp, A. Anagnostopoulos, M. Battegay, E. Bernasconi, J. Böni, D. L. Braun, H. C. Bucher, A. Calmy, M. Cavassini, A. Ciuffi, G. Dollenmaier, M. Egger, L. Elzi, J. Fehr, J. Fellay, H. Furrer, C. A. Fux, H. F. Günthard (President of the SHCS), D. Haerry (deputy of "Positive Council"), B. Hasse, H. H. Hirsch, M. Hoffmann, I. Hösli, M. Huber, C. R. Kahlert (Chairman of the Mother & Child Substudy), L. Kaiser, O. Keiser, T. Klimkait, R. D. Kouyos, H. Kovari, B. Ledergerber, G. Martinetti, B. Martinez de Tejada, C. Marzolini, K. J. Metzner, N. Müller, D. Nicca, P. Paioni, G. Pantaleo, M. Perreau, A. Rauch (Chairman of the Scientific Board), C. Rudin, K. Kusejko (Head of Data Centre), P. Schmid, R. Speck, M. Stöckle (Chairman of the Clinical and Laboratory Committee), P. Tarr, A. Trkola, P. Vernazza, G. Wandeler, R. Weber, S. Yerly.

We thank Melissa Robbiani for help with editing the manuscript.

Funding

This study has been financed within the framework of the Swiss HIV Cohort Study, supported by the Swiss National Science Foundation (grant #177499 and grant #179571 to H. F. G.); by the Swiss HIV Cohort Study research foundation; and by the Yvonne Jacob Foundation (to H. F. G.). The data are gathered by the Five Swiss University Hospitals, two Cantonal Hospitals, 15 affiliated hospitals and 36 private physicians (listed in http://www.shcs.ch/180-health-care-providers).

Competing interests

I have read the journal's policy and the authors of this manuscript have the following competing interests: The institution of E. B. received fees for E. B. participation in advisory boards and travel grants from Gilead Sciences, MSD, ViiV Healthcare, Pfizer, Abbvie, and Sandoz. K.J.M. has received travel grants and honoraria from Gilead Sciences, Roche Diagnostics, GlaxoSmithKline, Merck Sharp & Dohme, Bristol-Myers Squibb, ViiV and Abbott; and the University of Zurich received research grants from Gilead Science, Novartis, Roche, and Merck Sharp & Dohme for studies that Dr Metzner serves as principal investigator, and advisory board honoraria from Gilead Sciences. H. F. G. has received unrestricted research grants from Gilead Sciences and Roche; fees for data and safety monitoring board membership from Merck; consulting/advisory board membership fees from Gilead Sciences, Merck and ViiV Healthcare; and grants from SystemsX, and the National Institutes of Health. The institution of H. F. G. received educational grants form Gilead Sciences, ViiV, MSD, Abbvie and Sandoz. All other authors report no potential conflicts of interest.

Authors' contributions

H. F. G., R. D. K., K. J. M., and S. E. C. conceived the study, performed the analysis and wrote the first draft of the manuscript. C. L. performed the HIV-1 NFVGS. A. M., K. K., N. T., and H. N. contributed to the analysis of the results. M. H., J. B., M. P., T. K., S. Y., A. R., H. H. H., A. R., A. C., P. V., E. B., and M. C. collected and contributed data. All authors read and approved the final manuscript.

Corresponding authors:

Sandra E. Chaudron, PhD, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Rämistrasse 100, CH-8091 Zurich,
Current address: Old Road Campus OX3 7LF, Oxford, United Kingdom
Switzerland (Sandra.Chaudron@ndm.ox.ac.uk)

(Alternate) Huldrych F. Günthard, MD, Division of Infectious Diseases and Hospital Epidemiology, University Hospital Zurich, Rämistrasse 100, CH-8091 Zurich,
Switzerland (huldrych.guenthard@usz.ch)

References

1. Unaids. Fact sheet - Latest global and regional statistics on the status of the AIDS epidemic, **2019**.

2. Vernazza P, Bernard EJ. HIV is not transmitted under fully suppressive therapy: The Swiss Statement – eight years later. Swiss Med Wkly **2016**.

3. Garnett GP, Gazzard B. Risk of HIV transmission in discordant couples. Lancet **2008**; 372:270-1.

4. Federal Office of Public H. VIH, syphilis,gonorrhée et chlamydiose en Suisse en 2018: survol épidémiologique, **2019**.

5. Vernazza PL, Bernasconi E, Hirschel B. HIV superinfection: myth or reality? Schweiz Med Wochenschr **2000**; 130:1101-5.

6. Chohan B, Lavreys L, Rainwater SMJ, Overbaugh J. Evidence for Frequent Reinfection with Human Immunodeficiency Virus Type 1 of a Different Subtype. J Virol **2005**; 79:10701-8.

7. Ramos A, Hu DJ, Nguyen L, et al. Intersubtype human immunodeficiency virus type 1 superinfection following seroconversion to primary infection in two injection drug users. J Virol **2002**; 76:7444-52.

8. Jost S, Bernard M-C, Kaiser L, et al. A patient with HIV-1 super-infection. N Engl J Med **2002**; 347:731-6.

9. Smith DM, Richman DD, Little SJ. HIV superinfection. J Infect Dis **2005**.

10. Gottlieb GS, Nickle DC, Jensen MA, et al. HIV Type 1 Superinfection with a Dual-Tropic Virus and Rapid Progression to AIDS: A Case Report. Clin Infect Dis **2007**; 45:501-9.

11. McCutchan FE, Hoelscher M, Tovanabutra S, et al. In-Depth Analysis of a Heterosexually Acquired Human Immunodeficiency Virus Type 1 Superinfection: Evolution, Temporal Fluctuation, and Intercompartment Dynamics from the Seronegative Window Period through 30 Months Postinfection. J Virol **2005**; 79:11693-704.

12. Piantadosi A, Ngayo MO, Chohan B, Overbaugh J. Examination of a second region of the HIV type 1 genome reveals additional cases of superinfection. AIDS Res Hum Retroviruses **2008**; 24:1221.

13. Casado C, Pernas M, Alvaro T, et al. Coinfection and superinfection in patients with long-term, nonprogressive HIV-1 disease. J Infect Dis **2007**; 196:895-9.

14. Redd AD, Quinn TC, Tobian AAR. Frequency and Implications of HIV Superinfection. Lancet Infect Dis **2013**; 17:622-8.

15. Smith DM, Wong JK, Hightower GK, et al. Incidence of HIV Superinfection Following Primary Infection. JAMA **2004**; 292:1177-8.

16. Cortez V, Wang B, Dingens A, et al. The Broad Neutralizing Antibody Responses after HIV-1 Superinfection Are Not Dominated by Antibodies Directed to Epitopes Common in Single Infection. PLoS Pathog **2015**; 11:e1004973-e.

17. Serwanga J, Ssemwanga D, Muganga M, et al. HIV-1 superinfection can occur in the presence of broadly neutralizing antibodies. Vaccine **2018**; 36:578-86.

18. Luan H, Han X, Yu X, et al. Dual infection contributes to rapid disease progression in men who have sex with men in China. J Acquir Immune Defic Syndr **2017**; 75:480-7.

19. De Azevedo SSD, Delatorre E, Cô Rtes A FH, et al. HIV controllers suppress viral replication and evolution and prevent disease progression following intersubtype HIV-1 superinfection. AIDS **2019**; 33:399-410.

20. Günthard HF, Huber M, Kuster H, et al. HIV-1 superinfection in an HIV-2-infected woman with subsequent control of HIV-1 plasma viremia. Clin Infect Dis **2009**; 48:e117-20.

21. Yerly S, Jost S, Monnat M, et al. HIV-1 co/super-infection in intravenous drug users. AIDS (London, England) **2004**; 18:1413-21.

22. van der Kuyl AC, Kozaczynska K, van den Burg R, et al. Triple HIV-1 Infection. N Engl J Med **2005**; 352:2557-9.

23. Ronen K, McCoy CO, Matsen FA, et al. HIV-1 Superinfection Occurs Less Frequently Than Initial Infection in a Cohort of High-Risk Kenyan Women. PLoS Pathog **2013**; 9.

24. Bartha I, Assel M, Sloot PMA, et al. Superinfection with drug-resistant HIV is rare and does not contribute substantially to therapy failure in a large European cohort. BMC Infect Dis **2013**; 13:537.

25. Schoeni-Affolter F, Ledergerber B, Rickenbach M, et al. Cohort Profile: The Swiss HIV Cohort Study. Int J Epidemiol **2010**; 39:1179-89.

26. Yang WL, Kouyos R, Scherrer AU, et al. Assessing the paradox between transmitted and acquired HIV type 1 drug resistance mutations in the Swiss HIV Cohort Study from 1998 to 2012. J Infect Dis **2015**; 212:28-38.

27. Puller V, Neher R, Albert J. Estimating time of HIV-1 infection from next-generation sequence diversity. PLoS Comput Biol **2017**; 13:e1005775-e.

28. Carlisle LA, Turk T, Kusejko K, et al. Viral Diversity Based on Next-Generation Sequencing of HIV-1 Provides Precise Estimates of Infection Recency and Time since Infection. J Infect Dis **2019**; 220:254-65.

29. Li G, Piampongsant S, Faria NR, et al. An integrated map of HIV genome-wide variation from a population perspective. Retrovirology **2015**; 12:18.

30. Wymant C, Hall M, Ratmann O, et al. PHYLOSCANNER: Inferring Transmission from Within-and Between-Host Pathogen Genetic Diversity The Maela Pneumococcal Collaboration, and The BEEHIVE Collaboration ‡ 1 Big Data Institute. Mol Biol Evol **2017**.

31. Reichmuth ML, Chaudron SE, Bachmann N, et al. Using longitudinally sampled viral nucleotide sequences to characterize the drivers of HIV-1 transmission. HIV Med **2020**; 22:346-59.

32. Ragonnet-Cronin ML, Shilaih M, Günthard HF, et al. A Direct Comparison of Two Densely Sampled HIV Epidemics: The UK and Switzerland. Sci Rep **2016**; 6:32251.

33. Kouyos RD, Von Wyl V, Yerly S, et al. Ambiguous nucleotide calls from population-based sequencing of HIV-1 are a marker for viral diversity and the age of infection. Clin Infect Dis **2011**; 52:532-9.

34. Caetano DG, Côrtes FH, Bello G, et al. A case report of HIV-1 superinfection in an HIV controller leading to loss of viremia control: a retrospective of 10 years of follow-up. BMC Infect Dis **2019**; 19:588.

35. Casado C, Pernas M, Rava M, et al. High-Risk Sexual Practices Contribute to HIV-1 Double Infection Among Men Who Have Sex with Men in Madrid. AIDS Res Hum Retroviruses **2020**; 36:896-904.

36. Günthard HF, Aberg JA, Eron JJ, et al. Antiretroviral Treatment of Adult HIV Infection. JAMA **2014**; 312:410.

37. Saag MS, Benson CA, Gandhi RT, et al. Antiretroviral Drugs for Treatment and Prevention of HIV Infection in Adults. JAMA **2018**; 320:379.

38. Redd AD, Helleberg M, Sievers M, et al. Limited anti-HIV neutralizing antibody breadth and potency before and after HIV superinfection in Danish men who have sex with men. Infect Dis **2019**; 51:56-61.

39. Courtney CR, Mayr L, Nanfack AJ, et al. Contrasting antibody responses to intrasubtype superinfection with CRF02_AG. PLoS One **2017**; 12:e0173705-e.

40. Liu K, Linder CR, Warnow T. RAxML and FastTree: Comparing Two Methods for Large-Scale Maximum Likelihood Phylogeny Estimation. PLoS One **2011**; 6:e27731-e.

Table 1: Basic epidemiology of the different study population.

		Superinfection Confirmed	Superinfection Not Confirmed	Remaining Potential Superinfection	SHCS Individuals with ≥ 2 HIV-1 pol sequences
	Total	52	76	273	4250
Sex	Male (%)	37 (71)	53 (70)	182 (67)	2968 (70)
	Female (%)	15 (29)	23 (30)	91 (33)	1282 (30)
Risk	MSM (%)	22 (42)	25 (33)	92 (34)	1594 (38)
	HET (%)	15 (29)	28 (37)	94 (34)	1559 (37)
	IVD (%)	15 (29)	19 (25)	71 (26)	916 (22)
Ethnicity	White (%)	45 (86)	56 (74)	210 (77)	3299 (78)
	Black (%)	4 (8)	10 (13)	34 (12)	574 (13)
	Other Ethnicity (%)	3 (6)	10 (13)	29 (11)	377 (9)
Age	Median [IQR]	57 [53 - 64]	56 [50 - 60]	55 [49 - 60]	55 [49 - 60]
Number of subtype	1 subtype (%)	40 (77)	55 (72)	181 (66)	4007 (94)
	2 subtypes (%)	12 (23)	20 (26)	92 (34)	236 (6)
Other infections	Ever had Syphilis (%)	11 (21)	15 (20)	65 (24)	963 (23)
	Having CMV (%)	45 (87)	65 (86)	233 (85)	3614 (85)
	Ever had HCV (%)	21 (40)	28 (37)	90 (33)	1221 (29)

Note: Comparison of the sex, risk, ethnicity, age, number of subtypes assigned and coinfections for the confirmed and not confirmed HIV-1 superinfections, the remaining potential HIV-1 superinfections and the control group of SHCS individuals with ≥ 2 longitudinal sequences in the drug resistance database.

Table 2. Odd ratios for HIV-1 superinfection.

		Univariable OR (95% CI)	Univariable p-value	Multivariable OR (95% CI)	Multivariable p-value
Sex	Male	1		1	
	Female	0.939 (0.498-1.681)	0.837	1.192 (0.562-2.518)	0.644
Risk	MSM	1		1	
	HET	0.697 (0.353-1.681)	0.284	0.608 (0.244-1.428)	0.268
	IVD	1.186 (0.601-2.281)	0.612	0.695 (0.243-2.048)	0.502
Ethnicity	White	1		1	
	Black	0.511 (0.153-1.264)	0.200	0.64 (0.176-1.857)	0.446
	Other Ethnicity	0.583 (0.141-1.605)	0.368	0.622 (0.149-1.737)	0.432
Number of subtype	1 subtype	1		1	
	≥ 2 subtypes	5.094 (2.528-9.546)	0	5.409 (2.667-10.225)	0
Other infections	Ever had Syphilis	0.916 (0.446-1.726)	0.797	0.788 (0.359-1.623)	0.533
	Having CMV	1.131 (0.542-2.757)	0.763	1.201 (0.547-3.029)	0.671
	Ever had HCV	1.681 (0.949-2.918)	0.068	1.763 (0.737-3.853)	0.177

Note: Univariable and multivariable logistic regression for different risk factors that could be associated with the outcome of being HIV-1 superinfected. The sex, risk group, ethnicity, number of subtype and coinfections were considered for the regression for the 52 confirmed cases against the 4,250 control patients with ≥ 2 longitudinal sequences in the SHCS DRDB.

Figure 1. Study Population. Overview of the selection process of the study population in the SHCS. We considered 4,575 individuals, with ≥ 2 longitudinally sampled HIV-1 *pol* sequences (protease (PR) and reverse transcriptase (RT), to **further** study HIV-1 superinfection.

Figure 2. Sensitivity of the screening criteria and selection potential HIV-1 superinfections in the SHCS. Two criteria were considered to identify potential HIV-1 superinfections in the SHCS. The sensitivity of the maximum patristic distance was assessed with 10 different threshold values chosen from 0.01 to 0.10 and the one for the smallest cluster size, with 10 threshold values chosen from 5 to 250 sequences. (A) Overlap between the 10 values of the two selection criteria. The blue boxes circle the number of identified HIV-1 superinfections for the two final criteria separately. (B) Overlap between the patients having a maximum patristic distance ≥ 0.05, the ones whose sequences in the phylogeny do not create a monophyletic subtree (cluster) and the ones whose smallest cluster of sequences contain ≥ 20 sequences. (C) Representation of the maximum patristic distance against the smallest cluster size in log scale for 972 focal patients having non-monophyletic clusters (empty circles). The selection criteria thresholds are set at 0.05 and $\log_{10}(20)$ for x and the y axis respectively. The potential 325 HIV-1 superinfections are shown as blue circles.

Figure 3. Categorising the potential HIV-1 cases in the SHCS. 325 potential HIV-1 superinfections classified in 3 categories, representing the likelihood of HIV-1 superinfection (likely to most likely superinfection: category 1 (green box), 2 (yellow box) and 3 (red box) respectively). Category 1 individuals only had 2 longitudinally sampled partial HIV-1 *pol* sequences, distant in the phylogeny. Category 2: individuals have > 2 longitudinally sampled partial HIV-1 *pol* sequences with one sequence distant from the others in the phylogeny. Category 3 individuals have > 2 longitudinally sampled partial HIV-1 *pol* sequences with

sequences clustering with each other at different position in the tree. N represents the number of individuals identified per category.

Figure 4. Validation of HIV-1 Superinfection with near-full-length HIV-1 sequencing. of the 325 potential HIV-1 superinfections (SI) identified in the SHCS, 128 (44 in category 1 (green box), 69 in category 2 (yellow box) and 15 in category 3 (red box)) were near-full-length HIV-1 next-generation sequenced. We reconstructed the phylogeny of the HIV-1 partial *pol* (PR-RT) genomic area and applied the 2 selection criteria for HIV-1 superinfection. The *Confirmed SI* are the cases where superinfection could be validated with near-full-length HIV-1 sequencing and our method. The *Lack of Evidence* are either the cases for which we only have one time point or no time points matched between initial screen and validation analyses; or the cases with ≥ 2 time points matching that are not informative enough to identify superinfection in the initial screen or to validate it with near-full-length HIV-1 based analysis. The *Discrepant Phylogenetic Pattern* are the cases where we have ≥ 2 matching time points between the initial screen and validation analyses. However, for these cases there is a discrepancy between the initial screen and validation phylogeny resulting in them not being confirmed as superinfections with our analysis but still identified as superinfection with the matching time points. The average pairwise diversity (APD) was calculated for every case (see also Figure S1) and the number in blue are the total number of validated superinfections.

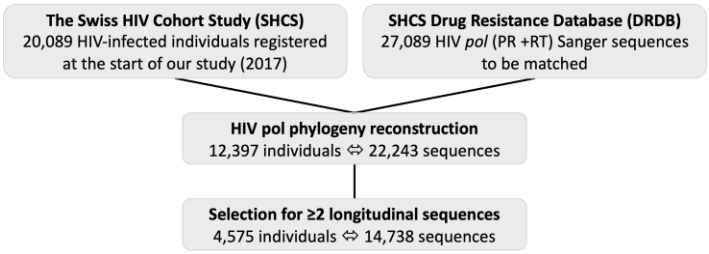

Figure 1
93x34 mm (.34 x DPI)

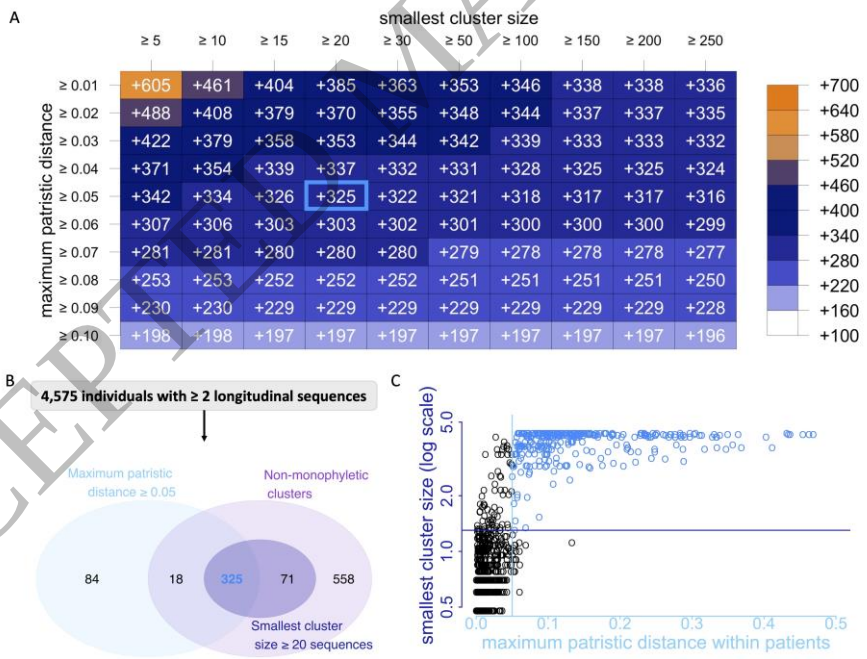

Figure 2
118x87 mm (.34 x DPI)

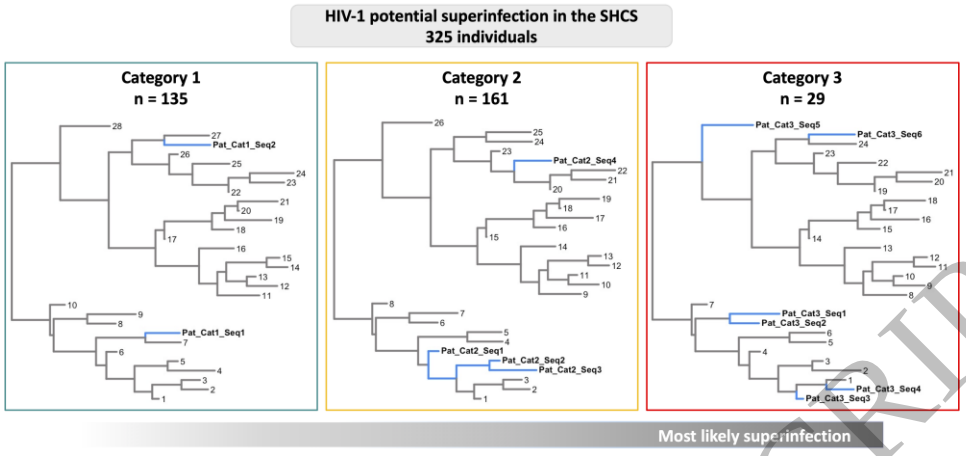

Figure 3
126x61 mm (.34 x DPI)

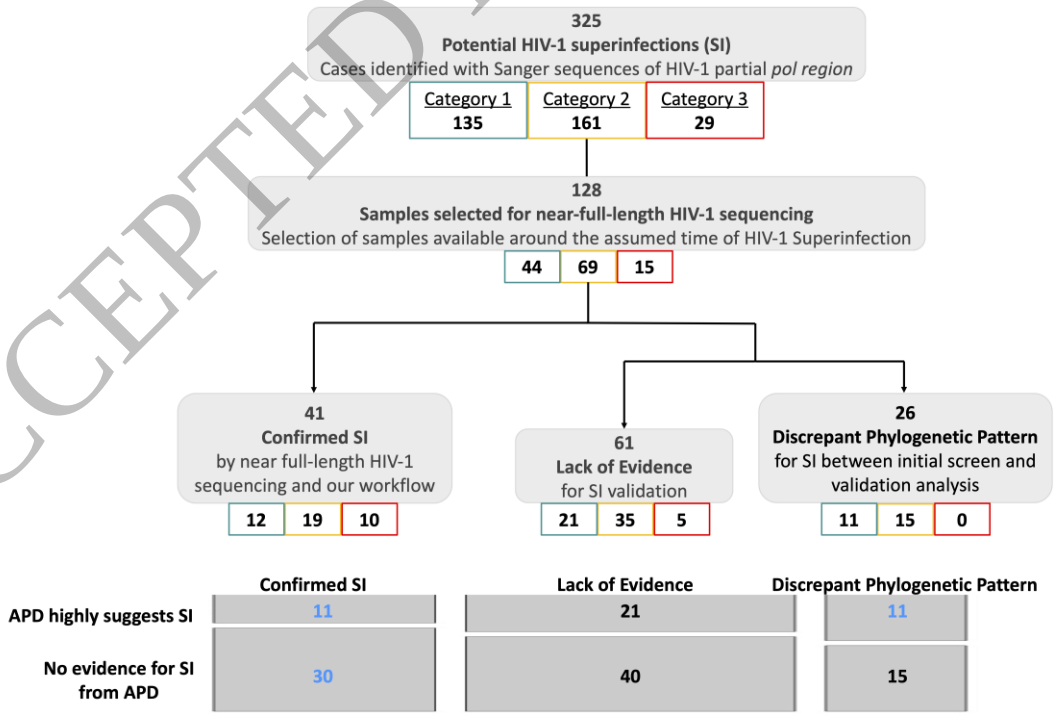

Figure 4
140x94 mm (.34 x DPI)