HHS Public Access

Author manuscript

Virology. Author manuscript; available in PMC 2019 April 01.

Published in final edited form as:

Virology. 2017 November ; 511: 249–255. doi:10.1016/j.virol.2017.08.031.

Decoding Noises in HIV Computational Genotyping

MingRui Jia[a,b,1], **Timothy Shaw**[b,1], **Xing Zhang**[c], **Dong Liu**[b,d], **Ye Shen**[b], **Amara E. Ezeamama**[e], **Chunfu Yang**[f], and **Ming Zhang**[b,*]

[a]Department of Pain Management, Shandong Provincial Hospital Affiliated to Shandong University, 324 Jingwu Road, Jinan, Shandong Province, 250021, China

[b]Department of Epidemiology and Biostatistics, University of Georgia, GA 30602, USA

[c]Diagnostic Imaging center, MD Anderson Cancer Center, Houston, TX 77030, USA

[d]Laboratory of Ichthyology, Shanghai Ocean University, Shanghai, 201306, China

[e]Department of Psychiatry, Michigan State University, MI 48824, USA

[f]International Laboratory Branch, Division of Global HIV/TB, Center for Global Health, Centers for Diseases Control and Prevention, GA 30333, USA

Abstract

Lack of a consistent and reliable genotyping system can critically impede HIV genomic research on pathogenesis, fitness, virulence, drug resistance, and genomic-based healthcare and treatment. At present, mis-genotyping, i.e., background noises in molecular genotyping, and its impact on epidemic surveillance is unknown. For the first time, we present a comprehensive assessment of HIV genotyping quality. HIV sequence data were retrieved from worldwide published records, and subjected to a systematic genotyping assessment pipeline. Results showed that mis-genotyped cases occurred at 4.6% globally, with some regional and high-risk population heterogeneities. Results also revealed a consistent mis-genotyping pattern in *gp120* in all studied populations except the group of men who have sex with men. Our study also suggests novel virus diversities in the mis-genotyped cases. Finally, this study reemphasizes the importance of implementing a standardized genotyping pipeline to avoid genotyping disparity and to advance our understanding of virus evolution in various epidemiological settings.

Keywords

HIV; molecular genotyping; epidemic surveillance

*Correspondence to: Dr. Ming Zhang, Department of Epidemiology and Biostatistics, University of Georgia, 101 Buck Road, Athens, GA 30602. USA. Tel: 706-542-2194, mzhang01@uga.edu.

[1]Equal contribution

Conflict of interest

none.

INTRODUCTION

The extraordinary diversity of HIV-1 is exemplified by nine subtypes and over seventy recombinant clades within the M group alone (1 and www.hiv.lanl.gov). In addition, elevated global epidemic surveillance and advances in biotechnology have immensely expanded our sampling and sequencing capability, leading to discoveries of more complex HIV-1 genomes (3–5). As much of the virus diversity information has been implemented in various aspects of HIV basic and translational research, epidemic surveillance, and genomic-based healthcare and prevention, a basic yet important question remains unanswered: to what degree are we sure of the virus genotyping information, a frontline reflection of the viral genetic diversity, is correct?

To address this question, a few small-scale retrospective studies have been performed. In one case, re-analysis of HIV isolates from Cyprus and Greece revealed that subtype I, rather than a pure subtype, actually was a mosaic clade that contains an unique complex A/G/H/K/? pattern (6). In another case, the absence of subtype E full-length genomes (7, 8) literally contradicts with the HIV subtype nomenclature (1), with which the subtype E was designated as a subtype, and subsequently has been utilized inappropriately as a full-genome genotyping reference.

In the past few years, we also have made notable progress in the investigation of HIV genotyping quality. Utilizing a stringent genotyping pipeline we have developed (3, 9–11), we re-examined several epidemiologically important clades (3, 10). For instance, the prevalence of CRF02, responsible for over 9 million HIV-1 infections worldwide (12), has entitled itself becoming one of most important clades considered in both global and regional vaccine design in West and West Central Africa (13). Of a total of 30 published, epidemiologically unlinked full-genome CRF02 strains, we found that 16 strains did not resemble the CRF02 prototype reference IBNG strain in neither genomic structure nor compositional sequences (3). In China, where BC recombinants are integral to studying local epidemics and development of regional antiviral therapies (14), we found that mis-genotyping rate was as high as 60% among all published full-genome CRF08_BC strains (3).

Leveraged by these small-scale and regional studies, it is desirable to gauge the quality of HIV genotyping on a broader scale at the global level, and ideally, to provide improved clade references for global epidemiologic surveillance and genomic-based treatment and prevention. We thus applied interdisciplinary expertise and experience in bioinformatics and molecular epidemiology to systematically investigate the quality of HIV genotyping data published worldwide. Results of this study not only advance our understanding in HIV evolutionary patterns in different epidemiological settings, but also shed light on improved HIV genomic-based healthcare and treatment.

MATERIALS AND METHODS

HIV sequences included in this study were retrieved from the GenBank and Los Alamos HIV Sequence Database (15) as of Sept 2014. Sequences shorter than 400 nucleotides and

redundant sequences were filtered out according to the criteria described in our prior study (3).

All qualified sequences were subjected to genotyping firstly by using the jumping profile hidden Markov model (jpHMM) method (3, 9–11). The jpHMM algorithm combines both the profile hidden Markov model (16) and jumping alignment (17). In this method, distinct HIV-1 subtypes can be discerned based on hidden Markov model-based subtypes profiles, which are associated with estimated probabilities for staying within one subtype or jumping to different subtypes. Through a Viterbi path search (18), the jpHMM computes the most probable subtype pattern for a given query sequence. A prominent feature that distinguishes the jpHMM method from other HIV genotyping methods is its high accuracy in predicting parental subtypes of recombinants and locating breakpoint positions (3, 9–11). Therefore we are confident with the application of the jpHMM method in this study.

Following the jpHMM genotyping procedure, sequences with conflicting genotype definitions between their original documented assignments and the jpHMM genotyping result were further examined by phylogenetic analyses. This was performed by the F84-based neighbor-joining method implemented in the PHYLIP program (version 3.69) (19). In brief, DNADIST and NEIGHBOR were used for constructing phylogenetic trees, followed by SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE analyses to assess reliability of clade clustering. Five-hundred iterations of nonparametric bootstrapping were utilized to provide statistical supports of clade classification and we used bootstrap of 75% to determine if a sequence belongs to a cluster. HIV subtype references (20) were used to infer correct clade clustering. A mis-genotyped case was defined when the jpHMM and phylogenetic analyses produced a consensus result, while it was different from the original GenBank and/or publication genotype assignment. The genotyping algorithm is summarized in Figure 1.

To assess the global prevalence of HIV mis-genotyped cases, we selected mis-genotyped cases by using one sequence per individual, followed by stratifications based on information of transmission route and sampling geographic region as reported in the original report. To estimate the regional prevalence of mis-genotyped cases, all mis-genotyped sequences were normalized by a total number of available HIV-1 sequences in the same geographic region. Mis-genotyped cases at both global and regional levels were also stratified pertaining to sampling years, in order to assess the temporality of mis-genotyped cases.

To investigate the distribution of mis-genotyped occurrences across the viral genome, counts of mis-genotyped sequences were normalized by the total number of viral sequences in the same genomic region in order to avoid unequal sampling bias across the genome. The HXB2 strain (K03455) was used as the genomic numbering reference when sequences were mapped on the HIV-1 genome.

RESULTS

We retrieved 459,881 global HIV-1 sequences from the GenBank and Los Alamos HIV Sequence Database as of Sept 2014. These sequences were filtered as described in the

Methods and Materials section. A total of 49,175 sequences were used for genotyping quality analyses, which revealed 2,236 mis-genotyped cases, each representing one patient source. As detailed in the Methods and Materials section, a mis-genotyped case was defined when the jpHMM and phylogenetic analyses produced a consensus result, while it was different from the original GenBank and/or publication genotype assignment. Collectively, we identified a total of 4.6% mis-genotyped cases from worldwide HIV-1 sequences analyzed in this study. Numbers of retrieved data and mis-genotyping cases are also summarized in Table 1.

We further assessed the global and regional prevalence of HIV mis-genotyping occurrences. As depicted in Figure 2A, West and West Central Africa, and South America, each bears a high level (>10%) of mis-genotyped HIV cases. A total of 965 mis-genotyped cases, representing one sequence per individual, were identified with confirmed transmission history based on literature curation. These cases were then stratified based on four mostly common routes of HIV transmission: heterosexual transmission, men who have sex with men (MSM), intravenous drug use (IDU), and children infected by mother-to-child transmission (MTCT, the data of the pairing mothers was inadequate for analysis). As shown in Fig. 2B, West Central Africa contained the highest count of mis-genotyped cases within the heterosexual population and MSM. In the IDU population, Argentina topped the mis-genotyping list. And frequent mis-genotyping cases in the MTCT population were observed in East Africa, Argentina, and China.

We also investigated the distribution of mis-genotyped cases across the viral genome. Both raw count and normalized count are summarized in Figure 3. Overall, mis-genotyped cases occurred at a frequency of 3–6% across the genome, which was mostly contributed by the heterosexual population and the least by the MSM population. Mis-genotyped cases are more enriched in regions of *gag, gp120* and *nef*. In the heterosexual population, mis-genotyped cases in *gag, pol,* and *gp120* regions were higher than the average. Similar trends were also observed in the *gp120* region in populations of IDU and MTCT. Furthermore, a consistent pattern of mis-genotyped cases in the *gp120* region was identified in all populations except MSM. This is not surprising, giving the extraordinary mutation capability of *gp120* under influences of immunological selection. This result underscores a remaining challenge of accurate genotyping the *gp120* region.

We further examined the temporality of mis-genotyped cases to determine the extent that the advanced sequencing technology and improved genotyping capability have contributed to improving genotyping accuracy since the beginning of HIV/AIDS epidemics. Our data spans across 25 years, from 1985 to 2009, during which we were able to obtain adequate sampling information. In general, the prevalence of mis-genotyped cases, each represent one sequence per individual, is approximately 4–5% over the time frame analyzed, with a short disruption during 1997–2002, for which a higher number (>5%) of mis-genotyped cases were identified (Fig. 4A). This temporary increase is likely attributed to elevated epidemic sampling efforts in the non Sub-Saharan Africa region (Fig. 4B) and to a larger number of heterosexual cases identified worldwide (Fig. 4C). Mis-genotyped cases in the IDU population were mostly from China, in particular a surge of such cases in 2004. Within the children infected by MTCT, the highest mis-genotyping portion was observed in 2008 and

that was mostly related to subtype B from Argentina. Brazil, a South American epicenter besides Argentina, was found to contribute to a high frequency of mis-genotyped subtype B in South America during 1992–1996. In addition, a mis-subtyping peak was noted in the Caribbean region during 1999, primarily associated with subtype D. These results indicate an inadequate genotyping accuracy remained over the past decades through 2009.

Discussion

The accuracy of HIV genotyping has plagued HIV research since the advent of HIV/AIDS epidemics. Historically, uncertainty of HIV genetic diversity, lack of genotyping tools, genetic similarity between viral clades, such as subtypes B vs. D, and subtypes A vs. G, all contributed to mis-genotyped cases prior to 2000, when the HIV nomenclature was established (1). In addition, a disparity in the genotyping procedure, result interpretation, and an increasing complexity of circulating strains together impose greater challenges on molecular genotyping. As shown in Figure 4, we have observed limited evidence supporting an improved genotyping accuracy along with advanced sequencing biotechnology and enhanced genotyping capability over the past decades through 2009. It might be the case that mis-genotyping will persist in future molecular HIV research, unless a standardized genotyping pipeline is developed to avoid the genotyping disparity and a much advanced understanding of virus evolution is gained to better inform clade references.

The consequence of mis-genotyped cases should not be overlooked. Here we showed that mis-genotyped cases occurred more frequently in West and West Central Africa, and South America (Figure 2A), more specifically, in populations of heterosexual and MSM in West Central Africa and the IDU population in Argentina (Figure 2B). HIV epidemics in these two geographic regions have been proven to be complex (3). Circulating strains in the regions are likely evolving toward more complicated decedents, thus making genotyping more challenging. It is also possible, as suggested in our prior studies (3), the diversity scope of HIV epidemics in these two regions has been underestimated. As a result, the genotyping accuracy could not be achieved due to lack of appropriate clade references representing epidemics in these two regions.

Our study also identified a consistent mis-genotyped pattern in the genomic region *gp120* (Figure 3). This occurs in all studied populations except MSM. Given the importance of *gp120* in virus entry and neutralizing antibody escape (24–27), mis-genotyped cases identified in this report suggest a possible knowledge gap pertaining to the *gp120* diversity, which might contribute to the incompetence of vaccine designs targeting *gp120*.

Globally speaking, we found mis-genotyped cases occurred at 4.6% level, with a much higher occurrence (>30% prevalence. Figure 2B) in the heterosexual population in West Central Africa, the intravenous drug use population in South America, the population of men who have sex with men from West Central Africa, and populations of children from mother-to-child transmission in East Africa and China. Mis-genotyped cases may not simply suggest that a standardized genotyping procedure is desired, but also suggest a closer surveillance of regional epidemics is needed in the aforementioned populations.

Of note, the patient epidemiological data, especially the risk factors, may subject to bias and errors. For example, in countries/regions where MSM and/or IDU is illegal, the patients often under-report their risk activity out of fear of prosecution or shame. Therefore, the results we presented here should be interpreted with caution.

Last but not least, the study method used in this study - the jpHMM and the genotyping gold standard phylogenetic analysis – was chosen based on our extensive experience in HIV genotyping (3, 10, 28, 29) and a careful consideration of HIV genotyping programs and tools. Some of the programs and tools include: i) Context-based Modeling for Expeditious Typing (COMET). It can be accessed via an on-line interface hosted at https://comet.lih.lu/. Unlike the jpHMM program, the COMET program does not provide the recombination breakpoint analysis (30). Additionally, COMET only provides an online version, making it difficult for large-scale data analyses, such as the one described in this study. ii) The SCUEAL program: It is accessible from http://www.datamonkey.org/dataupload_scueal.php. Currently only *pol* sequences can be subtyped by this program (noted in the above site), while the jpHMM has a broad application on other HIV genomic regions. iii) The European-based Subtype Analyzer Program (STAR): It is available at http://www.vgb.ucl.ac.uk/starn.shtml. Unlike the jpHMM program, the current version does not provide the recombination breakpoint predictions (30, 31). iv) The RIP program: It is available from https://www.hiv.lanl.gov/content/sequence/RIP/RIP.html. This tool is not feasible for large-scale data analyses due to its online-only capability. Additionally, the recombinant breakpoints are based on manual check with no reliable support for breakpoint prediction. v) The REGA tool: It is available from regatools.med.kuleuven.be/typing/v3/hiv/typingtool. Behind this tool, it is a phylogenetic-based analysis, specifically, the NJ analysis with bootstrapping. Although this algorithm is different from the jpHMM, which is a probabilistic-based genotyping approach, we did capture its feature by utilizing the phylogenetic method as described in the Materials and Methods session. Comparisons between our jpHMM method, the phylogenetic analysis and some other genotyping tools are described by both our studies (9–11) and others (32, 33).

In summary, through the largest scale study of its kind, we investigated the HIV genotyping quality of HIV at both regional and global levels, as well as within high-risk infection populations and across the viral genome. Our results suggest that mis-genotyped cases, defined as noises in accurate viral genotyping, actually contain informative messages that can be used to refine clade references in epidemic surveillance and help improve HIV basic and translational research. Finally, our results reemphasize the importance of implementing a standardized genotyping pipeline to avoid genotyping disparity and to advance our understanding of virus evolution in various epidemiological settings.

Acknowledgments

This study was supported by ARCS Foundation Scholars Award (TS) and NIH R03AI104258, R03AI120203 and UGA Faculty Research Grant 1025GR793002 (MZ). We also thank the comments and critiques from the anonymous reviewers that help improve this paper.

References

1. Robertson DL, Anderson JP, Bradac JA, Carr JK, Foley B, Funkhouser RK, Gao F, Hahn BH, Kalish ML, Kuiken C, Learn GH, Leitner T, McCutchan F, Osmanov S, Peeters M, Pieniazek D, Salminen M, Sharp PM, Wolinsky S, Korber B. 2000; HIV-1 nomenclature proposal. Science. 288:55–56. [PubMed: 10766634]
2. Hu WS, Temin HM. 1990; Retroviral recombination and reverse transcription. Science. 250:1227–1233. [PubMed: 1700865]
3. Zhang M, Foley B, Schultz AK, Macke JP, Bulla I, Stanke M, Morgenstern B, Korber B, Leitner T. 2010; The role of recombination in the emergence of a complex and dynamic HIV epidemic. Retrovirology. 7:25. [PubMed: 20331894]
4. Taveira, IBaN. HIV-1 Diversity and Its Implications in Diagnosis, Transmission, Disease Progression, and Antiretroviral Therapy. InTech; 2012.
5. UNAIDS. UNAIDS Report on the Global AIDS Epidemic. UNAIDS; 2012.
6. Paraskevis D, Magiorkinis M, Vandamme AM, Kostrikis LG, Hatzakis A. 2001; Reanalysis of human immunodeficiency virus type 1 isolates from Cyprus and Greece, initially designated 'subtype I', reveals a unique complex A/G/H/K/? mosaic pattern. J Gen Virol. 82:575–580. [PubMed: 11172098]
7. Gao F, Robertson DL, Morrison SG, Hui H, Craig S, Decker J, Fultz PN, Girard M, Shaw GM, Hahn BH, Sharp PM. 1996; The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. J Virol. 70:7013–7029. [PubMed: 8794346]
8. Murphy E, Korber B, Georges-Courbot MC, You B, Pinter A, Cook D, Kieny MP, Georges A, Mathiot C, Barre-Sinoussi F, et al. 1993; Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the central African Republic. AIDS Res Hum Retroviruses. 9:997–1006. [PubMed: 8280481]
9. Schultz AK, Zhang M, Bulla I, Leitner T, Korber B, Morgenstern B, Stanke M. 2009; jpHMM: improving the reliability of recombination prediction in HIV-1. Nucleic Acids Res. 37:W647–651. [PubMed: 19443440]
10. Schultz AK, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B, Stanke M. 2006; A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes. BMC bioinformatics. 7:265. [PubMed: 16716226]
11. Zhang M, Schultz AK, Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B, Stanke M. 2006; jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. Nucleic Acids Res. 34:W463–465. [PubMed: 16845050]
12. McCutchan FE. 2000; Understanding the genetic diversity of HIV-1. Aids. 14(Suppl 3):S31–44. [PubMed: 11086847]
13. Gaschen B, Taylor J, Yusim K, Foley B, Gao F, Lang D, Novitsky V, Haynes B, Hahn BH, Bhattacharya T, Korber B. 2002; Diversity considerations in HIV-1 vaccine selection. Science. 296:2354–2360. [PubMed: 12089434]
14. Foley, B, Leitner, T, Apetrei, C, Hahn, B, Mizrachi, I, Mullins, J, Rambaut, A, Wolinsky, S, Korber, B. HIV Sequence Compendium 2013. Los Alamos National Laboratory; 2013.
15. Los Alamos HIV Sequence Database Group. Los Alamos HIV Sequence Database. http://www.hiv.lanl.gov
16. Eddy SR. 1998; Profile hidden Markov models. Bioinformatics. 14:755–763. [PubMed: 9918945]
17. Spang R, Rehmsmeier M, Stoye J. 2002; A Novel Approach to Remote Homology Detection: Jumping Alignments. Journal of Computational Biology. 9:747–760. [PubMed: 12487762]
18. Viterbi A. 1967Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inform Theory IT-13. :260–269.
19. J F. 1989PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. :164–166.
20. Los Alamos HIV Sequence Database Group. Los Alamos HIV Sequence Alignment Page. https://www.hiv.lanl.gov/content/sequence/NEWALIGN/align.html

21. Lihana RW, Ssemwanga D, Abimiku A, Ndembi N. 2012; Update on HIV-1 diversity in Africa: a decade in review. AIDS Rev. 14:83–100. [PubMed: 22627605]
22. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM. 2008; The challenge of HIV-1 subtype diversity. N Engl J Med. 358:1590–1602. [PubMed: 18403767]
23. Vidal N, Mulanga C, Bazepeo SE, Lepira F, Delaporte E, Peeters M. 2006; Identification and molecular characterization of subsubtype A4 in central Africa. AIDS Res Hum Retroviruses. 22:182–187. [PubMed: 16478401]
24. Acharya P, Lusvarghi S, Bewley CA, Kwong PD. 2015; HIV-1 gp120 as a therapeutic target: navigating a moving labyrinth. Expert Opin Ther Targets. 19:765–783. [PubMed: 25724219]
25. Araujo LA, Almeida SE. 2013; HIV-1 diversity in the envelope glycoproteins: implications for viral entry inhibition. Viruses. 5:595–604. [PubMed: 23389465]
26. Merk A, Subramaniam S. 2013; HIV-1 envelope glycoprotein structure. Curr Opin Struct Biol. 23:268–276. [PubMed: 23602427]
27. Flores A, Quesada E. 2013; Entry inhibitors directed towards glycoprotein gp120: an overview on a promising target for HIV-1 therapy. Curr Med Chem. 20:751–771. [PubMed: 23278399]
28. Zhang M, Wilbe K, Wolfe ND, Gaschen B, Carr JK, Leitner T. 2005; HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2. AIDS research and human retroviruses. 21:955–960. [PubMed: 16386113]
29. Zhang M, Schultz AK, Calef C, Kuiken C, Leitner T, Korber B, Morgenstern B, Stanke M. 2006; jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. Nucleic acids research. 34:W463–465. [PubMed: 16845050]
30. Struck, D; Perez-Bercoff, D; Devaux, C; Schmit, JC. COMET: a novel approach to HIV-1 subtype prediction. 8th European HIV Drug Resistance Workshop; Sorrento, Italy. 2010.
31. Liu RWS TF. 2006; Web resources for HIV type 1 genotypic-resistance test interpretation. Clinical Infectious Diseases. 42:1608–1618. [PubMed: 16652319]
32. Pineda-Peña A, Faria N, Imbrechts S, Libin P, Abecasis A, Deforche K, Gómez-López A, Camacho R, de Oliveira T, Vandamme A. 2013; Automated subtyping of HIV-1 genetic sequences for clinical and surveillance purposes: performance evaluation of the new REGA version 3 and seven other tools. Infect Genet Evol. 19:337–348. [PubMed: 23660484]
33. Switzer, W; Saduvala, N; Zhang, T; Hernandez, A; P, L; Struck, D; Oliveira, T; Vandamme, A; Wertheim, J; Oster, A. Comparing Three HIV-1 Subtyping Tools and a Novel Phylogenetic-Based Method, abstr The Conference on Retroviruses and Opportunistic Infections (CROI); Boston, Massachusetts. 2016.

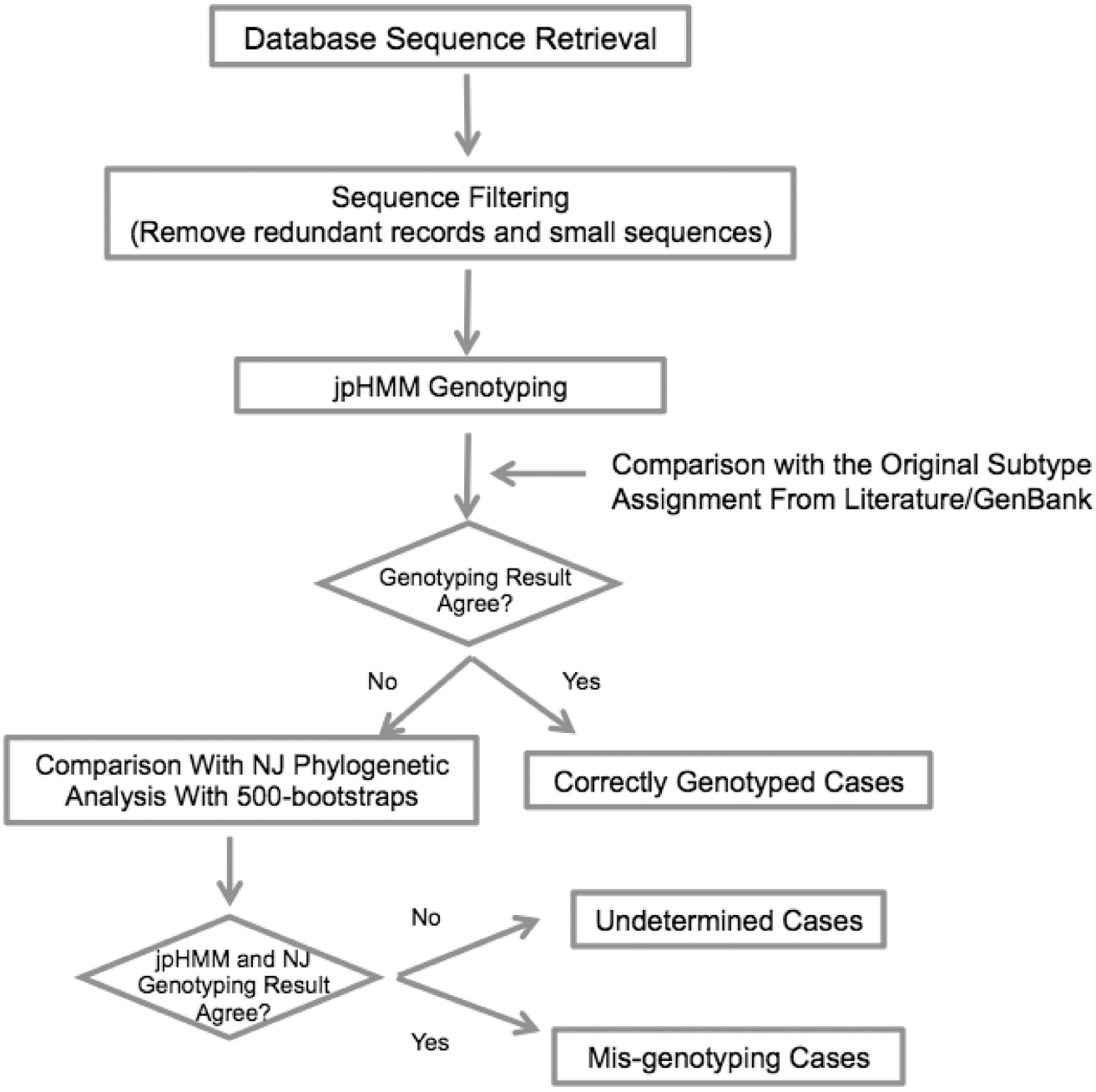
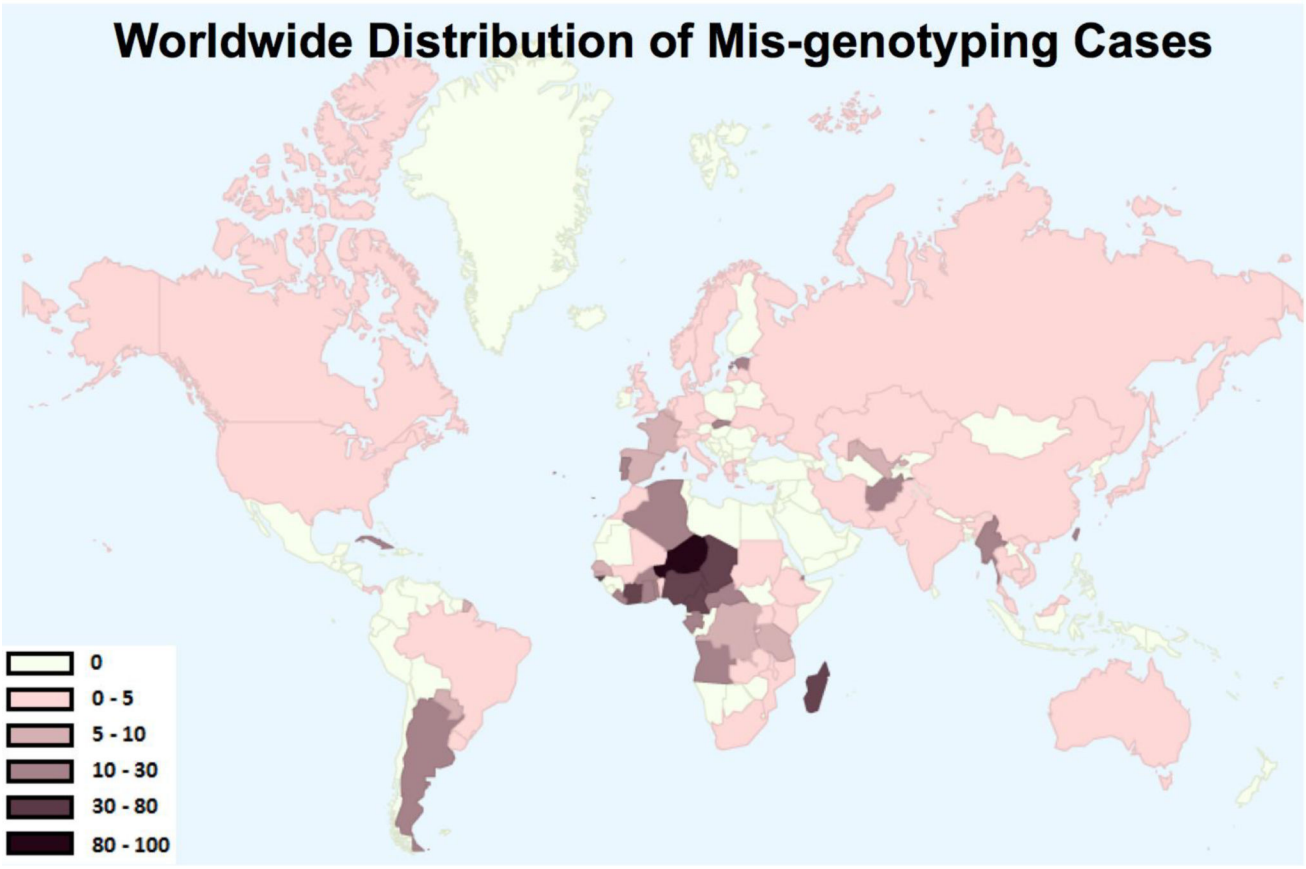


Figure 1. Algorithm of the genotyping process used in this study



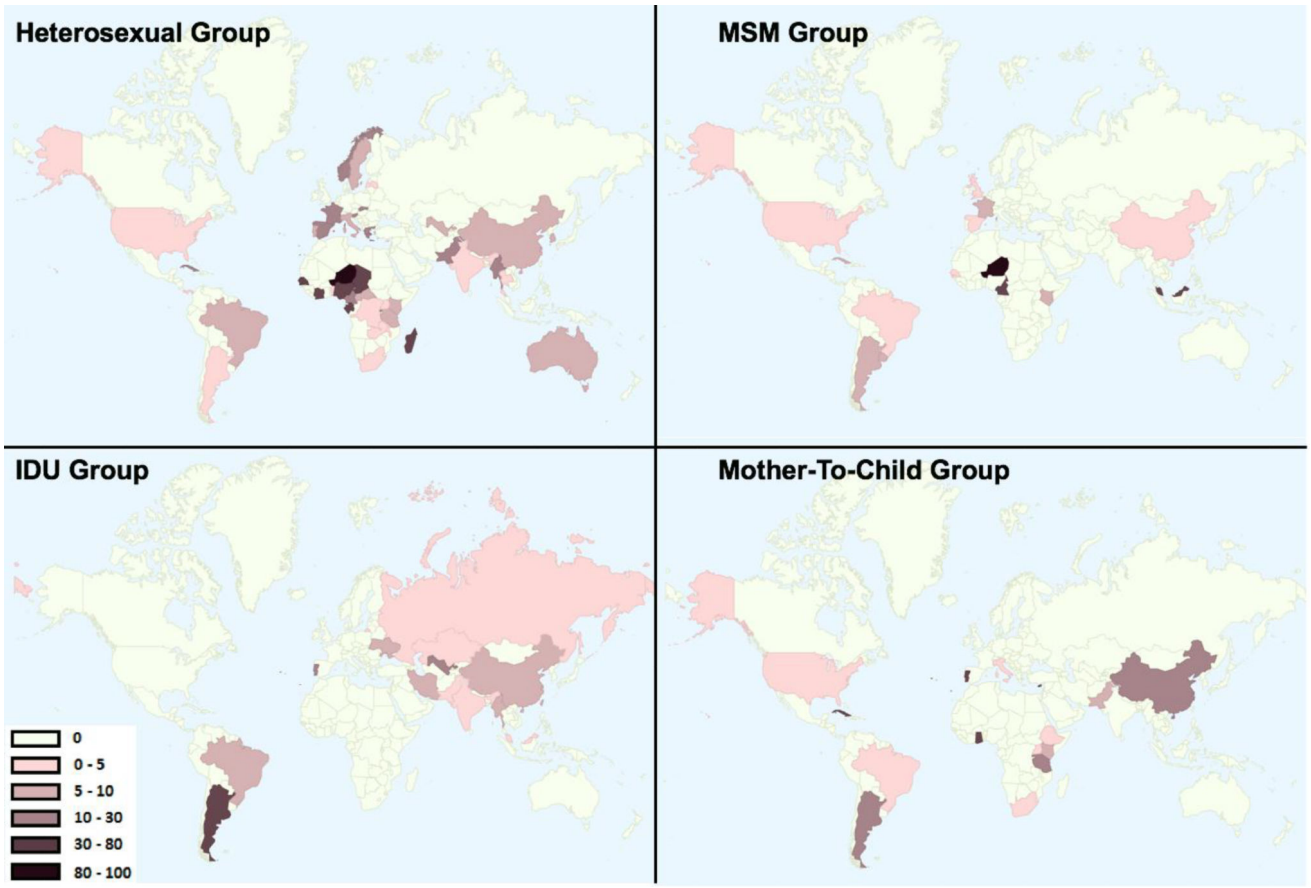


Figure 2. Global prevalence of HIV-1 mis-genotyped cases

Each case represents one sequence per individual. All mis-genotyped cases were normalized by a total number of available HIV-1 sequences in the same geographic region. (A) Worldwide overview of genotyping quality. (B) Genotyping quality within four high-risk infection groups. Color scale: percentage of mis-genotyped cases in individual geographic regions.

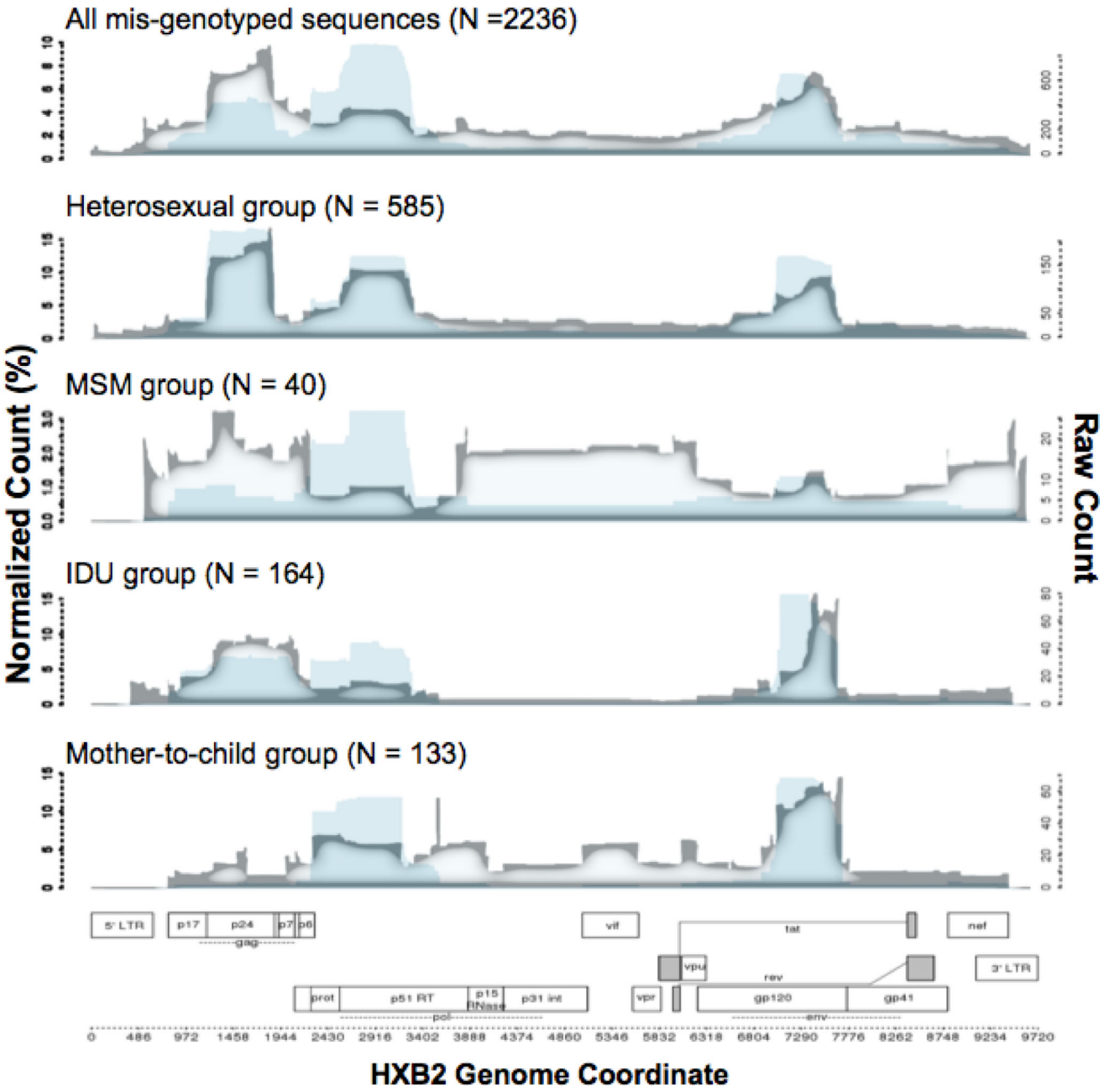
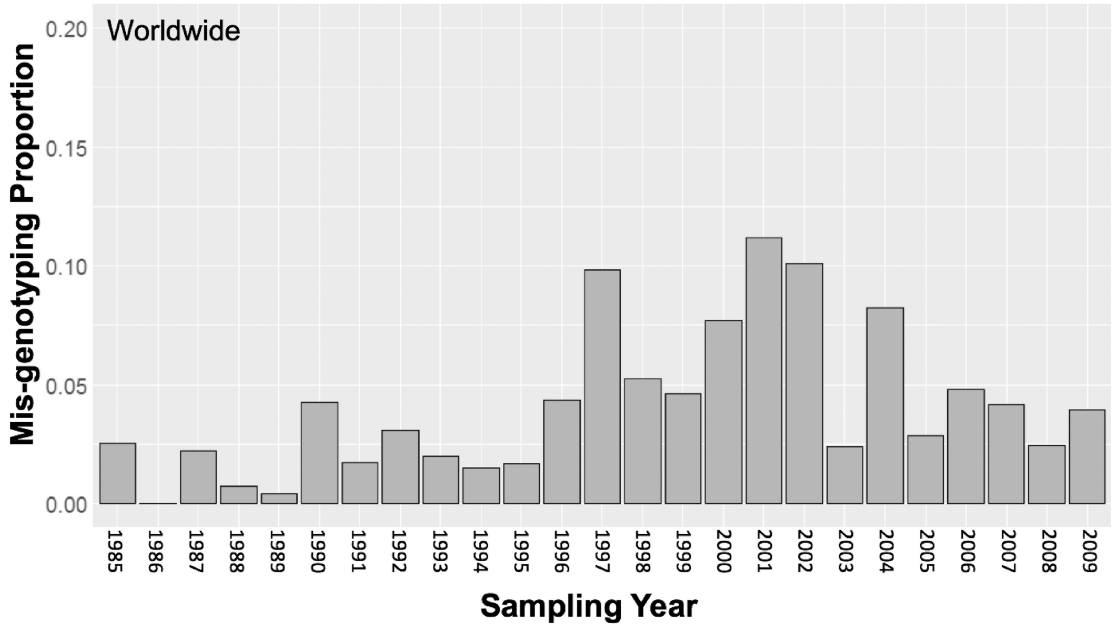
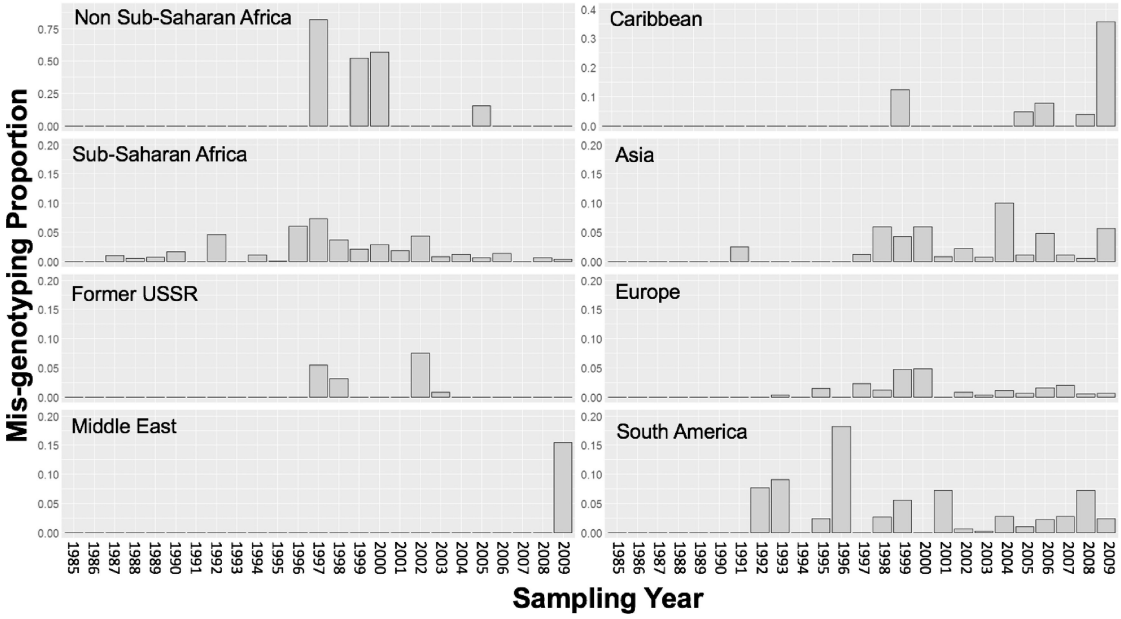


Figure 3. Distribution of mis-genotyped cases across the HIV-1 genome

Each case represents one sequence per individual. To avoid unequal sampling bias across the viral genome, counts of mis-genotyped cases were normalized by the total number of viral sequences in the same genomic region. Both raw count (in blue shade) and normalized count (in black shade) are summarized. Genome map at the bottom: reference genome of HXB2 strain (accession number: K03455).





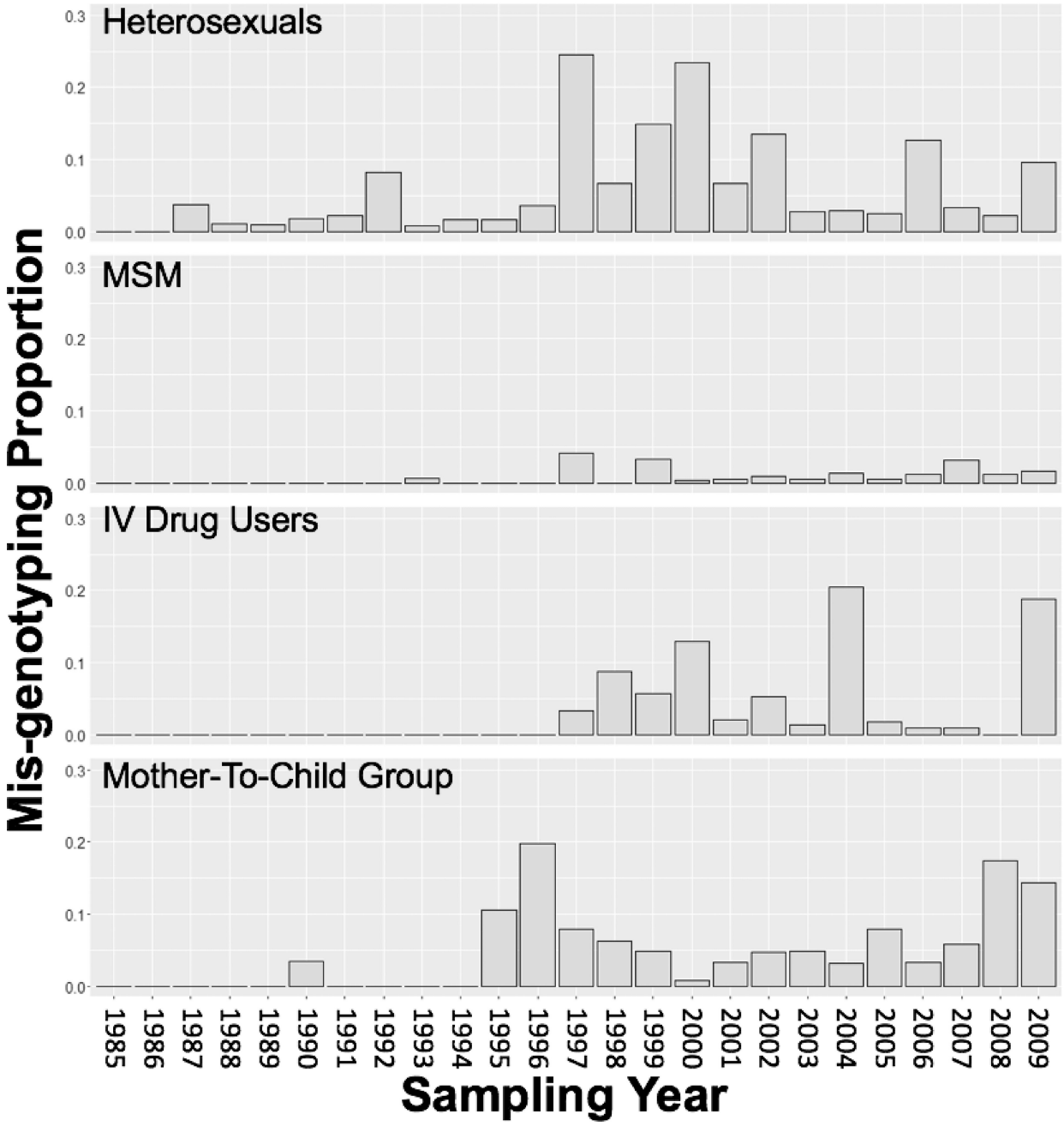


Figure 4. Temporality of mis-genotyped cases

Each case represents one sequence per individual. All cases were normalized by patient and geographic region as described in the Method section. The data spans across 25 years, from 1985 to 2009, during which we were able to obtain adequate sampling information. (A) Worldwide prevalence of mis-genotyped cases during 1985 – 2009. (B) The prevalence of mis-genotyped cases in different geographic regions during 1985 – 2009. (C) The prevalence of mis-genotyped cases within four high-risk infection populations during 1985 – 2009.

TABLE 1

Summary of Sequence Total Retrieved From The Public Domain And Mis-genotyped cases.

Number of sequences retrieved from the public domain	459,881
Number of sequences subject to genotyping	49,175
Number of identified mis-genotyped cases	2,236
Number of mis-genotyped case with transmission history info.	965
Number of mis-genotyped case with sampling country info.	2,234