Robustness of phylogenetic analysis for detecting clusters of new HIV infections

August Guang[1], Mark Howison[2], Mia Coetzer[3], Lauren Ledingham[3], Matt D'Antuono[3], Philip A. Chan[3], Charles Lawrence[4], Casey W. Dunn[5], Rami Kantor[3]

[1] Computing and Information Services, Brown University, Providence, RI, USA

[2] Research Improving People's Lives, Providence, RI, USA

[3] Division of Infectious Diseases, The Alpert Medical School, Brown University, Providence, RI, USA

[4] Division of Applied Mathematics, Brown University, Providence, RI, USA

[5] Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

1 Abstract

2 **Background:** Phylogenetic analysis of HIV sequences obtained as part of clinical care is increasingly applied to

3 detect clustering of new HIV infections and inform public health interventions to disrupt transmission. Conventional

4 approaches summarize the within-host HIV diversity with a single consensus sequence per individual of only the *pol*

5 gene, obtained from Sanger or next-generation sequencing (NGS).

6 **Methods:** We evaluate the robustness of the consensus approach and the potential benefits of considering the within-

7 host diversity in phylogenetic analysis and cluster inference for all newly HIV-diagnosed individuals in the first half

8 of 2013 at the largest HIV center in Rhode Island, USA. We compare Sanger and NGS-derived *pol* and near-whole

9 genome consensus sequences to an alternate approach that samples many sequences per individual from a profile

10 hidden Markov model of their NGS data.

11 **Results:** The space of phylogenies inferred through sampling is multi-modal, suggesting that a consensus-inferred

12 phylogeny is not an appropriate summary of within-host variation. Cluster inference differs in phylogenies from

13 consensus sequences from Sanger and NGS data, and across gene regions.

14 **Discussion:** The choice of sequencing and summarization methods affects the detection of clusters, and should be

15 considered carefully in public health applications of phylogenetic analysis to disrupt HIV transmission.

Background

Clinicians are generally interested in inferring transmission links between HIV-infected individuals to improve HIV treatment and prevention. In the absence of reliable patient contact histories, phylogenetic analysis of viral sequence data can be used to infer transmission clusters [1], under the assumption that two individuals sharing a most recent common ancestor in a phylogeny are more likely to share a transmission link in the real, unobservable transmission network. The application of phylogenetic analysis and cluster inference techniques in public health interventions to disrupt transmission was delineated as one of the four key pillars for achieving the Department of Health and Human Services' recently-announced plan for ending the HIV epidemic in the US [2].

While historically the phylogenetic informativeness of the *pol* region of the HIV genome was initially contested [3, 4], its use is now widespread in phylogenetic analysis and cluster inference, often due to the availability of *pol* sequences from routine clinical genotyping by commercial Sanger sequencing. In a meta study of published studies on HIV cluster infereance, 98 out of 105 (93%) analyzed the *pol* region 5.

The increasing availability of NGS technology has led to longer and deeper sequencing of HIV, and data sets that cover nearly the whole genome with thousands or more reads at each site. Recent evidence suggests improvements in both phylogenetic analysis and cluster inference from near-whole genome HIV sequences obtained with NGS. For example, Yebra *et al.* [6] found that the accuracy of phylogenetic reconstruction and cluster inference on simulated sequences improved with longer genomic regions (with the best accuracy from a gag-pol-env concatenation). Novitsky *et al.* [7] similarly studied the effects on cluster inference of using longer genomic regions from real near-whole genome Sanger sequences, and found that the proportion of sequences in clusters increased with longer sequence regions.

Most phylogenetic methods require a single fully resolved sequence for each individual included in the phylogeny. Accordingly, researchers often summarize the within-host variation present in NGS data sets with a consensus sequence. In the larger context of phylogenetic methods (e.g. beyond their application to HIV sequences), this consensus approach carries an underlying statistical assumption of *low relative entropy* [8]. In the context of HIV, tihs assumption is that the consensus sequence adequately captures the relevant information about the distribution of sequences that might be constructed given the within-host variation measured in deep sequencing through NGS. Some previous studies of HIV transmission dynamics have accounted for this variation with coalescent within-host evolutionary models [9, 10], but such models still assume a consensus sequence as the observed data.

In this study, we examine how well this assumption underlying the consensus approach holds on a data set comprising all newly HIV-diagnosed individual in the first six months of 2013 from the largest HIV center in Rhode Island, USA.

45   We present a new analysis method called *profile sampling* that uses the within-host variation in the NGS data to assess

46   the robustness of the consensus approach for phylogenetic analysis and cluster inference.

Methods

Data collection and sequencing

49   Our study and data collection were approved by Lifespan's Institutional Review Board. We collected viral sequences

50   from 37 individuals newly diagnosed with HIV during 2013 and treated at The Miriam Hospital Immunology Center

51   in Providence, Rhode Island, USA. Inclusion criteria were: (i) HIV-infected adults, 18 years of age or older; (ii)

52   diagnosed with HIV during the first six months of 2013; and (iii) available *pol* sequence from routine drug resistance

53   testing. Patient identifiers were removed and all analysis was conducted with de-identified sequence data.

54   We obtained near-whole genome viral sequences for the 37 participants using both Sanger and NGS sequencing meth-

55   ods. Blood specimens were obtained from participants with their consent and processed to isolate peripheral blood

56   mononuclear cells (PMBC), buffy coats, and plasma. From the collected plasma, total nucleic acid was extracted

57   for genotyping. An in-house genotyping assay was used to generate the near-whole genome based on previously

58   published methods [11, 12]. For each sample, two cDNA templates were generated by SuperscriptIII First Strand

59   Synthesis System (Thermofisher, Carlsbad, CA), followed by eight separate nested PCR reactions; these eight ampli-

60   cons span the near-whole genome of HIV. Final amplicon products were sequenced by the Sanger method using 3100

61   Genetic Analyzer (Applied Biosystems, Foster City, CA) and sequenced by NGS using Nextera XT DNA Library Prep

62   chemistry (Illumina, San Diego, CA) to generate multiplexed libraries for Illumina's MiSeq platform with 250 base

63   paired-end reads. Sanger consensus sequences were generated manually using Sequencher version 5.2.4 (Gene Codes,

64   Ann Arbor, MI) to confirm degenerate nucleotides. NGS data were processed and demultiplexed using BaseSpace

65   cloud based application? (Illumina, San Diego, CA).

Profile sampling

67   We introduce a new approach for incorporating within-host variation into phylogentic analysis, called *profile sampling*.

68   We start by aligning each individual's NGS reads using the hivmmer pipeline [13], which we extended to support

69   near-whole genome HIV data and to perform codon-aware alignment within each gene (pending release as hivmmer

70   version 0.3.0). Briefly, hivmmer performs quality control and error correction in overlapping regions of the read

71 pairs using PEAR version 0.9.11 [14], translates them into each possible reading frame, aligns them in amino acid

72 space to all group M reference sequences from the Los Alamos National Lab HIV Database with the probabilistic

73 multiple sequence aligner HMMER version 3.1b2 [15], and produces a codon frequency table across the near-whole

74 HIV genome, which we refer to as the individual's HIV *profile*.

75 We construct fully-resolved sequences by sampling codons at each site in the genome using the codon frequencies

76 from the profile. The collection of sampled sequences captures the empirical distribution of within-host variation at

77 the codon level. We sample 500 sequences from each individual's profile, then collate the sampled sequences into 500

78 profile-sampled data sets, each having one sampled sequence per individual. We perform phylogenetic inference on

79 each of the 500 profile-sampled data sets by estimating a multiple sequence alignment with mafft version 7.313 [16]

80 and a maximum-likelihood phylogeny with the GTRCAT model and 100 rapid bootstrap replicates using RAxML

81 version 8.2.12 [17]. In addition to the 500 profile-sampled phylogenies, we infer two additional phylogenies from

82 the NGS consensus sequences and Sanger consensus sequences. We perform cluster inference on these phylogenies

83 using Cluster Picker [18] with thresholds of 80% bootstrap support and 4.5% genetic distance. Analysis source code

84 is available from `https://github.com/kantorlab/hiv-profile-sampling`.

85 In addition to performing these analyses on the near-whole genome sequences ("wgs"), we also perform them on

86 subsets of the sequences in three cliniically relevant regions: the protease and reverse transcriptase regions at the

87 beginning of the *pol* gene ("prrt"), the *int* gene, and the *env* gene. The prrt region and *int* are routinely sequenced in

88 clinical care to detect drug resistance mutations and inform clinical choices of anti-retroviral therapy. The *env* region

89 is... Rami, why is this routinely sequenced??

Results

Profile sampling estimates within-host diversity

92 We estimate the within-host diversity for individuals as the average percent difference across all pairwise comparisons

93 of their 500 profile-sampled nucleotide sequences. These pairwise differences are calculated using the Hamming

94 distance [19] (which is also called the *p*-distance in the literature on HIV genetic diversity [5, 20]). Figure 1 shows the

95 estimated percent diversity in each region across individuals, ordered by *env*, which we expected *a priori* to be the most

96 variable region. The largest estimated diversity is in *env* for individual MC28 (4.0%), and *env* has the largest range

97 in estimated diversity (0.2% to 4.0%). The other regions have ranges of 0.2% to 1.9% (prrt), 0.1% to 2.0% (*int*) and

98 0.2% to 2.6% (wgs). These ranges are comparable to prior studies on within-host diversity [21, 22]. These estimates

99 of within-host diversity from the HIV profiles suggest that the consensus approach is discarding potentially-relevant

100 information that is present in the NGS data.

Phylogenetic estimates are sensitive to within-host diversity

102 Next, we investigate the impact of within-host diversity on phylogenetic topology and evolutionary distance estimates.

103 To examine the variation in topology, we calculate the pairwise geodesic distance [23, 24] among the 500 phylogenies

104 from the profile samples, as well as the phylogenies from the NGS and Sanger consensus sequences. Then we perform

105 multi-dimensional scaling (MDS) on the resulting distance matrix to visualize the topological space in two dimensions

106 (Figure 2). The results show that the topolical space is multi-model, and the consensus phylogenies inferred from the

107 consensus sequences do not fully summarize the clustering that occurs in topological space. The prrt region displays

108 four topological clusters, and the NGS and Sanger consensus phylogenies are located in difference clusters. There

109 are two clusters in *int*, and the NGS consensus phylogeny lies at the center of one of the clusters, while the Sanger

110 consensus phylogeny is an outlier. The consensus phlyogenies lie at opposite ends of the primary MDS axis in *env*,

111 and of the secondary MDS axis in the wgs region.

112 To examine the variation in estimated evolutionary distance, we sum the branch lengths within each phylogeny across

113 all branches and across only the tip branches. We visualize the distribution of these branch length sums in Figure

114 3. Overall, the estimates are larger in *env* and the wgs region, and smaller when restricting to only the tip branches.

115 In some cases, the consensus phylogenies provide an adequate summary of the distribution (as in Sanger consensus

116 phylogeny for all branches in *env*, or the NGS consensus phylogeny for tip branches in *env*). In other cases, the

117 consensus phylogenies have estimates that are outliers in the distribution (as in both consensus phylogenies for either

118 all or tip branches in the prrt region).

119 Taken together, these results demonstrate that within-host diversity is associated with variation in phylogenetic esti-

120 mates, and that the consensus approach provides only a point estimate that does not always adeuqatley summarize

121 the underlying variation. For example, the variation can be multi-modal and the point estimates from the consensus

122 phylogenies can be outliers.

Inferred clusters differ by sequencing method and genomic region

124 Finally, we examine the clusters that are detected in the NGS vs. Sanger consensus phylogenies, and across the four

125 regions. The NGS consensus phylogeny detects 8 clusters in prrt, 9 in *int*, 4 in *env*, and 6 in wgs. Each cluster appears

as a colored bar in the phylograms in Figure 4. We use the proportion of times a cluster appears across the 500 profile-sampled phylogenies as a measure of support for that cluster, and report it next to the colored bar. Some clusters are robustly detected across all regions, such as the cluster (MC25, MC26, MC52), which has support between 98% and 100%. Other clusters are detected at lower support, and are not detected consistently across regions. For example, cluster (MC14, MC59) only appears in the wgs phylogeny, and at 62% support.

The Sanger consensus phylogeny detects 8 clusters in prrt, 7 in *int*, 6 in *env*, and 7 in wgs (Figure 5). As with the NGS consensus phylogeny, the support and consistency of cluster detection varies across regions. Cluster detection is mostly consistent with the NGS consensus, except for the cluster (MC17, MC21), which only occurs in *env* in the Sanger consensus, and has zero support from the profile samples. Three clusters are consistently detected across NGS and Sanger, and across all regions: (MC23, MC24); (MC25, MC26, MC52); and (MC27, MC28).

Discussion

Because previous studies of the HIV epidemic have shown high intra-patient variation, and because we estimate similar within-host diversity using HIV profiles of our NGS data,

A limitation of our study is the small number of participants. The participants do have a dense temporal sampling, and comprise all newly HIV-diagnosed individuals in a six month period at the largest HIV center in Rhode Island, USA, who met the inclusion criteria. However, the overall size of the HIV epidemic in Rhode Island is much larger, estimated as 2,396 individuals in 2016 [25]. In future work, we hope to apply the profile sampling method to larger NGS data sets, to assess if cluster inference from Sanger versus NGS data agrees or differs especially for clusters larger than three individuals, which was the largest cluster size in our current data set.

Our construction of HIV profiles from NGS data is limited by the accuracy of the NGS assays themselves. The codon frequencies in the profiles may be biased measures of the true within-host diversity because of biases in PCR amplification during sample preparation. Sequencing protocols such as Primer ID [26] have been introduced to reduce and correct for these biases. Unfortunately, the Primer ID protocol supports only a limited region of the HIV genome, and has not yet been extended to support near-whole genome sequencing of HIV, but future work could study this limited region to determine if cluster inference changes with the implementation of a Primer ID protocol.

The true HIV transmission network is unknown, but phylogenetic analysis and cluster inference are promising tools for aiding clinicians and public health officials [2]. Current phylogenetic approaches do not fully utilize the information on within-host diversity available in near-whole genome NGS data. As NGS data sets are increasingly available and

more fully sample from the current HIV epidemic,

References

1. Leitner, T, Escanilla, D, Franzen, C, Uhlen, M, and Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences 1996;93:10864–10869.

2. Fauci, AS, Redfield, RR, Sigounas, G, Weahkee, MD, and Giroir, BP. Ending the HIV Epidemic: A Plan for the United States. JAMA 2019;321:844.

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

4. Stürmer, M, Preiser, W, Gute, P, Nisius, G, and Doerr, HW. Phylogenetic analysis of HIV-1 transmission: pol gene sequences are insufficient to clarify true relationships between patient isolates. AIDS 2004;18:2109–2113.

5. Hassan, AS, Pybus, OG, Sanders, EJ, Albert, J, and Esbjörnsson, J. Defining HIV-1 transmission clusters based on sequence data: AIDS 2017;31:1211–1222.

6. Yebra, G, Hodcroft, EB, Ragonnet-Cronin, ML, et al. Using nearly full-genome HIV sequence data improves phylogeny reconstruction in a simulated epidemic. Scientific Reports 2016;6:39489.

7. Novitsky, V, Moyo, S, Lei, Q, DeGruttola, V, and Essex, M. Importance of Viral Sequence Length and Number of Variable and Informative Sites in Analysis of HIV Clustering. AIDS Research and Human Retroviruses 2015;31:531–542.

8. Guang, A, Zapata, F, Howison, M, Lawrence, CE, and Dunn, CW. An Integrated Perspective on Phylogenetic Workflows. Trends in Ecology & Evolution 2016;31:116–126.

9. Giardina, F, Romero-Severson, EO, Albert, J, Britton, T, and Leitner, T. Inference of Transmission Network Structure from HIV Phylogenetic Trees. PLOS Computational Biology 2017;13:e1005316.

10. Romero-Severson, E, Skar, H, Bulla, I, Albert, J, and Leitner, T. Timing and Order of Transmission Events Is Not Directly Reflected in a Pathogen Phylogeny. Molecular Biology and Evolution 2014;31:2472–2482.

11. Nadai, Y, Eyzaguirre, LM, Constantine, NT, et al. Protocol for Nearly Full-Length Sequencing of HIV-1 RNA from Plasma. PLoS ONE 2008;3:e1420.

12. Di Giallonardo, FD, Töpfer, A, Rey, M, et al. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. Nucleic Acids Research 2014;42:e115–e115.

7

13. Howison, M, Coetzer, M, and Kantor, R. Measurement error and variant-calling in deep Illumina sequencing of HIV. Bioinformatics 2019;35:2029–2035.

14. Zhang, J, Kobert, K, Flouri, T, and Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 2014;30:614–620.

15. Eddy, SR. Accelerated Profile HMM Searches. PLOS Computational Biology 2011;7:e1002195.

16. Katoh, K and Standley, DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 2013;30:772–780.

17. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics 2014:btu033.

18. Ragonnet-Cronin, M, Hodcroft, E, Hué, S, et al. Automated analysis of phylogenetic clusters. BMC Bioinformatics 2013;14:317.

19. Allam, O, Samarani, S, and Ahmad, A. Hammering out HIV-1 incidence with Hamming distance: AIDS 2011;25:2047–2048.

20. Maldarelli, F, Kearney, M, Palmer, S, et al. HIV Populations Are Large and Accumulate High Genetic Diversity in a Nonlinear Fashion. Journal of Virology 2013;87:10313–10323.

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

22. Zanini, F, Brodin, J, Thebo, L, et al. Population genomics of intrapatient HIV-1 evolution. eLife 2015;4:e11282.

23. Billera, LJ, Holmes, SP, and Vogtmann, K. Geometry of the Space of Phylogenetic Trees. Advances in Applied Mathematics 2001;27:733–767.

24. Owen, M and Provan, JS. A Fast Algorithm for Computing Geodesic Distances in Tree Space. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2011;8:2–13.

25. Rhode Island Department of Health. HIV Progress [Internet]. Available from: https://health.ri.gov/data/hiv/. [Accessed 12 Dec 2019]. 2019.

26. Jabara, CB, Jones, CD, Roach, J, Anderson, JA, and Swanstrom, R. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proceedings of the National Academy of Sciences 2011;108:20166–20171.

Figure 1: Intra-patient genetic diversity (defined as the average percent difference across all pairwise comparisons of the 500 profile-sampled nucleotide sequences for a patient) is highest in the env region for most individuals, and lies within the range of previously reported values.
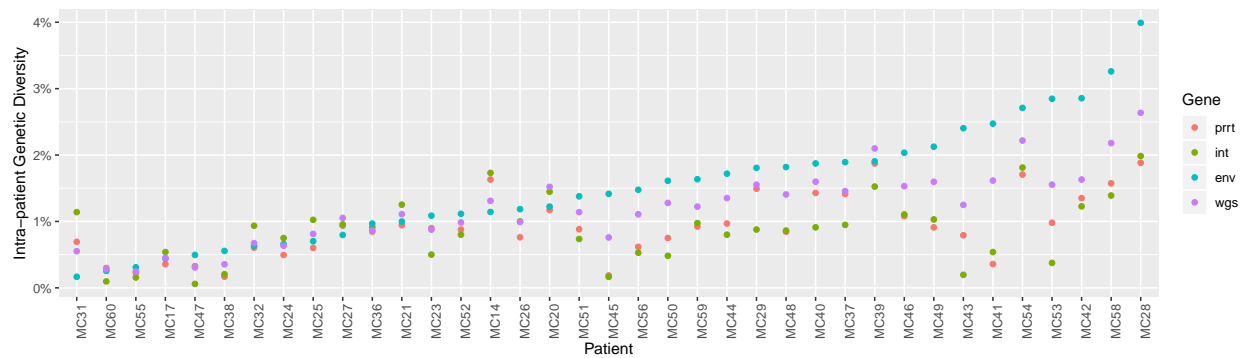

Figure 2: Multi-dimensional scaling of pairwise geodesic distance among maximum-likelihood phylogenies from the profile-sampling approach show that the space of phylogenies inferred for the prrt and int regions are multi-modal. The phylogenies from consensus sequences (black dots) are point estimates that do not capture the full variation in phylogenies that can be inferred from deeply-sequenced NGS data.
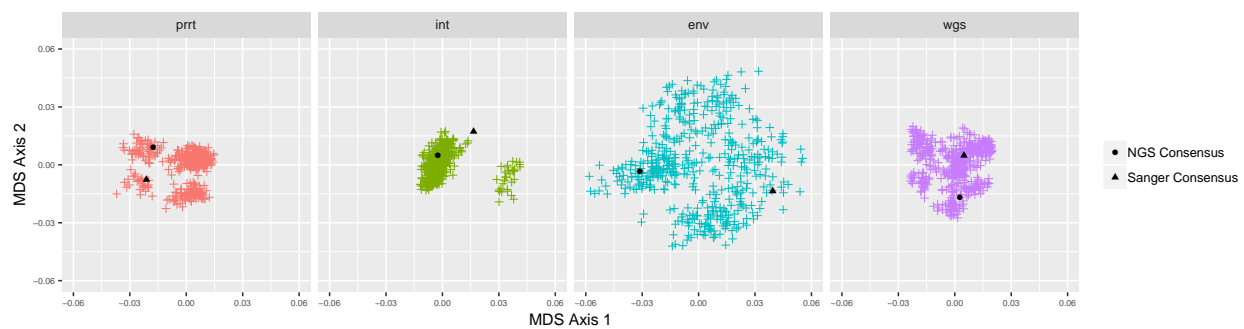
Figure 3: The total branch length in each of the profile-sampled phylogenies also varies. The phylogenies from consensus sequences (black dots) can lie at extreme values within these distribution, both when considering the lengths across all branches and the lengths across only the branches at the tips.
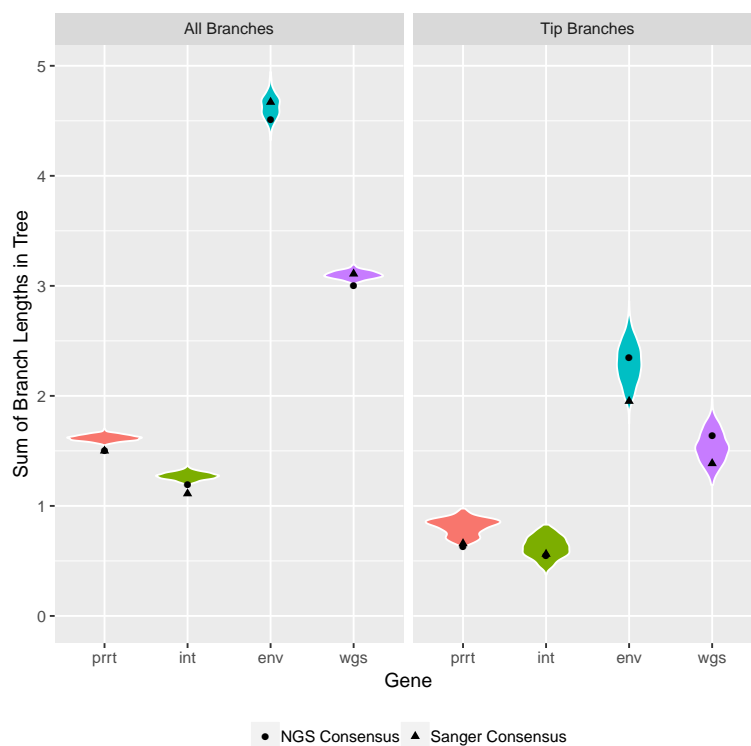
Figure 4: Clusters (vertical colored bars) inferred from the phylogenies of NGS consensus sequences differ across genomic regions. The largest number of clusters was inferred from the int region, and the smallest number from the env region. Profile sampling provides a bootstrapped measure of cluster support (annotation to bars).
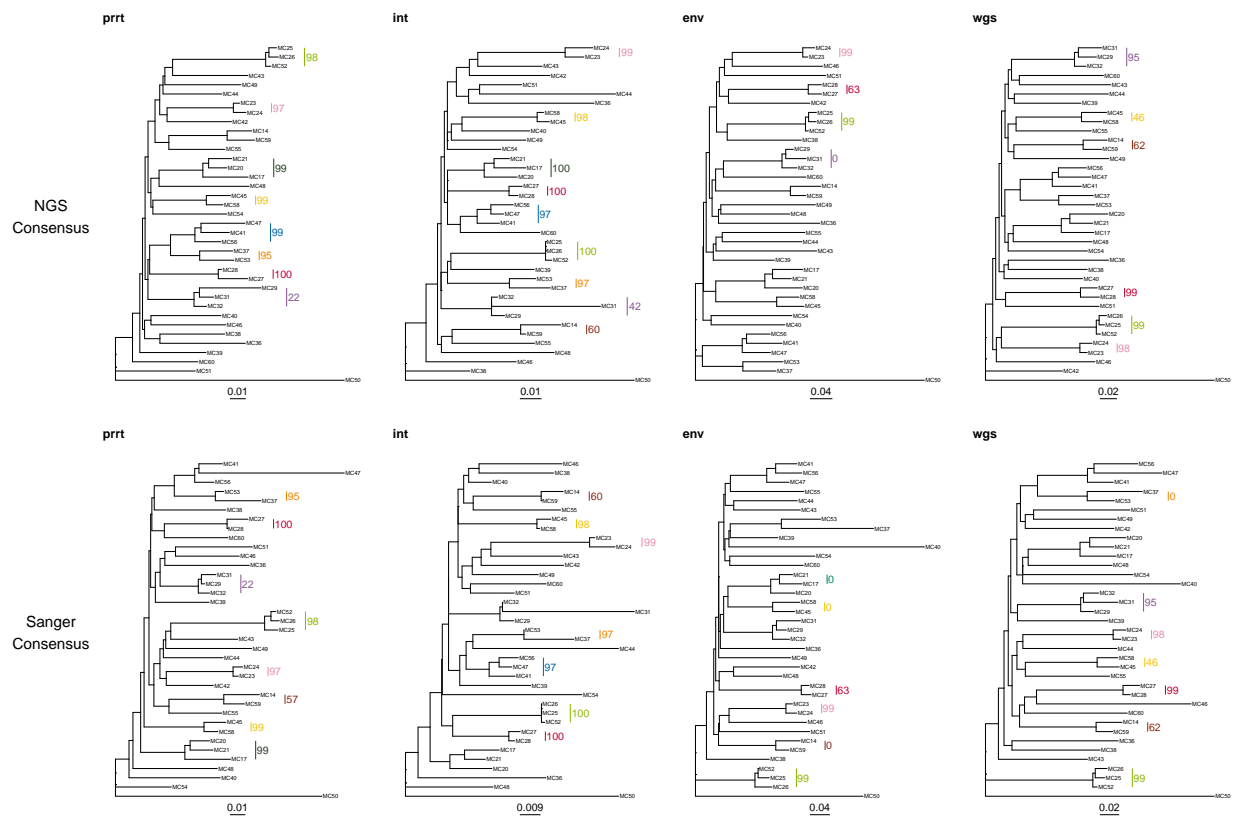
11

Figure 5: Summary of clusters identified by Sanger versus NGS consensus sequences across genomic regions. Numeric values indicate bootstrapped cluster support from the profile sampling method.