

# Robustness of phylogenetic analysis for detecting clusters of new HIV infections

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

<sup>1</sup> Computing and Information Services, Brown University, Providence, RI, USA

<sup>2</sup> Research Improving People's Lives, Providence, RI, USA

<sup>3</sup> Division of Infectious Diseases, The Alpert Medical School, Brown University, Providence, RI, USA

<sup>4</sup> Division of Applied Mathematics, Brown University, Providence, RI, USA

<sup>5</sup> Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA

## Abstract

**Background:** Phylogenetic analysis of HIV sequences obtained as part of clinical care is increasingly applied to detect clustering of new HIV infections and inform public health interventions to disrupt transmission. Conventional approaches summarize the within-host HIV diversity with a single consensus sequence per individual of only the *pol* gene, obtained from Sanger or next-generation sequencing (NGS).

**Methods:** We evaluate the robustness of the consensus approach and the potential benefits of considering the within-host diversity in phylogenetic analysis and cluster inference for all newly HIV-diagnosed individuals in the first half of 2013 at the largest HIV center in Rhode Island, USA. We compare Sanger and NGS-derived *pol* and near-whole genome consensus sequences to an alternate approach that samples many sequences per individual from a profile hidden Markov model of their NGS data.

**Results:** The space of phylogenies inferred through sampling is multi-modal, suggesting that a consensus-inferred phylogeny is not an appropriate summary of within-host variation. Cluster inference differs in phylogenies from consensus sequences from Sanger and NGS data, and across gene regions.

**Discussion:** The choice of sequencing and summarization methods affects the detection of clusters, and should be 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 recent meta study, Novitsky *et al.* (forthcoming) found that 102 out of 107 published studies of HIV cluster inference between 2016 and 2019 used the *pol* region alone.

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.* [5] 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.* [6] 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* [7]. In the context of HIV, this 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 [8, 9], 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. We present a new analysis method called *profile sampling* that uses the within-host variation in the NGS data to assess the robustness of the consensus approach for phylogenetic analysis and cluster inference.

## Methods

### Data collection and sequencing

Our study and data collection were approved by Lifespan's Institutional Review Board. We collected viral sequences from 37 individuals newly diagnosed with HIV during 2013 and treated at The Miriam Hospital Immunology Center in Providence, Rhode Island, USA. Inclusion criteria were: (i) HIV-infected adults, 18 years of age or older; (ii) diagnosed with HIV during the first six months of 2013; and (iii) available *pol* sequence from routine drug resistance testing. Patient identifiers were removed and all analysis was conducted with de-identified sequence data.

We obtained near-whole genome viral sequences for the 37 participants using both Sanger and NGS sequencing methods. Blood specimens were obtained from participants with their consent and processed to isolate peripheral blood mononuclear cells (PMBC), buffy coats, and plasma. From the collected plasma, total nucleic acid was extracted for genotyping. An in-house genotyping assay was used to generate the near-whole genome based on previously published methods [10, 11]. For each sample, two cDNA templates were generated by SuperscriptIII First Strand Synthesis System (Thermofisher, Carlsbad, CA), followed by eight separate nested PCR reactions; these eight amplicons span the near-whole genome of HIV. Final amplicon products were sequenced by the Sanger method using 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) and sequenced by NGS using Nextera XT DNA Library Prep chemistry (Illumina, San Diego, CA) to generate multiplexed libraries for Illumina's MiSeq platform with 250 base paired-end reads. Sanger consensus sequences were generated manually using Sequencher version 5.2.4 (Gene Codes, Ann Arbor, MI) to confirm degenerate nucleotides. NGS data were processed and demultiplexed using BaseSpace cloud based application? (Illumina, San Diego, CA).

### Profile sampling

We introduce a new approach for incorporating within-host variation into phylogenetic analysis, called *profile sampling*. We start by aligning each individual's NGS reads using the hivmmer pipeline [12], which we extended to support near-whole genome HIV data and to perform codon-aware alignment within each gene (pending release as hivmmer

version 0.3.0). Briefly, hivmmer performs quality control and error correction in overlapping regions of the read pairs using PEAR version 0.9.11 [13], translates them into each possible reading frame, aligns them in amino acid space to all group M reference sequences from the Los Alamos National Lab HIV Database with the probabilistic multiple sequence aligner HMMER version 3.1b2 [14], and produces a codon frequency table across the near-whole HIV genome, which we refer to as the individual's HIV *profile*.

We construct fully-resolved sequences by sampling codons at each site in the genome using the codon frequencies from the profile. The collection of sampled sequences captures the empirical distribution of within-host variation at the codon level. We sample 500 sequences from each individual's profile, then collate the sampled sequences into 500 profile-sampled data sets, each having one sampled sequence per individual. We perform phylogenetic inference on each of the 500 profile-sampled data sets by estimating a multiple sequence alignment with mafft version 7.313 [15] and a maximum-likelihood phylogeny with the GTRCAT model and 100 rapid bootstrap replicates using RAxML version 8.2.12 [16]. In addition to the 500 profile-sampled phylogenies, we infer two additional phylogenies from the NGS consensus sequences and Sanger consensus sequences. We perform cluster inference on these phylogenies using Cluster Picker [17] with thresholds of 80% bootstrap support and 4.5% genetic distance. Analysis source code is available from <https://github.com/kantorlab/hiv-profile-sampling>.

In addition to performing these analyses on the near-whole genome sequences ("wgs"), we also perform them on subsets of the sequences in three clinically relevant regions: the protease and reverse transcriptase regions at the beginning of the *pol* gene ("prtr"), the *int* gene, and the *env* gene. The prtr region and *int* are routinely sequenced in clinical care to detect drug resistance mutations and inform clinical choices of anti-retroviral therapy. The *env* region is... Rami, why is this routinely sequenced??

## Results

### Profile sampling captures within-host diversity

We examine how much

Profile sampling allows us to visualize the degree of polymorphism within a patient (Figure 1) and sample genomic sequences in proportion to the within-host variation. It can also build a consensus sequence for easy comparison of new approaches to established methods. By visualizing the degree of polymorphism, we can assess whether genomic variation exists at significantly high enough frequencies such that neither consensus genomes nor any other point

estimator are sufficient summaries.

Figure 1 shows substantial variation exists in the HIV profiles of all patients. Across all patients, we found that 5.63 percent of sites in the HIV genome are polymorphic, with 11.4 percent of those sites in the gag region, 21.4 percent of those sites in the pol region, and 29.5 percent of those sites in the env region of the genome. Here we define polymorphism to mean that a particular site has greater than 1

## Phylogenetic estimates are sensitive to inter-host diversity

Given that previous studies on individuals infected with HIV have shown high intra-patient variation, and given that we found similar results in our set of data (Figure 1), we investigated its impact on downstream topological variation in the phylogenetic tree. The results show that trees inferred from the profile-sampled sequences do not cluster around any of the trees inferred from a genome point estimate (Figure 2).

When an MDS plot is generated for only the profile-sampled trees (Figure 3), they appeared to form two clusters. This is likely due to zoom on the profile-sampled trees, as can be seen from the axes scales for Figure 2 and Figure 3. Supertrees of the two clusters reveal that most of the clades in the two subtrees are the same, with only patients out of 37 having a different topological placement on the two trees (Figure 4). This suggests that distinct clusters of phylogenetic trees exist given the intra-patient variation, but that the differences between them are not large. However, the consensus genome appears to be an inadequate estimate summary for the intra-patient variation, as can be seen in Figure 5.

## Inferred clusters differ by sequencing method and genomic region

## Discussion

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 [18]. 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 19 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.

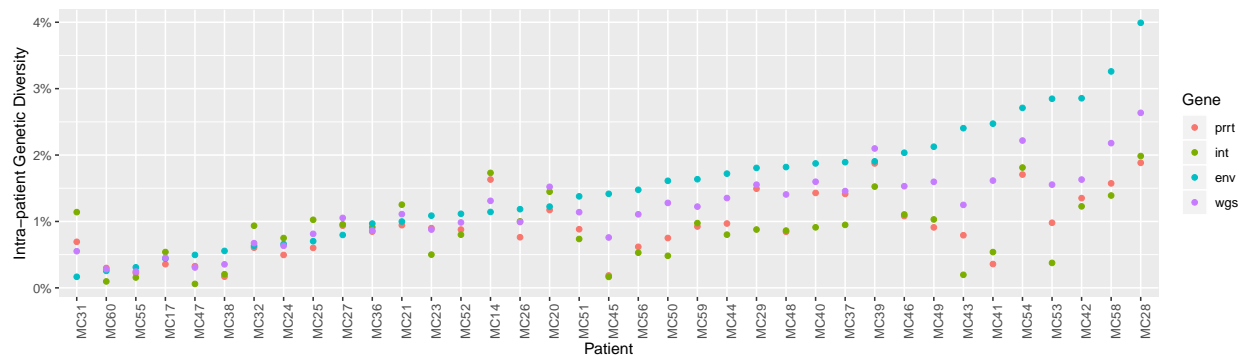
The true HIV transmission network is unknown, but phylogenetic analysis and cluster inference are promising tools for aiding clinicians and public health officials. Current phylogenetic approaches do not fully utilize the information on within-host diversity available from deeper and longer NGS sequencing technologies. As NGS data sets are increasingly available and more fully sample from the current HIV epidemic,

## References

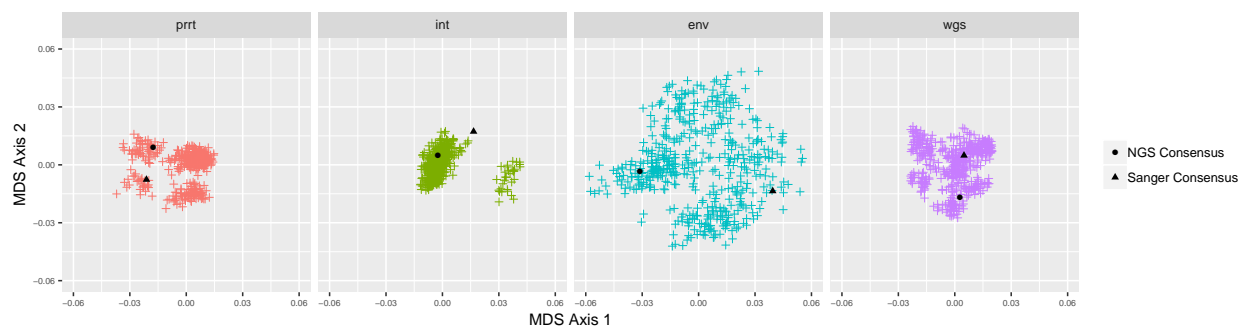
1. Leitner, T, Escanilla, D, Franzen, C, Uhlen, M, and Albert, J. Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. *Proceedings of the National Academy of Sciences* 1996;93:10864–10869.
2. Fauci, AS, Redfield, RR, Sigounas, G, Weahkee, MD, and Giroir, BP. Ending 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. 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.
6. 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.
7. 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.

- 150 8. Giardina, F, Romero-Severson, EO, Albert, J, Britton, T, and Leitner, T. Inference of Transmission Network  
151 Structure from HIV Phylogenetic Trees. *PLOS Computational Biology* 2017;13:e1005316.
- 152 9. Romero-Severson, E, Skar, H, Bulla, I, Albert, J, and Leitner, T. Timing and Order of Transmission Events Is  
153 Not Directly Reflected in a Pathogen Phylogeny. *Molecular Biology and Evolution* 2014;31:2472–2482.
- 154 10. Nadai, Y, Eyzaguirre, LM, Constantine, NT, et al. Protocol for Nearly Full-Length Sequencing of HIV-1 RNA  
155 from Plasma. *PLoS ONE* 2008;3:e1420.
- 156 11. Di Giallonardo, FD, Töpfer, A, Rey, M, et al. Full-length haplotype reconstruction to infer the structure of  
157 heterogeneous virus populations. *Nucleic Acids Research* 2014;42:e115–e115.
- 158 12. Howison, M, Coetzer, M, and Kantor, R. Measurement error and variant-calling in deep Illumina sequencing of  
159 HIV. *Bioinformatics* 2019;35:2029–2035.
- 160 13. Zhang, J, Kobert, K, Flouri, T, and Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR.  
161 *Bioinformatics* 2014;30:614–620.
- 162 14. Eddy, SR. Accelerated Profile HMM Searches. *PLOS Computational Biology* 2011;7:e1002195.
- 163 15. Katoh, K and Standley, DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Per-  
164 formance and Usability. *Molecular Biology and Evolution* 2013;30:772–780.
- 165 16. Stamatakis, A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.  
166 *Bioinformatics* 2014:btu033.
- 167 17. Ragonnet-Cronin, M, Hodcroft, E, Hué, S, et al. Automated analysis of phylogenetic clusters. *BMC Bioinfor-*  
168 *matics* 2013;14:317.
- 169 18. Rhode Island Department of Health. HIV Progress [Internet]. Available from: <https://health.ri.gov/data/hiv/>. [Ac-  
170 cessed 12 Dec 2019]. 2019.
- 171 19. Jabara, CB, Jones, CD, Roach, J, Anderson, JA, and Swanstrom, R. Accurate sampling and deep sequencing of  
172 the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences* 2011;108:20166–  
173 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.

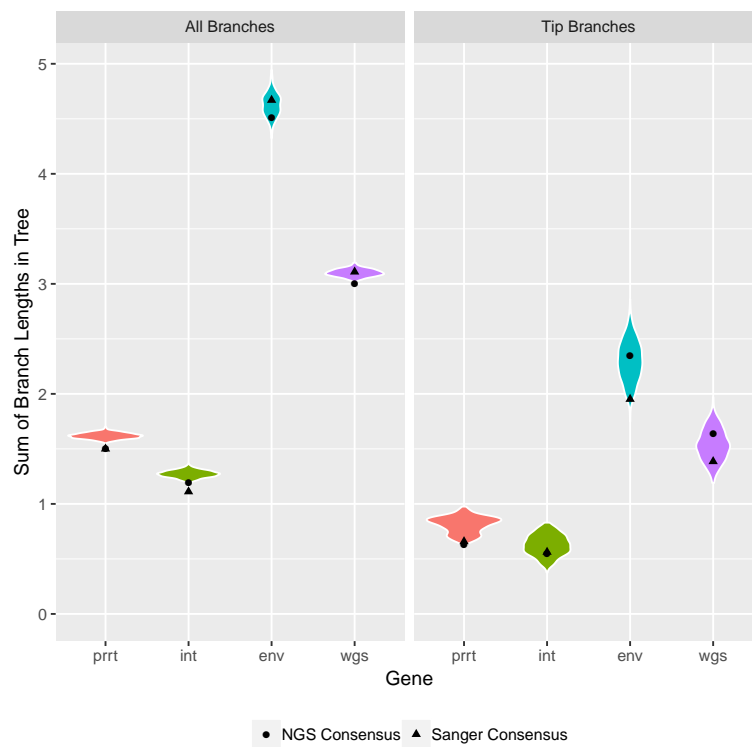


**Figure 2:** Multi-dimensional scaling of pairwise geodesic distance among maximum-likelihood phylogenies from the profile-sampling approach show that the space of phylogenies inferred for the prrt and int regions are multi-modal. The phylogenies from consensus sequences (black dots) are point estimates that do not capture the full variation in phylogenies that can be inferred from deeply-sequenced NGS data.

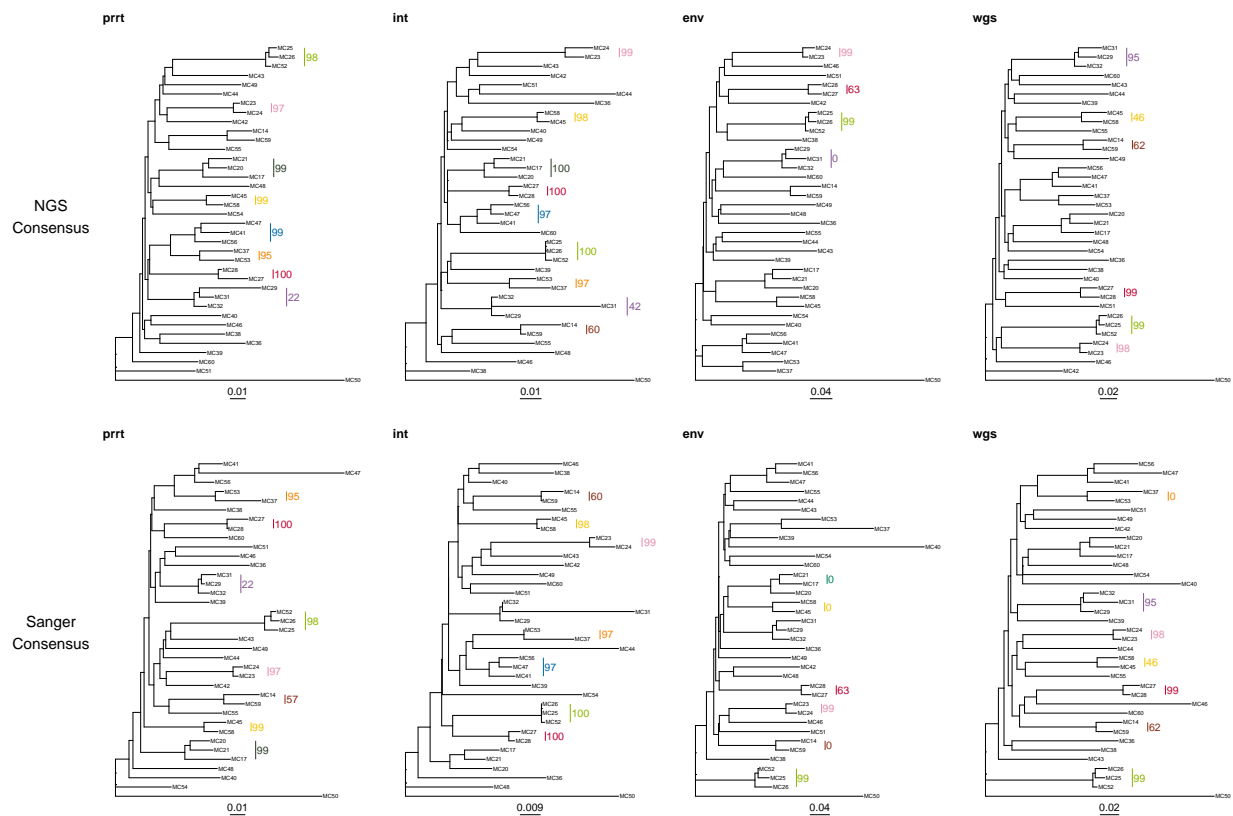




**Figure 3:** The total branch length in each of the profile-sampled phylogenies also varies. The phylogenies from consensus sequences (black dots) can lie at extreme values within these distribution, both when considering the lengths across all branches and the lengths across only the branches at the 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).



**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.

	NGS Consensus				Sanger Consensus			
	prrt	int	env	wgs	prrt	int	env	wgs
MC17,MC21 -							0	
MC29,MC31,MC32 -	21.8	42.2	0.4	95.2	21.8			95.2
MC14,MC59 -		59.6		62	56.6	59.6	0	62
MC37,MC53 -	94.6	96.8			94.6	96.8		0
MC23,MC24 -	96.6	99.4	99.4	98.4	96.6	99.4	99.4	98.4
MC25,MC26,MC52 -	97.8	99.6	99.4	99	97.8	99.6	99.4	99
MC17,MC20,MC21 -	99	99.8			99			
MC41,MC47,MC56 -	99.2	97.4				97.4		
MC45,MC58 -	99.2	98.4		46	99.2	98.4	0	46
MC27,MC28 -	100	100	63.4	99	100	100	63.4	99