An Improved Circular Consensus Algorithm with an Application to Detect HIV-1 Drug-Resistance Associated Mutations (DRAMs)

N Lance Hepler[1*], Nigel Delaney[1*], Michael Brown[1], Melissa Laird Smith[1], David Katzenstein[2], Ellen E. Paxinos[1], David Alexander[1]
[1] PacBio, 1380 Willow Road, Menlo Park, CA 94025
[2] Stanford University Medical Centre, Division of Infectious Disease Stanford, CA 94305-5107

Introduction

- Confident resolution of heterogeneous, complex populations continues to rely on cost- and labor-intensive Sanger sequencing methods.
- Long, single-molecule sequencing reads can now be analyzed using a redesigned circular consensus sequencing algorithm (CCS2) to generate high-quality reads across longer insert lengths.
- Here we validate this approach with the characterization of the HIV-1 K103N drug resistance associated mutation (DRAM), which has proven challenging to traditional detection methods.

Materials and Molecular Methods

- A region spanning ~1300 bp of the HIV *RT* gene was PCR-amplified from non-mutated (“wild”) and mutant (“K103N”) plasmids and several primary samples.

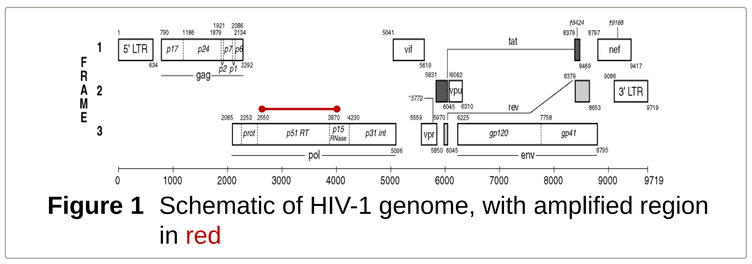

Figure 1 Schematic of HIV-1 genome, with amplified region in red

- An A-to-C mutation was introduced at the 3rd position of residue 103 in the RT gene of a clonal HIV-1 strain (“wild”) using site-directed mutagenesis (Fig 1).
- Sequence surrounding position 103 of the RT is difficult to resolve due to a 6-A homopolymer.
- Detection of this variant was further tested by sequencing sequential primary samples from an individual who had failed an NNTRI containing therapy.
- Subject 24’s virus contained an even longer homopolymer of KKKK, instead of KQKK) at baseline, which mutated to KKNK at therapy failure.

SAMPLE	100		101		102		103		104		105		RT aa Position
"wild" control	T	T	A A A A	C	A	G	A A A	A	A A	A	T	C	A	base
	L		K		Q		K		K		S			amino acid
K103N control	T	T	A A A A	C	A	G	A A	C	A A A		T	C	A	base
	L		K		Q		N		K		S			amino acid
p24-t1	T	T	A A A A A	A	G		A A A	A	A A	A	T	C	A	base
	L		K		K		K		K		S			amino acid
p24-t3	T	T	A A A A A	A	G		A A	C	A A A		T	C	A	base
	L		K		K		N		K		S			amino acid

Table 1 Table showing the homopolymeric region surrounding the clinically important K103N variant in the *RT* gene from samples in this study.

- Sequencing was performed on each sample using a PacBio RS II with P6-C4 chemistry and 4 hour movies.

Algorithm Development

- Sequencing data were analyzed using the new CCS2 algorithm, which uses a hybrid discriminative/generative probabilistic model of the SMRT Sequencing process (“Arrow”) to polish consensus sequences to high accuracy.
- The Arrow model achieves this result by adjusting the model parameters using fixed covariates taken from each ZMW, thus accounting for the differences in the SMRT Sequencing process for each molecule.

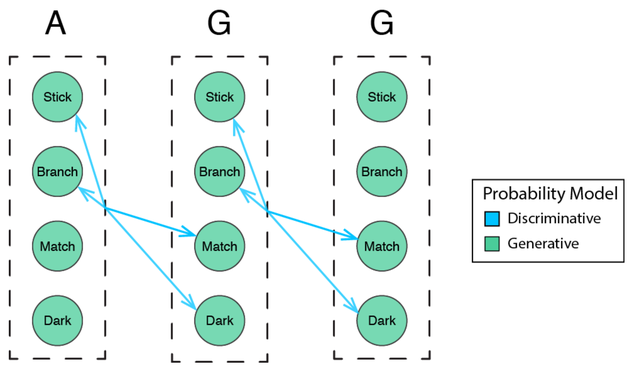

Figure 2 Diagram of the “Arrow” model over an example AGG template context. Branch and stick refer to homopolymeric and non-homopolymeric insertion events, respectively, while dark is synonymous with deletion. Transition parameters vary based on fixed per-reaction covariates.

- Improvements in empirical accuracy using the hybrid discriminative/generative model in CCS2, achieves a per-read empirical quality of QV30 with just 15X coverage.

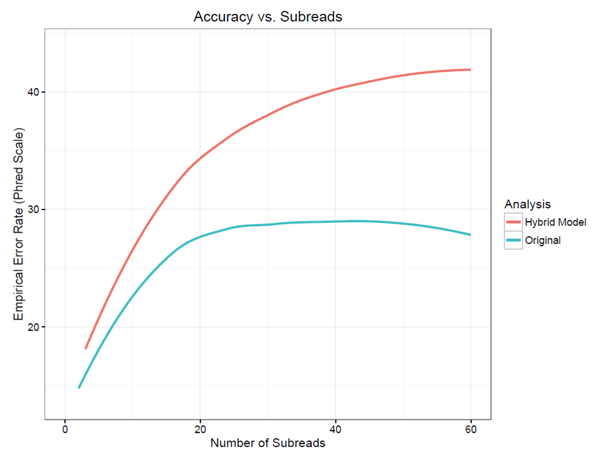

Figure 3 The original Quiver-based algorithm was unable to model the sequencing reaction on a per-molecule basis, and was therefore unable to achieve QV30 at any coverage. The new hybrid Arrow-based CCS2 achieves high accuracy as coverage increases.

Results

- More than 5000 1.3 kb consensus sequences with a collective empirical quality of ~QV40 were generated for each sample.
- We demonstrate a 0% miscall rate in both unmixed samples, and estimate a 49.7% frequency for the K103N mutation in the mixed sample, consistent with data produced by orthogonal platforms.

		% K103N detected			
Sample	Time point	PacBio RSII Quiver	PacBio RSII Arrow	Illumina MiSeq	Sanger detection?
50/50 Mix		50.4%	49.7%	56.0%	Y
Subject 24	T1	0	0	0	N
	T3	15.3%	15.3%	19.1%	2

Table 2 Results of K103N quantification from synthetic 50/50 mixture and primary samples from subject 24.

- Topology was generated using a GTR+gamma evolutionary model in `FastTree`.
- Tertiary analysis of more than 10,000 sequences comprising the two time points show evolution of the viral population, as well as the emergence of the resistance mutation in several well-supported clusters, suggesting the mutation arose in more than one backbone.

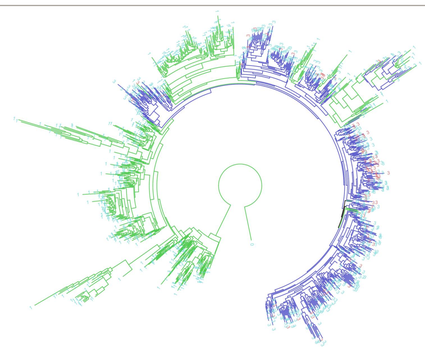

Figure 4 Phylogenetic relationships of sequences at baseline (t=1 in Green, and after K103N failure (t=3, in blue) in subject p24.

Conclusions

- We demonstrate high-quality consensus for single molecules by modeling the sequencing reaction using the new Arrow modeling framework.
- We demonstrate accurate quantitation of mixtures of the K103N DRAM both in a well-characterized mixture and primary samples, as compared to data from other platforms.
- Single-molecule resolution of viral populations reveals parallel emergence of the K103N mutation in response to drug therapy.

Acknowledgements

The authors would like to thank Justin De La Cruz, Benjamin Pinsky, Colleen Ludka, and Yan Guo for sample processing.