#418 June 19th, 2016

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

Melissa Laird Smith[1], Nigel Delany[1], N. Lance Hepler[1], David Alexander[1], David Katzenstein[2], Michael Brown[1] and Ellen Paxinos[1]
[1]PacBio, 1380 Willow Road, Menlo Park, CA 94025; [2]Stanford University School of Medicine, 291 Campus Drive, Stanford, CA

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.
- This new approach was validated with the characterization of the HIV-1 K103N drug resistance-associated mutation in both clonal and patient samples.
- The K103N DRAM is clinically relevant, but has proven technically challenging to resolve due to regional sequence context.

Materials

- A region spanning ~1.3 kb of the RT region of the pol gene of HIV was PCR-amplified from clonal and several primary samples.

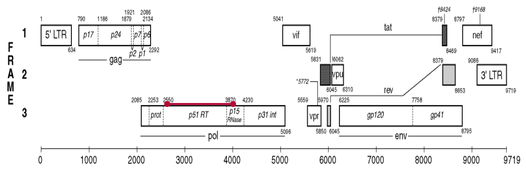


Figure 1. Schematic of HIV-1 genome, with amplified region in red.

- Detection of this variant was tested using two sample sets **(Table 1):**
  - Clonal wild-type and K103N variant amplicons
  - Sequential primary samples from an individual who had failed an NNTRI-containing therapy

Methods

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

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	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	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

- > 5000 1.3 kb consensus sequences with a collective empirical quality of ~QV40 were generated for each sample.

Algorithm Development

- Sequencing data were analyzed using the new CCS2 algorithm, called Arrow.
- Arrow uses a hybrid discriminative/generative probabilistic model to polish consensus sequences for high accuracy.
- Arrow model achieves high accuracy by adjusting the model parameters using fixed covariates taken from the sequencing data of each molecule(Fig **2).**
- The new hybrid-Arrow-based CCS2 algorithm achieves high accuracy as coverage increases **(Fig 3)**

Arrow Model & Enhanced Accuracy

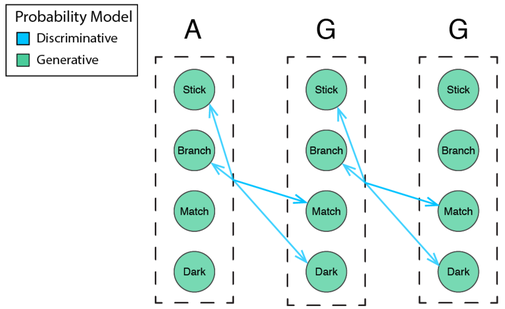


Figure 2. Diagram of the "Arrow" model over an example AGG template context.

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.

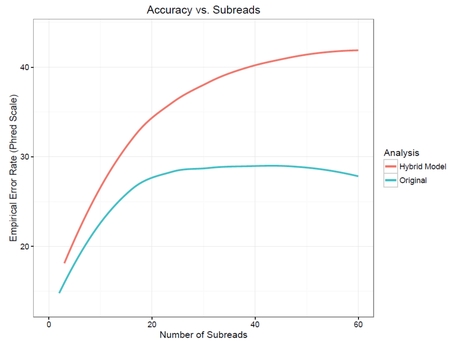


Figure 3. Arrow vs. Quiver per-read empirical QV by coverage

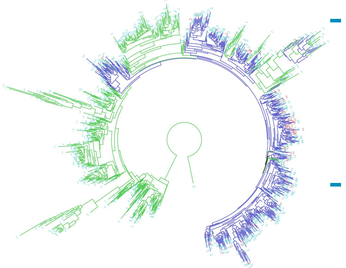
- Arrow achieves a per-read empirical QV30 at 15-fold coverage.
- Previous Quiver-based algorithm unable to model sequencing reactions on a per-molecule basis, and therefore could not achieve QV30 single-molecule consensus accuracy at any coverage.

HIV K103N DRAM Detection

- 0% miscall rate observed in both unmixed samples
- K103N mutation frequency estimated at 49.7% in the mixed sample, consistent with orthogonal data

		% K103N DETECTED			
SAMPLE	TIME POINT	PACBIO RSII QUIVER	PACBIO RS II 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



- Tertiary analysis of >10,000 sequences shows viral population evolution and the K103N emergence in several well-supported clusters
- Suggests K103N 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. Topology generated using a GTR+gamma evolutionary model in `FastTree`.

Conclusions

- We demonstrate high-quality consensus for single molecules by modeling the sequencing reaction using the new Arrow modeling framework.
- We achieve accurate quantitation of K103N DRAM mixes, in both a well-characterized mix and primary samples, and tracks parallel evolution of the K103 mutation in response to therapy.

Acknowledgements

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

For Research Use Only. Not for use in diagnostics procedures. © Copyright 2016 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences.
BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. All other trademarks are the sole property of their respective owners.