Analysis of PCR Bias Using Primer IDs and Illumina Sequencing of HIV RNA Populations

Valerie F. Boltz[1], Wei Shao[2], Mary F. Kearney [1], John W. Mellors[3], Frank Maldarelli[1] and John M. Coffin[4]

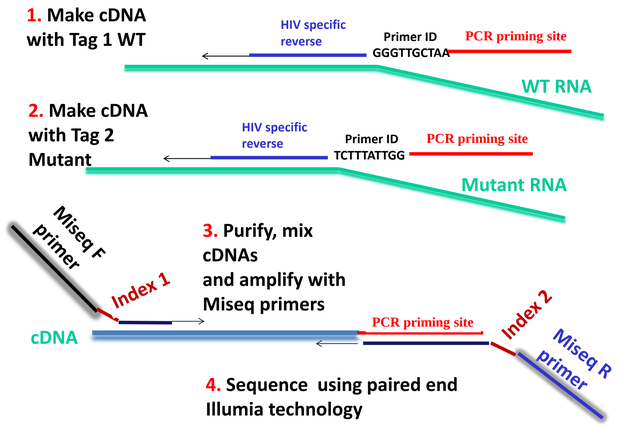
[1]National Cancer Institute, NIH, Maryland, USA; [2]Advanced Biomedical Computing Center, Maryland, USA, [3]University of Pittsburgh, Pennsylvania, USA, [4]Tufts University, Massachusetts, USA,

Introduction

Illumina sequencing is widely used to study HIV-1 populations. To achieve sequences that accurately represent the virus population, primer IDs have been used to label each molecule with a unique tag during cDNA synthesis. However, their use reveals an apparent PCR bias in which a large fraction of the resulting sequences obtained are represented only once in the data set. Here we show that carry-over into the PCR reaction of unused primer IDs and short incompletely transcribed but tagged cDNA amplicons appear as PCR bias but are actually artifacts. In addition, using these primer IDs, in a method we call the "Filibuster Correction", we were able to reduce the PCR and sequencing error rates to levels comparable to SGS.

Methods 1: cDNA with known IDs

cDNA from WT *pol* transcripts was synthesized using a primer tagged with a Primer ID of known sequence to tag all molecules in the reaction identically. Separately and using a different known primer ID sequence, cDNA was synthesized from transcripts containing drug resistance mutations. The two reactions were purified, mixed and amplified together with Illumina primers.



Build consensus sequences from reads that share identical primer IDs

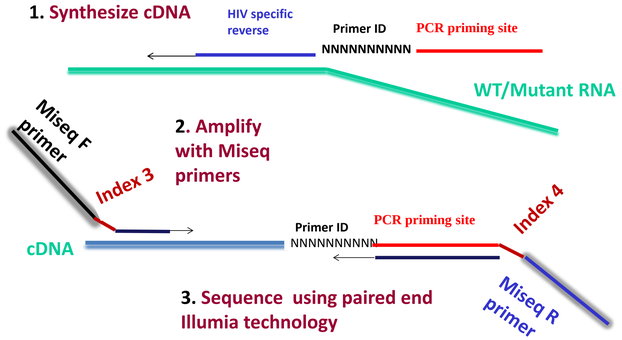
Results For Methods 1

Out of a total of 234680 sequences, 23.8% were labeled with the wrong primer ID.

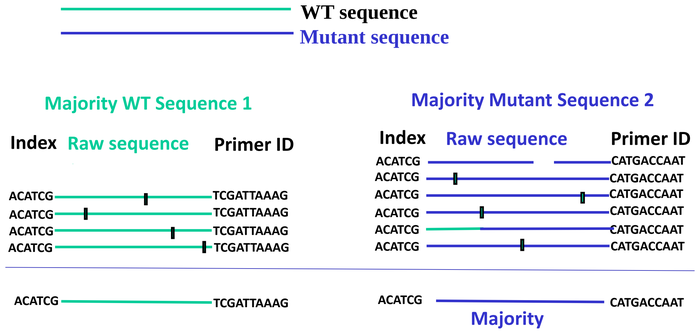
Tag	% WT Sequences	% Mutant Sequences	% Recombinant Sequences	% Sequences Tagged Incorrectly
GGGTTGCTAA wt tag	17	18.5	64.5	18.5
TCTTTATTGG mut tag	5.3	82	7.2	5.3
Total				23.8

Methods 2: cDNA with IDs of random bases

Cloned WT and Mutant *pol* transcripts containing drug resistance mutations were mixed at 50/50 ratios and used as template. Primers tagged with Primer IDs of 10 random nucleotides were then used to synthesize cDNA



Build consensus sequences from reads that share identical primer IDs

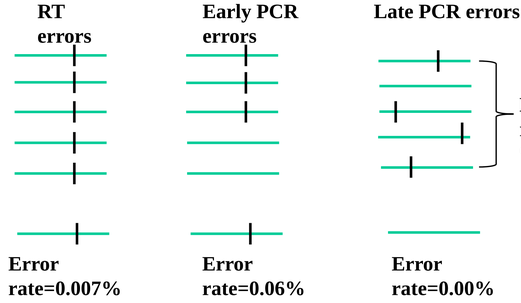


Acknowledgements

We wish to thank Jason Rausch, Ann Wiegand, and Jon Spindler for helpful discussions. We acknowledge with gratitude Vinay Pathak for helping to purchase the Miseq

Results For Methods 2

1. Substitution Error Rate and Indel Error Rate was Reduced on average to 0.04%



2. Censor divergent positions in each primer ID consensus where the minor base is >1/5. Or in other words 80% of the bases at that position have to be in agreement.

Filibuster Correction (Super Majority)					
Cutoff	# Consensus Seq	%WT	% Mutant	% Recombinants	% Error
>0.5	4767	43.89	48.14	7.93	0.006
≥0.6	4767	46.49	48.81	4.36	0.005
≥0.7	4767	49.23	49.07	1.07	0.003
≥0.8	4767	49.7	48.79	0.42	0.003
≥0.9	4767	49.15	48.37	0.4	0.003

Conclusions

Our results show that up to 25% of sequences derived from HIV RNA may be artifacts which limits the accuracy of the current primer ID approach. Steps to avoid these artifacts must be employed when using primer IDs. However, to achieve results that accurately represent the virus population, the unique tagging of individual templates is essential and using the "Filibuster Correction" method for analysis, error rates comparable to the gold standard of SGS can be achieved.