Accurate Identification of HCV Sequence Variants Using Deep Sequencing With a Primer ID

Don Cooper[1], Abha Chopra[1], John Blinco[1], Cassandra Jabara[2], Stan Lemon[3], Richard Barnard[4], Silvana Gaudieri[1,5], Mark Watson[1], Simon Mallal[1,6]

[1]Institute for Immunology and Infectious Diseases, Murdoch University, Western Australia, [2]Department of Biology, & Lineberger Comprehensive Cancer Center, University of North Carolina, USA, [3]Division of Infectious Diseases, Department of Medicine, & Lineberger Comprehensive Cancer Center, University of North Carolina, USA, [4]Antiviral Research, Merck and Co., West Point, PA, USA, [5]School of Anatomy, Physiology and Human Biology, University of Western Australia, [6]Department of Clinical Immunology and Immunogenetics, Royal Perth Hospital, Western Australia

Background:

Viral Sequencing

Deep sequencing techniques have revolutionised the way rare viral variants are detected. However, determining how many viral templates were amplified and sequenced, discriminating between true variants and errors introduced during amplification and sequencing, and calculating the ratios of variants in the population are all error prone.

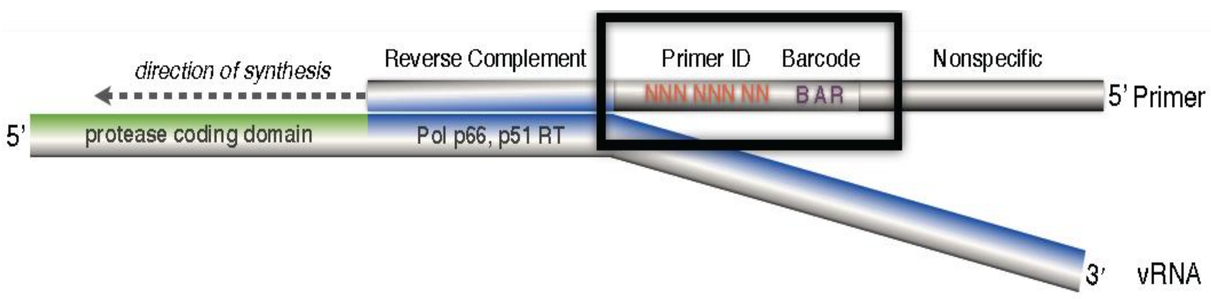
Sources of error

Sources of error often overlooked are PCR amplification bias, sequencing errors and recombination during the PCR, particularly with a high concentration of similar templates[1].

These errors can completely obscure true variants and falsely represent sequences, via recombination, that were not present in the original sample.

Primer ID concept:

Primer ID (PID) is a technique that allows errors in PCR and sequencing to be observed and controlled on the 454 FLX sequencing platform by labelling sequences arising from individual viral RNA molecules with a primer identification tag[2].



- In the 5′ tail of the primer, a degenerate string of eight nucleotides created a Primer ID, allowing for 65,536 unique combinations.
- Three nucleotide barcode was designed for the sample ID.
- A heterologous string of nucleotides with low affinity to the Viral genome was included in the far 5′ end for use as the priming site in the PCR amplification.
- By limiting the amount of viral template at cDNA conversion, most primer IDs will represent the reverse transcription of a single viral template.
- Most amplified products with the same PID can be grouped as originating from the same viral template. Any variation seen from sequences with the same PID must represent errors in PCR amplification and sequencing.

References:

[1] Daniel J.G. Lahr & Laura A. Katz Biotechniques 2009 47:857-866

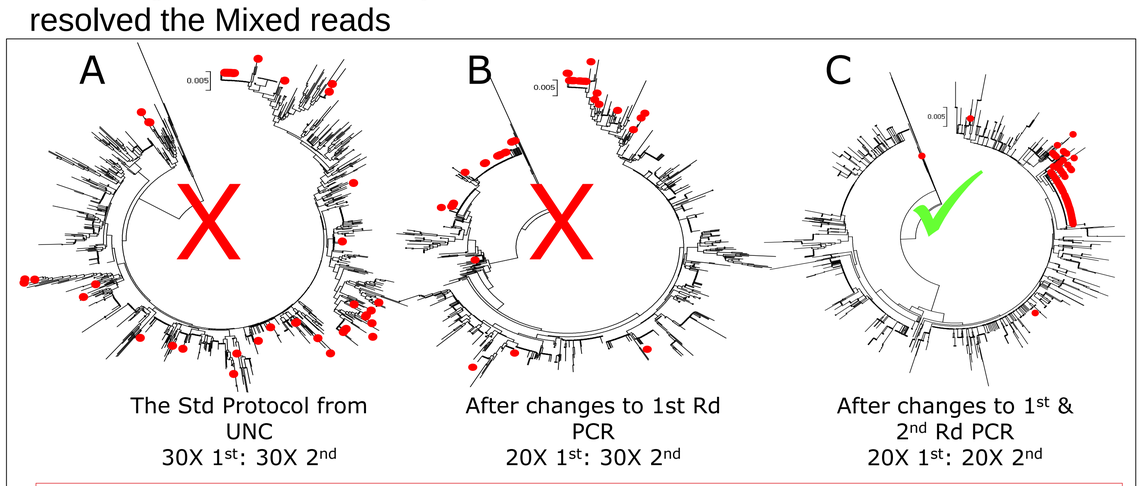
[2] Jabara CB, et al. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. Proc Natl Acad Sci U S A. 2011 108:20166-71

Results:

Applying the protocol for primer ID from Jabara *et al.*[2] to the NS3 region of the HCV resulted in:

1. **An unexpectedly high proportion of PIDs with seemingly unrelated or mixed reads.**

Reducing the number of cycles in the first round and second round PCR resolved the Mixed reads



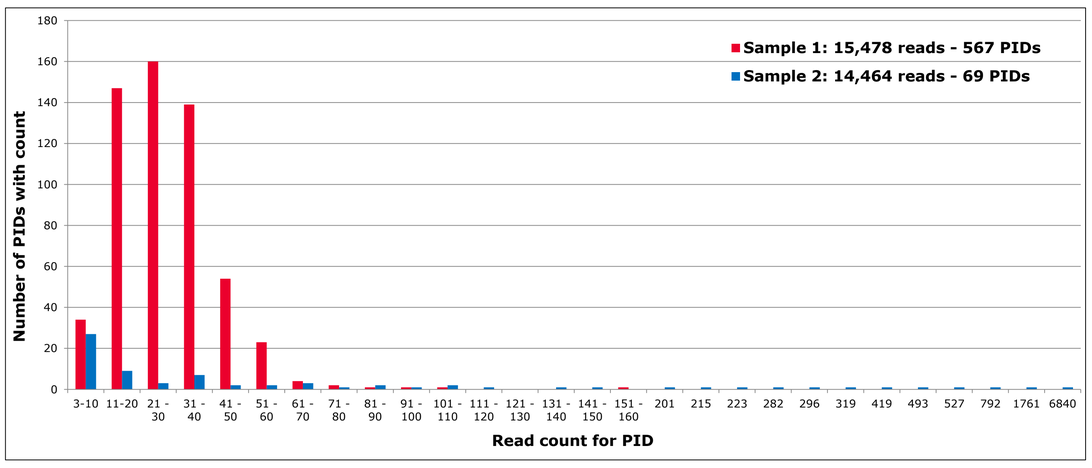
The Std Protocol from UNC
30X 1st: 30X 2nd

After changes to 1st Rd PCR
20X 1st: 30X 2nd

After changes to 1st & 2nd Rd PCR
20X 1st: 20X 2nd

Spread of reads from a single PID (red dots) on whole sample variation, for different PCR cycle numbers

2. **In some samples the distribution of reads between PIDs was grossly uneven. Repeated samples show the same pattern, i.e. it appears sample dependent.**

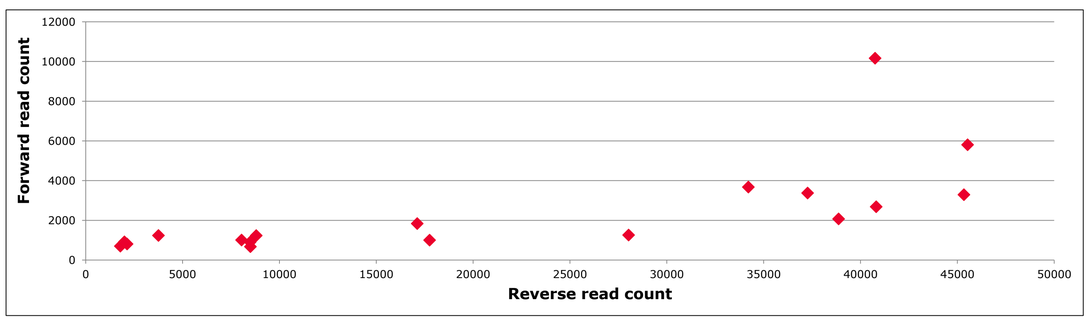


3. **In some samples there was a large number of short reads (160-190 bases).**

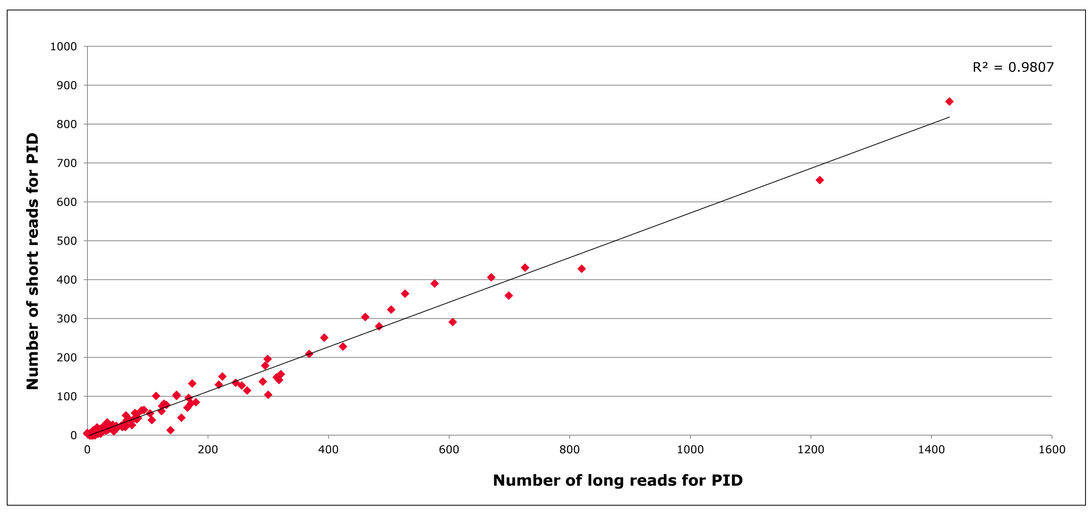
The percentage of short reads appeared to be run dependant.

Run	Samples	Mean short %	Std dev short %
1	16	28.1	22.8
2	22	5.3	3.5
3	30	34	16.3

Most of the short reads are reverse reads.



The short reads did not appear to be sequence related as all PIDs were similarly affected. Additionally, there was little correlation for short read percentage between samples and their repeats.



Conclusions:

When using the primer ID method it is important to consider template load, PCR cycling number, and sample dependant effects, and how changes to these will effect data quality.

Once optimised however, primer ID is a valuable tool for current generation deep sequencing methods that require reverse transcription and amplification of viral templates.

Acknowledgments:
We would like to thank our co-authors for their ongoing assistance with developing this project, and all staff at IIID for their assistance and input.
This work has been funded by Merck and Co.