O-128

Conference on Retroviruses and Opportunistic Infections

March 3-6, Atlanta, GA

Impact of experimental conditions on the utility of "primer ID" tagging for next-generation HIV sequencing

Richard H. Liang, Theresa Mo, Winnie Dong, Guinevere Q. Lee, Luke C. Swenson, Rosemary M. McCloskey, Conan K. Woods, Chanson J. Brumme, P. Richard Harrigan, Art F.Y. Poon

BC Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

PRESENTING AUTHOR:

Richard H. Liang
BC Centre for Excellence in HIV/AIDS
608-1081 Burrard Street
Vancouver, BC V6Z 1Y6
Canada

rliang@cfenet.ubc.ca
P: 604-682-2344 x62466
F: 604-806-9044

Abstract

Background

Primer IDs (pIDs) are "tags" used in next-generation sequencing consisting of random nucleotides attached to sequencing templates via degenerate primers during reverse transcription of HIV RNA molecules into cDNA. These pIDs (1) help track the number of templates in a sample, and (2) allow variation of sequences sharing a pID to be averaged out to correct for sequencing error.

Three potential issues complicate the above applications. First, multiple cDNAs may share a pID by chance. Second, a pID must be observed in at least 3 sequences to allow error correction; as such, pIDs with read depth 1 or 2 must be rejected. Third, the resulting sequence could influence amplification and sequencing – for example, by adding stretches of homopolymers. We assess how the utility of pIDs depends on experimental conditions such as the initial number of HIV RNA molecules.

Results

The number of pIDs required to reliably achieve 95%-good labelling is drastically lower than that required for perfect labelling For example, 1,193,698 pIDs are required to assure perfect labelling of 502 HIV RNA molecules, whereas only 6,121 pIDs are enough to assure 95%-good labelling.

Our model also predicts that discarding pIDs represented by fewer than 3 sequences rejects over 40% of reads in samples with 10000 templates and 20000 reads. In practice, fewer reads are rejected than predicted, though still a large proportion. Results were consistent across pID designs.

Methods

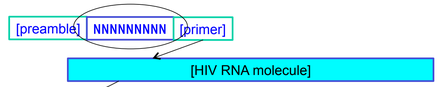
The number of possible pIDs is determined by the degeneracy of the starting primer. Assuming all possible pIDs are equally likely to attach to any given template, we calculated the numbers of pIDs necessary for a 90% probability of (1) perfect labelling of the input HIV RNA molecules (i.e. every template is assigned a unique pID) and (2) 95%-good labelling (95% of templates receive unique pIDs) as functions of template count. We also computed confidence intervals for the number of pIDs represented by fewer than 3 sequences as a function of the numbers of reads and templates, and compared the model predictions to data obtained via Roche/454 pyrosequencing of HIV populations using four pID designs (NNNNNNNNN, NNDNNHNNV, NBDHVBDHV, RYRYRYRYR).

Conclusions

While perfect labelling is unrealistic, achieving 95%-good labelling in sequencing with pIDs is far more practical. In sequencing samples with high template counts, rejecting pIDs with read depth 1 or 2 may be unadvisable because it will discard too much data depending on the number of reads.

What is a primer ID? (Jabara *et al.* [1])

string of random nucleotides that labels an RNA template



reverse transcription

ACCTCCTGG	[primer]	[cDNA sequence 1]
CCTTCTAAG	[primer]	[cDNA sequence 2]
TTCATGAAC	[primer]	[cDNA sequence 3]

PCR amplification and sequencing

Resulting cDNA molecules are marked with these labels **so we can tell them apart** (e.g. if two RNA molecules are the same)

ACCTCCTGG	[primer]	[noisy sequence 1]
ACCTCCTGG	[primer]	[noisy sequence 1]
ACCTCCTGG	[primer]	[noisy sequence 1]
ACCTCCTGG	[primer]	[noisy sequence 1]

CCTTCTAAG	[primer]	[noisy sequence 2]
CCTTCTAAG	[primer]	[noisy sequence 2]
CCTTCTAAG	[primer]	[noisy sequence 2]

After amplification the labels allow us to **identify where the molecules came from**; we can use this to

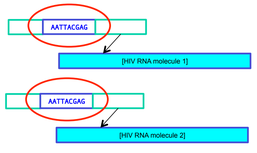
- track how many virions we actually sequenced, i.e. **correct for oversampling**
- correct for noise in sequences

Acknowledgements

This work was supported in part by an Operating Grant from the Canadian Institutes for Health Research (CIHR; grant # HOP115700) to AFYP. AFYP is supported by a Michael Smith Foundation for Health Research (MSFHR)/St. Paul's Hospital Foundation-Providence Health Care Research Institute (SPHF-PHCRI) Career Investigator Scholar Award. PRH is supported by a CIHR/GSK Research Chair in Clinical Virology. RMM is the recipient of a VPR USRA from Simon Fraser University.

Primer ID Collisions

What can go wrong



The same pID can attach to two different RNA molecules by chance – now we can't tell these two molecules apart anymore, and our dataset is *corrupted* (Sheward *et al.* [2]). We can reduce this chance by *increasing the complexity of the design* of the pID to **increase their diversity**, but this may introduce other problems.

However, **this may not be a major concern** if *most* of the RNA molecules are still uniquely labelled – i.e. if the data is still "**pretty clean**".

Can we ensure that we have pristine data free of this problem? If not, **how clean can we make it**?

Background

If there are 30 students in a classroom, what is the probability that no students share a birthday?

$$1 \times \frac{364}{365} \times \frac{363}{365} \times \ldots \times \frac{336}{365} = 0.2937$$

Similarly, if there are N primer IDs and m HIV RNA templates, all of which are equally likely to be tagged by any pID, what is the probability that no two molecules share a tag?

$$1 \times \frac{N-1}{N} \times \frac{N-2}{N} \times \ldots \times \frac{N-m+1}{N}$$

So how many pIDs do we need to have a good probability (0.9) of *perfectly labelling* the HIV molecules, i.e. so that all have their own unique pID? What about just *most* of them having unique pIDs?

Methods

Using these mathematical assumptions, we found the number of pIDs required to give a 90% chance of:

- *perfect labelling* (every HIV molecule is labelled by its own pID that is not shared with any other; i.e. data is **pristine**)
- *95%-good labelling* (the number of pIDs used is 95% the number of input HIV molecules, so *most* molecules have a unique label; i.e. data is **pretty clean** and most HIV RNA molecules that were successfully sequenced are distinguishable from one another)

for a given template count. These 95%-good labelling results became computationally difficult so we used an approximation for the higher template counts.

Results

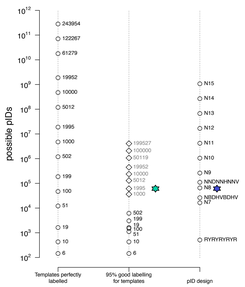


Figure 1: The number of pIDs required for 90% chance of:

- perfect labelling for the given template count (the numbers on the plot)
- 95%-good labelling

as well as the number of pIDs provided by different pID designs.

We found that

- 90% chance of perfect labelling is unrealistic for even moderately high template counts and would require much too long of a pID design
- 90% chance of 95%-good labelling is much more realistic

estimated template count of Jabara *et al.* [1]

pID design used by Jabara *et al.* [1]

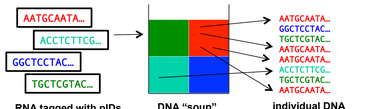
Conclusions

Obtaining **perfect labelling is impractical** for even moderate numbers of templates with reasonable-length pID designs. However, 95%-good labelling is easily obtained for up to 10,000 templates with an 8-bp length pID design. Thus **we can aspire to have "pretty clean" data** coming out of a typical NGS run even if we can't get "pristine" data.

This must be taken into account when designing the pID.

Shallow Read Depth

What can go wrong



All but the red sequences get rejected

For error correction using pIDs, **at least 3 sequences with the same pID are needed**. Anything with shallower read depth (we call these singletons and doubletons) is rejected. **How much data does this toss out**?

Methods

Assume:

- all HIV templates are uniquely tagged with a pID (say there are m of them)
- all HIV templates are amplified via PCR to produce a large (effectively unlimited) number of DNA molecules; this creates a "soup" of m ingredients in equal proportions
- all DNA molecules are equally likely to be sequenced

We tested the above assumptions to see how well they fit against experimental 454 data, and then used this model to predict the proportion of reads that one rejects in the above error-correction scheme as a function of the number of reads and the number of HIV templates.

Experimental data details:

- two clinically-derived plasma samples (pVL 312209 and 177998) were extracted
- RT-PCR of HIV RT with pID-tagged primers was performed in triplicate on undiluted, 10x diluted, and 100x diluted extracts
- primer ID designs considered were NNNNNNNNN, NNDNNHNNV, NBDHVBDHV, RYRYRYRYR
- batches were sequenced on a Roche/454 GS-FLX platform
- data collected using Python scripts

Results

Comparing theoretical predictions to experimental data (**Figure 2**), we found:

- for undiluted samples, most observed rejected sequences fall within estimated error bounds
- model fit is not robust against decreases in input template number, likely due to changes in RT efficiency as the sample becomes more dilute
- poor labelling can become a factor

Theoretical predictions (**Figure 3**) suggest that we throw out a great deal of data with our typical run parameters (read depth ~3000), though this issue did not affect the original studies of Jabara *et al.* [1].

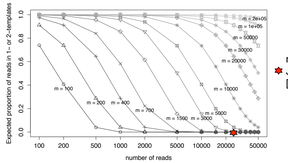


Figure 3: Here, m is the number of extracted RNA templates that survive RT and are represented in the sample. The y-axis represents the proportion of data rejected:

- As the number of templates increases, more data must be rejected;
- as the number of reads increases, less data is rejected.

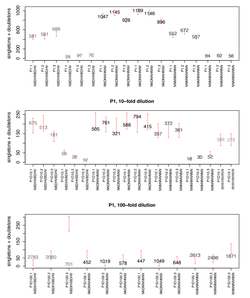


Figure 2: Experimental data (the numbers are the number of reads produced for that replicate; their position is the number of pIDs rejected) compared to theoretical prediction (error bars). pVL for this sample is 312209.

Conclusions

If pID-based error correction is desired, the **number of reads must be large enough to handle the number of templates in your sample**. If this is impossible then **consider retaining the data with shallow read depth**.

Sequencing/Amplification Issues

What can go wrong

The **pID may introduce bias** in reverse transcription, PCR amplification, or sequencing. Roche/454 sequencing in particular may be significantly affected *if the pID introduces homopolymers*. Also, different designs may be used; for example, to limit the occurrence and length of homopolymers. **Do different pID designs introduce biases**?

Methods

We sought to consider two scales: **microscopic**, or effects of individual pIDs on individual reads; and **macroscopic**, or the effects of using different pID designs on the sample as a whole.

Microscopic

Using data from the same experimental data as in the last, we fitted generalized linear models to gauge what factors have effects on

- "success" of a pID: how many times it gets sequenced
- error probability of a pID: how prone it is to indel sequencing error

We fitted against several *intrinsic* factors and several *extrinsic* factors:

- intrinsic: homopolymer "score" (ad hoc measure of how many homopolymers are in the pID), nucleotide content
- extrinsic: concentration of sample, pID design, sequencing barcode, sequencer plate region

Macroscopic

We restricted to the data from one sample (pVL 312209), rejecting batches that produced fewer than 500 reads. We used a data collection scheme that used pIDs to correct for oversampling but did *not* discard singleton and doubleton reads. We then focused on a 100bp portion of RT and tallied up the frequency of all variants occurring across all batches. We tracked the prevalence within undiluted batches (3 using design NBDHVBDHV; 6 using NNDNNHNNV; 3 using N9) of the two most prevalent variants overall (variant 1, 33.7% overall; variant 2, 15.4% overall).

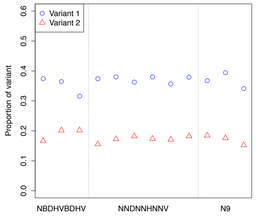


Figure 4: prevalence within undiluted batches of the two overall most frequent HIV variants. Overall, variant 1 represented 33.7% of all observed HIV sequences, and variant 2 represented 15.4% of all observed HIV sequences.

Results

Microscopic

pID success:

- no significant intrinsic factors
- dominated by external factors

Indel error rate:

- more homopolymers in the pID make for more indel errors
- pIDs with lots of Gs are more error prone
- some external factors also significant, including pID design

Macroscopic

Figure 4 shows the prevalence of variants 1 and 2 in all undiluted batches. This plot shows no apparent bias coming from different primer ID designs. Dilution appears to have been a factor (not shown) – more analysis will follow.

Conclusions

At the level of individual reads, **sequencing success is not strongly influenced by the pID sequence itself or by the pID design, but indel error rate can be**.

At the macroscopic level, **pID design does not appear to introduce bias**.

References

1. C.B. Jabara, C.D. Jones, J. Roach, J.A. Anderson, and R. Swanstrom. Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID. *Proceedings of the National Academy of Sciences*, 108(50):20166–20171, 2011.
2. D.J. Sheward, B. Murrell, and C. Williamson. Degenerate Primer IDs and the birthday problem. *Proceedings of the National Academy of Sciences*, 109(21):E1330–E1330, 2012.

BRITISH COLUMBIA CENTRE *for* EXCELLENCE *in* HIV/AIDS

UBC

Providence HEALTH CARE
How you want to be treated.