Tropism Testing by MiSeq Is Comparable to 454-Based Methods but Exhibits Contamination Issues

BRITISH COLUMBIA CENTRE for EXCELLENCE in HIV/AIDS

Luke C Swenson, Winnie Dong, Celia KS Chui, Conan K Woods, Marjorie Robbins, Alejandro Gonzalez-Serna, Aram Karakas, Art FY Poon, Chanson J Brumme, P Richard Harrigan

British Columbia Centre for Excellence in HIV/AIDS, Vancouver, Canada

BC Centre for Excellence in HIV/AIDS
680-1081 Burrard St
Vancouver, BC, Canada
V6Z 1Y6
Email: lswenson@cfenet.ubc.ca
prharrigan@cfenet.ubc.ca

Abstract

(Note: The poster contains updated information but the original abstract is reproduced without modification. There is now an option for applications requiring sensitive detection of low level variants which can decrease carry-forward contamination)

Background: Bench-top next-generation sequencing (NGS) has become increasingly popular for HIV applications within the past five years. While "454"-based methods have dominated HIV research, the Illumina MiSeq is gaining popularity based on cost, ease of use, read depth and accuracy through homopolymer regions.

Methodology: A total of 514 plasma samples were run on two NGS platforms, and sequencing data from the HIV V3 region were analyzed. Stored first-round PCR products were obtained from the samples which were originally sequenced on a Roche/454 GS-FLX.

The samples underwent further amplification with adaptors appropriate for the Illumina MiSeq. Barcoded samples were sequenced in multiplexed runs of 96 samples on both instruments. Samples were excluded if they had <750 reads by GS-FLX or <10,000 by MiSeq. CXCR4-using tropism was assigned if ≥2% of sequences had a geno2pheno false-positive rate <3.5 - parameters which were previously optimized for the 454, though not the MiSeq.

Results: Both platforms were well correlated in terms of the proportion of X4 reads per sample (slope= 0.91, R2=0.90, Kappa=0.78). Median read depth was ~89,000 by MiSeq and ~2500 by GS-FLX. For tropism classifications, the platforms were 89% concordant (459 of 514 samples).

Of the 55 samples with discordant classifications, 62% were X4 by MiSeq but R5 by 454, indicating potential over-sensitivity for minor variants by MiSeq. The median difference in %X4 reads between platforms was 0.2% higher by MiSeq (interquartile range [IQR]: -1.1%-1.3%). Where discordant, samples differed by an absolute value of 2.5% X4 (IQR: 1.3%-8.3%).

We also noted systematic low-level carry-forward contamination on the MiSeq platform: on average, a minimum of 0.04% of the total number of recovered reads (standard deviation 0.02%) were obvious contaminants (non-V3 sequences carried over from previous runs). Contaminant sequences persist in subsequent runs, appearing in as many as 7 runs after they were originally processed.

Conclusions: The MiSeq and GS-FLX systems were well correlated with each other and gave similar tropism classifications in 89% of samples. Where discordant, MiSeq tended to call more samples as X4. Laboratories using MiSeq platforms should account for carry-over contamination where possible in their data processing.

Background

- Arguably, the most commonly used next-generation sequencing platform in HIV research has been the 454 Life Sciences Genome Sequencer FLX
- The Illumina MiSeq is gaining popularity but has not been as extensively evaluated as 454-based methods
- Next-generation sequencing of the V3 region of HIV *env* can be used to identify CXCR4-using HIV variants in clinical samples
- Previously, a 454-based version of the next-generation sequencing method was shown to have good concordance with phenotypic tropism assays and was also a good predictor of treatment response to the CCR5 antagonist, maraviroc
- We sought to evaluate whether a similar method using the MiSeq platform is comparable to the previously validated 454-based method

Methodology

- HIV V3 was previously deep-sequenced by 454 GS-FLX in **802 plasma samples** We obtained stored first-round PCR products from these samples and amplified them further using primers and multiplex indices adapted for the MiSeq
- All samples were **processed by both 454 and MiSeq**
- V3 sequences were interpreted using the geno2pheno algorithm (3.5 False-Positive Rate cutoff)
- The percentage of X4 sequences per sample was calculated
- R5 tropism was defined as having <2% X4 sequences

Quality control

- 454 sequences were required to have a geno2pheno alignment score of 97.5% or higher
- MiSeq sequences having bases with quality scores <Q10 within V3 were excluded
- For both methodologies, sequences were excluded if they contained stop codons, had nucleotide lengths indivisible by 3, or lengths <99 or >114 nucleotides (i.e., 33-38 amino acids long)
- Samples with fewer than 750 reads by 454 or fewer than 10,000 reads by MiSeq were excluded from the analysis

Results

- Median read depth was ~2500 by 454 and ~110,000 by MiSeq
- The proportions of X4 reads per sample were similar between platforms (**Figure 1**), with slope=0.98, and **R^2=0.89** (p<0.001)

- Samples were classified as being R5 samples (with <2% X4 variants) or as X4 samples (≥2% X4 variants),
- Between methodologies, the overall **concordance was 88%** (Kappa=0.75)
- Relative to 454 sequencing, tropism testing by MiSeq had **95% sensitivity** (399/421 X4 samples) and **80% specificity** (303/381 R5 samples)

- Samples with concordant tropism results by either sequencing method also had similar %X4 sequences
- **Concordant R5 samples** had a median of **0.0% X4** by 454 (interquartile range: 0-0%) and **0.5% X4** by MiSeq (interquartile range: 0.3-0.9%)
- **Concordant X4 samples** had a median of **41.4% X4** by 454 (interquartile range 14.1-87.4%) and **41.9% X4** by MiSeq (interquartile range: 9.6-93.5%)
- Where samples were discordant, the percentage of X4 variants was generally low (**Figure 2**)

Figure 1: Scatterplot of the %X4 sequences per sample by 454 vs MiSeq

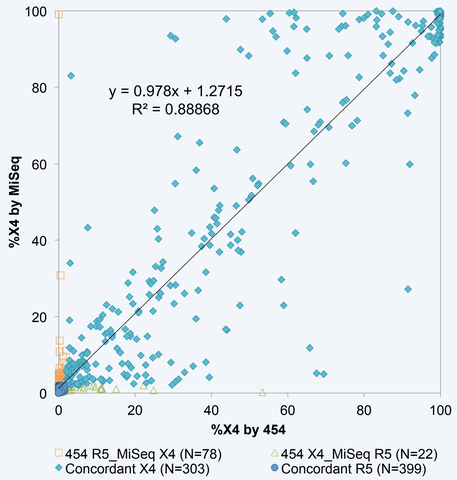
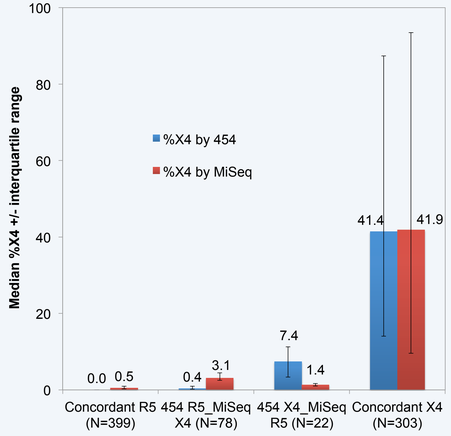


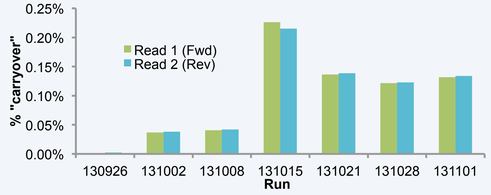
Figure 2: Average %X4 for concordant and discordant samples



Contamination

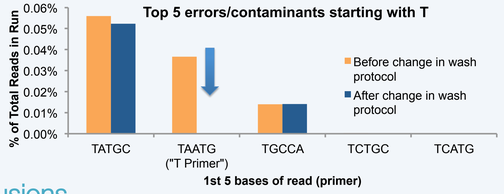
- Our previous analyses identified low levels of carry-forward contaminants
- These reads were remnants from previous runs which appeared to have been insufficiently washed away between runs
- The contaminants persisted for as many as 7 runs after they had been initially processed
- **Figure 3:** New modified PCR primer containing an additional 5' "T" is introduced after run 130926 (1st run on graph)
- Subsequent runs alternated between using the original (AATG) and the modified (TAATG) PCR primer
- Carry-forward contamination was defined as sequencing reads beginning with "TAATG" that were observed in runs using the unmodified primer

Figure 3: Percentage carry-forward contamination



- An identical library was sequenced both before (orange) and after (blue) the institution of a modified was protocol (Figure 4)
- A modified wash protocol was able to remove carry-forward contaminants as assessed by absence of the TAATG primer (**Figure 4**). This protocol is available from Illumina for laboratories requiring sensitivity at similarly low levels

Figure 4: Wash protocol decreases carry-forward contamination



Conclusions

- The MiSeq and 454 GS-FLX systems were well correlated and gave similar tropism classifications in 88% of samples
- Contamination must be considered when assaying for low-level HIV variants