# Studying the error for HCV amplicon sequencing with the Illumina MiSeq platform

Zoya Dimitrova[1], Lilia Ganova-Raeva[1], Amanda Sue[1], Pavel Skums[1], Natalia Saveleva[1], and Yuri Khudyakov[1]

CDC, DVH, Molecular Epidemiology and Bioinformatics , Atlanta, GA 30329,
*izd7@cdc.gov*

**Abstract.** Next Generation Sequencing (NGS) is a well established technology for acquisition of vast amount of nucleic acid sequence data at very low cost and high throughput. We aim to provide solid framework for the accurate determination of the viral haplotypes of Hepatitis C virus (HCV) using the Illumina MiSeq platform. To do that, we studied the error rates of sequences generated with the longest available read kits for the MiSeq-v3 (600), applied to bar-coded HCV amplicon libraries of low complexity.

**Keywords:** NGS, amplicon, low complexity, bar-coding

## 1 Introduction

NGS has found many application for the detection of low frequency variants and rare mutations that add great value to medical diagnostics. Most of NGS applications have been developed to work with high complexity samples as eucaryotic genomes, bacterial genomes, or multiple loci[1]. The use of NGS in studies of viruses like the Hepatitis C virus (HCV) present unique challenges [2].

HCV mutates rapidly and exists in the infected host as a population of multiple variants referred to as quasi species or haplotypes. This unique dynamic population can be used to identify transmission linkages to other individuals and study the viral evolution. HCV surveillance and outbreak tracking is done best by precise determination and comparison of multiple haplotypes from amplicons of the Hyper-variable Region 1 of the viral Envelope gene (HVR1) [3].

When NGS is used for amplicon sequencing, low sample complexity can cause cross cluster hybridization and result in unreliable data [4]. In addition, sequencing errors occur that are specific to the Illumina MiSeq platform. We have generated data sets to address these issues of sequence quality. The data will provide basis for the development of appropriate data error correction algorithms tailored to the specifics of the HCV intra-host population and the sequencing platform.

## 2 Methods

*Clones for NGS libraries*

Part of the HCV HVR1 region, 309 nucleotides (nt) long, from 15 different patients was cloned in *E.coli* to ensure genetic homogeneity of the target gene sequence. The genetic distance between the clones selected for the experiment varied 1.9-67.2%. Eight different forward primers were synthesized to contain the HVR1 specific sequence at the 3' end followed by 8 different 10-mer barcodes and by adapter sequence needed to incorporate the required Illumina sequencing primer. The same strategy was used for the reverse primers. After the first round of PCR amplification using a single clone as template, the resulting product incorporates both primers and thus acquires a unique 20-mer barcode, 10 nt long, on each end. The product was then used in a second round of PCR amplification, an index PCR, which attaches additional 8-mer barcode (index) that is read in the first index read and the needed Illumina clustering sequences. This index is utilized by the MiSeq (Illumina, Inc.) instrument for de-multiplexing. For our study we sequenced two libraries with different labeling configurations. In Library 1 we used 8 different indices with two unique pairs of barcodes for each index, e.g. index1-barcode1-clone1-barcode2 and index1-barcode2-clone9-barcode1.This allowed us to include 16 samples in the first library: 13 clones, 2 outbreak samples and a negative control. In Library 2 an unique pair of barcodes was used for each unique index, e.g. index1-barcode1-clone1-barcode1 and index2-barcode2-clone2-barcode2. The second library consisted of 8 clones samples. Both libraries were run for 55 hrs on MiSeq Instrument with v3 chemistry providing 2x300 nt paired reads.

*NGS data processing*

All reads longer than 200 nt were pre-screened for matching the expected index, barcodes and gene-specific primers. Only reads with at most 1 nt difference form the index, at most 1 nt difference from the bar code and up to 3 nt from the primer sequence in both the forward and reverse read were selected.

To study the sequencing error, all selected reads were aligned to the clone sequence of the respective sample. Reads that aligned better to a different clone than to the expected one were considered to be result from incorrect clustering and not included in the error analysis. Based on these alignments, we calculated the rates of the substitutions, insertions and deletions for the forward (R1) and reverse reads (R2) for the 8 clones that were used in both libraries. The quality scores for each type of error were analyzed. The processing and analysis were performed using MATLAB R2015a (The MathWorks, Inc.)

The paired reads R1 and R2 were merged using the software package Context-Aware Scheme for Paired-End Reads (CASPER) [5]. This is a five step algorithm that includes pre-processing, constructing a table of k-mer counts, finding the best overlap position, resolving mismatches in the overlap by using quality scores and context base correction, and merging forward and reverse reads. The threshold rate for allowed mismatches in the best overlapping region of R1 and R2 direction of the reads was set to 0.01 and only reads with less mismatches were merged.

## 3 Results and Discussion

*Run characteristics*

The MiSeq run parameters for Library 1 were as follows: 997 clusters/sqrmm with 90.25% of the clusters passing filter, phasing/pre-phasing 0.188/0.043 for R1 and 0.186/0.014 for R2, 26.11M reads passing filter with 82.6% greater then Q30. For Library 2 the run had 975 Clusters/sqrmm and 94.6% clusters passing filter, phasing/pre-phasing 0.198/0.261 for R1 and 0.079/0.020 for R2, 22.24M reads passing filter with 80.2% greater then Q30.

For Library 1, 99.82% of the reads were longer than 200nt and for Library 2 95.64% were longer than 200nt. 67.88%of the reads in Library 1 had the expected index and after de-multiplexing by index, 79.4% of the reads had the expected barcodes and primers. For library 2, 92.41% of the reads de-multiplexed by the index and 90.14% of them had the expected configuration of barcodes and primers.

*Error rates*

The R1 reads and the paired R2 reads from both libraries were examined for the following types of errors: insertions, deletions and substitutions. All three types of errors were found and the values were comparable for both libraries. On average, the substitution rates were found to be significantly higher than the insertion rates (32.5x for Library 1 and 40x for Library 2) or the deletion rates (26x for Library 1 and 30x for Library 2). Boxplots of the compiled data for all types of errors in Library 1 and 2 are shown in Figure 1.

*Quality scores of the sequencing errors*

All the errors were examined in relation to the quality scores at the corresponding positions. Figure 2 shows the distribution of quality scores for the different types of errors in R1 and R2 directions. Both substitution and insertions have median quality scores below 12, however, substitution errors with high quality scores can be found very frequently in R1: 32.66% of the substitutions in library 1 and 26.25% for library 2 had quality scores greater than 32. This percentage is much lower for the substitutions in R2: 7.29% and 8.87% respectively. This finding is consistent with previously published data [6]. Notably, the deletions in both directions have average quality scores well above Q32.

*Merged reads and haplotypes*

Due to the stringent mismatch threshold of 0.01, only 40% of the pre-screened reads in library 1 and 42% of the reads in library 2 were merged and error-corrected by CASPER.The heterogeneity of the resulting sequences was evaluated for clones 1-8 only that were common for both libraries. For Library 1 57.8% of the data were a perfect match to the expected haplotype, i.e. the sequence of the corresponding biological clone as determined by Sanger sequencing and confirmed by the NGS consensus. The percentage varied between the different clones from 30.72 to 70.6. For Library 2, this value was 61.6% and also varied from 30.99 to 68.68. The remaining sequence represent multiple haplotypes found from each sample.

*Conclusion* The data analyzed here suggest that there is a need for custom algorithm for merging and error correction of viral amplicon sequences. The sig-

nificant presence of substitution errors in general, and in particular substitutions with high quality scores in R1 indicate that amplicon reads resulting from the V3 (600) sequencing chemistry need to be corrected by an algorithm that treats R1 and R2 data differently. This requirement will be very important for the accurate interpretation of viral quasi-species. The current data do not indicate that the occurrences of insertions or deletions are areas of significant concern. Existing marketed tagging methods have approx 0.3% rate of sample miss-identification due to cluster intermixing [4]. Analysis is in progress to evaluate if this particular multiplex tagging of the samples is adequate to address issues of miss labeling.

## References


1. Archer et al., Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II, BMC Bioinformatics. 2012 Mar 23;13:47.
2. N. Beerenwinkel, H.F. Gunthard, V. Roth, K.J. Metzner (2012)Challenges and opportunites in setimating viral genetic diversity from next-generation sequencign data, Frontiers in Microbiology, v3,329
3. Campo DS, Dimitrova Z, Yamasaki L, Skums P, Lau DT, Vaughan G, Forbi JC, Teo CG, Khudyakov Y. 2014. Next-generation sequencing reveals large connected networks of intra-host HCV variants. BMC Genomics 15 Suppl 5:S4.
4. Martin Kircher, Susanna Sawyer and Matthias Meyer (2012) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform, Nucleic Acids Research, 2012, Vol. 40, No. 1 e3
5. S. Kwon, B. Lee, S. Yoon (2014) CASPER: context-aware scheme for paired-end reads from high-throughput amplicon sequencing, BMC Bioinformatics. 2014; 15(Suppl 9): S10.
6. M. Schirmer, U. Z. Ijaz, R. D'Amore, N. Hall, W.T.Sloan, Ch.Quince (2015) Insight into biases and sequencign errors for amplicon seqencign wiht the Illumina MiSeq platform, NAR, v.43(6),e34

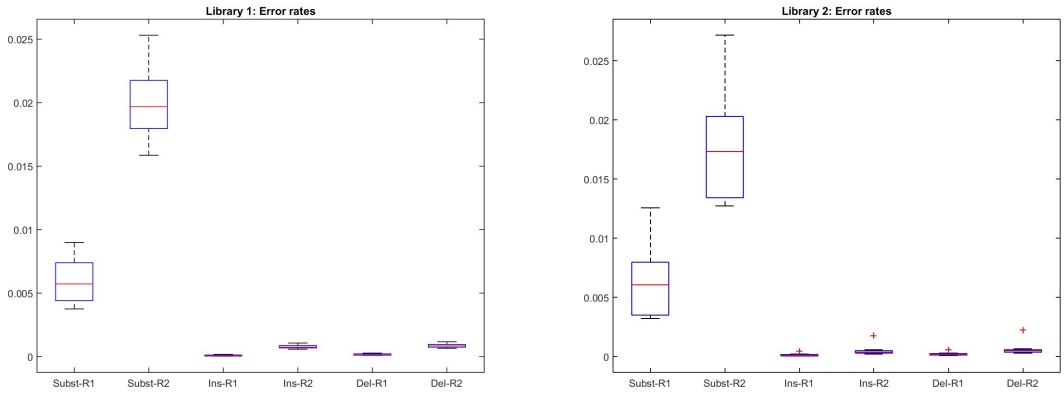


**Fig. 1.** Boxplot of the errors rates detected in Library 1 (left panel) and Library 2 (right panel) by the type of error and direction of the reads

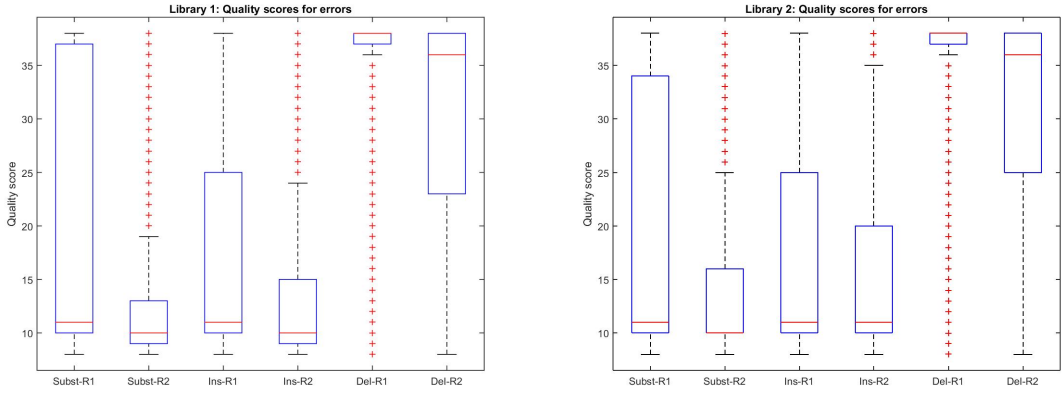


**Fig. 2.** Boxplot of the quality scores of the errors detected in Library 1 by the type of error and direction of the reads (left panel) and Library 2 (right panel) by the type of error and direction of the reads