Bioinformatics Advance Access published August 31, 2014

VirVarSeq: a low frequency Virus Variant detection pipeline for Illumina Sequencing using adaptive base-calling accuracy filtering.

Bie M.P. Verbist [1,*], Kim Thys [2], Joke Reumers [2], Yves Wetzels [2], Koen Van der Borght [2], Willem Talloen [2], Jeroen Aerssens [2], Lieven Clement [3], Olivier Thas [1,4 *]

[1]Ghent University, Department of Mathematical Modeling, Statistics and Bioinformatics, Coupure Links 653, 9000 Gent, Belgium.
[2]Janssen R&D, Turnhoutseweg 30, 2340 Beerse, Belgium.
[3]Ghent University, Applied Mathematics, Informatics and Statistics, Krijgslaan 281 S9, 9000 Gent, Belgium.
[4]University of Wollongong, National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, NSW 2522, Australia.

Associate Editor: Dr. Inanc Birol

ABSTRACT

Motivation: In virology, massively parallel sequencing (MPS) opens many opportunities for studying viral quasi-species, e.g. in HIV-1 and HCV-infected patients. This is essential for understanding pathways to resistance, which can substantially improve treatment. Although MPS platforms allow in-depth characterization of sequence variation, their measurements still involve substantial technical noise. For Illumina sequencing, single base substitutions are the main error source and impede powerful assessment of low-frequency mutations. Fortunately, base calls are complemented with quality scores that are useful for differentiating errors from the real low-frequency mutations.

Results: A variant calling tool, Q-cpileup, is proposed, which exploits the quality scores of nucleotides in a filtering strategy to increase specificity. The tool is imbedded in an open source pipeline, VirVarSeq, which allows variant calling starting from fastq files. Using both plasmid mixtures and clinical samples we show that Q-cpileup is able to reduce the number of false positives. The filtering strategy is adaptive and provides an optimized threshold for individual samples in each sequencing run. Additionally, linkage information is kept between single nucleotide polymorphisms as variants are called at codon level. This enables virologists to have an immediate biological interpretation of the reported variants with respect to their anti-viral drug responses. **A comparison with existing SNP-callers tools reveals that calling variants at codon level with Q-cpileup results in an outstanding sensitivity while maintaining a good specificity for variants with frequencies down to 0.5%.**

Availability: The VirVarSeq is available, together with a users guide and test data at sourceforge:
http://sourceforge.net/projects/virtools/?source=directory

Contact: bie.verbist@ugent.be

Supplementary Information Supplementary Information is available at Bioinformatics online

1 INTRODUCTION

RNA viruses such as HIV-1 and HCV exist in their host as complex populations composed of several closely related subgroups. They are referred to as quasi-species and originate from high and error prone replication rates (Domingo *et al.*, 2012). This heterogeneous mixture of genomes allows a viral population to rapidly adapt to changing environments. The fittest mutants out compete the others, allowing the virus to develop resistance to antiviral therapy. The characterization of sequence variation within the viral population is key for understanding pathways to resistance, but the identification of low-frequency variants remains challenging (Codoner *et al.*, 2011; Gianella *et al.*, 2010).

Until recently, the genetic diversity of a virus population could only be assessed through genotyping by Sanger sequencing, which provides information on the most abundant viral variants only. Massive parallel sequencing (MPS) technologies allow for a more in-depth characterization of sequence variation, including low-frequency viral strains. However, their measurements still involve substantial technical noise, complicating the analysis (Dohm *et al.*, 2008; Beerenwinkel *et al.*, 2012). Pyrosequencing, commercialized by Roche 454, was the most common sequencing method for viral population sequencing (Beerenwinkel and Zagordi, 2011). The recent announcement by Roche to retract the 454 technology from the market by mid-2016, illustrates the pressing need to evaluate and implement alternative technologies. Recently Illumina's sequencing technique has strengthened its position in this field (Thys *et al.*, 2014). Illumina also complements the sequenced nucleotides with quality scores (Q) (Ewing and Green, 1998) that reflect the base

[*] to whom correspondence should be addressed

calling substitution error probability. 454 quality scores, however, do not have such an intuitive interpretation. (De Beuf *et al*., 2012) Filtering based on quality scores (Reumers *et al*., 2012) has already proven valuable to reduce false positives. It often involves the use of a hard quality threshold. Unfortunately, this does not account for variation in quality between runs resulting in too stringent or too relaxed thresholds.

Most variant calling tools focus either on the detection of single nucleotide polymorphisms (SNPs) (Macalalad *et al*., 2012; Wilm *et al*., 2012) or perform haplotype reconstruction (Zagordi *et al*., 2011; Shirmer *et al*., 2012). **Haplotype assembly has its weakness in the detection of low-frequency variants (Prosperi, M.C.F. *et al*.) whereas the latter is our main interest. Instead, we prefer to call variants within the read length to avoid challenges encountered in haplotype reconstruction. On the other hand, linkage between the nucleotides is lost when calling variants at SNP level. In this contribution we introduce a novel strategy for calling variants at codon level (nucleotide triplets), which facilitates immediate biological interpretations, particularly in virology applications where drug target regions are of interest.**

In this article we present an innovative approach for variant calling at codon level, named Q-cpileup, that reduces the number of false positive findings by exploiting the quality scores of the nucleotides generated by sequencing. Our thresholding strategy is adaptive so as to provide an optimized threshold for individual samples in each sequencing run. Q-cpileup is imbedded in a pipeline, called VirVarSeq, which starts from fastq files.

2 METHODS

Several samples were sequenced using Illumina's genome analyzer (GA) IIx according to manufacturing protocols (see supplementary material). The VirVarSeq pipeline proceeds as follows:

1. The sequenced reads are aligned against a reference sequence using the Burrows-Wheeler Aligner Tool (BWA) (Li and Durbin, 2009).
2. Based on this alignment a consensus sequence is defined.
3. A realignment is performed against this consensus. This strategy of iterative mapping will increase coverage especially in samples where the consensus strongly deviates from the reference (see supplementary material Figure 1).
4. After alignment, Q-cpileup is executed which consists of a three step analysis:

   a. **In a first step the quality scores of the codons in the reading frame of interest are retrieved.**
   b. Next the threshold is determined dependent on the quality of the run.
   c. Finally the filtered codon table is constructed.

The VirVarSeq pipeline, which runs from fastq to filtered codon table, is available at http://sourceforge.net/projects/virtools/?source=directory together with a users guide (see supplementary material 2). **All reads containing indel errors are removed prior to run Q-cpileup. It is hereby assumed that indels will result in non-viable virus. In some rare occasions, however, there might be an insertion mutation at codon level, which can be investigated in a separate analysis (see indel Table supplementary material).** Below, the different steps from Q-cpileup will be explained in more detail.

2.1 Quality of Codons

A pileup of read bases is generated using the alignments to a consensus sequence. In analogy with mpileup of samtools, for which the base-pair information at each reference position is described, cpileup describes the codon information at each amino acid position of the reference genome. For each position in the reference genome, the different codons are reported together with one quality score for each codon. This requires that the quality scores of the three nucleotides within a codon have to be summarized. A comparative analysis of different summarizations revealed that the weakest link, i.e. minimum quality score of the three nucleotides in the codon, provided the best separation between low and high quality codons (see supplementary material Figure 2). This minimum quality score represents the codon's nucleotide with the highest probability of being a sequencing error. **A codon table is build based on the pileup where for each codon position of the reference the different codons within a sample are reported together with their frequency (Table 1). The minimum quality scores of the codons at a particular position are averaged to give a rough idea about the overall quality. However the individual minimum quality scores of the codons itself are used in subsequent analyses.**

2.2 Q-intersection threshold (QIT)

The distribution of the minimum quality scores was checked and compared for one particular sample sequenced in two different runs and three different lanes reaching average coverages around 30,000 (Figure 1). The shape of the distributions can be approximated by a mixture distribution with 3 truncated normal components (**see supplementary material for model selection and goodness of fit in supplementary figure 3**). Truncation is performed at the lower and upper ends of the quality score range. The first mode represents a point probability at quality score 2, which is the lowest Illumina quality score. This is due to an artifact created by Illumina's base caller. Read ends with a segment of mostly low quality (Q15 or below) are given a quality score of 2. The second component distribution is a distribution of low quality scores, reflecting the sequencing error distribution. Finally, the highest mode, close to 40, originates from a distribution of reliable calls. Note, that the mixture of 3 normal components for the errors and reliable calls should only be considered as a working assumption and that neither trimming nor filtering of the data is required

Table 1. Example of a **frequency table** at codon level. The different codons observed in a sample are counted at each codon position of the reference genome and their frequencies are calculated using the coverage at that particular position. The quality scores are summarized by averaging the minimum quality scores of the codons.

Position	Ref Codon	Codon	Count	Coverage	Frequency (%)	Mean Q
001	GGG	AGG	167	18,966	0.88	35
001	GGG	GTG	83	18,966	0.44	16
001	GGG	TGG	15	18,966	0.08	33
001	GGG	GGG	18,693	18,966	98.6	37
002	CGT	CAG	461	19,217	2.4	30
002	CGT	GGG	20	19,217	0.1	5

Position: amino acid position of the reference. Ref: codon of the reference genome at a particular position. Codon: codon present in a sample at a particular position. Count: the number of times a particular codon occurs in the cpileup at a particular position. Coverage: the number of reads that fully cover a particular codon position. Frequency: Count/Coverage. Mean: Average of the minimum quality scores of a particular codon at a particular position.

prior to fitting the mixture models. The EM algorithm of McLachlan (McLachlan and Jones, 1988) will be applied for fitting these normal mixture models. We have written a R-wrapper to run the original Fortran code of McLachlan which is embedded in the pipeline VirVarSeq. The EM algorithm was initialized by setting the three modes at 2, 10 and 35 and the variances at 0.8 for the point probability and 40 for the other two distributions. The marginal error probability, the sum of mixing proportion of the distributions at 2 and 10, was set to 15%.

The bulk of quality scores were high, indicating a majority of reliable calls in the dataset (green distribution in Figure 1). At the other end a clear point probability at the quality score of 2 was seen. The red distribution in Figure 1 corresponds to low quality codons which are likely to be sequencing error. There are several criteria how to define a threshold for filtering the low quality codons and for distinguishing between errors and reliable calls. An approach is chosen that is adaptive and robust. The intersection point of the two component distributions was used and is referred to as the Q-intersection threshold (QIT), which is indicated with vertical dashed lines in Figure 1. **The distribution of the minimum quality scores of the codons and hence the QIT varies between different runs for the same sample**, confirming the need for an adaptive filtering strategy.

2.3 Filtering of Codon Tables

Once the QIT is determined, an updated codon table can be constructed. All codons with a minimum quality score below the threshold will be filtered from the analysis. The influence of trimming is negligible as it mainly affects low quality nucleotides, which are removed by the filter. The three step analysis returns a codon table with different variants and their frequencies at each codon position of the reference, which is robust to sequencing errors.

3 RESULTS

3.1 HCV plasmids mixtures

To assess the filtering accuracy of Q-cpileup, we made use of two plasmids that carry HCV NS3 amino acids 1 to 181 and that differ only at two codon positions, 36 and 155. The two plasmids were mixed in four different proportions: 1:10, 1:50, 1:100 and 1:200 and sequenced with average coverages of 96,211, 81,179, 95,590 and 74,820 respectively. (see supplementary material for sample preparation; sequencing data are available at the European Nucleotide Archive, accession number PRJEB5028). A minor variant is defined as a codon that differs from the consensus. Hence, only 2 minor variants are expected at codon positions **36 (GTC (consensus) → ATG) and 155 (CGG (consensus) → AAA), all others can be considered as false positives**. For each of the four sequenced mixes the QIT was determined (Table 2). Comparison of the number of variants when no filtering but trimming (QIT=0) is applied and after Q-cpileup reveals that adaptive quality filtering can reduce the number of false positives by 20% to 50% (third and fourth column of Table 2). With Q-cpileup, no false positives are reported with frequencies above 1%, a reporting limit defined in Thys et al. (Thys *et al.*, 2014). This is in strong contrast with the 7% to 12% FDR without filtering, for discoveries with frequencies above 1%. **Q-cpileup is able to reduce the number of false positives while the frequencies of the true minor variants (last columns of Table 2) remain unaltered.**

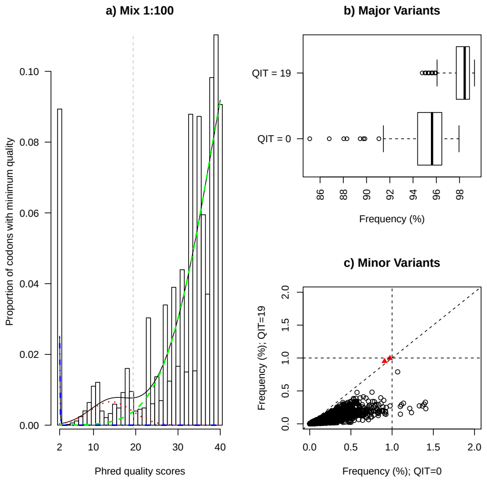

Fig. 2. a) **Distribution of the minimum quality scores of the codons present in the HCV plasmids mixed 1:100. The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold (QIT=19) is indicated with a vertical black dashed line.** b) Boxplot of the major variant frequencies when no filtering is applied (QIT=0) and after Q-cpileup with QIT=19. c) Scatterplot of minor variants (frequencies before filtering (QIT=0) on the X-axis and after Q-cpileup on the Y-axis). The true minor variants at codon positions 36 and 155 are indicated with red triangles, all others can be regarded as false positives. The 1% reporting limit is indicated with dotted lines.

The results for the mixing proportion of 1% are shown in more detail in Figure 2. A QIT of 19 was used for filtering in this sample (Figure 2a). In panel b, the codons equal to the consensus (called major variants) are investigated before (QIT=0) and after filtering (QIT=19). Note that the plasmid data have only two real variants, meaning that the codons investigated here should have frequencies close to 100%. After applying Q-cpileup the frequencies of the codons are indeed closer to hundred, indicating again that the number of false positives is reduced. The minor variants are compared in panel c. Without filtering, several minor variants are reported with frequencies above the reporting limit of 1%. Their frequencies are strongly reduced after filtering, while the estimates of the true minor variant frequencies remain (red triangles). This suggests again that quality score filtering provides effective noise reduction while still retaining the reliable calls at low frequency. Figure 2c reveals that our new filtering method allows for lowering the reporting limit, although, the discovery of true variants below 0.5% remains challenging.

3.2 Comparison with LoFreq, V-Phaser 2 and ShoRAH

The performance of Q-cpileup is compared with three other methodologies: LoFreq (v0.5.0) (Wilm *et al.*, 2012), V-phaser 2 (v2.0) (Macalalad *et al.*, 2012), and ShoRAH (v0.8) (Zagordi *et al.*, 2011). They were run in their default settings and using the previously described plasmid mixtures. With ShoRAH we were unable to use the original bam file since unresolvable problems (even after discussion with the developers) were encountered when extracting the reads from the desired region. Therefor the ShoRAH results are based on a bam file with remapped reads against the reference region of interest. None

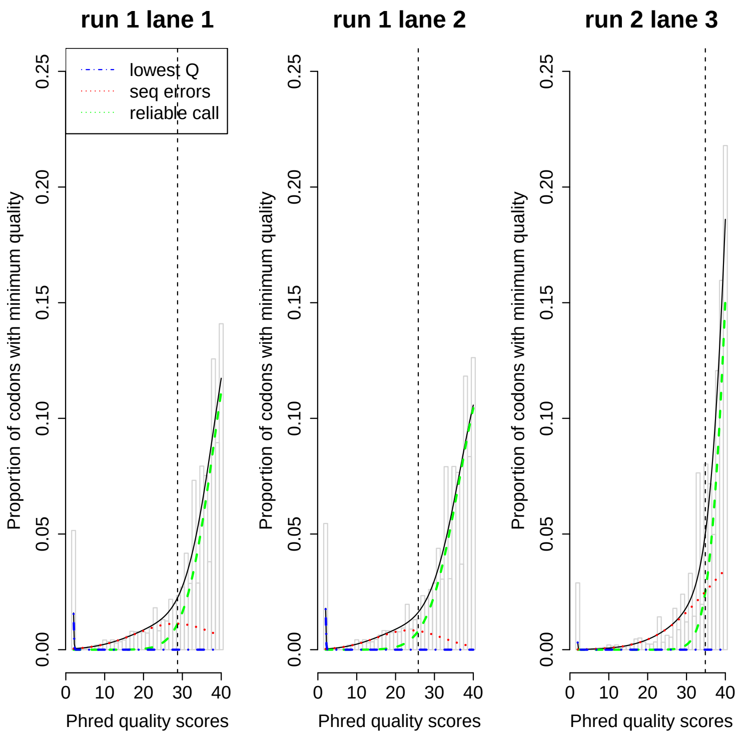

Fig. 1. Distribution of the minimum quality scores of the codons present in a HCV sample which was sequenced twice in the same run (run1) but in different lanes (lane 1 and 2), and which was sequenced yet another time in another run (run2). The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold (QIT) is indicated with a vertical black dashed line.

Table 2. HCV plasmids results for 181 codon positions sequenced with average depth of 87000. For each of four mixing proportions, the QIT is reported that was used for filtering. The number of reported codons, the false discovery rate for discoveries with frequency above 1% and the estimated frequencies of the true variants are compared before filtering is applied and after Q-cpileup.

Mix	QIT	N° Variants		FDR (%)		Freq Codon 36 (%)		Freq Codon 155 (%)	
		no filtering	Q-cpileup	no filtering	Q-cpileup	no filtering	Q-cpileup	no filtering	Q-cpileup
1:10	20	22,405	10,583	12	0	10.4	11.6	10.3	11.6
1:50	20	12,724	10,488	7	0	2.28	2.37	2.25	2.37
1:100	19	14,886	10,631	6	0	0.97	1.00	0.91	0.95
1:200	19	15,692	10,813	8	0	0.50	0.49	0.43	0.45

of the existing methods calls variants immediately at codon level. Hence, three SNP callers were chosen based on their capabilities for inferring diversity within viral populations. The two codon variant positions present in the mixtures, differ at 5 nucleotide positions which should be discovered by the SNP-callers. The results are presented in top part of Table 3. None of the methods could discover the 5 SNPs at 0.5% level and only LoFreq could retrieve all SNPs present at 1%. Hence, LoFreq is the most sensitive SNP-caller in this comparison, but it also detects some false positives with frequency above 1% (bottom rows of Table 3). Comparison with Table 2 reveals that the sensitivity and specificity for discoveries above 1% with Q-cpileup is outstanding. Especially its sensitivity is of utmost importance: the method is initially developed for finding resistance-associated mutations and missing important variants might mislead further treatment.

3.3 Clinical HCV sample and comparison with 454

Subsequently, Q-cpileup was applied on a clinical HCV sample (see supplementary material for sample preparation). The fit of

Table 3. Frequency estimates of the 5 SNPs present in the mixture of plasmids with mixing proportions 1:200, 1:100 and 1:50. In case of ShoRAH, the frequency is estimated from three overlapping windows, but often the SNP is only retrieved in two out of the three windows (denoted with *). The bottom rows of the table report the number of codons detected in the NS3 region which in theory should be 548 (543 WT + 5 SNPs). The number of false discoveries with frequencies above 1% is expressed using the false discovery rate calculated as the number of false discoveries with frequencies above 1% divided by the total number of discoveries with frequency above 1%.

		LoFreq			V-Phaser 2			ShoRAH		
	SNP (WT)	1:200	1:100	1:50	1:200	1:100	1:50	1:200	1:100	1:50
36	A (G)	/	1.03	2.41	0.59	1.06	2.37	/	/	2.22
	T (T)									
	G (C)	0.54	1.01	2.38	/	0.94	2.33	/	/	2.22
155	A (C)	0.66	1.03	2.16	/	/	/	0.44*	0.80*	1.78*
	A (G)	0.48	0.91	2.10	0.52	1.04	2.11	0.44*	0.80*	1.78*
	A (G)	/	0.89	2.05	0.48	/	2.07	/	/	1.28
	FDR (%)	0.18	0.18	0.18	0	0.18	0	0	0	0
	N° Variants	549	553	550	565	578	571	549	546	549

the mixture distribution returned a QIT of 26 (Figure 3a). In Figure 3b the frequencies of the codons before and after filtering were plotted on the log scale, which allows a better comparison of low-frequency variants. The frequencies of the variants that were removed after filtering were indicated in gray at the bottom. As the truth is unknown in clinical samples one cannot reliably separate true variants from sequencing errors. However, one could compare the discovered variants with 454 sequencing results. Since 454 sequencing chemistry is different, another error profile can be expected. Hence, variants that were not discovered with 454 are more likely to be Illumina sequencing errors (and vice versa) and were indicated with red triangles in Figure 3b. Above 1% a good correlation between the filtered and the unfiltered variant frequencies is observed. The variants that were not detected in the 454 experiment, likely to be false discoveries, drop in frequency after applying Q-cpileup. Hence, our approach seems to control the false discovery rate at a reasonable level up to variant frequencies of 0.5%. Note, that a lower coverage depth of the 454 experiment did not allow for comparing Illumina and 454 sequencing for frequencies below 0.1%.

3.4 Effect of inter-/intra-run variability on QIT

An equimolar pool of forty-two clinical HCV samples was sequenced three times so as to investigate the variability in sequencing quality and hence the variability of the QIT. The 42 samples were sequenced twice within the same run (R1) but on two different lanes (L1 and L2) and they were all sequenced again in another run (R2 L3). For one of these samples the mixture distribution of each of the three sequencing runs is shown in Figure 1. **A clear difference in quality between the runs is observed, resulting in different QIT's.** Boxplots of the QIT's of 42 samples for each of the sequencing runs shows that the inter-run variability of the QIT is larger than the variability between lanes of the same run (Figure 4).

First, the effect of the intra-run variability **of the QIT's** on the final codon table was investigated. The number of reported codons, with frequency above 1% are plotted for both lanes in Figure 5a. Seventy-four percent of the samples report at most 1 additional codon, depending on the lane on which it is sequenced. The maximum difference in reported number of codons is four.

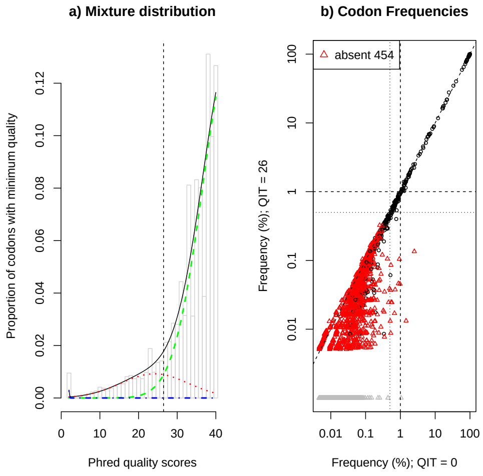

Fig. 3. a) **Distribution of the minimum quality scores of the codons present in a HCV clinical sample. The black line shows the overall fit of the mixture distribution, which consists of the blue, the red and the green component distributions. The blue and the red distributions correspond to codons that likely result from sequencing errors, and the green distribution represents reliable calls. The quality intersection threshold is indicated with a vertical black dashed line.** b) Scatterplot of estimated codon frequencies before (QIT=0) and after filtering (QIT=26) on the x-axis and y-axis, respectively. The reporting limit of 1% and 0.5% are indicated with dashed lines and dotted lines, respectively. The codons that were not reported after Q-cpileup are indicated in gray at the bottom. Codons not detected with 454 sequencing were indicated with red triangles and are likely to be false positives.

We further explored frequency differences for all codons. The distribution of the differences in frequency between the two lanes is plotted for each sample in Figure 5b. Overall, the frequencies are similar with some deviations up to 3%. These maxima, however, are mainly originating from codons located in a GC-rich region, where a coverage drop is observed. The sample where the maximum absolute difference is observed (indicated in red), is investigated in

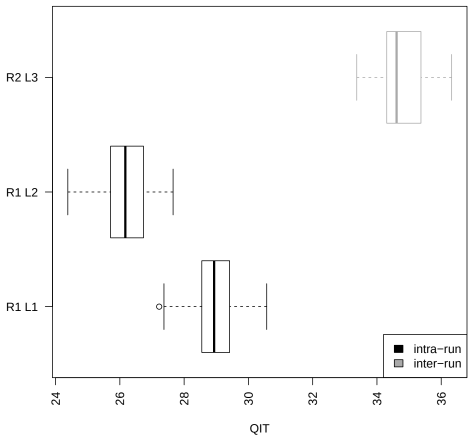

Fig. 4. Boxplot of the QIT's of 42 clinical samples for each of the sequencing runs denoted by run (R) and lane (L) to investigate inter- and intra-run variability. The black boxplots are part of the intra-run comparison, while the gray boxplot represents the QIT's of the other run. The inter-run variability is larger compared to the intra-run variability.

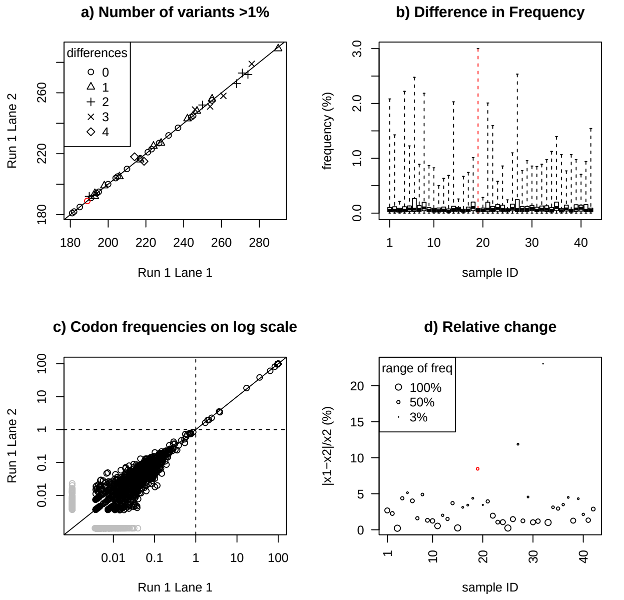

Fig. 5. Effect of the intra-run variability of the QITs on the final reported codon frequencies. a) Plot of the number of **codons** with a frequency greater than 1% for **the 42 samples sequenced on lane 1 (x-axis) and lane 2 (y-axis)**. The plotting symbols indicate the number of codons that differ between the two lanes. b) Boxplots of differences in codon frequency between two lanes for each of the 42 samples. For each of the samples, 75% of the codons show differences in frequencies close to zero**, while the upper whiskers range roughly between 0.5% and 3% difference in reported codon frequency, depending on the lane where the sample was sequenced.** c) Comparison of all codon frequencies on the log scale between two lanes for the sample where the maximum frequency difference is observed. The frequencies of codons not present in the other lane are plotted in gray. d) Relative change for codons with maximum absolute difference plotted for each sample. The relative change is calculated $[x_1\text{-}x_2]/x_2$ with x_1 and x_2 the codon frequencies for lane 1 and lane 2 respectively (Sample with the maximum absolute difference in red). The sizes of the dots are scaled according to the estimated frequency in lane 1, indicating that most maximum differences occur at higher frequencies.

detail in Figure 5c. The frequencies for all codons discovered on both lanes are plotted against each other on the log scale. Above 1% no substantial differences can be observed. Finally, the maximum differences, for each of the 42 samples, are reported on a relative scale in Figure 5d. The sizes of the dots are scaled according to the absolute frequency obtained in lane 1, which teaches us that most deviations occur for frequencies above 50%. Overall, the reported variants and their frequencies are comparable on both lanes after applying Q-cpileup, despite slightly different QITs. This is in strong contrast with the raw data (see Figure 4 in supplementary material). These raw data were only trimmed, which is partially based on quality scores. Without Q-cpileup the comparison of the samples sequenced on both lanes reveal that (a) the number of reported variants differ up to 16 variants, (b) deviations of the frequencies go up to 6% and (c) even for the variants with frequency above 1% some clear deviations between the two lanes can be observed. This suggests that Q-cpileup is able to reduce the number of false positives while retaining the true signal and that the adaptive approach is able to account for differences in quality between the lanes.

In the next step the inter-run variability of the QITs was investigated. The second run was a very high quality run with less error (Figure 1 comparing run 1 lane 2 and run 2 lane 3). The overall good quality in run 2 makes the estimation of the error component of the mixture challenging and results in a large QIT. The effect of these high QIT's on the final codon table was investigated. The number of reported variants, with frequency above 1%, are again very similar even after applying these high thresholds (Figure 6a). 83% of the samples report at most 1 additional variant and only 1 sample report 4 additional variants, which is again the maximum difference in reported number of variants. The distribution of the frequency differences remains small overall, but the maximum difference in frequency rises as well as the relative change for these maxima (Figure 6b and 6d). But these maxima occur again at rather high frequencies. When focusing on a particular sample, high QITs seem to have no negative impact on the final codon table (Figure 6c). Despite the questionable working assumption that the second component separates the error distribution from the distribution of the reliable calls, the resulting QITs provide reliable codon tables. Hence, our filtering approach seems robust to deviations from the working assumption. Q-cpileup is adaptive and thereby can cope with differences in quality between runs.

3.5 Robustness of the method

Approximately 400 samples, from both HCV and HIV infected subjects, were sequenced in 7 different runs and analyzed with Q-cpileup. Some examples of the HCV results are displayed in Figure 3 and 5. The sequenced amplicons of HCV samples cover GC-rich

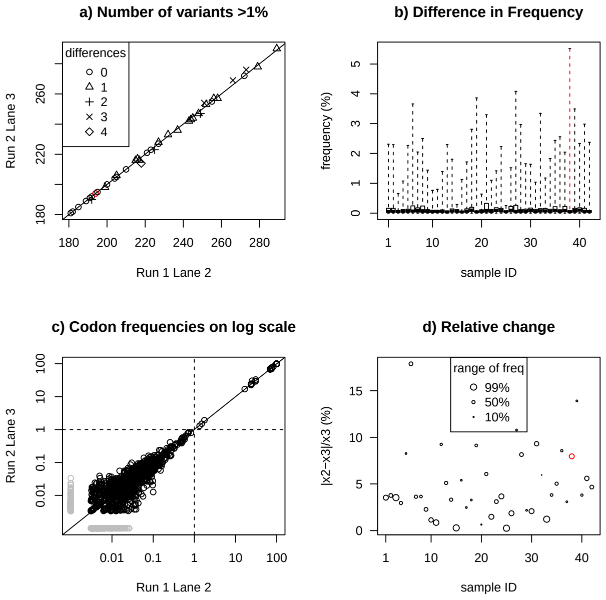

Fig. 6. Effect of the inter-run variability of the QITs on the final reported codon frequencies. a) Plot of the number of **codons** with a frequency greater than 1% **for the 42 samples sequenced on run 1 lane 2 (x-axis) and on run 2 lane 3 (y-axis)**. The plotting symbols indicate the number of codons that differ between the two lanes. b) Boxplots of differences in codon frequency between two runs for each of the 42 samples. For each sample, 75% of the codons show differences in frequencies close to zero **, while the upper whiskers range roughly between 1% and 5% difference in reported codon frequency, depending on the run where the sample was sequenced.** c) Comparison of all codon frequencies on the log scale between the two runs for the sample where the maximum frequency difference is observed. The frequencies of variants that are absent in the other run are plotted in gray. d) Relative change for codons with maximum absolute frequency difference plotted for each sample and scaled according to the estimated frequency in condition 2. The relative change is calculated as [x_2-x_3]/x_3 with x_2 and x_3 the codon frequencies for run 1 lane 2 and run 2 lane 3 respectively (sample with maximum difference is indicated in red).The sizes of the dots are scaled according to the estimated frequency in run 1, indicating that most maximum differences occur at higher frequencies.

regions. It is known that Illumina is error prone in these regions (Dohm *et al.*, 2008). This is reflected in the distribution of the minimum quality scores where more low values are observed than with amplicons of HIV samples for which no GC-rich regions were covered (Supplementary material Figure 5). **It is especially in these GC-rich regions where you expect that false positives exist with frequencies above 1% as seen in Figure 3b on the x-axis. When applying Q-cpileup, the frequencies of these false positives could be reduced.** In Figure 5 of the supplementary material, we illustrate that the proposed strategy also similarly reduces the noise for HIV samples.

Although this new variant calling tool was primarily developed based on Illumina's GAIIx sequencing data, it can also be a valuable tool for other Illumina sequencing platforms such as HiSeq. Indeed, HiSeq uses the same sequencing-by-synthesis technique and suffers from the same error types (Minoche *et al.*, 2011). HiSeq data, analyzed with Q-cpileup, are presented in Figure 6 in supplementary material. The tool, however, is not applicable for pyrosequencing techniques such as Roche 454. Their quality scores do not reflect substitution error probabilities, but probabilities of calling homopolymers of particular length (De Beuf *et al.*, 2012). For these types of data, the tool can however still generate a codon table but no threshold determination or QIT-based filtering should be applied.

4 DISCUSSION

Many sequence variant identification tools have been described in literature to call variants at SNP level. Most approaches are tailored to call SNPs in human resequencing projects (Nielsen *et al.*, 2011) where SNPs can be either heterologous (50%) or homologous (100%). However, in viral deep sequencing projects, the SNP frequency can vary between 100% and 0%, whereas also SNPs present in less than 1% of the reads may be of interest. **Further, our main focus is the detection of drug resistant variants for which a specific drug target region has to be investigated. Hence, it would be beneficial to report variants at codon level per amino acid position so as to enable an immediate biological interpretation of the variants with respect to their anti-viral drug responses.** Variant calling tools used by the virologists in this research field are either not fully described in-house tools (Noguera-Julian *et al.*, 2013) or build based on SNPs (Parameswaran *et al.*, 2012) at DNA or RNA level only. In the latter case one needs to retrieve the linkage information between the neighboring nucleotides in order to deduce effects at coded amino acid level. This process is not always straightforward. Some of the variant callers developed for viral population sequencing have add-on tools, like V-profiler (Henn *et al.*, 2012) for V-phaser (Macalalad *et al.*, 2012), to convert the list of SNPs to a list of codon variants. However, these are mainly developed for 454 data. None of the available tools report the variants immediately at codon level. Therefor Q-cpileup was initially developed and imbedded in a pipeline. By taking the weakest link as a representative for the quality of the codon, the filtering remains at individual nucleotide level but reporting is done at codon level. The approach is adaptive to allow for differences in quality between the runs.

The intersection point between the distributions of the errors and reliable calls is suggested as a threshold (QIT). However, other criteria could also be considered. For instance, the number of false discoveries can be controlled by defining the QIT as a certain quantile of the reliable call component distribution. By using the 5% quantile as a QIT, 5% of the codons, which are truly present in the population, are falsely considered as being error. Note that this statement relies heavily on the interpretation of the distributions as errors and reliable calls. However, the three component mixture distribution should only be considered as a working assumption. Moreover, it is impossible to check the actual interpretation of the mixture distributions. Hence, the interpretation of the different components as error and reliable calls can not always be warranted, particularly when low quality scores are underrepresented. Therefore we advise users to assess the distribution plots as diagnostic tool to critically judge whether the chosen threshold is acceptable and/or meaningful.

Instead of applying hard filtering with a QIT, as suggested in this paper, the posterior probabilities of the reliable call distribution could be used. The counts of a codon at a particular position can be weighted with these probabilities. By doing so, codons with a low quality score will have a low contribution in the final frequency estimates. Hence, data are not filtered but weighted with the probabilities of being truly present in the reliable codon population. This method also relies on the interpretation of the different component distributions as error distribution and reliable call distribution.

Importantly, we have shown that the filtering strategy using hard threshold QIT is robust to runs where the distributions deviate from the working assumption. The impact on the final codon table frequencies was minimal. Overall our proposed filtering strategy controls the false positive rate at reasonable levels with no false discoveries above a reporting limit of 1%. The noise is effectively reduced while retaining the reliable calls even at low frequency. This suggests that the reporting limit of detection at 1% could be lowered, although distinguishing true variants from error at 0.5% remains challenging. **Depending on the risk one is willing to take to include a small number of false positives, either one of the cut-offs can be used. More importantly, Q-cPileup shows a splendid sensitivity which could not be achieved by the SNP-callers while maintaining a good specificity for variants with frequencies down to 0.5%. This sensitivity will allow further investigation of the reported variants above one of the cut-offs defined by the specificity to search for resistance-associated mutations, or in the next step to monitor drug resistance and guide treatment (Dierynck *et al*., 2014). Currently the clinical cut-off is not yet defined for minority drug-resistant virus variants and it is still a subject of open debate (Schneider *et al*., 2014). Some studies have found no significant association between the presence of low-frequency variants and subsequent virological failure, while others report clear correlations (Vandenhende *et al*., 2014). The availability of methods that detect low-frequency variants at codon level with high sensitivity and good specificity, can help in defining the clinical benefit of low-frequency resistance testing.**

5 CONCLUSION

A variant calling tool is proposed for identifying true variants at codon level within a viral population using Illumina sequencing. The variants are filtered using base-calling quality scores for reducing false positive findings. The lowest quality score of the three nucleotides of the codon is taken as representative for the codon. An adaptive strategy is developed to provide an optimized threshold for individual samples in each sequencing run. The intersection point of the component distributions of the mixture is suggested as a valuable threshold, QIT. Codons with a quality score below this threshold are not reported. The robustness against deviations of the working assumption justifies the utilities of our method for low and high quality sequencing runs. It is shown that the generated filtered codon table is reporting far less false positives **compared to codon table based on the raw data. Moreover, VirVarSeq has a superb sensitivity compared with existing SNP-callers while maintaining a good specificity for codon variants with frequencies down to 0.5%. This suggests that the current reporting limit of detection at 1% can even be lowered.** The tool is implemented in a user friendly open source pipeline, VirVarSeq, which allows virologist to call variants at codon level starting from the fastq files.

ACKNOWLEDGEMENT

The authors wish to thank the scientists at Janssen Pharmaceutica who collected and produced the data as well as Tobias Verbeke and Joris Meys who gave the necessary IT support. Herwig Van Marck is kindly acknowledged for the insightful discussions. We are grateful to Prof McLachlan and his team for providing the original Fortran code and for giving support to write the R-wrapper. Thanks to Osvaldo, developper of ShoRAH, who was always very helpful when problems were encountered while running ShoRAH.

Funding: Part of this research was supported by Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT) who provided a Baekeland mandatory to BV [BM100679], by IAP research network grant [no. P7/06] of the Belgian government (Belgian Science Policy) and by Multidisciplinary Research Partnership Bioinformatics: From Nucleotides to Networks Project [01MR0310W] of Ghent University. We thank Luc Bijnens for the financial support given to the project at Janssen Pharmaceutica.

REFERENCES

Beerenwinkel, N.*et al*. (2012) Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Frontiers in Microbiology*, doi: 10.3389/fmicb.2012.00329.

Beerenwinkel, N. and Zagordi, O. (2011) Ultra-deep sequencing for the analysis of viral populations. *Current Opinion in Virology*, **1**, 413-418.

Codoner, F.M. *et al*. (2011) Added value of deep sequencing relative to population sequencing in heavily pre-treated HIV-1-infected subjects. *PLoS One*, doi: 10.1371/journal.pone.0019461.

De Beuf, K. *et al*.(2012) Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model *BMC Bioinformatics*, **13** doi:10.1186/1471-2105-13-303.

Dierynck I, *et al* Deep Sequencing of the HCV NS3/4A Region Confirms Low Prevalence of Telaprevir-resistant Variants Both at Baseline and End of Study. *Journal of Infectious Diseases*, accepted.

Dohm, J.C.*et al*. (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, doi: 10.1093/nar/gkn425.

Domingo, E. *et al*. (2012) Viral quasispecies evolution. *Microbiology and molecular biology reviews*, **76**, 159-216.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.*, **8**, 186-194.

Gianella, S.*et al*. (2010) Minority variants of drug-resistant HIV. *J Infect Dis.*, **202**, 657-666.

Henn M.R. *et al*. (2012) Whole Genome Deep Sequencing of HIV-1 Reveals the Impact of Early Minor Variants Upon Immune Recognition During Acute Infection. *PlOS Pathogens*, doi: 10.1371/journal.ppat.1002529.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-1760. *Bioinformatics*, **25**, 1754-1760.

Macalalad, A.R. *et al*. (2012) Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Computational Biology*, doi: 10.1371/journal.pcbi.1002417 .

McLachlan, G.J. and Jones, P.N. (1988) Fitting Mixture Models to Grouped and Truncated Data via the EM Algorithm *Biometrics*, **44**, 571-578.

Minoche, A.E.*et al*. (2011) Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biology*, doi:10.1186/gb-2011-12-11-r112.

Nielsen, R.*et al*. (2011) Genotype and SNP calling from next-generation sequencing data *Nat Rev Genet*, **12**, 443-451.

Noguera-Julian M.*et al*. (2013) Stable HIV-1 integrase diversity during initial HIV-1 RNA Decay suggests complete blockade of plasma HIV-1 replication by effective raltegravir-containing salvage therapy. *Virology journal*, doi:10.1186/1743-422X-10-350.

Parameswaran P.*et al*. (2012) Genome-Wide Patterns of Intrahuman Dengue Virus Diversity Reveal Associations with Viral Phylogenetic Clade and Interhost Diversity. *Journal of Virology*, doi: 10.1128/JVI.00736-12.

Prosperi, M.C.F. *et al*. (2013) Empirical validation of viral quasispecies assembly algorithms: state-of-the-art and challenges. *Sci. Rep. 3*, doi:10.1038/srep02837.

Reumers, J. (2012) Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. *Nature Biotechnology*, **30**, 61-68.

Schneider, M.D., *et al*. (2014) Antiviral therapy of hepatitis C in 2014: Do we need resistance testing? *Antiviral Research*, Doi: 10.1016/j.antiviral.2014.02.011.

Shirmer, M. *et al*. (2012) Benchmarking of viral haplotype reconstruction programmes: an overview of the capacities and limitations of currently available programmes *Brief Bioinform*, doi: 10.1093/bib/bbs081 .

Thys, K. *et al*. (2014) Evaluating the use of the Illumina deep sequencing platform for the detection of minority variants in HIV and HCV. *Journal of Virological Methods*, under review.

Vandenhende, *et al*. (2014) Prevalence and Evolution of Low Frequency HIV Drug Resistance Mutations Detected by Ultra Deep Sequencing in Patients Experiencing First Line Antiretroviral Therapy Failure. *PLoS ONE*, doi:10.1371/journal.pone.0086771.

Wilm, A. *et al*. (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets *Nucleic Acids Research*, **40**, 11189-11201.

Zagordi, O. *et al*. (2011) ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data *BMC Bioinformatics*, doi:10.1186/1471-2105-12-119.