


SOFTWARE ARTICLE

Open Access



# Estimation of genetic diversity in viral populations from next generation sequencing data with extremely deep coverage

Jean P. Zukurov<sup>1</sup>, Sieberth do Nascimento-Brito<sup>2,5</sup>, Angela C. Volpini<sup>6\*</sup> , Guilherme C. Oliveira<sup>6</sup>, Luiz Mario R. Janini<sup>1,2</sup> and Fernando Antoneli<sup>3,4</sup>

## Abstract

**Background:** In this paper we propose a method and discuss its computational implementation as an integrated tool for the analysis of viral genetic diversity on data generated by high-throughput sequencing. The main motivation for this work is to better understand the genetic diversity of viruses with high rates of nucleotide substitution, as HIV-1 and Influenza. Most methods for viral diversity estimation proposed so far are intended to take benefit of the longer reads produced by some next-generation sequencing platforms in order to estimate a population of haplotypes which represent the diversity of the original population. The method proposed here is custom-made to take advantage of the very low error rate and extremely deep coverage per site, which are the main features of some neglected technologies that have not received much attention due to the short length of its reads, which precludes haplotype estimation. This approach allowed us to avoid some hard problems related to haplotype reconstruction (need of long reads, preliminary error filtering and assembly).

**Results:** We propose to measure genetic diversity of a viral population through a family of multinomial probability distributions indexed by the sites of the virus genome, each one representing the distribution of nucleic bases per site. Moreover, the implementation of the method focuses on two main optimization strategies: a read mapping/alignment procedure that aims at the recovery of the maximum possible number of short-reads; the inference of the multinomial parameters in a Bayesian framework with smoothed Dirichlet estimation. The Bayesian approach provides conditional probability distributions for the multinomial parameters allowing one to take into account the prior information of the control experiment and providing a natural way to separate signal from noise, since it automatically furnishes Bayesian confidence intervals and thus avoids the drawbacks of preliminary error filtering.

**Conclusions:** The methods described in this paper have been implemented as an integrated tool called *Tanden* (Tool for Analysis of Diversity in Viral Populations) and successfully tested on samples obtained from HIV-1 strain NL4-3 (group M, subtype B) cultivations on primary human cell cultures in many distinct viral propagation conditions. *Tanden* is written in C# (Microsoft), runs on the Windows operating system, and can be downloaded from: <http://tanden.url.ph/>.

**Keywords:** Viral diversity, Bayesian inference, Dirichlet distribution

\*Correspondence: [avolpini@cpqrr.fiocruz.br](mailto:avolpini@cpqrr.fiocruz.br)

<sup>6</sup> Genomics and Computational Biology Group, Centro de Pesquisas René Rachou (CPqRR), Fundação Oswaldo Cruz (FIOCRUZ), Belo Horizonte, Brazil

Full list of author information is available at the end of the article

## Background

Viruses with RNA genomes are recognized to generate particularly mutant-rich populations called *quasispecies*. The genetic heterogeneity characteristic of viral quasispecies is largely due to high mutational rates combined with an elevated population size [1]. Human immunodeficiency virus 1 (HIV-1), as an example, has a mean substitution rate of order  $10^{-5}$  per nucleotide position [2]; that is by far higher than those of cellular organisms [3, 4] and assures a constant viral mutant production.

Next-generation sequencing (NGS) platforms have been used mainly for the de novo sequence assembly of viruses. However, more recently, new interest arose in re-sequencing known virus genomes using NGS to study the diversity of viral populations. All NGS platforms produce short segments of DNA, called *reads*, which provide only imperfect and incomplete information about the structure of the viral population. Sequencing errors and length of reads are factors that must be taken into account in the analysis of data obtained from NGS viral quasi-species. In addition, reverse transcription and PCR amplification are procedures prone to errors. The impact of these errors on studies of viral diversity could be huge (see below), therefore one wants to separate true genetic variation from methodological noise and if both are of the same order of magnitude the task becomes virtually impossible.

Regarding the development of tools to estimate genetic diversity of viral populations (total number of genetic characteristics in a viral ensemble), the most commonly used NGS platforms are the 454<sup>TM</sup> (Life Sciences/Roche)—since Roche's announced in October 2013 it will shut-down the 454<sup>TM</sup> platform [5] its use has been fading—and the Illumina<sup>TM</sup> (Solexa), mainly due to their capacity to produce longer reads. The ability to produce relatively long sequences favors the development of methods aiming at the haplotype (the genetic variants in a viral population) reconstruction of the representative viral particles in population [6, 7]. However, the propagation of sequencing errors is a serious problem in these methods, requiring the development of procedures for error correction, which may introduce unwanted biases. In general, the fraction of wrong reads increases with the error rate per base and the average length. The expected proportion of reads with at least one sequencing error as a function of the error rate per base  $\epsilon$  and the average length  $L$  of reads is given by the relation  $1 - (1 - \epsilon)^L$  [8]. As the estimated error rate of 454<sup>TM</sup> is about 0.1–0.5 % and Illumina<sup>TM</sup> error rates are in the range of 0.1–1 % [9], with an average length of reads from 400 bp up to 1000 bp, the proportion of reads with at least one error is in the range 35–90 %. The platform SOLiD<sup>TM</sup> (Life Technologies), for instance, is at the other end of the spectrum. With reads of short length, of at

most 50 bases (the main limitation for the construction of haplotypes) and estimated error rate of 0.06 % [9], the proportion of reads with at least one error is around 2 %. Recently, a different solution to the problem of sequencing errors has been proposed [10], based on the development of high-fidelity sequencing protocols [11].

A more serious challenge associated with the assembly of all possible haplotypes is the *NP-hardness* of the corresponding combinatorial optimization problems [12]. In fact, some approximate solution must be employed and a crucial hindering factor is the ratio between the size of the reads and the size of the genomic region being reconstructed. For instance, it has been reported [10] that short read lengths (less than 100 base pairs) dramatically inhibit reconstruction of genomes with more than 3400 bp, evidenced by the failing to produce any complete genome. Another major shortcoming of all existing methods for haplotype reconstruction is that they are unable to handle large insertions or deletions (indels), only very recently this problem seems to have been overcome [13].

As mentioned before, the ability of the other NGS platforms to produce relatively long sequences have been a great stimulus to the development of methods for building *haplotype representatives* of viral particles in the population and the vast majority of softwares for viral diversity estimation that have been proposed until very recently adopt this perspective [6]. The aim of this work is to propose a different approach to measure genetic diversity that does not demand any kind of length assumption on the short reads, but takes advantage of the low error rate and the high depth of coverage per site inherent to some NGS platforms. Therefore, we shall considerably depart from the most traditional developments aiming at haplotype reconstruction, since not every one has access to the NGS platforms appropriate for that purpose. Indeed, although the short length of the reads produced by these platforms essentially hinders haplotype reconstruction, it is possible to measure genetic diversity through probability distributions along the genome (one per site) and this approach is enhanced by the highly deep coverage provided by these NGS platforms.

A recent study [14] comparatively assessed the performance of some NGS platforms (including 454<sup>TM</sup> and Illumina<sup>TM</sup>) and reported an average (range) coverage of ~23,000 reads (5000–47,000) for the Illumina<sup>TM</sup> and ~7000 reads (2000–22,000) for the 454<sup>TM</sup>. We used the SOLiD<sup>TM</sup> platform and were able to achieve an average (range) coverage of ~50,000 reads (10,000–150,000), for instance (see Fig. 2). In addition, the low error rate of 0.06 % provided by the SOLiD<sup>TM</sup> platform virtually eliminates the necessity of any error correction procedure. Instead, we use the estimated probability distributions to separate signal from noise.

The first step in nucleotide sequence analysis is read mapping/alignment. This is important for many bioinformatics applications, as exemplified by nucleic acid conformational structure prediction and phylogeny studies [15, 16]. As expected, this is also an important aspect for NGS data analysis involving all the different platforms as Ion Torrent™ (Life Technologies), SOLiD™, 454™ and Illumina™ [17, 18] and others. Nowadays, users can choose from a panoply of tools for mapping and indexing NGS reads, available on-line and for download. MAQ (Mapping and Assembly with Qualities) [19], BWA (Burrows–Wheeler alignment tool) [20], BFAST (Blat-like fast accurate search tool) [21], Bowtie [22] and MOSAIK [23] are examples of such alternatives. Those tools allow the fast mapping and alignment of reads belonging to genomes up to  $10^9$  bp in length [20, 24, 25].

After the read mapping is finished, the following step consists in the choice of a strategy for statistical inference. There is a wide variety of methods depending on the scope and the goals of the analysis: (1) consensus generation, (2) *single nucleotide variant* (SNV), also called *single position diversity estimation*, (3) *local diversity estimation* and (4) read graph-based haplotype reconstruction, also known as *global diversity estimation*, see [8, 26] for a thorough explanation of these concepts. Existing tools for genetic diversity evaluation of viral NGS sequences, intended for 454™ and Illumina™ platforms [8, 26–33], are based on several techniques aiming at haplotype reconstruction [30, 34–39].

In order to estimate the genetic diversity without resorting to haplotype reconstruction, we propose to represent the genetic diversity of a sample population through a family of multinomial probability distributions indexed by the sites of the virus genome, each one representing the distribution of nucleic bases per site. Moreover, the inference of the multinomial parameters is done in a Bayesian framework using smoothed Dirichlet estimation inspired by a method for modeling text data [40].

Inference of multinomial parameters is a challenging problem in statistics. For the simplest case, i.e., the Bernoulli model or binomial estimation, the history traces back to Thomas Bayes [41]. Karl Pearson [42] called this seemingly simple problem the “fundamental problem of practical statistics”. In the frequentist context the problem is called “interval estimation of a binomial proportion” and there is a textbook solution based on a confidence interval for this problem, which however has several drawbacks [43]. In both frameworks the frequency of occurrence of a category plays a crucial role, leading to the “sufficient statistics” in the Bayesian context and the “estimator of proportion in a sample” in the frequentist context.

The choice of a Bayesian framework is motivated by two features that are not present in the frequentist framework: (1) it allows one to obtain conditional (posterior) probability distributions for the multinomial parameters and thus interpret the point estimates as probabilities—this interpretation conceptually incorrect when applied to pure relative frequencies (which is not the same thing as adopting frequentist framework)—even though the law of large numbers implies that they converge to the point estimates obtained from the Dirichlet distributions when the number of observations goes to infinity; (2) one may take into account the prior information of the control experiment (whose genome sequence is known) within the inference of a posterior experimental condition by means of *Bayes’ formula* and thus relate two temporally connected events. Finally, it provides a natural way to separate signal from noise through credible (or Bayesian confidence) intervals—another problem with the use of pure relative frequencies is that it is not possible to associate “error bars” to them. Therefore, in our approach, the errors introduced during the sequencing process are not filtered before the inference, but after it, when we identify the relevant signal—this allows us to avoid the drawbacks of preliminary error filtering [10, 44].

In summary, we sought to build an analysis platform suitable to address the problem of estimation of the populational genetic diversity of RNA viruses. Due to high mutational rates and accelerated replicative kinetics, RNA viruses constitute ensembles of variants, known as *quasispecies*, which, instead of a collection of viral particles, behave as a single and coherent organism which is act on by the host’s pressures [45]. Furthermore, our mathematical and computational approach allows for a better understanding of virtually all the viral diversity present in a clinical sample. By saying this, we mean that our method gives the user an idea about the real structure of a viral population contained in a clinical sample at a given time. When a quasispecies population is challenged by selective pressures, it responds as a sole organism due to the mutational link between the genetic variants it contains. Following this train of thoughts, at any given moment this distribution of mutants may bear variants with resistance mutations (which could minimize therapeutic success) or virus with genomic compositions that were sufficiently close to therapeutical resistance. In this manner, as far as clinical applications are concerned, our method offers an opportunity to the clinician to observe the entire viral mutational landscape in a clinical sample. This comprehensive view could help the clinician when deciding the best therapeutic approach.

Based on the aforementioned assumptions and that a NGS platform generates an extremely high number of

reads of short length allowing for a deep and extensive coverage of the data, and with a very low error rate, we propose an approach to the estimation of genetic diversity of viral populations that does not make requirements on the form of the sequenced data (such as [10], which works only with Illumina™) and does not assume any statistical model for filtering errors [8]. Despite its apparent mathematical and computational involvement the approach proposed here is one of the simplest conceptually correct possible choices.

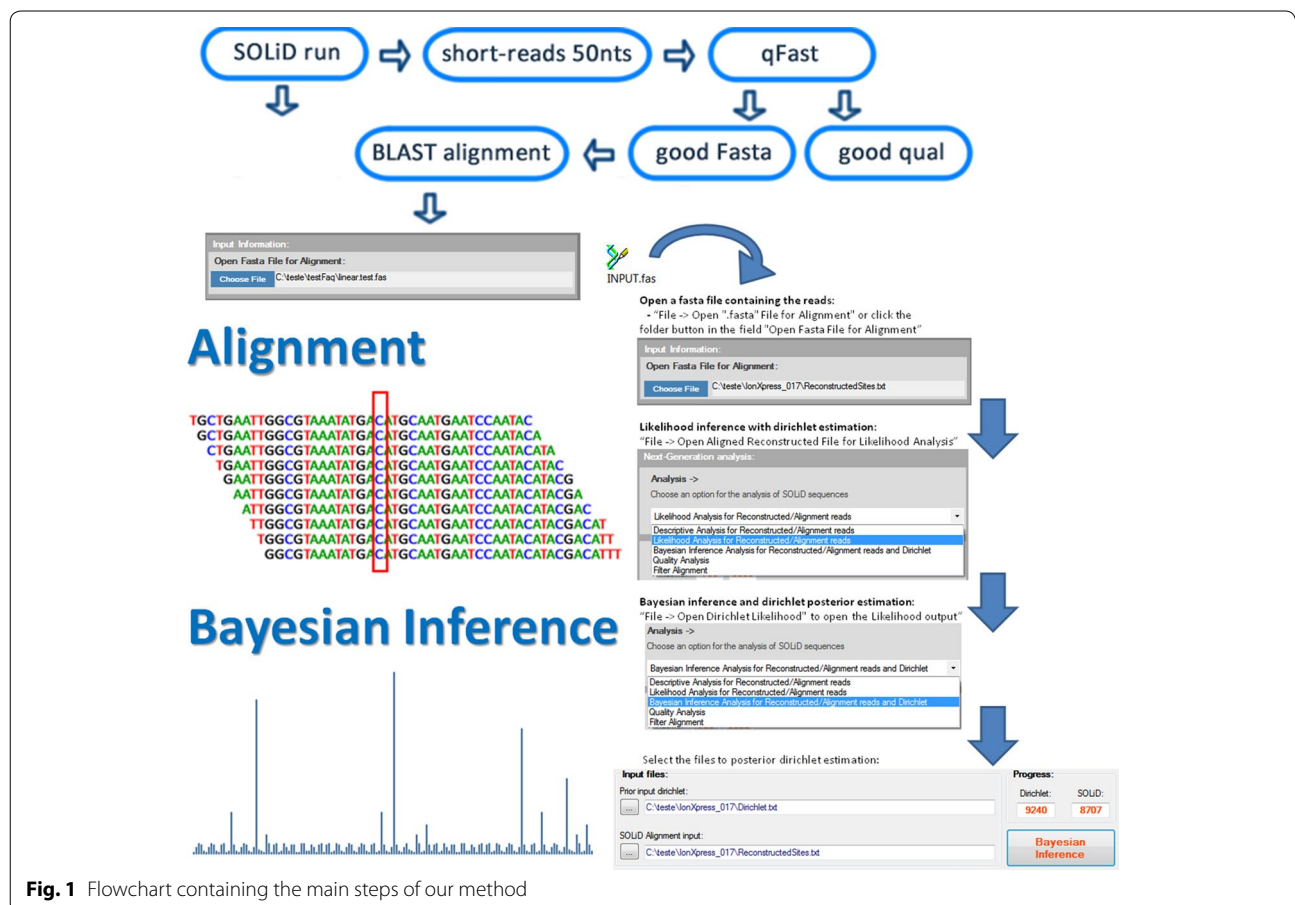
Unfortunately, we could not find any other method or software in the literature, which uses a similar form to represent the viral genetic diversity as a family of distributions indexed by the genome and does not need sufficiently long reads—all proposals in the literature that we were able to find are aimed at haplotype reconstruction and require longer reads (more than 100 bases). Any attempt to make comparison between such different approaches would be misleading, therefore it is not our aim here to make the point if the method presented here is an improvement over (non-)existing similar ones. In fact, we believe that the approach proposed here should not be considered alternative or rival, but complementary, to haplotype reconstruction.

### Implementation

Here we describe the main steps of our method. There are two stages, the first is the read mapping/alignment and the second is the nucleic bases inference (see flowchart in Fig. 1). The method presented here works, in principle, with data generated in any NGS machine, as long it is (converted to and) stored in the *FASTA file format*. Even several distinct outputs from different platforms (with distinct read lengths) may be combined into one file and used as input.

### Experimental procedure and preparation

The computational tool developed here assumes that there is one, or more, different viral propagation situations (such as, varying the cellular activation status, the co-receptor usage and the target cells, etc.), all of which have the same viral population as infecting source (see Table 1). A sample from the viral population prior to any infection experiment must be sequenced and the raw sequence data obtained will be referred to as the *control experiment*. The importance of the control experiment is that it will be used to infer the noise of the system, since it is essentially a clonal population. The raw sequence



**Fig. 1** Flowchart containing the main steps of our method



**Table 1 Summary of sequence data analyzed**

Sequenced data	Reads	Nucl.	Mapped (%)	Time (hour)
HIV1—non-stimulated CD4/R5	$1.11 \times 10^7$	$5.53 \times 10^8$	91.81	34.88
HIV2—stimulated CD4/R5	$1.04 \times 10^7$	$5.19 \times 10^8$	90.84	34.18
HIV3—non-stimulated CD4/X4	$1.10 \times 10^7$	$5.50 \times 10^8$	91.22	29.05
HIV4—stimulated PBMC/X4	$9.41 \times 10^6$	$4.71 \times 10^8$	91.14	13.63
HIV5—stimulated PBMC/R5	$1.14 \times 10^7$	$5.73 \times 10^8$	90.49	24.09
HIV6—non-stimulated PBMC/X4	$1.12 \times 10^7$	$5.62 \times 10^8$	91.42	27.54
HIV7—non-stimulated PBMC/R5	$1.19 \times 10^7$	$5.94 \times 10^8$	93.27	15.34
HIV8—control (pNL4-3kfs)	$1.01 \times 10^7$	$5.05 \times 10^8$	93.10	30.14
Total	$8.65 \times 10^7$	$4.33 \times 10^9$	91.67	208.85

The first column lists all the experimental conditions and the control experiment that where sequenced, the second column (Reads) displays the number of reads sequenced in each condition, the third column (Nucl.) displays the number of nucleotides in each condition, the fourth column (Mapped) displays the percentage of reads that have been mapped and the fifth column (Time) displays the time elapsed in each mapping procedure

data obtained from samples extracted from infected cells, after a fixed number of replication cycles, will be collectively called *experimental conditions*.

Raw sequence data from the sequencing must be treated according to the standard procedures of the specific NGS platform [46] up to the generation of FASTA files, which are the standard type of input file adopted in our implementation.

The data used in this paper for testing the method is the subject of another publication [47], where the details of the experiments and the biological implications to the HIV replication are discussed. The method and the results described here do not depend on the experimental details and the results reported in [47].

### Read mapping/alignment

The main goal of this step is the mapping of reads with 50 nucleotides or more originating from the NGS platform to a database of reference sequences. The database may contain several sequences, which must be aligned amongst themselves. The read mapping is performed using a local executable of BLAST [48]. The criteria for retaining the reads are the following: (1) it must align at least 45 nucleotides and (2) have the lowest  $e$  value score. A first alignment attempt is made with sequences from reads in the forward sense; in case of no match, a second attempt with the reverse complementary sequence is performed. Moreover, since we are using several references, the output can, in principle, display the same number of matches as there are reference sequences. The criteria for the selection of

the most suitable alignment option are the following (in this order): (1) the lowest  $e$  value score and (2) the lowest *Hamming distance* from the consensus sequence obtained from the sequencing of the control experiment. The alignment strategy described above is set as default, but the user, according to some specific purposes or simply for increasing processing speed, can change some of its parameters. Finally, it is possible to create suitable reference databases for specific research purposes.

### Nucleic base estimation

The probability distributions of nucleic bases (A, T, C, G) at each position of the genome are estimated from the aligned data. In this respect, our approach may be classified as a *diversity estimation in single positions*. The idea is that at each position in the genome the probability distribution is given by a *multinomial distribution*, determined by four probabilities ( $p_A, p_T, p_C, p_G$ ) satisfying  $p_A + p_T + p_C + p_G = 1$ . These conditional probabilities represent the fraction of the population that has each of the four associated nucleic bases at the corresponding site given the observed sequence data. Thus, one has a family of multinomial probability distributions indexed by the sites of the genome, where at each position the four probabilities ( $p_A, p_T, p_C, p_G$ ) should be estimated from the data.

The Bayesian framework for the inference of categorical data is based on the notion of *conjugate prior*, which in the case of categorical data is given by the Dirichlet distributions [49, 50]. A *Dirichlet distribution* is characterized by a  $n$ -tuple of positive numbers  $\alpha = (\alpha_1, \dots, \alpha_n)$  called *hyper-parameters*—however, unlike the multinomial parameters that must sum to one, the hyper-parameters are unconstrained. In our case, the Dirichlet distribution of each site is parametrized by the quadruple  $(\alpha_A, \alpha_T, \alpha_C, \alpha_G)$ . The fact that the Dirichlet distribution is the conjugate prior of the multinomial distribution amounts to saying that the *Bayes' formula* for the posterior distribution takes a very simple form in terms of the hyper-parameters: if  $(\alpha_1, \dots, \alpha_n)$  is a vector of hyper-parameters of a Dirichlet prior distribution and the counts of each of the  $k$  categories in an experiment are  $(c_1, \dots, c_n)$  then the posterior distribution is also a Dirichlet distribution with hyper-parameters  $(\alpha_1 + c_1, \dots, \alpha_n + c_n)$ . Within this context, the first step in the estimation of the probability distributions of nucleic bases consists in using the sequenced data from the *control experiment* as the input for the determination of prior hyper-parameters. Then, in the second step, one considers this distribution together with the sequenced data form the experimental conditions one uses *Bayes' formula* to compute the *posterior hyper-parameters*.

A  $n$ -dimensional *Dirichlet distribution* is defined in by a smooth probability density function on the set  $\Delta$  of  $n$ -dimensional multinomial distributions, which is parametrized as  $\Delta_n = \{(p_1, \dots, p_{n-1}) : p_1 + \dots + p_{n-1} \leq 1\}$ , here  $n$  is the number of distinct categories (states) that can be observed and  $p_k$  is the probability of observing the  $k$ -th category, for  $k = 1, \dots, n$  with  $p_n = 1 - p_1 - \dots - p_{n-1}$ . The *Dirichlet probability density function* is given by

$$\text{Dir}(p_1, \dots, p_{n-1} | \alpha_1, \dots, \alpha_n) = \frac{1}{B(\alpha)} \prod_k p_k^{\alpha_k - 1}$$

where  $B(\alpha)$  is a normalizing factor defined in terms of the *gamma function*  $\Gamma$  as

$$B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma\left(\sum_k \alpha_k\right)}$$

for a vector  $\alpha = (\alpha_1, \dots, \alpha_n)$ . Note that the choice  $(\alpha_1, \dots, \alpha_n) = (1, \dots, 1)$  gives the uniform distribution (the *flat or uninformative prior*) on  $\Delta_n$  with mass equal to the volume of  $\Delta_n : B(1, \dots, 1) = 1/\Gamma(n) = 1/(n-1)!$ .

The Dirichlet hyper-parameters associated to the sequenced data from the control experiment represent the “noise” of the system and can be obtained by maximum likelihood estimation (MLE) through the Newton–Raphson method [51]. The *log-likelihood function*  $g$  of the Dirichlet distribution is given by  $g = N \log L$  and

$$\begin{aligned} \log L(\alpha_1, \dots, \alpha_n | p_1, \dots, p_n) \\ = \log \Gamma\left(\sum_k \alpha_k\right) - \sum_k \log \Gamma(\alpha_k) + \sum_k (\alpha_k - 1) \log p_k, \end{aligned}$$

where  $N$  is the *sample size* and  $\log p_k = 1/N \sum_j \log p_{jk}$  ( $j = 1, \dots, N, k = 1, \dots, n$ ) is called the *sufficient statistics* associated to a sample of  $n$ -categorical vector observations  $\{p_1, \dots, p_N\}$  of sample size  $N$ . Thus each vector  $p_j = (p_{j1}, \dots, p_{jn})$  has  $n$  components, each component  $p_{jk}$  is the frequency of the  $k$ th category at the  $j$ th sample.

The *Newton–Raphson method* for the maximum likelihood estimation of Dirichlet hyper-parameters amounts to the iteration of the following fixed-point scheme [49], which converges to the unique maximum value of  $g$ :

$$\alpha^{\text{new}} = \alpha^{\text{old}} + [H^{-1} \nabla g] (\alpha^{\text{old}}),$$

where  $\alpha^{\text{new}}$  and  $\alpha^{\text{old}}$  are vectors of Dirichlet hyper-parameters. The function  $\nabla g(\alpha)$  is the *gradient vector* (derivative) of the log-likelihood function  $g$ , with components

$$[\nabla g(\alpha)]_k = \Psi\left(\sum_k \alpha_k\right) - \Psi(\alpha_k) + \log p_k,$$

where  $\Psi = (\log \Gamma)'$  is the *digamma function*. The function  $H^{-1}(\alpha)$  is the inverse of the *Hessian matrix* of the log-likelihood function  $g$  and the product  $(H^{-1} \nabla g)(\alpha) = H^{-1}(\alpha) \nabla g(\alpha)$  has components  $[(H^{-1} \nabla g)(\alpha)]_k$  given by

$$\begin{aligned} & \left[ (H^{-1} \nabla g)(\alpha) \right]_k \\ & = (\Psi'(\alpha_k))^{-1} ((\nabla g)_k \\ & + \left( \sum_l \frac{(\nabla g)_l}{\Psi'(\alpha_k)} \right) / \left( \frac{1}{\Psi'(\sum_l \alpha_l)} - \sum_l \frac{1}{\Psi'(\alpha_l)} \right)), \end{aligned}$$

where  $\Psi'$  is the *trigamma function* ( $k, l = 1, \dots, n$ ). Several suggestions for the initialization step (that is, the initial value of  $\alpha^{\text{old}}$ ) of the iteration scheme described above have appeared in the literature [50, 52, 53]. The proposal of Ronning [53] is the most suitable for the modified iteration scheme adopted here.

Since we are dealing with a sparse estimation problem in the sense that one of the categories occur with much higher frequency than the other categories, we shall employ the *smoothed sufficient statistics* defined by introducing a small parameter  $\eta$  and setting  $p_{jk} = M_{jk}/M$ , where  $M_{jk}$  is the number of occurrences of the  $k$ th category at the  $j$ th sample and  $M$  is total number of observations at the  $j$ th sample and  $p_{jk} = \eta$  if there is no observation of the  $k$ th category at the  $j$ th sample. The *smoothing parameter*  $\eta$  acts as “background noise” representing sequencing and PCR errors that can not be removed. However it can be suitably tuned in order to account for the true variability of the data. When this procedure is applied to the sequenced data from the control experiment (a “clonal” population) one would expect no diversity at all. However, that is not completely true and, in fact, even the sequenced data from the control experiment should display some variability (mainly due to sequencing errors). The smoothing parameter  $\eta$  should be same (or smaller than) of the order of magnitude of expected error rate  $\epsilon$ . In the smoothed version of the Newton–Raphson iteration scheme, Ronning’s initialization step is given by setting  $\alpha^{\text{old}} = (\eta, \dots, \eta)$ .

The sufficient statistics is computed by a simple re-sampling procedure [54, 55] in order to generate sequences of categorical observations from the raw sequenced data, by randomly sampling nucleotides from each aligned position. Here, the imperfect clonality of the sequenced data from the control experiment is useful, since it ensures that the re-sampled ensemble has some variability, which is consistent with having a small non-zero smoothing parameter. The re-sampling procedure has one parameter that can be adjusted by the user: the relative size of observations given as a fraction  $0 < z < 1$  of the size  $C$  of the set of nucleotides covering the given site. If the number of

bases covering the given site is denoted by  $C$  then  $M = zC$  is the number of observations used to compute one sample vector  $\mathbf{p}_j = (p_{j1}, \dots, p_{jn})$  and the corresponding sample size  $N$  is given by (the integer part of) the logarithm of the total number of all possible sample vectors:

$$N = \lceil \log \Gamma(C) - \log \Gamma(zC) - \log \Gamma((1-z)C) \rceil.$$

*Stirling's formula* gives the following approximation in terms of  $C$ :

$$N \approx C(-z \log z - (1-z) \log(1-z)) / \log 2.$$

For instance, for the default value of  $z$ , which is 80 %, one has a sample of size  $N \approx 0.7C$ , each sample vector computed from 0.8  $M$  nucleotides. On the other hand, the value  $z = 50$  % gives a sample of size  $N \approx C$ , each sample vector computed from 0.5  $M$  nucleotides.

Once the hyper-parameters of the prior distribution are estimated, they must be used together with the sequenced data of the other experimental conditions in order to compute the hyper-parameters of the posterior distributions by *Bayes' formula*, as a result, one obtains a family of Dirichlet probability distributions indexed by the genome of the organism for every sequenced experimental condition, including the control experiment.

In order to obtain point estimates of categorical probabilities per site for each experimental condition ( $p_A, p_T, p_C, p_G$ ), one may use a central tendency measure of the corresponding Dirichlet distribution (see [49]). Let  $\mathbf{x} = (x_1, \dots, x_n)$  be a random vector distributed according to a Dirichlet distribution with corresponding hyper-parameters  $(\alpha_1, \dots, \alpha_n)$  then the number  $s = \alpha_1 + \dots + \alpha_n$  is called the *concentration parameter* of the corresponding Dirichlet distribution. It provides a measure of the "quality" of the inference: the greater the value of  $s$  the better is the "precision" of the inference (see [49]). The *mean value* of  $\mathbf{x}$  is

$$\langle x_k \rangle = \alpha_k / s.$$

The *maximum a posteriori* (MAP) estimate, which is given by the *mode* of  $\mathbf{x}$ , has become a very popular method of point estimation [10]. Moreover, the coordinates  $\langle \langle x_k \rangle \rangle$  of the mode of  $\mathbf{x}$  may be directly calculated in terms of the hyper-parameters  $\alpha_k$  only when  $\alpha_k > 1$  ( $k = 1, \dots, n$ ):

$$\langle \langle x_k \rangle \rangle = \alpha_k / (s - n).$$

This is much simpler than the contrived expectation-maximization (EM) approximate schemes usually employed to obtain the MAP estimate from a log-likelihood function, in which case approximations are unavoidable, since this function is non-convex.

Note that both the mean  $\langle x_k \rangle$  and the mode  $\langle \langle x_k \rangle \rangle$  converge to the same value when the number of

observations goes to infinity. In particular, when the number of observed nucleic bases at certain site is very large then the relative frequencies of the different nucleic bases are very close to the values of the Dirichlet mean and mode.

Confidence values associated to the point estimates may be defined in terms of a dispersion measure of the corresponding Dirichlet distribution. The *variance* of  $\mathbf{x}$  is given by

$$\text{Var}(x_k) = \alpha_k s - \alpha_k^2 / (s^2 (s + 1)).$$

Since the marginal distribution of each  $x_k$  is a one-dimensional Dirichlet distribution, also known as *Beta distribution*, the *standard deviation of the mean*  $\sigma(\mathbf{x}) = \text{square-root}(\text{Var}(\mathbf{x}))$  may be used to construct *Bayesian credible intervals* about the expectation value.

The *standard deviation of the mean*  $\sigma(x_k)$  may be used to define credible intervals about the mode as well. Since the Beta distribution is unimodal, when all  $\alpha_k > 1$  ( $k = 1, \dots, n$ ), and has finite variance, a 3-sigma interval around the mean or the mode would provide about 95 % of confidence in the prediction (this is a general consequence of the Gauss-Vysochanskij-Petunin inequality, see [56]).

Finally, we should note that the inference procedure explained above is clearly not restricted to the case of four nucleotides (A, T, C, G), that is  $n = 4$ . It is trivial to modify it in order to account for insertions and deletions, or to work with codons and amino-acids.

### Selection criteria and error filtering

Once the inference has been completed it is desirable to filter the errors and extract some subset of the data—for instance, most conserved sites, most variable sites, etc. In order to do so we have implemented two selection criteria based on simple quantities: (1) complementary probability per site and (2) variational distance per site.

The *complementary probability per site* is defined as  $p_{\text{comp}} = 1 - \max\{p_A, p_T, p_C, p_G\}$  and it depends only on the probability distribution of each site. It provides a measure of how much the distribution is concentrated in one state. For instance, if the complementary probability at a site is high it means that there was variation in the site prior to the experiment.

The *variational distance per site* is a positive number between 0 and 2 defined by  $vd = |p_A - p'_A| + |p_{AT} - p'_{AT}| + |p_C - p'_C| + |p_G - p'_G|$ , where  $(p_A, p_T, p_C, p_G)$  is the probability distribution per site of the control experiment and  $(p'_A, p'_T, p'_C, p'_G)$  is the probability distribution at the corresponding site of the experimental condition. It is a measure of the relative variation per site from the clonal population before and after the infection. If it is very low

at a site it means that the site did not undergo significant changes in relation to the sequenced data from the control experiment.

The complementary probabilities and the variational distance can work as filters and the user must specify the thresholds for them. By using these two criteria in combination one may easily obtain some qualitative information about the behavior at a site.

An example of how to combine our method with any haplotype reconstruction procedure is the following. In a haplotype reconstructed population the fraction of a nucleic base  $X$  at a given position could be computed by summing the proportions of all haplotypes that have the nucleic base  $X$  at the given position. Using these proportions one can construct a family of distributions  $(f_A, f_T, f_C, f_G)$ , with  $f_A + f_T + f_C + f_G = 1$ , per position. Since in practice it is very difficult to obtain the low-frequent variants the variational distance between the distribution  $(f_A, f_T, f_C, f_G)$  and the distribution  $(p_A, p_T, p_C, p_G)$  can be used to estimate how far one is from having obtained all the variants up to the lowest-frequent ones.

Another application, is to use the distributions  $(p_A, p_T, p_C, p_G)$  to generate a population of “random haplotypes” with the correct nucleic base distribution and compare with a population of reconstructed haplotypes in order to study the correlations between the sites.

## Results and discussion

The method presented here was tested on samples obtained after the HIV-1 strain NL4-3 (group M, subtype B) cultivation on primary human cell cultures. Different viral propagation conditions were used—varying the cellular activation status, the co-receptor usage and the target cells. The pseudo-typed viruses produced in these experiments were able to perform exactly one round of the replicative cycle. As a whole, there were seven experimental conditions in addition to the *control experiment* (Table 1).

### Experimental procedure and preparation

The experimental procedure was performed in accordance with the standard procedures of the NGS platform SOLiD™ [46], up to the generation of FASTA files, which are the input data of our computational tool. Standard Life Technologies guidelines were used during sample preparation and sequencing while using the SOLiD™ platform v. 3.0. The size of the FASTA files containing the reads of each condition is around 700 Mb, consisting of about  $10^7$  reads.

### Read mapping/alignment

The use of BLAST to perform the read mapping has two reasons: first the multiple reference sequences allowed by BLAST makes it advantageous for analysis of viral

populations classically described as quasispecies, since we can include several variant genomes belonging to the same phylogenetic branch, and second, it is easier to control the parameters of the alignment and multi-threading inside our program.

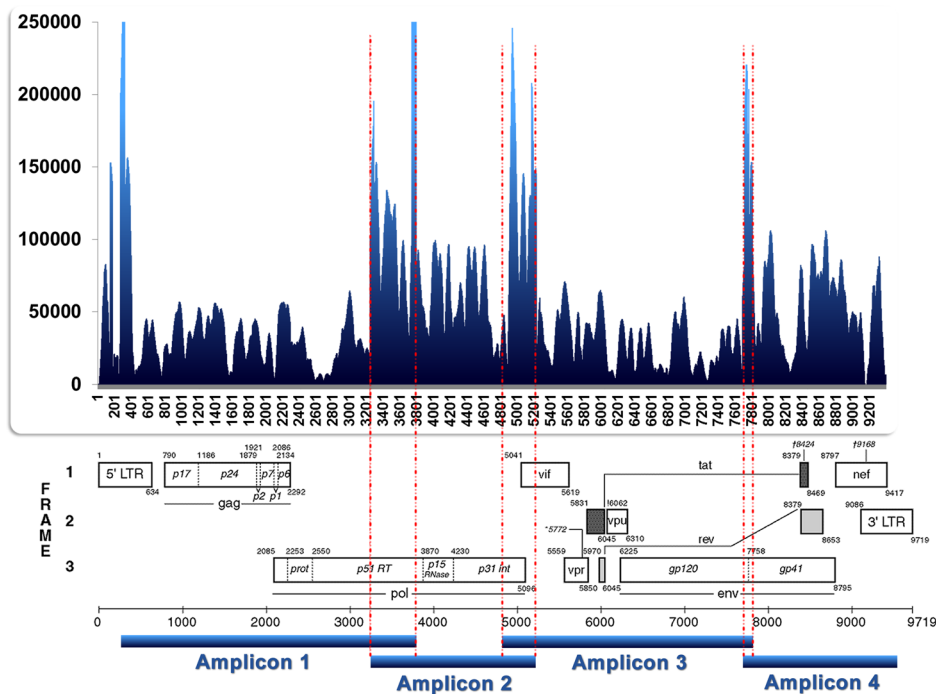
We used a reference database composed by 1258 sequences, properly aligned, all representatives of HIV-1 epidemics, most of them belonging to major group (group M) and its subtypes A, B, C, D, F, G, H, J, K, and some of their recombinants (<http://www.hiv.lanl.gov/content/sequence/HelpDocs/subtypes-more.html>). Some of the sequences are of complete HIV-1 genomes while others represented virtually (more than 90 % complete) complete genomes. All sequences are available at NCBI/Genbank. The reference database also contains sequences from HIV-1 group O, SIVcpz and HIV-1 strain NL4-3 and is packaged together with the software.

Even though it is known that BLAST is not the fastest aligner when compared with next-generation aligners, we were able to achieve a reasonable speed and control by setting the parameters and implementing some optimizations. For instance, even though BLAST is capable of identifying alignments in both the forward and the reverse complementary senses, we have found that manually doing this significantly increases the retrieval of reads—we had 15 % increase in the retrieval of reads. BLAST can be sensitive if the right parameters are chosen (small word size, in particular). It can find an alignment of a 42-mer with a multiple mismatches and gaps. On the other hand, some next generation aligners may fail to find an alignment if a mismatch or gap (or more than one of these) occurs within the beginning of the read, as this portion is used as a seed. An important feature of BLAST is that all alignments are returned. If a read has 1000 alignments, 1000 alignments are reported. Another advantage is the ability to perform sub-string alignments. Next generation aligners tend to be focused on aligning the entire read length. BLAST will find an alignment and report what position within the read that the alignment start and ends. Finally, BLAST is a more sensible treatment of N's. Some of the next-generation aligners store bases in 2-bit format. Meaning they can only internally represent A, T, C, G. The solution is to randomly assign N's to one of the other bases, a solution that some may find imperfect.

For each experimental condition, comprising the alignment of around  $10^7$  reads of 50 bases against  $10^3$  HIV-reference sequences with  $10^4$  sites each, we were able to map around 90 % of the reads, since the estimated fraction of reads with at least one error is around 2 %, we have achieved an almost optimal retrieval of reads.

For instance, Fig. 2 shows the result of the alignment of the sequenced data from the control experiment and the





**Fig. 2** Depth and coverage of one SOLiD™ sequencing of the HIV-1 genome. The *major peaks* in the *middle* representing the most deeply covered regions coincide with the overlapping primers from the PCR step, an evidence that there is in fact some influence of pre-sequencing phases on the frequency of the short-reads retained in the alignment. The *major peak* in the *beginning* is related to the difficulties in mapping the LTR region. Other *significant peaks* maybe due to PCR artifacts, as well

corresponding site coverage. The average site coverage is around 50,000 reads with some peaks going beyond 150,000 reads. The running time on each experimental condition was around 30 h on a Intel i7 (12 cores, clock of 3.30 GHz) with 32 GB of RAM memory and 2 TB of disk space. It is worthwhile mentioning that the program uses at most three cores and requires 2.8 GB of RAM memory to handle files with 700 MB, thus it is conceivable that the program could run on any computer matching this minimal configuration.

### Inference

An important difficulty that should be overcome in order to implement the inference procedure for Dirichlet hyper-parameters in the context of nucleotides is due to the *sparsity*. Even with the high mutation rate displayed by viruses, there is a fair amount of nucleotide conservation. From a populational point of view, most of individuals will present the same nucleotide at a specific genomic position, and only the less representative subgroups, if any, will present one of the three remaining possibilities.

The standard Bayesian method outlined in most textbooks, where one usually chooses an uninformative (uniform) prior distribution is appropriate for the general task of multinomial estimation [57], but generally

provides poor results when used for *sparse* multinomial distributions. This is primarily a consequence of the erroneous assumption that all categories should be considered as equally possible values for each site. Indeed, sparse multinomial distributions are characterized by the fact that only a few symbols actually occur (site conservation). In such cases, applying the standard method will give too much weight to symbols that never occur and consequently give a poor estimate of the true distribution. This issue becomes critical in our case when treating data obtained from the control experiment, which, in principle, is a *clonal population*, where one expects a uniquely well-defined nucleotide at each site and thus the Dirichlet likelihood function would be identically zero.

The sparsity problem, namely, the fact that one of the categories occur with much higher frequency than the other categories, is usually solved in the literature of text modeling (see [40]) by introducing a *smoothing parameter*  $\eta$  and modifying the Newton–Raphson method in such a way that the sufficient statistics does not have any zero entry. In practice, this may result in an over-smoothed distribution, but one can choose a small enough value for  $\eta$  in such a way that all the rare events do not have the same probability of appearing in all states.

### Check points and validation

The method described here contains some heuristic decisions that should be supervised and properly validated. We have added a checkpoint at the read mapping stage and performed a validation of the smoothed Newton–Raphson method, attaining very good concordance with the expected results.

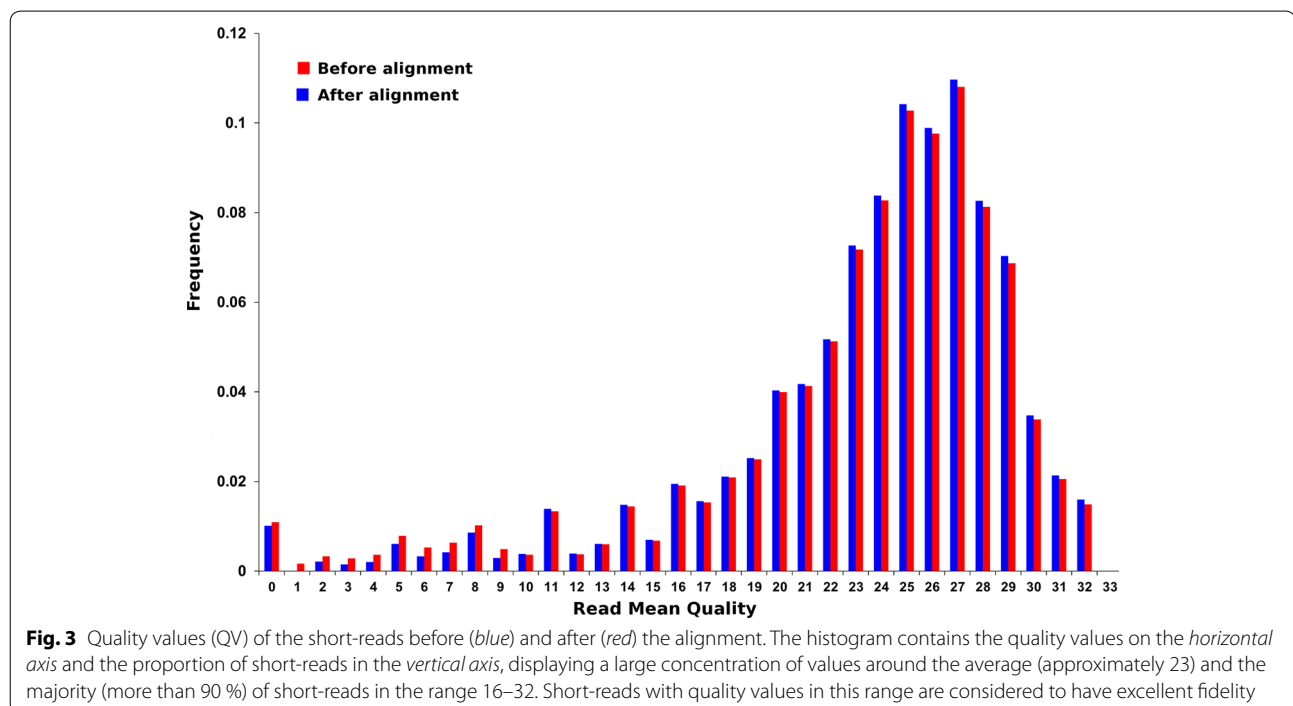
In order to assess the reliability of the read mapping procedure, the *quality values* (QV) of the reads have been used as a proxy. The SOLiD™ platform outputs two files after primary analysis [46, 58]: a sequence file in color-space and a quality file containing the corresponding quality values. The QV of a read is a positive integer ranging from 0 to 50 and is given by the logarithm of the inverse probability of the color call being inaccurate, i.e. the higher the QV the higher the confidence in the color call's accuracy. By computing the *distribution of quality values* of the reads from each experimental condition and the control experiment, prior and after the alignment, and comparing them, it is observed that they are almost identical (see Fig. 3). This shows that the alignment procedure does not introduce any bias towards higher or lower quality values. The reliability of the read mapping procedure is guaranteed by the stringency of the criteria for retaining the reads.

The checking of the read mapping procedure was done by computing the distributions of all quality values of each condition prior and after the alignment (see Fig. 3). The mean value of the QV distributions remained unchanged after alignment. Likewise, at both steps of the process

more than 80 % of reads had QV comprised between 20 and 32, assuring that the quality of the retrieved sequences was preserved and no bias was introduced.

The validation of the nucleotide inference step is performed at two points. The re-sampling procedure has been validated by comparing at each site, the nucleotide frequencies obtained from all the reads that cover the site, with the nucleotide frequencies obtained from the sampled reads that cover the site. It is observed that both frequencies agree with high precision (up to order  $10^{-4}$ ). This ensures that the sufficient statistics obtained with re-sampling is the correct one. The validation of the implementation of the Newton–Raphson scheme for the Dirichlet maximum likelihood is performed by computing the MLE for a standard data set that is not sparse. The data set for pollen counts analyzed in Mosiman is often used for testing Dirichlet maximum likelihood implementations (see [49]). Since our implementation has a smoothing parameter, it is expected that the obtained values converge to the known values when the parameter approaches zero. It is indeed observed that this convergence occurs, with perfect agreement occurring above the order of magnitude of the smoothing parameter.

We have included in the software the appropriate options for the user to perform validation procedures. In particular, it is possible to run the Dirichlet MLE on any data set (with 4 categories) given as a list of multinomial observations.



### Setting the smoothing parameter and separating the signal from the noise

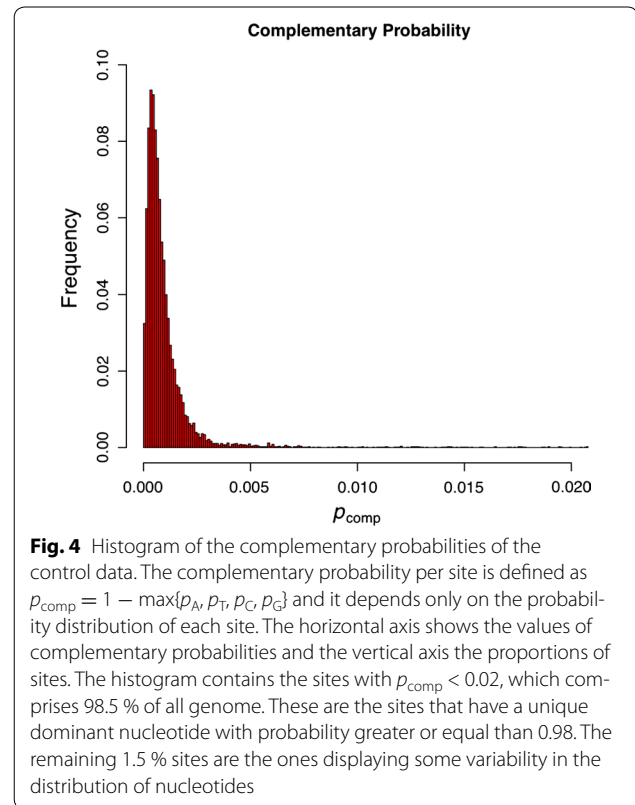
The complementary probability of an *ideal clonal population* is expected to be identically zero. However, in the smoothed inferential scheme proposed here it is expected that  $p_{\text{comp}} \approx \eta$ . The choice of smoothing parameter does not affect the running time of the Newton–Raphson method; it affects the values of the estimated hyperparameters. The effect is of the same order of magnitude of the smoothing parameter. This suggests the following guidelines for setting the smoothing parameter:  $\eta$  must be smaller than the error rate of the sequencing. Since the expected error rate  $\varepsilon$  in our case is around  $6 \times 10^{-4}$  a value of  $\eta = 10^{-5}$  is a reasonable choice for the smoothing parameter.

The concentration parameter  $s$  of the Dirichlet distributions for the sequenced data from the control experiment is a measure of the quality of the inference: when  $s > 1$  the inference may be considered meaningful. Sites with  $s \leq 1$  may be excluded from further analysis. Sites with low value of  $s$  may happen due to poor coverage and total conservation (all reads with the same nucleotide at that position).

Due to sequencing and PCR errors (and other events which may have occurred prior to cloning the initial genome),  $p_{\text{comp}}$  may, in fact, display a broad distribution over the genome (see Fig. 4). In any case, the complementary probability of the control experiment may be considered as an average error rate per site and its distribution over the genome may be used to set a cut off value for separating the signal from the noise (everything below this value should be considered noise). It is expected that  $\text{MODE}(p_{\text{comp}}) \approx \eta$ , that is, the majority of sites will behave as in a clonal population and this indeed is the case (see Table 2). Furthermore, it is expected to have a concentration of the distribution of  $p_{\text{comp}}$  near the error rate  $\varepsilon = 6 \times 10^{-4}$ . Since the distribution of  $p_{\text{comp}}$  is extremely skewed with a long tail, the median is a better measure of centrality than the mean value. In fact, we have found that  $\text{MEDIAN}(p_{\text{comp}}) \approx \varepsilon$  as expected (see Table 2).

The expectation  $\text{MEAN}(p_{\text{comp}})$ , which is very sensitive to the long tail, is a reasonable conservative choice of a cut off value for noise filtering, a more conservative choice would be  $\text{MEAN}(p_{\text{comp}}) + \text{SD} - \text{MEAN}(p_{\text{comp}})$ . While these choices provide uniform cut off along the genome it is possible to use the individual Dirichlet distributions at each site to construct a more refined cut off function. Finally, the cut off value for  $p_{\text{comp}}$  can be used to obtain a cut off value for the variational distance, since the cut off value of  $vd$  is twice the cut off value of  $p_{\text{comp}}$  ( $vd$  is a piece-wise linear function of the probabilities).

After the nucleotide probability distributions of the control experiment was computed we have found 40 genomic positions with concentration parameter  $s$  less or



**Fig. 4** Histogram of the complementary probabilities of the control data. The complementary probability per site is defined as  $p_{\text{comp}} = 1 - \max\{p_A, p_T, p_C, p_G\}$  and it depends only on the probability distribution of each site. The horizontal axis shows the values of complementary probabilities and the vertical axis the proportions of sites. The histogram contains the sites with  $p_{\text{comp}} < 0.02$ , which comprises 98.5 % of all genome. These are the sites that have a unique dominant nucleotide with probability greater or equal than 0.98. The remaining 1.5 % sites are the ones displaying some variability in the distribution of nucleotides

**Table 2** Summary statistics of the complementary probability ( $p_{\text{comp}}$ ) and the concentration parameter ( $s$ ) of the control experiment, after removal of the genomic positions with concentration parameter below 1

Statistics	$p_{\text{comp}}$	$s$
Mean	0.00260	3.43
Deviation	0.01824	0.50
Median	0.00066	3.60
Mode	0.00001	3.74
Minimum	0.00001	1.01
Maximum	0.49242	6.26

equal than one. These genomic positions correspond to portions of the genome where the coverage dropped substantially in comparison with the mean coverage ( $\sim 10^2$  reads). These sites were excluded from the remaining analysis. The expectation is  $\text{MEAN}(p_{\text{comp}}) = 0.002$  (that is, probabilities are considered significantly distinct if they differ by more than 0.02 %). The conservative choice of cut off value is given by  $\text{MEAN} + \text{SD} - \text{MEAN} = 0.002 + 0.018 = 0.02$  (see Table 2). The complementary probability may also be used to make sure that the variability observed in the experimental conditions is not a feature that has been transferred from the control experiment to

the experimental condition. The distribution of the complementary probabilities of the control experiment shows that 98.5 % of genomic positions have  $p_{comp} < 0.02$  (this means less than 2 % of nucleotide variation). The remaining 1.5 % genomic positions correspond to sites were the population acquired its variation prior to exposition to the experimental condition (see Fig. 5).

The posterior probability distributions of all 7 experimental conditions were computed. Figure 6 presents the values of the variational distance between the control experiment and one of the experimental conditions and its distribution is shown in Fig. 6, lower panel. Considering the same conservative cut off value of 0.04 for the variational distance (Fig. 6, upper panel), one has that 98 % of the genomic positions felt under this threshold, these are sites that did not display nucleotide variation after exposition to the experimental condition. The remaining 2 % genomic positions contain all the populational variation acquired after exactly one round of the replicative cycle.

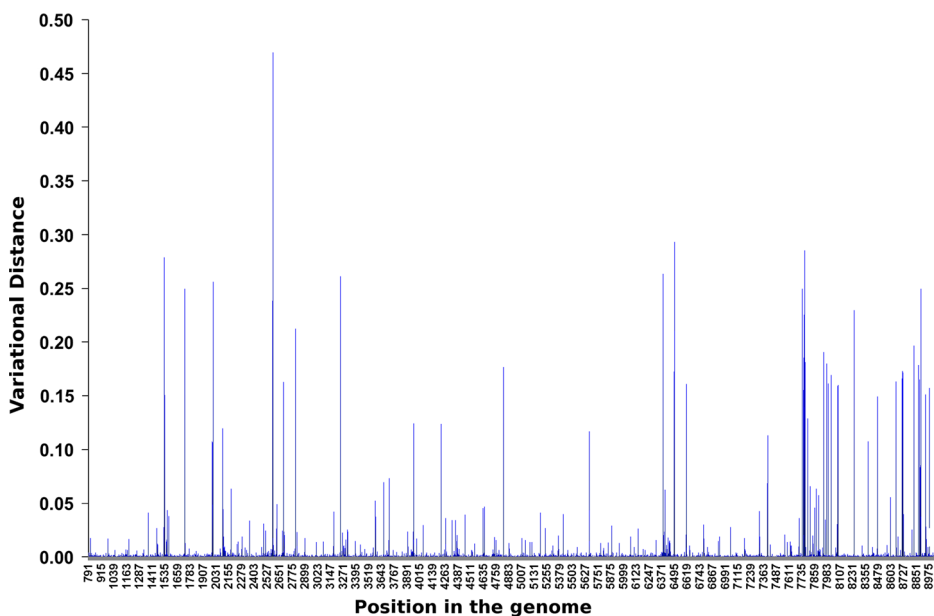
**Conclusions**

High throughput sequencing technologies are constantly evolving and new platforms and refinements in the chemistry and base calling algorithms are constantly improving. Recently the PacBio™ sequencer has been gaining space as it produces long reads, but with a large

number of randomly generated sequencing errors [59]. New approaches to sequencing using known technologies have been proposed, such as circle sequencing for Illumina [60]. We expect that the proposed approach, with slight modifications can be adopted for other technologies such as Ion Torrent™, Illumina™ (HiSeq, MiSeq and NextSeq) and PacBio™.

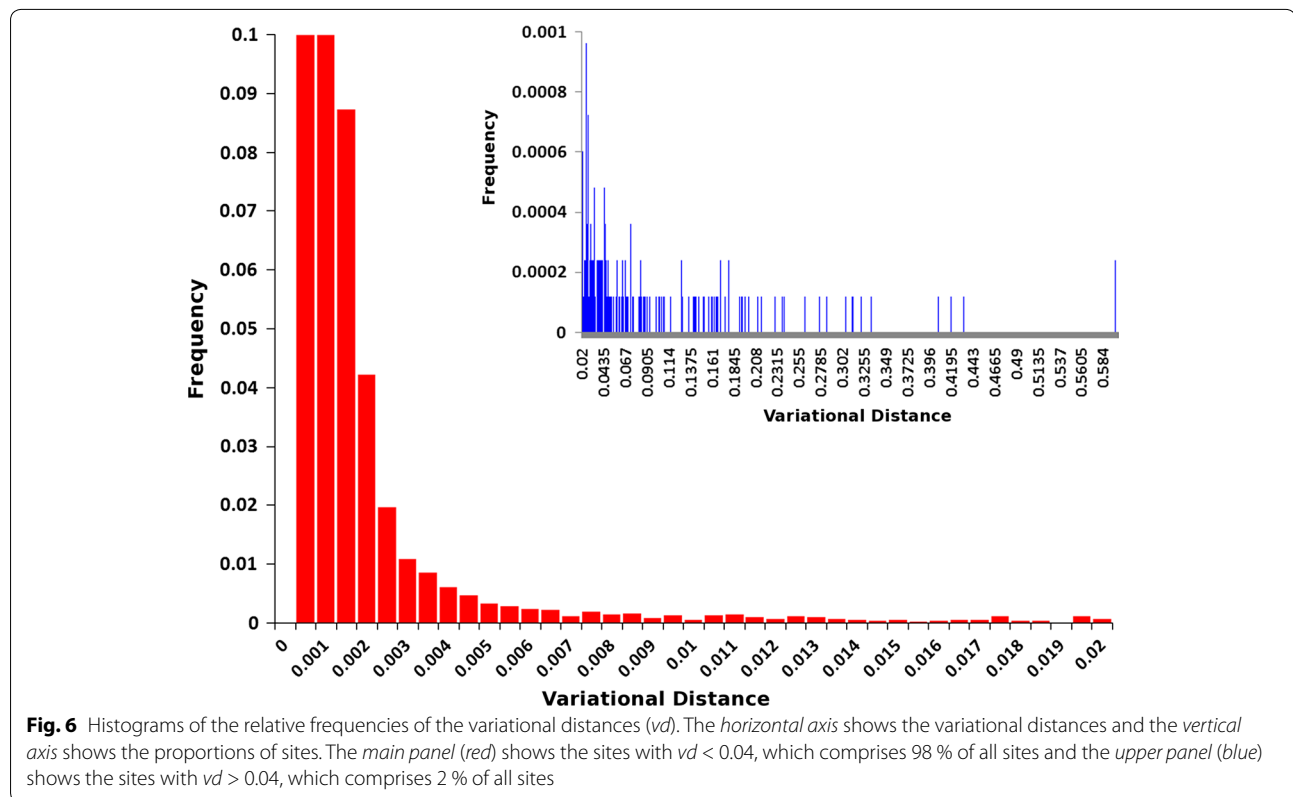
We have described a platform suitable to address the problem of estimation of populational diversity of RNA viruses. Based on the fact that the SOLiD™ sequencing platforms generate an extremely high number of reads allowing for a deep and extensive coverage of the data with very low error rate, we propose to measure the populational genetic diversity through a family of probability distributions indexed by the sites of the genome, each one representing the populational distribution of the diversity. This approach allowed us to avoid some very hard problems related to haplotype reconstruction (need of long reads, preliminary error filtering and assembly) and emphasize the main features of the sequencing technology used in this work, the SOLiD™ platform.

We have tested the method proposed here on samples obtained after the HIV-1 strain NL4-3 (group M, subtype B) cultivation on primary human cell cultures in many distinct viral propagation conditions, thus successfully demonstrating the capability of the method in handling large data-sets and delivering very clean



**Fig. 5** Variational distance ( $vd$ ) between the control data and an experimental condition along the genome. The variational distance per site is defined by  $vd = |p_A - p'_A| + |p_T - p'_T| + |p_C - p'_C| + |p_G - p'_G|$ , where  $(p_A, p_T, p_C, p_G)$  is the probability distribution per site in the control data and  $(p'_A, p'_T, p'_C, p'_G)$  is the probability distribution of the corresponding site in the experimental condition. The horizontal axis shows the sites of the genome (with the LTR regions removed) and the vertical axis shows the corresponding variational distances. Applying the conservative cut-off value of 0.04 for  $vd$  one obtains the sites with significant variation





results, suggesting that the software is a valuable tool for investigating the genetic diversity in viral populations. We have successfully demonstrated *Tanden's* capability of handling large data-sets and delivering very clean results, suggesting that the software is a valuable tool for investigating the genetic diversity in viral populations as a complementary to some haplotype reconstruction method.

### Availability and requirements

Project name: Tanden

Project web site: <http://tanden.url.ph/>

Operating systems: Windows

Programming language: Microsoft-C#

License: free to all users under the LGPL license

Minimum requirements: 4 GB RAM (16 GB recommended), 500 GB disk space

Third party software used: BLAST + standalone for windows.

(<http://www.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/>)

### Authors' contributions

JPZ: contributed to the statistical analysis, developed and implemented the software, drafted the manuscript; SNB: carried out the biological analysis, drafted the manuscript; ACV: contributed to NGS and bioinformatics analysis, drafted the manuscript; GCO: contributed to NGS bioinformatics analysis,

drafted the manuscript; LMRJ: conceived the statistical analysis, contributed to NGS and biological analysis, drafted the manuscript; FA: conceived the statistical analysis, developed and implemented the software, contributed to NGS and biological analysis, drafted the manuscript. All authors read and approved the final manuscript.

### Author details

<sup>1</sup> Departamento de Medicina, Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil. <sup>2</sup> Departamento de Microbiologia, Imunologia e Parasitologia, Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil. <sup>3</sup> Departamento de Informática em Saúde, Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil. <sup>4</sup> Laboratório de de Biocomplexidade e Genômica Evolutiva, Escola Paulista de Medicina (EPM), Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil. <sup>5</sup> Departamento de Microbiologia e Imunologia Veterinária, Universidade Federal Rural do Rio de Janeiro (UFRRJ), Rio de Janeiro, Brazil. <sup>6</sup> Genomics and Computational Biology Group, Centro de Pesquisas René Rachou (CPqRR), Fundação Oswaldo Cruz (FIOCRUZ), Belo Horizonte, Brazil.

### Acknowledgements

The ideas presented in this paper were developed in collaboration with Francisco A. R. Bosco, who suddenly passed away in December of 2012.

### Competing interests

The authors declare that they have no competing interests.

### Funding

Support was provided to: JPZ and SNB by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [<http://www.capes.gov.br/>]; ACV by Fogarty International Center National Institutes of Health [<http://www.fic.nih.gov/>] (TW007012); GCO by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) [<http://www.fapemig.br/>] (309,312/2012-4, 306362/2012-0), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)

[<http://www.cnpq.br>], Fogarty International Center National Institutes of Health [<http://www.fic.nih.gov>] (TW007012); LMRJ by Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [<http://www.fapesp.br>] (2009/14543-0); FA Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [<http://www.cnpq.br>] (PQ-306362/2012-0). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Received: 25 March 2015 Accepted: 25 February 2016

Published online: 11 March 2016

## References

- Duffy S, Shackleton LA, Holmes EC. Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet*. 2008;9:267–76.
- Mansky LM, Temin HM. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol*. 1995;69:5087–94.
- Fu Q, Mitnik A, Johnson PLF, Bos K, Lari M, Bollongino R, Sun C, Giemisch L, Schmitz R, Burger J, Ronchitelli AM, Martini F, Cremonesi RG, Svoboda J, Bauer P, Caramelli D, Castellano S, Reich D, Pääbo S, Krause J. A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr Biol*. 2013;23:553–9.
- Okoro CK, Kingsley RA, Connor TR, Harris SR, Parry CM, Al-Mashhadani N, Kariuki S, Musefula CL, Gordon MA, de Pinna E, Wain J, Heyderman RS, Obaro S, Alonso PL, Mandomando I, MacLennan CA, Tapia MD, Levine MM, Tennant SM, Parkhill J, Dougan G. Intracontinental spread of human invasive *Salmonella* Typhimurium pathovariants in sub-Saharan Africa. *Nat Genet*. 2012;44:1215–21.
- Nederbragt AJ. On the middle ground between open source and commercial software—the case of the Newbler program. *Genome Biol*. 2014;15:113.
- Beerenwinkel N, Zagordi O. Ultra-deep sequencing for the analysis of viral populations. *Curr Opin Virol*. 2011;1:413–8.
- Schopman NCT, Willemssen M, Liu YP, Bradley T, van Kampen A, Baas F, Berkhout B, Haasnoot J. Deep sequencing of virus-infected cells reveals HIV-encoded small RNAs. *Nucleic Acids Res*. 2012;40:414–27.
- Beerenwinkel N, Günthard HF, Roth V, Metzner KJ. Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol*. 2012;3:16.
- Mardis E. Next-generation sequencing technologies. In: Current topics in genome analysis. National Human Genome Research Institute. 2014. p. 1–26. <http://www.genome.gov/12514288>.
- Mangul S, Wu NC, Mancuso N, Zelikovsky A, Sun R, Eskin E. Accurate viral population assembly from ultra-deep sequencing data. *Bioinforma Oxf Engl*. 2014;30:i329–37.
- Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA*. 2011;108:9530–5.
- Huang A, Kantor R, DeLong A, Schreier L, Istrail S. QColors: an algorithm for conservative viral quasispecies reconstruction from short and non-contiguous next generation sequencing reads. *In Silico Biol*. 2011;11:193–201.
- Töpfer A, Marschall T, Bull RA, Luciani F, Schönhuth A, Beerenwinkel N. Viral Quasispecies Assembly via Maximal Clique Enumeration. *PLoS Comput Biol*. 2014;10:e1003515.
- Giallonardo FD, Töpfer A, Rey M, Prabhakaran S, Dupont Y, Leemann C, Schmutz S, Campbell NK, Joos B, Lecca MR, Patrignani A, Däumer M, Beisel C, Rusert P, Trkola A, Günthard HF, Roth V, Beerenwinkel N, Metzner KJ. Full-length haplotype reconstruction to infer the structure of heterogeneous virus populations. *Nucleic Acids Res*. 2014;42:e115.
- Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*. 2004;5:150–63.
- Wallace IM, O'Sullivan O, Higgins DG. Evaluation of iterative alignment algorithms for multiple alignment. *Bioinformatics*. 2005;21:1408–14.
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26:1135–45.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. Subgroup 1000 genome project data processing: the sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
- Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18:1851–8.
- Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25:1754–60.
- Homer N, Merriman B, Nelson SF. BFAST: an alignment tool for large scale genome resequencing. *PLoS One*. 2009;11:e7767.
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357–9.
- Wan-Ping Lee MPS. MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS One*. 2014;9:e90581.
- Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ. Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet*. 2011;56:406–14.
- Flicek P, Birney E. Sense from sequence reads: methods for alignment and assembly. *Nat Methods*. 2009;6:6–12.
- Goya R, Sun MG, Morin RD, Leung G, Ha G, Wiegand KC, Senz J, Crisan A, Marra MA, Hirst M, Huntsman D, Murphy KP, Aparicio S, Shah SP. SNVMix: predicting single nucleotide variants from next-generation sequencing of tumors. *Bioinformatics*. 2010;26:730–6.
- Prosperi MCF, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC, Capobianchi MR, Ulivi G. Combinatorial analysis and algorithms for quasispecies reconstruction using next-generation sequencing. *BMC Bioinform*. 2011;12:5.
- Willerth SM, Pedro HAM, Pachter L, Humeau LM, Arkin AP, Schaffer DV. Development of a low bias method for characterizing viral populations using next generation sequencing technology. *PLoS One*. 2010;5:e13564.
- Zagordi O, Däumer M, Beisel C, Beerenwinkel N. Read length versus depth of coverage for viral quasispecies reconstruction. *PLoS One*. 2012;7:e47046.
- Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res*. 2007;17:1195–201.
- Fischer W, Ganusov VV, Giorgi EE, Hraber PT, Keele BF, Leitner T, Han CS, Gleason CD, Green L, Lo CC, Nag A, Wallstrom TC, Wang S, McMichael AJ, Haynes BF, Hahn BH, Perelson AS, Borrow P, Shaw GM, Bhattacharya T, Korber BT. Transmission of single HIV-1 genomes and dynamics of early immune escape revealed by ultra-deep sequencing. *PLoS One*. 2010;5:e12303.
- Lataillade M, Chiarella J, Yang R, Schnittman S, Writz V, Uy J, Seekins D, Krystal M, Mancini M, McGrath D, Simen B, Egholm M, Kozal M. Prevalence and clinical significance of HIV drug resistance mutations by ultra-deep sequencing in antiretroviral-naïve subjects in the CASTLE study. *PLoS One*. 2010;5:e10952.
- Tsibris AMN, Korber B, Arnaout R, Russ C, Lo C-C, Leitner T, Gaschen B, Theiler J, Paredes R, Su Z, Hughes MD, Gulick RM, Greaves W, Coakley E, Flexner C, Nusbaum C, Kuritzkes DR. Quantitative deep sequencing reveals dynamic HIV-1 escape and large population shifts during CCR5 antagonist therapy in vivo. *PLoS One*. 2009;4:e5683.
- Bruselles A, Rozera G, Bartolini B, Prosperi M, Del Nonno F, Narciso P, Capobianchi MR, Abbate I. Use of massive parallel pyrosequencing for near full-length characterization of a unique HIV Type 1 BF recombinant associated with a fatal primary infection. *AIDS Res Hum Retroviruses*. 2009;25:937–42.
- Eshleman SH, Hudelson SE, Redd AD, Wang L, Debes R, Chen YQ, Martens CA, Ricklefs SM, Selig EJ, Porcella SF, Munshaw S, Ray SC, Piwowar-Manning E, McCauley M, Hosseinipour MC, Kumwenda J, Hakim JG, Chariyalertsak S, de Bruyn G, Grinsztejn B, Kumarasamy N, Makhema J, Mayer KH, Pilotto J, Santos BR, Quinn TC, Cohen MS, Hughes JP. Analysis of genetic linkage of HIV from couples enrolled in the HIV prevention trials network 052 trial. *J Infect Dis*. 2011;204:1918–26.
- Macalalad AR, Zody MC, Charlebois P, Lennon NJ, Newman RM, Malboeuf CM, Ryan EM, Boutwell CL, Power KA, Brackney DE, Pesko KN, Levin JZ, Ebel GD, Allen TM, Birren BW, Henn MR. Highly sensitive and specific detection of rare variants in mixed viral populations from massively parallel sequence data. *PLoS Comput Biol*. 2012;8:e1002417.
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, Wilson RK. VarScan 2: somatic mutation and copy

- number alteration discovery in cancer by exome sequencing. *Genome Res.* 2012;22:568–76.
38. Eriksson N, Pachter L, Mitsuya Y, Rhee S-Y, Wang C, Gharizadeh B, Ronaghi M, Shafer RW, Beerwinkler N. Viral population estimation using pyrosequencing. *PLoS Comput Biol.* 2008;4:e1000074.
  39. Westbrook K, Astrovskaya I, Campob D, Khudyakov Y, Bermanc P, Zelikovsky A. HCV quasispecies assembly using network flows. In: *Bioinformatics research and applications*. Berlin-Heidelberg: Springer-Verlag; 2008:159–70.
  40. Madsen R, Kauchak D, Elkan C. Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd international conference on machine learning*. Bonn; 2005:545–52.
  41. Bayes M, Price M. An essay towards solving a problem in the doctrine of chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philos Trans.* 1763;53:370–418.
  42. Pearson K. The fundamental problem of practical statistics. *Biometrika.* 1920;13:1–16.
  43. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Stat Sci.* 2001;16:101–33.
  44. Verbist B, Clement L, Reumers J, Thys K, Vapirev A, Talloen W, Wetzels Y, Meys J, Aerssens J, Bijnsens L, Thas O. ViVaMBC: estimating viral sequence variation in complex populations from illumina deep-sequencing data using model-based clustering. *BMC Bioinform.* 2015;16:59.
  45. Domingo E, Sheldon J, Perales C. Viral quasispecies evolution. *Microbiol Mol Biol Rev.* 2012;76:159–216.
  46. Biosystems A. *Applied biosystems SOLiD(TM) 3 system: instrument operation guide*. 2009.
  47. Nascimento-Brito S, Paulo Zukurov J, Maricato JT, Volpini AC, Salim ACM, Araújo FMG, Coimbra RS, Oliveira GC, Antoneli F, Janini LMR. HIV-1 tropism determines different mutation profiles in proviral DNA. *PLoS One.* 2015;10:e0139037.
  48. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215:403–10.
  49. Ng KW, Tian G-L, Tang M-L. *Dirichlet and related distributions: theory methods and applications*, vol. 889. New York: Wiley; 2011.
  50. Wicker N, Muller J, Kalathura RK, Pocha O. A maximum likelihood approximation method for Dirichlet's parameter estimation. *Comput Stat Data Anal.* 2008;52:1315–22.
  51. Press WHH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical recipes: the art of scientific computing*. 3rd ed. Cambridge: Cambridge University Press; 2007.
  52. Narayanan A. A note on parameter estimation in the multivariate beta distribution. *Comput Math Appl.* 1992;24:11–7.
  53. Ronning G. Maximum likelihood estimation of dirichlet distributions. *J Stat Comput Simul.* 1989;32:215–21.
  54. Efron B, Gong G. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *Am Stat.* 1983;37:36–48.
  55. Efron B. *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics; 1987.
  56. Pukelsheim F. The three sigma rule. *Am Stat.* 1994;48:88.
  57. Kendall MG, O'Hagan A, Foster, J. *Kendall's advanced theory of statistics, vol 2B-Bayesian Inference*, 2nd ed. London: Edward Arnold Press; 2004.
  58. Peckham HE, McLaughlin SF, Ni JN, Rhodes MD, Malek JA, McKernan KJ, Blanchard AP. SOLiD sequencing and 2-base encoding. Poster. USA: *Applied Biosystems*; 2007.
  59. Roberts RJ, Carneiro MO, Schatz MC. The advantages of SMRT sequencing. *Genome Biol.* 2013;14:405.
  60. Lou DI, Hussmann JA, McBee RM, Acevedo A, Andino R, Press WH, Sawyer SL. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc Natl Acad Sci.* 2013;110:19872–7.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

