METHODOLOGY ARTICLE Open Access

Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications

Ziwen He[1], Xinnian Li[1], Shaoping Ling[2], Yun-Xin Fu[3], Eric Hungate[4], Suhua Shi[1*] and Chung-I Wu[1,2,4*]

Abstract

Background: As the error rate is high and the distribution of errors across sites is non-uniform in next generation sequencing (NGS) data, it has been a challenge to estimate DNA polymorphism (θ) accurately from NGS data.

Results: By computer simulations, we compare the two methods of data acquisition - sequencing each diploid individual separately and sequencing the pooled sample. Under the current NGS error rate, sequencing each individual separately offers little advantage unless the coverage per individual is high (>20X). We hence propose a new method for estimating θ from pooled samples that have been subjected to two separate rounds of DNA sequencing. Since errors from the two sequencing applications are usually non-overlapping, it is possible to separate low frequency polymorphisms from sequencing errors. Simulation results show that the dual applications method is reliable even when the error rate is high and θ is low.

Conclusions: In studies of natural populations where the sequencing coverage is usually modest (~2X per individual), the dual applications method on pooled samples should be a reasonable choice.

Keywords: Next generation sequencing, DNA polymorphism, Sequencing error, Pooled sample, Dual sequencing applications

Background

The next generation sequencing (NGS) technologies have dramatically increased the throughput. The new technologies, including those being developed currently, improve on many aspects of DNA sequencing but a higher accuracy than the traditional Sanger sequencing does not appear to be one of them. The nature of the technology would result in specific types of sequencing errors inherent in each process. In general, the new sequencing methods have an error rate between 0.1% and 1.0% [1]. Due to the non-random distribution of errors across sites where some sites can be 10 times more error prone than the average, single nucleotide polymorphism (SNP) calling can often be difficult [2-4].

In this study, we are concerned with estimating a fundamental parameter of natural populations, namely, Watterson's θ of DNA polymorphism [5]. Briefly, θ is the number of nucleotide differences between two sequences of the same locus, randomly chosen from the population. It is a good measure of genetic diversity and a basic parameter for doing population genetic analysis (e.g. tests of positive selection, [6-8]). As polymorphism in natural populations is dominated by low frequency variants [9], which are often indistinguishable from sequencing errors, using the new sequencing technologies to estimate polymorphism will remain a challenge in the near future. A number of methods have been proposed to separate errors from rare polymorphisms [10-14]. Among them, Nielsen *et al.*'s approach [14] is most direct by filtering out errors from the raw read data. However, since error signals may vary from operation to

* Correspondence: lssssh@mail.sysu.edu.cn; cw16@uchicago.edu
[1]State Key Laboratory of Biocontrol and Guangdong Key Laboratory of Plant Resources, Sun Yat-sen University, 135 Xingang West Road, Guangzhou 510275, China
[2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China
Full list of author information is available at the end of the article

operation, its general applicability will need to be evaluated.

There are two ways to prepare samples for sequencing and polymorphism estimation. First, sequencing is done on individual samples, or at least on pooled samples with each sample individually barcoded [15]. We call this type of data “single-line data”. Second, DNA samples from multiple individuals are pooled in equal quantity for sequencing without individual identification [16]. It is referred to as “Pooled-line data”. We should note that sequencing each diploid sample individually is in fact a pooled-line approach as two haploid genomes are sequenced together. In order to call SNP accurately for both haploids, the diploid has to be sequenced to a sufficient depth (e.g. 20X) [3]. Since individual samples are generally not sequenced to such a depth (e.g. the 1000 Human Genome Project [17]), most methods cited above examine the aggregate properties of these individual sequences. In other words, although individuals may be sequenced separately, the data are pooled in the analysis. Hence, for many population genetic questions, little information would be lost by sequencing pooled samples and the efficiency would be greatly improved when the sample number is large. It would then be possible to sequence each pool with greater exactitude in order to filter out errors from the data.

We now propose a method which minimizes the confounding effects of sequencing errors by combining two different sequencing applications. Dual sequencing applications have previously been carried out on the Illumina GA and SOLiD platforms for the same samples [16,18,19]. It has been shown that the two technologies have nearly non-overlapping error distributions [4]. Dual platform is in fact a standard method as NGS sequencing, on whichever platform, needs to be backed up by another method, usually by Sanger sequencing or other genotyping tools [4,20,21]. Dual applications on two NGS platforms is simply a more systematic and large-scale method of error correction. Such dual applications can also be expected on newer and very different technologies such as HiSeq [22], Ion Protons [23], PacBio [24] and MspA nanopore [25]. When dual platform sequencing is not feasible, dual applications of the same platform on the same DNA sample, independently prepared for sequencing, may serve the same purpose. The correlation of error distribution between two applications on the same platform is slightly higher than those on different platforms but is often adequate for error corrections.

In this study, we first investigated a simple single-line method by extracting haploid information from individual diploids. We then propose dual sequencing applications to improve the pooled-line method for analyzing pooled samples of diploids.

Results

Single-line data

If the effort of data collection is not a limiting factor, the best method is to sequence each diploid individual to a sufficient depth such that true polymorphisms, with the variant frequency at 0.5, can be unambiguously separated from errors.

To ensure false positive error rate being less than 10%, it need more than 20X depth for most next generation sequencing platforms [3]. With a lower coverage, there would be many sites where the distinction between errors and polymorphisms is not possible. Therefore, when data are obtained with low coverage of diploid individuals (say, 2X), we suggest taking data from only one haploid genome per diploid individual. In this scheme, an average depth of 2X would ensure that 86% of individuals could be covered at each site, provided that the distribution of sequencing depth at each site follows a Poisson distribution, $p(depth > 0) = 1\text{-}e^{-2}$. Since we are interested in comparing various methods of estimating genetic diversity, all of them are applied to data with an average depth of 2X per diploid individual.

Theory

Define θ as the nucleotide diversity per site. Let S denote the number of segregating sites and l denote the total number of sites. Watterson showed that

$$E(S) = a_n \theta l, \tag{1}$$

where $a_n = \sum_{i=1}^{n-1} \frac{1}{i}$ and n is the sample size [26]. We assume n individuals with an average depth of 2X per individual. Hence, at site j, only n_j individuals would be sequenced ($n_j \le n$). Among these n_j individuals, we randomly select one read to represent a haploid genome of this individual. When this site is observed to be polymorphic among the n_j genomes, $S_j = 1$; otherwise, $S_j = 0$. In the absence of sequencing error, the estimate of θ is

$$\hat{\theta} = \frac{1}{l} \sum_{j=1}^{l} \frac{S_j}{a_{n_j}}, \tag{2}$$

where $a_{n_j} = \sum_{i=1}^{n_j-1} \frac{1}{i}$.

Because some variants observed among the n_j individuals would be sequencing errors, we need to consider a more reliable portion of the frequency spectrum in the estimation of θ. Given the current sequencing error rate [1], sequencing errors would usually appear as singletons (number of variant, b, being 1 or n_j-1) or doubletons ($b = 2, n_j$-2). Ewens showed that $\frac{1}{b} / \sum_{i=1}^{n-1} \frac{1}{i}$ is the probability that a mutant is represented b times in n samples and the estimate of θ should be

$$\hat{\theta} = \frac{1}{l}\sum_{j=1}^{l}\frac{S_j}{a'_{n_j}}, \tag{3}$$

where $a'_{n_j} = \sum_{i=1+z}^{n_j-1-z}\frac{1}{i}$ [9]. In this formula, $z = 1$ when singletons are removed and $z = 2$ when both singletons and doubletons are removed.

Simulations

We simulated a 100 kb region of different θ and sequencing error rate. The results are presented below the heading of “Single-line” in Table 1. $S_{>0}$ denotes that all segregating sites detected by reads are counted. $S_{>1}$ represents all segregating sites excluding singletons, while $S_{>2}$ excluding both singletons and doubletons.

When the error rate is set to 0, the estimates of θ using $S_{>0}$ are very close to the true values. When the error rate is 0.001 to 0.01, the estimates of θ using $S_{>0}$ become extremely unreliable, as expected, and the removal of singletons and doubletons becomes necessary. With an error rate of 0.001, the estimate of θ using $S_{>2}$ is 0.101, very close to the true value of 0.1. If the error rate is as high as 0.01, estimation by the single line method becomes unreliable even the singletons and doubletons are removed.

A serious problem in SNP calling is the non-random distribution of errors across sites [2,3]. In reality, some sites can be 10 times more error prone than the rest [4]. We hence conducted simulations with the assumption that the error rate is Beta distributed ($\varepsilon \sim Beta(a,\beta)$). We use different shape parameters ($a = 0.1, 0.2, 0.4$ *and* 0.8). It is clear from Table 2 that, when the error rate is non-constant, the single line method is not accurate for estimating θ even with the removal of singletons and doubletons.

Pooled-lines data from single platform

From the section above, it appears that the most efficient strategy for accurately estimating genetic diversity would not be single-line sequencing. Given the low coverage for each individual, variant frequencies, rather than the genotypes of individuals, are the quantities of interest. Pooling samples for bulk sequencing may be equally informative but at a lower cost and effort [16]. When pooled samples are sequenced, each haploid genome would not present equally in the final data and the coverage would vary from site to site. The statistics to correct for these fluctuations are given below. In this and the next sections, the pooled samples are sequenced by one single application or by dual applications. The ability to separate errors from true polymorphisms differs greatly between the two approaches.

Theory

Equal amount of DNAs from each individual are pooled and the pooled samples are sequenced on one sequencing platform. Assuming a segregating site with b mutants in a sample of size n is covered by r reads in an Illumina GA or SOLiD dataset, Jiang *et al.* [10] showed

Table 1 Estimating θ with constant sequencing error rate

Error rate	Sites used	θ = 0.1 / kb			θ = 1 / kb		
		Single line	Pooled-lines		Single line	Pooled-lines	
			single platform	dual applications		single platform	dual applications
0	$S_{>0}$	0.099 (0.007)	0.100 (0.003)	0.100 (0.007)	0.999 (0.023)	1.000 (0.009)	0.999 (0.022)
	$S_{>1}$	0.100 (0.012)	0.100 (0.005)	0.100 (0.010)	0.999 (0.041)	1.000 (0.016)	1.000 (0.034)
	$S_{>2}$	0.099 (0.016)	0.100 (0.007)	0.100 (0.013)	0.998 (0.050)	1.000 (0.023)	1.000 (0.041)
0.001	$S_{>0}$	5.992 (0.131)	22.054 (0.212)	0.323 (0.026)	6.884 (0.131)	22.872 (0.225)	1.226 (0.033)
	$S_{>1}$	0.129 (0.017)	0.507 (0.032)	0.100 (0.011)	1.032 (0.040)	1.409 (0.035)	1.003 (0.034)
	$S_{>2}$	0.101 (0.017)	0.105 (0.008)	0.100 (0.013)	0.999 (0.052)	1.009 (0.023)	1.002 (0.042)
0.005	$S_{>0}$	28.389 (0.271)	90.901 (0.357)	5.269 (0.116)	29.184 (0.275)	91.485 (0.358)	6.165 (0.118)
	$S_{>1}$	0.810 (0.054)	9.295 (0.149)	0.112 (0.012)	1.716 (0.064)	10.167 (0.152)	1.024 (0.035)
	$S_{>2}$	0.112 (0.017)	0.662 (0.039)	0.100 (0.014)	1.016 (0.053)	1.575 (0.046)	1.010 (0.042)
0.01	$S_{>0}$	53.883 (0.351)	146.007 (0.357)	18.820 (0.215)	54.615 (0.331)	146.329 (1.111)	19.684 (0.218)
	$S_{>1}$	2.861 (0.105)	31.994 (0.258)	0.257 (0.025)	3.777 (0.107)	32.823 (0.821)	1.179 (0.040)
	$S_{>2}$	0.180 (0.027)	4.061 (0.108)	0.102 (0.013)	1.091 (0.057)	4.993 (0.327)	1.017 (0.041)

The average depth is 2X per individual in single line method, 2X per haploid genome in single platform method and 1X per haploid genome in each application in dual applications method. The means (and the standard deviations) of θ are estimated from 1000 replicates. Error rate is per site.

Table 2 Estimating θ with Beta distributed sequencing error rate

Parameter α of Beta distribution	Sites used	θ = 0.1 / kb			θ = 1 / kb		
		Single-line	Pooled-lines		Single-line	Pooled-lines	
			single platform	dual applications		single platform	dual applications
	$S_{>0}$	19.917 (0.237)	38.335 (0.274)	2.534 (0.083)	20.759 (0.227)	39.095 (0.267)	3.444 (0.085)
0.1	$S_{>1}$	4.709 (0.139)	17.500 (0.197)	0.336 (0.032)	5.605 (0.133)	18.347 (0.197))	1.248 (0.043)
	$S_{>2}$	1.172 (0.073)	9.541 (0.159)	0.130 (0.017)	2.073 (0.081)	10.425 (0.156)	1.041 (0.042)
	$S_{>0}$	23.170 (0.255)	51.539 (0.290)	3.499 (0.098)	23.964 (0.230)	52.237 (0.306)	4.393 (0.097)
0.2	$S_{>1}$	3.343 (0.112)	18.415 (0.203)	0.250 (0.026)	4.243 (0.120)	19.249 (0.208)	1.160 (0.040)
	$S_{>2}$	0.534 (0.049)	7.555 (0.143)	0.109 (0.014)	1.443 (0.071)	8.444 (0.145)	1.015 (0.041)
	$S_{>0}$	25.398 (0.240)	64.217 (0.333)	4.243 (0.105)	26.210 (0.239)	64.855 (0.335)	5.134 (0.108)
0.4	$S_{>1}$	2.259 (0.095)	17.193 (0.201)	0.181 (0.019)	3.164 (0.097)	18.017 (0.201)	1.090 (0.038)
	$S_{>2}$	0.267 (0.034)	4.916 (0.114)	0.103 (0.013)	1.175 (0.059)	5.808 (0.114)	1.010 (0.042)
	$S_{>0}$	26.772 (0.270)	74.504 (0.355)	4.717 (0.112)	27.578 (0.262)	75.097 (0.340)	5.609 (0.111)
0.8	$S_{>1}$	1.591 (0.073)	14.860 (0.196)	0.143 (0.015)	2.492 (0.087)	15.706 (0.181)	1.054 (0.035)
	$S_{>2}$	0.171 (0.024)	2.870 (0.084)	0.101 (0.013)	1.076 (0.057)	3.773 (0.088)	1.010 (0.041)

The average error rate is 0.005 per site. The average depth is 2X per individual in single line method, 2X per haploid genome in single platform method and 1X per haploid genome in each application in dual applications method. The means (and the standard deviations) of θ are estimated from 1000 replicates.

that the probability $q_1(b)$ that this segregating site is detected by reads is

$$q_1(b,r) = 1 - (1-b/n)^r - (b/n)^r, \tag{4}$$

for $0 < b < n$, and the probability q_2 that a segregating site with an arbitrary b value is detected by reads is

$$q_2(r) = \sum_{b=1}^{n-1} q_{nb} q_1(b,r). \tag{5}$$

Ewens showed that $q_{nb} = (1/b)/a_n = \frac{1}{b}/\sum_{i=1}^{n-1}\frac{1}{i}$ is the probability that a mutant presents b times in n samples [9].

Let S_T denote the number of segregating sites detected by reads, and we can obtain

$$E(S_T) = \frac{S}{l}\sum_{j=1}^{l} q_2(r_j), \tag{6}$$

where r_j is the number of reads covering the site j. Hence the estimate of θ is

$$\begin{aligned}\hat{\theta} &= S/a_n l \\ &= \frac{E(S_T)}{a_n \sum_{j=1}^{l} q_2(r_j)}.\end{aligned} \tag{7}$$

Replacing q_2 with equation (5) yields

$$\hat{\theta} = \frac{E(S_T)}{\sum_{j=1}^{l}\sum_{b=1}^{n-1}\frac{q_1(b,r_j)}{b}}. \tag{8}$$

Now we shall consider a more realistic case with sequencing errors in the data. Let’s assume a case in which a site is covered by r reads in a single platform and has mismatches in x read(s) caused by sequencing error. The probability $P_\varepsilon(r,x)$ of its occurrence at a non-segregating site is

$$p_\varepsilon(r,x) = C_r^x \varepsilon^x (1-\varepsilon)^{r-x}, \tag{9}$$

where ε denotes the sequencing error at this site. Since the average raw error rate ranges from 0.1% to 1.0% [1], the sequencing error can cause severe problems when estimating polymorphism.

However, if using an observed segregating site only when the minor allele has more than z reads, we may obtain more accurate estimates. Instead of equation (4), the probability that a site with b mutants in a sample of size n is detected by r reads as a segregating site with more than z reads of each allele is

$$\begin{aligned}q_1(z,b,r) = 1 &- \sum_{x=0}^{z} C_r^x (1-b/n)^{r-x}(b/n)^x \\ &- \sum_{x=r-z}^{r} C_r^x (1-b/n)^{r-x}(b/n)^x.\end{aligned} \tag{10}$$

The estimate of θ is now

$$\hat{\theta} = \frac{E(S_{>z})}{\sum_{j=1}^{l}\sum_{b=1}^{n-1}\frac{q_1(z,b,r_j)}{b}}. \tag{11}$$

$S_{>z}$ denotes the number of segregating sites at which two different alleles are both detected by more than z reads. All segregating sites detected by reads are counted when $z = 0$. Hence, $S_{>0}$ is equal to S_T.

The procedure to estimate θ using pooled-lines data from single platform is as follows. For each site, we (1) treat the data as missing if the number of reads is less than $r_{\min}$ in this platform; (2) retain alleles having more than z reads. If there is only one allele in this platform, we treat this site as a nonsegregating site; if two, as a segregating site; if more than two, we treat the data as missing; (3) use equation (11) to calculate θ for the single platform. $r_{\min}$ should not be no lower than $(2z + 2)$. In the following simulations, we set $r_{\min} = 6$.

Simulations

We used the simulated data to test this method. The results are referred to as "single platform" in Table 1. When singletons (the minor depth allele is covered by only one read, $z = 1$) or both singletons and doubletons (the minor depth allele is covered by two reads, $z = 2$) are discarded (the row "$S_{>1}$" and "$S_{>2}$"), the standard deviation becomes larger if there is no sequencing error. In reality, sequencing error cannot be ignored.

We assume that the error rate is constant across sites. Different error rates (0.001, 0.005 and 0.01) are used in the estimation. The simulation results are displayed in Table 1. For example, when $S_{>0}$ is used with an error rate of 0.001, sequencing errors lead to very poor estimates of θ. The mean estimate of θ is 22.054 per kb if all segregating sites are used, which is many times higher than the true value of 0.1 per kb. The estimation becomes more accurate when $S_{>1}$ or $S_{>2}$ is used. Thus, when the error rate is low (e.g. 0.001), this method can be used to estimate θ with both singletons and doubletons discarded. However, when the error rate is high (e.g. 0.005 or 0.01), even excluding singletons and doubletons (the row "$S_{>2}$" in Table 1) does not lead to acceptable estimates. For simulations with the assumption that the error rate is Beta distributed and its mean is 0.005, the estimates are also unacceptable (shown in Table 2).

Pooled-lines data from dual sequencing applications

It is customary to validate calls of variants by another method. For example, variant calls on the Illumina platform are often validated by Sanger sequencing or by fast SNP genotyping methods, e.g. Sequenom genotyping [4,20]. Because validation is often laborious and incomplete, it may be more efficient and informative to deploy two sequencing methods fully and independently. If the two applications have distinctive error-distribution patterns, the errors could be identified and excluded by reciprocally correcting each other's errors. Indeed, several widely used sequencing methods (as well as the latest methods that are in development) are based on very different chemistry and protocols. As shown below, we analyzed the sequencing results obtained by Illumina based data and SOLiD, and as expected, we observed the two datasets showed non-overlapping errors.

Data on error correlation between sequencing applications

Dual platforms - We re-analyzed sequencing data from a species of mangrove trees, *Sonneratia alba*, known to be completely monomorphic within some populations [4]. DNA sequences for 71 genes from one such population were generated using the Illumina GA and SOLiD platforms at a depth of ~2500X and ~5400X, respectively. For sites with more than 2000X depth in both platforms, we called variants using a set of criteria more stringent than the previous study. As shown in Figure 1a, Illumina GA and SOLiD systems both call many false SNPs, few of which are called by both. Because the sample is known to be monomorphic by Sanger sequencing [4], the detected variants are all false SNPs, which fortunately do not show overlap between platforms. Pearson's correlation coefficient of the error rate distributions between the two platforms is only 0.054.

Single platform - For analyzing the correlation of two samples sequenced on the same platform, we use our own unpublished data from the Illumina HiSeq platform. A sample of 35 individuals from a mangrove species, *Avicennia marina,* was taken from each of two nearby populations in Thailand. Equal amount of DNAs from 35 individuals (or 70 haploid genomes) were pooled. 93 genes were amplified for both of the two pooled samples and sequenced on an Illumina HiSeq platform. For sites with more than 2000X depth in both samples, we called SNPs at the sites whose minor allele frequency (MAF) is lower than 0.01 in both samples. In total, 55,602 sites were retained and were plotted in Figure 1b. Almost all of these variants are sequencing errors as explained in Methods. Figure 1b shows the observed error rates on these sites. Pearson's correlation coefficient of the error rate distributions between these two samples is only 0.142, a little higher than that between platforms of Figure 1a. Therefore, for samples prepared and sequenced twice on one platform, sequencing errors also overlap only rarely.

Theory

If sequencing errors from two applications do not overlap, segregating sites detected by both should be true

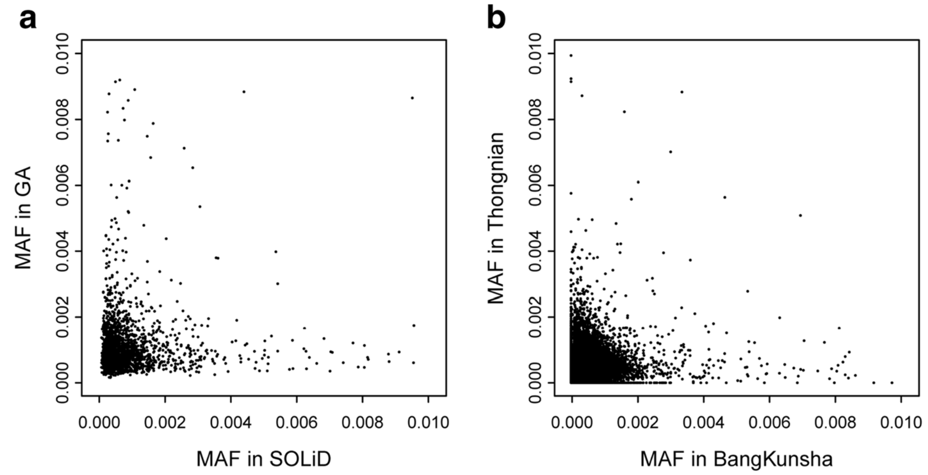

Figure 1 Error rate correlation patterns. a) MAF (minor allele frequency) of putative SNPs called by either SOLiD or Illumina GA. **b)** MAF in two samples (Bangkunsha and Thongnian) sequenced by Illumina HiSeq.

variants. The probability $q_1(b)$ that a segregating site with b mutants in a sample of size n is detected in both applications is

$$q_1(b, r_1, r_2) = \prod_{k=1,2} \left(1 - (1 - b/n)^{r_k} - (b/n)^{r_k}\right). \tag{12}$$

The overall estimate of θ by the combined data is

$$\hat{\theta} = \frac{E(S_T)}{\sum_{j=1}^{l} \sum_{b=1}^{n-1} \frac{q_1(b, r_{1j}, r_{2j})}{b}}, \tag{13}$$

where r_{1j} is the number of reads covering site j in the first dataset, while r_{2j} is the number of reads covering site j in the second dataset.

For non-overlapping errors, a site with b mutants in a sample of size n that is detected as a segregating site with more than z reads of each allele in both applications is associated with the probability

$$q_1(z, b, r_1, r_2) = \prod_{k=1,2} \left(1 - \sum_{x=0}^{z} C_{r_k}^{x} (1-b/n)^{r_k - x} (b/n)^{x} - \sum_{x=r_k-z}^{r_k} C_{r_k}^{x} (1-b/n)^{r_k - x} (b/n)^{x}\right). \tag{14}$$

The θ estimated by the dual applications method is

$$\hat{\theta} = \frac{E(S_{>z})}{\sum_{j=1}^{l} \sum_{b=1}^{n-1} \frac{q_1(z, b, r_{1j}, r_{2j})}{b}}. \tag{15}$$

Here $S_{>z}$ denotes the number of segregating sites in which two different alleles are both detected by more than z reads on both applications.

The procedure to estimate θ using data from dual sequencing applications is as follows. For each site, we (1) treat the data as missing if the number of reads is less than $r_{\min}$ on either applications; (2) retain alleles having more than z reads on both applications. If there is only one allele on either application, we treat this site as a non-segregating site. A site is considered segregating only when reads from both applications report segregation; (3) use equation (15) to calculate θ for the combined dataset. We set $r_{\min} = 6$ in the following simulations.

Simulations

The simulation procedure is almost the same as that for the single platform, but with data from an additional sequencing application. The means and the standard deviations of θ estimates using different parameters are reported in Table 1. For sequencing data without errors, the dual platform method can accurately estimates θ, although the standard deviation values are slightly larger than those obtained by the single platform method. However, with the increase of the error rate, the advantage of the dual platform method compared with other methods becomes obvious (Figure 2). With an error rate of 0.01, the mean estimate of θ is 0.102 per kb when using $S_{>2}$, which is only 2% higher than the real value (0.1 per kb). This estimate is dramatically better than the corresponding single platform estimate (4.061) or the single line estimate (0.180). This method is also better than the others when the error rate is Beta distributed as shown in Table 2.

In Figure 3, we used different region lengths to test the dual applications method. The estimations of θ is acceptable even when the region is small (e.g. 10 kb). For a 40 kb region (the real number of S is about 180), the standard deviation of θ estimates is account for only 5% of the real θ.

Discussion

While NGS has increased the power of DNA sequencing by orders of magnitude in the recent years, its accuracy per read is the one aspect that has not been improved. For example, 454 Pyrosequencing is susceptible to homopolymer indels [1]. The Illumina GA and SOLiD platforms are both PCR based systems and are prone to base substitution errors. The first glimpses of newer technologies do not offer promises for improving per read accuracy either. Nevertheless, the nature of the substitution errors may differ among platforms since major sources of errors, from library construction to base-pair determination, depend on different physical and chemical principles among these technologies. The method described herein takes advantage of the non-overlapping distributions to minimize error rates.

The error rate across all sites is platform-dependent and not constant (e.g. Beta distribution) [4]. When doing the simulation, we assume that a nucleotide has an equal probability of being read incorrectly as one of the three other nucleotides. However, the patterns of error rates for the real data are much more complex. The frequencies of base substitution error could vary by 10 to 11 fold, with A to C transversions being among the most frequent substitution errors and C to G transversions among the least frequent ones [27]. Therefore, if a non-segregating site (e.g. A) has two reads with sequencing errors, a doubleton error is more likely (e.g. two A to C errors) rather than two singleton errors (e.g. one from A to C and another from A to T). In other words, the unevenly distributed errors can cause severe problems in estimating polymorphism. In this situation, we strongly suggest using dual sequencing applications to avoid this kind of errors.

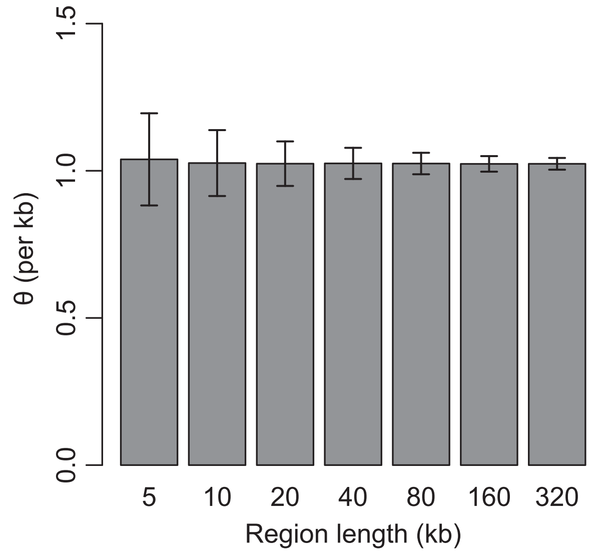

Figure 3 θ estimation of dual applications for different region length. The θ value of simulation data is set to 1 per kb. The sequencing error rate is set to 0.005. Singletons are discarded in the estimation ($S_{>1}$). The length of each error bar is 2 times the standard deviation. The means (and the standard deviations) of θ are estimated from 1000 replicates.

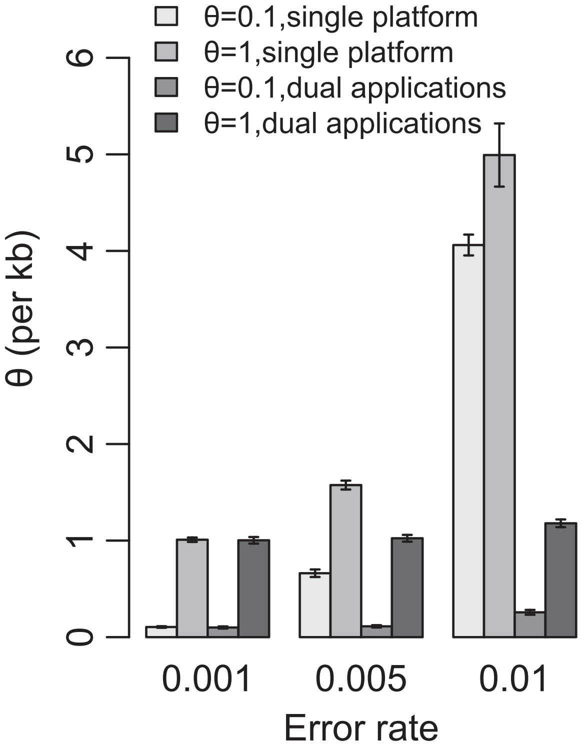

Figure 2 θ estimation of simulation data of pooled-lines sample with 3 different sequencing errors. The θ value of simulation data is set to 0.1 / 1 per kb. Singletons are discarded in dual applications method ($S_{>1}$). Singletons and doubletons are discarded in single platform method ($S_{>2}$). The length of each error bar is 2 times the standard deviation. The means (and the standard deviations) of θ are estimated from 1000 replicates.

Conclusions

Our model can estimate θ accurately by combining data from two different sequencing applications. The method is robust even when the error rate is extremely high and variable across sites. We also evaluated the relative merits of pooled-lines versus single-line data. If the coverage per line is low, dual sequencing application on pooled lines yields the best results. However, the inherent high error rates in the NGS technologies impose constraints on the estimation of polymorphisms. Even under the best of conditions with sequencing done on two platforms, singletons and doubletons still have to be removed. If the estimation requires accuracy in the low frequency portion of the variant spectrum, it will be necessary to carry out sequencing on each line individually with a high coverage of >20X. For many scientific questions, our strategy of dual sequencing applications on pooled samples with modest coverage can yield the most information for the same level of effort.

Methods

Sample preparing and sequencing

We sampled two *Avicennia marina* populations (Bangkunsha and Thongnian) in Thailand. Equal amount of DNAs from 35 individuals (diploids) in each population were pooled, respectively. 93 genes were respectively amplified for both of the two pooled samples and then sequenced on an Illumina HiSeq platform.

Reads alignment and SNPs calling

We use *MAQ* [28] to align reads to the known references. Nucleotides with base quality low than 20 are discussed. For sites with more than 2000X depth in both samples, we called candidate polymorphic sites whose minor allele frequency is lower than 0.01 in both samples. In total, 55,602 sites were retained and were plotted in Figure 1b.

In re-analyzing sequencing data of *Sonneratia alba,* Singletons are discarded and only the mutant alleles with at least one read aligned in forward strands and one read aligned in backward strands are retained for the following analyses. 2382 candidate sites were plotted in Figure 1a.

Searching sites with errors

Consider a singleton site with MAF being 1/70, if the sequencing depth of this site is 2000X, we can infer the probability of observing its MAF < 0.01 to be 0.0375, using the distribution function of a Binomial distribution. For a singleton with the same MAF being observed in both samples, the probability is 0.0014 (the square of 0.0375). If the site has more than two mutant alleles or the depth is more than 2000X, the probability will decrease. The total number of SNPs of these two populations is estimated no more than 500 for these 93 genes. Therefore, there should be no more than 0.7 (0.0014*500) true polymorphic sites in Figure 1b. Near all candidate polymorphic sites in Figure 1b are introduced by errors.

Simulation progress

We simulated sequencing progress with a Poisson distributed depth. Errors were added randomly for each site with the given error rate. We wrote Perl scripts to evaluate θ for single/dual applications method described in the main text. The means and the standard deviations of θ for each combination of parameters in Table 1 and Table 2 are estimated from 1000 replicates.

For single-line data, we simulated a 100 kb region for 25 diploid individuals with an average depth of 2X; hence, $Max(n_j) = 25$ as only one read is used per individual. We set θ to be 0.1 or 1 per kb and error rate to be 0, 0.001, 0.005 or 0.01 per site.

For pooled-lines data, a 100 kb region is simulated for 25 diploids (50 haploid genomes) using a single platform or dual platforms. We set different θ values (0.1 / 1 per kb) and used $S_{>0}$, $S_{>1}$ and $S_{>2}$ in the estimate. The average depth is 2X per haploid genome in single platform method and 1X per haploid genome in each application in dual applications method.

Abbreviations

NGS: Next generation sequencing; SNP: Single nucleotide polymorphism; MAF: Minor allele frequency.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ZH carried out the theoretical derivation, simulation and drafted the manuscript. XL and SL carried out the sequence alignment and the analysis of the correlation of sequencing errors between two platforms/samples. YXF participated in the theoretical derivation. EH participated in the simulation and helped to draft the manuscript. SS and CIW conceived using dual sequencing applications to correct sequencing errors and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Ziheng Yang, Tianqi Zhu, Weiwei Zhai and Suo Qiu for their helpful comments and suggestions. SS is supported by the grants from the National Natural Science Foundation of China (41130208) and Science Foundation of State Key Laboratory of Biocontrol (SKLB201224, SKLBC12K04, SKLBC2011K01). CIW is supported by the grants from the National Natural Science Foundation of China (31130069). ZH is supported by scholarship award for excellent doctoral student granted by ministry of education and Yat-sen innovation project of Sun Yat-sen University.

Author details

[1]State Key Laboratory of Biocontrol and Guangdong Key Laboratory of Plant Resources, Sun Yat-sen University, 135 Xingang West Road, Guangzhou 510275, China. [2]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, 1 Beichen West Road, Beijing 100101, China. [3]Human Genetics Center, University of Texas School of Public Health, 1200 Herman Presser Drive, Houston, TX 77030, USA. [4]Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA.

Received: 18 January 2013 Accepted: 3 August 2013
Published: 7 August 2013

References

1. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135–1145.
2. Druley TE, Vallania FLM, Wegner DJ, Varley KE, Knowles OL, Bonds JA, Robison SW, Doniger SW, Hamvas A, Cole FS, Fay JC, Mitra RD: **Quantification of rare allelic variants from pooled genomic DNA.** *Nat Methods* 2009, **6:**263–265.
3. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA: **Evaluation of next generation sequencing platforms for population targeted sequencing studies.** *Genome Biol* 2009, **10:**R32.
4. Zhou R, Ling S, Zhao W, Osada N, Chen S, Zhang M, He Z, Bao H, Zhong C, Zhang B, Lu X, Turissini D, Duke NC, Lu J, Shi S, Wu CI: **Population genetics in non-model organisms: II. Natural selection in marginal habitats revealed by deep sequencing on dual platforms.** *Mol Biol Evol* 2011, **28:**2833–2842.
5. Watterson GA: **On the number of segregating sites in genetical models without recombination.** *Theol Popul Biol* 1975, **7:**256–276.
6. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123:**585–595.
7. Fu YX: **Statistical properties of segregating sites.** *Theor Popul Biol* 1995, **48:**172–197.
8. Zeng K, Shi S, Wu CI: **Compound tests for the detection of hitchhiking under positive selection.** *Mol Biol Evol* 2007, **24:**1898–1908.
9. Ewens WJ: *Mathematical population genetics.* Berlin: Springer-Verlag; 1979.
10. Jiang R, Tavare S, Marjoram P: **Population genetic inference from resequencing data.** *Genetics* 2009, **181:**187–197.
11. Liu X, Maxwell TJ, Boerwinkle E, Fu YX: **Inferring population mutation rate and sequencing error rate using the SNP frequency spectrum in a sample of DNA sequences.** *Mol Biol Evol* 2009, **26:**1479–1490.
12. Li H: **A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.** *Bioinformatics* 2011, **27:**2987–2993.
13. Le SQ, Durbin R: **SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples.** *Genome Res* 2011, **21:**952–960.
14. Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J: **SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data.** *PLoS ONE* 2012, **7:**e37558.
15. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless AM, Wang J, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang J, Hudson ME, Bent AF: **Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean.** *Science* 2012, **338:**1206–1209.
16. He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, Greenberg AJ, Hudson RR, Wu CI, Shi S: **Two Evolutionary Histories in the Genome of Rice: the Roles of Domestication Genes.** *PLoS Genet* 2011, **7:**e1002100.
17. The 1000 Genomes Project Consortium: **A map of human genome variation from population-scale sequencing. *Nature*.** *Nature* 2010, **467:**1061–1073.
18. Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G: **Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing.** *Nucleic Acids Res* 2010, **38:**4755–4767.
19. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, *et al*: **The developmental transcriptome of Drosophila melanogaster.** *Nature* 2011, **471:**473–479.
20. Calvo SE, Tucker EJ, Compton AG, Kirby DM, Crawford G, Burtt NP, Rivas M, Guiducci C, Bruno DL, Goldberger OA, Redman MC, Wiltshire E, Wilson CJ, Altshuler D, Gabriel SB, Daly MJ, Thorburn DR, Mootha VK: **High-throughput, pooled sequencing identifies mutations in NUBPL and FOXRED1 in human complex I deficiency.** *Nat Genet* 2010, **42:**851–858.
21. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, Albers CA, Zhang ZD, Conrad DF, Lunter G, Zheng H, Ayub Q, DePristo MA, Banks E, Hu M, Handsaker RE, Rosenfeld JA, Fromer M, Jin M, Mu XJ, Khurana E, Ye K, Kay M, Saunders GI, Suner MM, Hunt T, *et al*: **A systematic survey of loss-of-function variants in human protein-coding genes.** *Science* 2012, **335:**823–828.
22. Minoche AE, Dohm JC, Himmelbauer H: **Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems.** *Genome Biol* 2011, **12:**R112.
23. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, *et al*: **An integrated semiconductor device enabling non-optical genome sequencing.** *Nature* 2011, **475:**348–352.
24. Coupland P, Chandra T, Quail M, Reik W, Swerdlow H: **Direct sequencing of small genomes on the Pacific Biosciences RS without library preparation.** *Biotechniques* 2012, **53:**365–372.
25. Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH: **Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase.** *Nat Biotechnol* 2012, **30:**439–353.
26. Watterson GA: **Heterosis or neutrality.** *Genetics* 1977, **85:**789–814.
27. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36:**e105.
28. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851–1858.

doi:10.1186/1471-2164-14-535
Cite this article as: He *et al.*: Estimating DNA polymorphism from next generation sequencing data with high error rate by dual sequencing applications. *BMC Genomics* 2013 **14**:535.