BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 22 2012, pages 2870–2874
doi:10.1093/bioinformatics/bts563

Genome analysis
Advance Access publication October 8, 2012

# COPE: an accurate *k*-mer-based pair-end reads connection tool to facilitate genome assembly

Binghang Liu[1,2,†], Jianying Yuan[2,†], Siu-Ming Yiu[1,3], Zhenyu Li[1,2], Yinlong Xie[1,2], Yanxiang Chen[2], Yujian Shi[2], Hao Zhang[2], Yingrui Li[1,2], Tak-Wah Lam[1,3,*] and Ruibang Luo[1,2,3,*]

[1]HKU-BGI BAL (Bioinformatics Algorithms and Core Technology Research Laboratory), The University of Hong Kong, Hong Kong, [2]BGI-Shenzhen, Shenzhen, Guangdong 518083, China and [3]Department of Computer Science, The University of Hong Kong, Hong Kong

Associate Editor: Inanc Birol

## ABSTRACT

**Motivation:** The boost of next-generation sequencing technologies provides us with an unprecedented opportunity for elucidating genetic mysteries, yet the short-read length hinders us from better assembling the genome from scratch. New protocols now exist that can generate overlapping pair-end reads. By joining the 3′ ends of each read pair, one is able to construct longer reads for assembling. However, effectively joining two overlapped pair-end reads remains a challenging task.

**Result:** In this article, we present an efficient tool called Connecting Overlapped Pair-End (COPE) reads, to connect overlapping pair-end reads using *k*-mer frequencies. We evaluated our tool on 30× simulated pair-end reads from *Arabidopsis thaliana* with 1% base error. COPE connected over 99% of reads with 98.8% accuracy, which is, respectively, 10 and 2% higher than the recently published tool FLASH. When COPE is applied to real reads for genome assembly, the resulting contigs are found to have fewer errors and give a 14-fold improvement in the N50 measurement when compared with the contigs produced using unconnected reads.

**Availability and implementation:** COPE is implemented in C++ and is freely available as open-source code at ftp://ftp.genomics.org.cn/pub/cope.

**Contact:** twlam@cs.hku.hk or luoruibang@genomics.org.cn

Received on June 4, 2012; revised on September 10, 2012; accepted on September 12, 2012

## 1 INTRODUCTION

With the rapid development of high-throughput short-read sequencing technologies, a laboratory can now assemble a genome within a few weeks using only a few thousand US dollars (Glenn, 2011). Despite the fact that existing genome assembly algorithms (Gnerre *et al.*, 2011; Li *et al.*, 2010) already designed to fully utilize the advantages of short reads, the short length of reads still deter us from a more comprehensive genome. Using longer reads will significantly improve the quality of not only genome assembly (Magoc and Salzberg, 2011) but also transcriptome (Martin and Wang, 2011) and meta-genome assembly (Rodrigue *et al.*, 2010).

To overcome this problem, new protocols for library preparation now exist that can generate pair-end reads with an insert distance shorter than the total length of both reads, i.e. with the 3′ ends of both reads overlapped. This enables us to extend the read length by authentically overlapping the ends of both reads to generate a longer overlapped read. These longer reads were shown to dramatically increase the quality of the assembled genome (Magoc and Salzberg, 2011).

However, the insert distance cannot be controlled to be the same for all read pairs, instead it follows a distribution. Together with the sequencing errors, it becomes difficult to determine the correct position that the two reads overlap and what bases are in the overlapping region in case the two reads do not agree. The correctness of the connected reads is extremely important as existing assembly algorithms are sensitive to errors. Many assemblers are based on the approach of *de Bruijn* graph in which a read is sheared into *k*-bp long substrings (called *k*-mers) (Li *et al.*, 2010). One single base error will introduce *k* spurious *k*-mer. The situation could be worse if a larger *k* is used. For Illumina sequencing technology specifically, the sequencing errors tend to intensify toward the 3′ end of the reads, which makes the connection task more challenging.

To deal with sequencing errors, a straightforward approach is to perform error correction before joining the reads. Many existing error correction algorithms (Kelley *et al.*, 2010) may require to trim the 3′ end of a read if the correction fails. This will significantly reduce the number of reads that can be successfully connected. One of the most recently published read connection tools, FLASH (Magoc and Salzberg, 2011), follows the idea of performing error correction before joining the reads, and it is found that quite a few read pairs cannot be connected as they are trimmed. Other existing tools such as PANDAseq (Masella *et al.*, 2012) and SHERA (Rodrigue *et al.*, 2010) also fail to provide a method to overcome the problem either.

In this article, we present Connecting Overlapped Pair-End (COPE) reads, a novel approach to connect pair-end reads by utilizing *k*-mer frequency information directly to authenticate possible overlaps of reads. We remark that *k*-mer frequency information was also used in error correction in reads.

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

We evaluated the correctness of our tool on 30-fold simulated 100 bp paired-end reads with 1% error and found it surmounted FLASH on both accuracy and the number of reads connected. We also applied COPE to real reads for genome assembly and showed that the N50 of the resulting contigs is 14 times longer than the contigs produced using unconnected reads and even with fewer errors.

## 2 METHODS

Before we present the details of COPE, we first describe the limitations of existing tools.

First, a minimum overlap is required. Since the insert distance is not exact, the length of overlapped region varies. Existing methods require typically at least 10 bp overlap at 3′ end because shorter overlaps often occur by chance. A higher requirement of minimum overlap reduces spurious connections but will miss quite a few true overlaps. For example, using a Poisson distribution with $\sigma = 180$ to simulate the insert distance distribution for reads with length 100 bp shows that 17.2% of read pairs are with insert distances from 190 to 200, i.e. the overlapping regions of these pairs are of length <10. These read pairs could not be connected if a minimum overlap of 10 bp is required.

Second, repetitive or low complexity patterns at 3′ end could significantly diversify the overlaps.

Third, base disconcordance exists in the overlapping region. For example, FLASH selects a base by comparing the two base quality values. Yet if the quality values are equal, the selection will be arbitrary, which makes the connected reads prone to substitution errors.

COPE was developed to tackle these three problems. COPE takes paired FASTQ files as input. COPE first uses an alignment-based connection algorithm similar to previous tools to connect those read pairs with relatively long overlap and few errors and then utilizes $k$-mer frequency and ‘auxiliary’ reads to further connect the remaining read pairs (Fig. 1). It is worth mentioning that the latter step no longer requires an overlap to be at least a certain minimum length. Details are as follows.

### 2.1 Alignment-based connection algorithm

The foremost task in connecting pair-end reads is to obtain an authentic overlap and to select the right bases in the overlap region. We align both reads at their 3′ ends to check all possible overlaps with length longer than $k$, where $k$ *is* the length of a $k$-mer. Those with overlap shorter than $k$ will be examined in the next stage. Due to the fact that indels would hardly occur in the Illumina platform (Nakamura *et al.*, 2011), no gap was allowed in COPE during alignment. For each read pair, we calculate the matching rate for each possible overlap using formula (1), where $l_{\text{overlap}}$ is the length of the overlapping region and $n_{\text{mismatch}}$ is the number of nucleotides in the overlapping region that do not agree

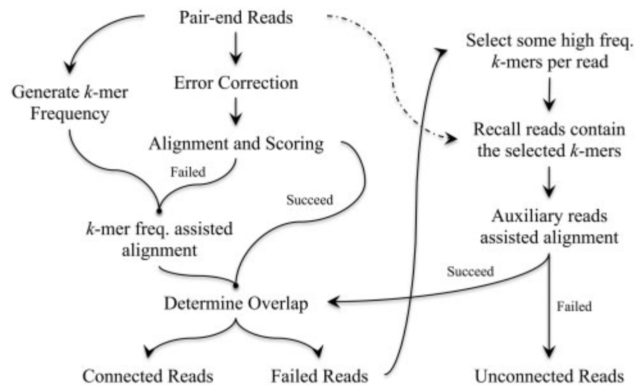

**Fig. 1.** The workflow of COPE. Error correction without trimming is suggested but optional before connection

$$\text{Rate}\,(l_{\text{overlap}}) = (l_{\text{overlap}} - n_{\text{mismatch}})/l_{\text{overlap}} \qquad (1)$$

If the matching rate is larger than a threshold, we mark the overlap as a candidate and it will be assigned a ‘combination score’ computed as follows. Consider a mismatch position in the overlapped region; assume that the two reads have base $b_0$ and $b_1$, with error probabilities $e_0$ and $e_1$ (inferred from the quality values of the bases), respectively. If we choose $b_0$, which means that $b_0$ is authentic and $b_1$ is a sequencing error, then the probability for this base selection to be right is ‘1 – $e_0$’. For an overlap with multiple mismatches, we multiply the probability for each mismatch based on the base selection. The combination score of the overlap is defined as the highest possible value for this multiplied probability.

When there is no mismatch in the overlapping region, the combination score is set to 1 (the highest possible score). The score will decrease with an increasing number of mismatches and selected bases with high error probabilities. If more than one combination has the same score, then we name each pair-end read a ‘dilemma read’ and put it aside for next stage.

Unlike previous tools, COPE is more stringent in considering the alignment scores so as to obtain higher accuracy. We define the overlap with the highest score as optimal, and the one with the second highest score as sub-optimal (Fig. 2). It is required that the score difference between the optimal and sub-optimal overlaps should be higher than a predefined threshold set by user. COPE takes advantage of the $k$-mer assisted stage to re-consider those reads that fail this separation criterion. This largely avoids inauthentic connection of reads with polyrun or tandem repeats at 3′ end at this stage.

### 2.2 $k$-Mer frequency assisted connection

A $k$-mer frequency table is constructed for all reads and is used to handle read pairs that fail to be connected in the previous stage. The basic idea is as follows. Consider an overlap with a wrong base being chosen, the $k$-mers spanning the base are inauthentic and usually of very low frequency. Thus, we can generalize the combination score of an overlap using $k$-mer frequency information by modifying the probability of base selection at a certain mismatch position from (1–$e$) to $\eta$(1–$e$), where $\eta$ is the percentage of all possible $k$-mers spanning the mismatch with high frequency. Theoretically, $\eta$ will be large if an authentic base has been selected, since the $k$-mers spanning the base also exist in other reads.

The threshold for classifying $k$-mers to be high frequency is determined automatically from the frequency table using the standard technique derived from genome assembly (Li *et al.*, 2010). Generally, the new combination score, which takes $k$-mer frequency information into consideration, can resolve the majority of the dilemma reads from the previous stage. More importantly, we find that the frequency information of spanning $k$-mers is very effective in validating short overlaps (down to length 3 in the current implementation). This is a major improvement because

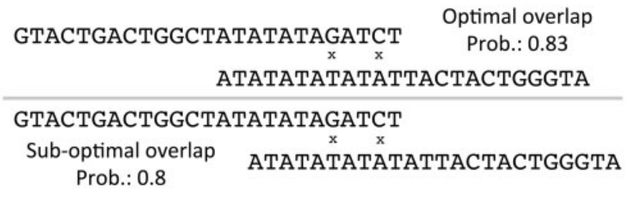

**Fig. 2.** An illustration of near optimal and sub-optimal probability caused by tandem repeats. The sub-optimal should be the right answer. But due to two bases longer overlapped region in optimal, its probability is 0.03 higher than the sub-optimal

read pairs that cannot be connected due to too short overlap can account to over 15% in a typical PE100 library with a 180-bp insert distance.

Note that checking the frequency information of all spanning *k*-mers is time-consuming. COPE uses a trick to speed up this stage by avoiding examining all spanning *k*-mers. The idea is that a spurious overlap usually has some *k*-mers with frequency as low as 1 and seldom exceeding 3. Thus, once we have seen a few spanning *k*-mers with frequency lower than the threshold of 3, we stop checking other spanning *k-m*ers and immediately conclude that the overlap is spurious. In addition, since *k*-mers from repetitive sequences are with extremely high frequency (say over two times the average), if there are more than three spanning *k*-mers of extremely high frequency, the read pair will be put aside for next stage.

Most of the dilemma reads will be resolved in this stage, but a small portion still remains unconnected. This may be due to the fact that these reads contain some polyrun, tandem repeats or low complexity sequence near the 3′ ends that are longer than the *k*-mer length. Notably, increasing the *k*-mer length would not necessarily improve the number of reads connected, but will drastically increase the computational resources. Thus, we need another way to tackle those reads.

### 2.3 Auxiliary reads and cross connection

In the final stage, COPE uses other reads to handle the read pairs unconnected in previous stages. We call these reads auxiliary reads, which are defined as follows.

First, one or more *k*-mers with normal frequency are extracted from both reads in an unconnected read pair. Then all the reads containing these *k*-mers are recalled. These reads are the auxiliary reads for validating a possible overlap of the read pair in concern. More specifically, we require that all recalled reads to be concordant with the overlap, or the read pair will remain unconnected. Mismatches between the ends of both reads were solved as in the previous stage.

Lastly, for the overlapping region, we assign quality values to the bases as follow. For a mismatch, we assign the base quality of the selected allele to the nucleotide. For a match, we assign the smaller quality value to the nucleotide.

## 3 RESULTS

We tested the efficacy of COPE using both simulated and real data and illustrated our advantages by comparing to FLASH (Magoc and Salzberg, 2011), one of the most recently published utility to join pair-end reads, for correctness and sensitivity. We also demonstrated that COPE outperforms FLASH in terms of both length and base accuracy using the dataset published with FLASH.

### 3.1 Connection of simulated data

We simulated 30-fold 100 bp pair-end reads with an insert distance of 180 bp and a standard deviation of 9 bp from *Arabidopsis thaliana*. The imbalanced error distribution profile along the Illumina reads was derived by pIRS (Hu *et al*., 2012) from a set of human genome reads (Wang *et al*., 2008) sequenced by Illumina HiSeq 2000 (downloaded from http://yh.genomics.org.cn). In the simulated data, 0.9% of pair-end reads had an actual insert distance $\geq$200 and were unable to be connected, whereas 11% had a 3′ end overlapped with <10 bp. The simulated data are available to the public together with the source code.

COPE was designed to fully utilize all sequenced pair-end reads with an overlap, no matter whether the overlapped region is only 1 bp or as long as a read. Thus, we do not need to set a minimum overlap length. In contrast, FLASH requires a minimum overlap of 10 bp. Both COPE and FLASH were set to allow at most 25% of bases in the overlapping region to be mismatches. The results for COPE and FLASH for simulated reads with error rates ranging from 0 to 3% are shown in Table 1. Notably, in the previous study, FLASH considered a connection as correct as long as the overlap is correct, even if the wrong base is selected as the consensus base at one of the overlapping positions (Magoc and Salzberg, 2011). We continue to use this criterion as ‘Correctly Overlapped’ in Table 1. However, due to the fact that a single base error will invalidate all *k*-mers spanning the bases (as referred to in the ‘Introduction’ to this article), and that this will in turn cut down the effective data depth and increase the computational resources consumed drastically in assembly, we exploit a more stringent criterion called ‘Authentic Connection’ in Table 1, which requires not only a correct overlap but also that all bases in the overlapped region should be selected correctly.

**Table 1.** Results of COPE and FLASH on simulated reads

| Program | Error rate (%) | Connected (%) | Correctly overlapped (%) | Authentic connection (%) |
|---|---|---|---|---|
| FLASH | 0.0 | 89.12 | 99.74 | – |
| | 0.5 | 88.89 | 99.74 | 98.14 |
| | 1.0 | 88.10 | 99.74 | 96.57 |
| | 3.0 | 79.55 | 99.67 | 85.55 |
| COPE | 0.0 | 99.73 | 99.98 | – |
| | 0.5 | 99.47 | 99.95 | 99.54 |
| | 1.0 | 99.13 | 99.95 | 98.75 |
| | 3.0 | 92.56 | 99.92 | 93.24 |

With the ability to connect pair-end reads with an overlap <10 bp, COPE outperforms FLASH with at least 10% more reads connected when the error rate is <1%, which signifies the state-of-the-art sequencing quality. In connected reads, rate of reads that are correctly overlapped is very high for both COPE and FLASH, with COPE averaging 0.2% higher than FLASH. When considering the correctness of bases in the overlapped region, COPE was better than FLASH with ~1.4–2.2% higher correctness on low error rates. Notably, with a relatively higher error rate (3%), COPE can still maintain 93.2% correctness of connected reads, which is ~8% higher than FLASH.

Further investigation into the error connected reads by COPE shows that there are two possible situations where COPE cannot reach the right answer: (1) where both reads from a read pair are perfect polyrun or tandem repeats and (2) where both bases at a position in an overlapped region are simultaneously wrong due to a high error rate.

### 3.2 Connection of real data

To make a direct comparison with FLASH as to how authentically connected reads affect genome assembly, we used short-read data from *Staphylococcus aureus*, a bacteria genome, which was sequenced by the Broad Institute and used by FLASH for evaluation. The data are available at http://gage.cbcb.umd.edu/data. They comprise 45-fold 101 bp error corrected [using Quake

Table 2. Assembly result using reads of *S.aureus*

| Program | Contig N50 (bp) | Genome covered (%) | Mismatch (bp) | Mismatch rate | Indel (bp) | Indel rate |
|---|---|---|---|---|---|---|
| Original | 809 | 96.3 | 628 | $2.18 \times 10^{-04}$ | 50 | $1.74 \times 10^{-05}$ |
| FLASH | 5171 | 95.8 | 361 | $1.29 \times 10^{-04}$ | 105 | $3.75 \times 10^{-05}$ |
| COPE | 11 257 | 95.8 | 139 | $5.00 \times 10^{-05}$ | 30 | $1.08 \times 10^{-05}$ |

(Kelley *et al.*, 2010)] paired-end reads with 180 bp insert distance (SRA no.: SRX07714) plus 45× 37 bp mate-pair reads with 3.5 kb insert distance (SRX007111). Note that COPE is able to determine the authentic base in an overlapped region using *k*-mer frequency when encountering a sequencing error, error correction will not significantly improve the connection quality of COPE, but it would be essential to FLASH.

COPE successfully connected 79.2% of read pairs from the 180 bp insert distance library, which is over a quarter higher than FLASH (52.6%). Unconnected reads were used as normal pair-end reads with 180 bp insert distance for assembly. Original reads, reads connected by FLASH and COPE were assembled by SOAPdenovo, respectively, using the same parameters with *k*-mer length of 31 bp.

As shown in Table 2, using COPE-connected reads, the contig N50 is over two times longer than FLASH and two orders of magnitude longer than original reads. The number of mismatches and indels using COPE reads is lower than those of FLASH and Original, this may be due to the reason that FLASH has inauthentically connected some reads and in turn introduced errors to the assembly.

To evaluate the mismatch rate of the assembly, we used LAST (Kielbasa *et al.*, 2011) to compare all contigs to the reference genome. Noteworthily, with a much higher base accuracy at overlapped regions, the mismatch rate of COPE is ~2.6 times lower than FLASH. On the other hand, the mismatch rate of FLASH is about twice that of using the original reads, this can be explained by its arbitrary selection of discordinate bases in overlapped regions.

## 3.3 Selecting larger *k*-mer for genome assembly

The quality of genome assembled by *de Bruijn* graph-based assemblers largely depends on the length of *k*-mer used. The longer the *k*-mer used, the greater the number of repetitive sequences that can be resolved. However, with a constant read length, using a longer *k*-mer will drastically decrease the effective depth of *k*-mer. The depth of *k*-mer is calculated using formula (2).

$$d_{\text{kmer}} = d_{\text{base}} \times (L - K + 1)/L \qquad (2)$$

Note: $d$ is depth, $L$ and $K$ is the length of read and *k*-mer, respectively.

To obtain a reasonable genome assembly, higher *k*-mer depth is recommended to avoid insufficient transition between adjacent *k*-mers (Li *et al.*, 2012), a depth <20-fold may result in insufficient transition between adjacent *k*-mers, especially when the sequencing error is high. The relationship between *k*-mer depth, *k*-mer length and read length is shown in Table 3.

Table 3. With a 30-fold constant read depth, the *k*-mer depth under different read lengths and *k*-mer lengths is shown

| Read length | *k*-Mer length | | | |
|---|---|---|---|---|
| | 31 | 51 | 71 | 91 |
| 100 | 21 | 15 | 9 | 3 |
| 180 | 25 | 22 | 18 | 15 |

Table 4. Assembling 30-fold error free reads of *A.thaliana* using different read lengths and *k*-mer with SOAPdenovo

| | *k*-Mer length | Number of contig (lower the better) | Contig N50 (larger the better) | Contig N90 (larger the better) |
|---|---|---|---|---|
| Original | 47 | 50 663 | 26 390 | 1647 |
| COPE | 47 | 46 230 | 29 441 | 2117 |
| Original | 63 | 48 307 | 29 153 | 1943 |
| COPE | 63 | 40 691 | 38 414 | 3038 |
| Original | 81 | 118 919 | 1782 | 484 |
| COPE | 81 | 32 242 | 56 193 | 4878 |

As shown in Table 3, when using 91-mer, 100 bp pair-end reads, they only provide a 3-fold effective *k*-mer while 180 bp connected reads can still provide 15-fold. To test the performance of different length of reads and *k*-mers, we assembled 30-fold error free pair-end reads from *A.thaliana* with a *k*-mer length range from 47 to 81, using both original and overlapped reads. The results are shown in Table 4.

COPE-connected reads show its potential to increase Contig N50 and N90 with both short and long *k*-mer. When using 81-mer, the Contig N50 using original reads significantly dropped to an unacceptable level due to the effective *k*-mer depth being only 6-fold in this case. However, the improvement is tremendous for overlapped reads when using 81-mer, which is almost twice the contig N50 length of 47-mer.

## 3.4 Time requirements

COPE runs in two steps. The first step is building a *k*-mer frequency table. This step is multi-threaded and requires 16 GB memory constantly using *k*-mer size 17. The second step is connecting reads. This step can be parallelized in per lane granularity without extra memory consumption. We ran both COPE and FLASH with 30× 100 bp pair-end reads from *A.thaliana* on a quad-core Intel i7 3.06 GHz desktop computer with 24 GB of memory. Run times are summarized in Table 5.

As shown in Table 5, the running time of COPE is about double that of FLASH, but still acceptable. The longer running time is due to the computation required to determine the connection based on *k*-mer frequency information. The memory

**Table 5.** Time consumption on 30× *A.thaliana* reads with FLASH and COPE. To avoid expensive I/O operations, we make use of a SSD with an average speed of 550 MB/s for read/ write operations to store and access the files

| Program (Quad-Core) | Real time (s) | Max memory (M) |
|---|---|---|
| FLASH | 374 | 3.44 |
| COPE align | 237 | 3.44 |
| COPE Freq. | | |
| Freq. loading | 157 | 4109 |
| Connect | 426 | |
| COPE total | 820 | 4109 |

consumption of COPE mainly depends on the *k*-mer frequency table and is independent of the depth of sequencing. We set *k* to 17, so it requires ~4 GB memory and the memory consumption is similar even if we increase the depth of sequencing.

The overall running time for COPE is linearly proportional to the read length multiplied by the number of reads. Bad sequencing quality may also increase the running time proportionally.

## 4 CONCLUSION AND DISCUSSION

In order to obtain a more comprehensive genome assembly, the bioinformatics community is always willing to adopt longer reads without increasing sequencing costs. Several attempts have been made to make use of the new type of library to generate longer reads but older methods have failed to maintain correctness and sensitivity simultaneously. To tackle the problem, we developed COPE, an accurate tool to make the maximum use of overlapped pair-end reads. As shown in our experiments, highly accurate overlapped reads can drastically improve the quality of genome assembly.

Besides assembly, overlapped reads can also be used for resequencing. Longer reads will enable the discovery of larger indels using the split-read method and also facilitate single-nucleotide polymorphism discovery in mutation concentrated regions like rearrangement hotspots. The use of longer overlapped reads in RNA-seq experiments will allow more precise splicing junctions and expression levels to be defined.

Due to the limitation of read length, COPE may fail to deal with read pairs from repetitive genome patterns longer than the insert size if the overlapping region is filled with tandem repeats or excessive amount of errors. In the future, we may incorporate longer reads from other sequencing platform (for example, Roche 454 FLX+) to facilitate the connection.

## ACKNOWLEDGEMENTS

The authors thank the reviewers for their useful feedback and constructive comments.

*Funding*: T.-W.L. and R.L.: partially supported by the Hong Kong GRF grant HKU-713512E. The project was also supported by the State Key Development Program for Basic Research of China-973 Program [2011CB809203]; National High Technology Research and Development Program of China-863 program [2012AA02A201]; the National Natural Science Foundation of China [90612019]; the Shenzhen Key Laboratory of Trans-omics Biotechnologies [CXB201108250096A]; and the Shenzhen Municipal Government of China [JC201005260191A and CXB201108250096A].

*Conflict of Interest*: none declared.

## REFERENCES

Glenn,T.C. (2011) Field guide to next-generation DNA sequencers. *Mol. Ecol. Resour*., **11**, 759–769.

Gnerre,S. *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl Acad. Sci. USA*, **108**, 1513–1518.

Hu,X. *et al.* (2012) pIRS: profile-based Illumina pair-end reads simulator. *Bioinformatics*, **28**, 1533–1535.

Kelley,D.R. *et al.* (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol*., **11**, R116.

Kielbasa,S.M. *et al.* (2011) Adaptive seeds tame genomic sequence comparison. *Genome Res*., **21**, 487–493.

Li,R. *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*., **20**, 265–272.

Li,Z. *et al.* (2012) Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct. Genomics*, **11**, 25–37.

Magoc,T. and Salzberg,S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, **27**, 2957–2963.

Martin,J.A. and Wang,Z. (2011) Next-generation transcriptome assembly. *Nat. Rev. Genet*., **12**, 671–682.

Masella,A.P. *et al.* (2012) PANDAseq: PAired-eND assembler for Illumina sequences. *BMC Bioinformatics*, **13**, 31.

Nakamura,K. *et al.* (2011) Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*., **39**, e90.

Rodrigue,S. *et al.* (2010) Unlocking short read sequencing for metagenomics. *PLoS One*, **5**, e11840.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.