DESCRIPTION

TITLE OF INVENTION: [METHOD FOR GENERATING EXTENDED SEQUENCE READS]

[0001] CROSS-REFERENCE TO RELATED APPLICATION

[0002] This application claims benefit of, and priority from, Singapore patent application No. 201302940-0, filed on 17 April 2013, the contents of which are hereby incorporated herein by reference. This application claims benefit of, the earlier filing date of PCT application No. PCT/SG2014/000172 entered national phase in Singapore as patent application No. 11201508193T, and divisional Singapore patent application No. 10201708498V the contents of all of which are hereby incorporated herein by reference.

[0003] FIELD OF THE INVENTION

[0004] The present invention relates to a method for generating extended sequence reads and in particular, but not exclusively, to a method for generating extended sequence reads of large genomes.

[0005] BACKGROUND

[0006] The current maximum read length of next-generation sequencing technologies, such as those developed by Illumina® and Life Technologies™, is around 250 bases. The read length is one of the main factors that determine the quality of a genome assembly. In general, longer reads make better assemblies because they span more repeats.

[0007] Furthermore, increasing the read length of next-generation sequencing technologies enables broader applications such as being able to sequence larger genomes, generate extended sequence reads and be useful for long-range haplotype analysis on targeted genomic regions.

[0008] Therefore, there is a need for an approach to increase the read length of these commercially available sequencing platforms to several kilobases.

[0009] SUMMARY OF THE INVENTION

[0010] In accordance with a first aspect of the present invention, there is provided a method for generating extended sequence reads of long DNA molecules in a sample, comprising the steps of:

(i) assigning a specific barcode sequence to each template DNA molecule in a sample to obtain barcode-tagged molecules;

(ii) amplifying the barcode-tagged molecules;

(iii) fragmenting the amplified barcode-tagged molecules to obtain barcode-containing fragments;

(iv) juxtaposing the barcode-containing fragments to random short segments of the original DNA template molecule during the process of generating a sequencing library to obtain demultiplexed reads; and

(v) assembling the demultiplexed reads to obtain extended sequence reads for each DNA template molecule.

[0011] Preferably, the method further comprises the step of labelling the amplified barcode-tagged molecules with biotin.

[0012] Preferably, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the step of subjecting the amplified barcode-tagged molecules to unidirectional deletion from the barcode-distal end of the barcode-tagged molecules to obtain barcode-containing fragments.

[0013] Preferably, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the steps of:

(i) creating a nick at the barcode-distal end of the amplified barcode-tagged molecules;

(ii) performing a nick translation towards the barcode-proximal end; and

(iii) treating with endonuclease the resulting molecules to generate blunt ends, to obtain barcode-containing fragments.

[0014] Preferably, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the step of performing random fragmentation by a mechanical method or an enzymatic method to obtain barcode-containing fragments.

[0015] Preferably, the barcode-containing fragments have lengths ranging from about 300 base pairs to N base pairs, wherein N equals to the length of the original DNA template molecule.

[0016] Preferably, the method further comprises the step of purifying the barcode-containing fragments using streptavidin-coated paramagnetic beads.

[0017] Preferably, wherein the step of purifying the barcode-containing fragments comprises dissociating the biotin-labelled molecules from the streptavidin-coated paramagnetic beads.

[0018] Preferably, wherein the step of purifying the barcode-containing fragments comprises dissociating the biotin-labelled molecules from the streptavidin-coated paramagnetic beads, further comprises the step of circularizing the purified barcode-containing fragments by intramolecular ligation.

[0019] Preferably, the method further comprises the step of ligating sequencing adaptors onto the ends of the barcode-containing fragments prior to the step of juxtaposing the barcode-containing fragments to random short segments of the original DNA template molecule.

[0020] Preferably, the specific barcode sequence is assigned by linker ligation to each template DNA molecule.

[0021] Preferably, wherein the step of amplifying the barcode-tagged molecules is by circularizing the barcode-tagged molecules and performing rolling circle amplification.

[0022] Preferably, the extended sequences reads are compatible for sequencing on sequencing platforms.

[0023] In accordance with a second aspect of the present invention, there is provided a system for obtaining extended sequence reads from template molecules of a DNA sequence, comprising:

(i) a quality filtering module for filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences;

(ii) a barcode analysis module for identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;

(iii) a demultiplexing module for using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and

(iv) an assembly module for assembling demultiplexed reads to obtain extended sequence reads for each template molecule.

[0024] The system in accordance to the second aspect of the present invention wherein the DNA sequence is a known sequence, further comprising:

(i) a sequence alignment module for performing paired-end alignment to a reference sequence and removing disconcordant alignments;

(ii) a demultiplexing module for using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs in place of the demultiplexing module according to the second aspect of the present invention; and

(iii) a haplotyping module for obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule in place of the assembly module according to the second aspect of the present invention.

[0025] In accordance with a third aspect of the present invention, there is provided a computer-readable medium with an executable programme stored thereon, the programme comprising instructions for obtaining extended sequence reads from template molecules of a DNA sequence, wherein the programme instructs a microprocessor to perform the following steps:

(i) filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences;

(ii) identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;

(iii) using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and

(iv) assembling demultiplexed reads to obtain extended sequence reads for each template molecule.

[0026] The computer-readable medium in accordance to the third aspect of the present invention wherein the DNA sequence is a known sequence, wherein the programme instructs the microprocessor to further perform the following steps:

(i) performing paired-end alignment to a reference sequence and removing disconcordant alignments at the step of identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;

(ii) replacing the step of using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs with the step of using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs; and

(iii) replacing the step of assembling demultiplexed reads with the step of obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule.

[0027] The present invention provides an approach that can: 1) increase the effective read length of these commercially available sequencing platforms to several kilobases and 2) be broadly applied to obtain long sequence reads from mixed template populations.

[0028] The present invention applies the concept of barcoding to generate long sequence reads by providing a technical advance in juxtaposing the assigned barcode to

random overlapping segments of the original template.

[0029] The present invention relies on assigning barcodes to individual template molecules, allowing for unambiguous assembly of template sequences even for molecules with high sequence similarity. This also means that the present invention will work for sequencing targeted genomic regions or viral genomes.

[0030] Accordingly, aspect(s) of the present invention also includes:-

a) A method to assign unique DNA barcodes, i.e., a random string of DNA nucleotides, to individual long (>3 kilo bases (kb)) template molecules.

b) A method to juxtapose the assigned DNA barcode to random short segments of the original template molecule during the process of generating a sequencing library.

c) A method of using the DNA barcode associated with each molecule of a sequencing library to identify the template of origin.

d) A method of using DNA barcodes to substantially reduce the error rate of massively parallel sequencing.

e) A method for barcode-directed assembly of short sequence reads to obtain individual template sequences.

[0031] Other aspects and advantages of the present invention will become apparent to those skilled in the art from a review of the ensuing description, which proceeds with reference to the following illustrative drawings of preferred embodiments.

[0032] BRIEF DESCRIPTION OF THE DRAWINGS

[0033] Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

[0034] **Figure 1: Barcode assignment and mutation identification**. (A) Each template molecule is assigned to a random barcode, which will uniquely identify each template molecule from a heterogeneous sample. Barcodes also perform another function of removing the vast majority of errors introduced by next-generation sequencing. Briefly,

barcode assignment primers carrying universal sequences (green rectangle; Uni-A or Uni-B) on their 5'-end will anneal to opposite ends of template. One of the primers will also carry a barcode (blue rectangle). After two rounds of linear amplification, each template molecule will be tagged with a unique barcode. Next, barcode assignment primers are removed by exonuclease digestion, followed by up to 40 cycles of PCR using universal primers to generate whole-genome amplicons. After PCR amplification, mutations that pre-existed on the template (red cross) or errors introduced during barcode assignment cycle #1 or #2 (blue cross) will be found in all molecules associated with a particular barcode (hereafter referred to as "daughter" molecules). In contrast, errors introduced in subsequent steps of library preparation, sequencing, or base-calling (green cross) can be easily identified because they will only be present in a minority of daughter molecules. (B) To generate uniform sequence coverage across each template molecule, each barcode-tagged genome will be subjected to a series of molecular biologic steps that will ultimately ligate sequencing adaptors (red rectangles) onto the ends of overlapping fragments of the genome. The resulting sequence reads will be assigned to the template molecule identified by the barcode.

[0035] Figure 2: Methodology to generate library of overlapping fragments tagged with the same barcode. After PCR amplification with universal primers, the PCR amplicons are subjected to unidirectional deletion from the barcode-distal end to achieve a broad size distribution of fragments ranging from ~300 bp to N bp, where N equals the length of the original template molecule. Barcode-containing fragments are purified using streptavidin-coated paramagnetic beads. Next, the biotinylated DNA is dissociated from the bead/DNA complex and circularized by intramolecular ligation. After circularization, different regions of the template will be juxtaposed to its barcode. The double-stranded DNA circles are subjected to library preparation by adapting a commonly used transposome-based method (Illumina®) to generate molecules that are compatible for sequencing on the Illumina® platform.

[0036] Figure 3: Software package for obtaining extra-long sequence reads by reference-assisted assembly.

[0037] Figure 4: Software package for obtaining extra-long sequence reads from template molecules of unknown sequence.

[0038] **Figure 5: Average coverage depth per HBV genome.** Average coverage depth per base position across the HBV genome. Data shown is the average coverage from 4,294 unique viral genomes.

[0039] **Figure 6: Genome coverage per HBV genome**. Percentage of each genome covered by at least 5 unique reads. 2,717 high coverage genomes—defined as ≥5 reads across ≥85% of the genome—were recovered.

[0040] **Figure 7: Allele frequency of mixed clones.** Two HBV clones with 17 known SNPs between them were mixed at (A) 1:99 or (B) 1:9 ratios prior to barcoding. BAsE-Seq was performed for each mixed-clone pool. Barcodes were removed from each read-pair prior to alignment and the resulting data was treated as a "bulk" sequencing experiment to determine overall allele frequencies at the SNP positions. The allele frequencies in both libraries were very close to the mixing ratio (indicating that pipetting error and PCR bias were negligible), and were used as control libraries to test the sensitivity and accuracy of our methodology.

[0041] **Figure 8: Predicted SNV frequencies and background error rate.** Single nucleotide variants (SNVs) in Lib_1:99 (control library containing minor haplotype present at 1% frequency) were estimated by calculating the frequency of Clone-1 genotype calls present in 2881 high coverage genome sequences. In this control library containing a minor haplotype present at 1% frequency, true-positive SNVs were significantly separated from background error ($p < 0.0001$).

[0042] **DETAILED DESCRIPTION**

[0043] The present invention applies the concept of barcoding to generate long sequence reads by providing a technical advance in juxtaposing the assigned barcode to random overlapping segments of the original template.

[0044] The present invention relies on assigning barcodes to individual template molecules, allowing for unambiguous assembly of template sequences even for molecules with high sequence similarity. This also means that the present invention will work for sequencing targeted genomic regions or viral genomes.

[0045] The current maximum read length of next-generation sequencing technologies,

such as those developed by Illumina® and Life Technologies™, is around 250 bases. The present invention, also known as "Barcode-directed Assembly for Extra-long Sequences (BAsE-Seq)" provides an approach that can: 1) increase the effective read length of these commercially available sequencing platforms to several kilobases and 2) be broadly applied to obtain long sequence reads from mixed template populations. In brief, our method relies on assigning random DNA barcodes to long template molecules (Figure 1A), followed by a library preparation protocol that juxtaposes the assigned barcode to random short segments of the original template (Figures 1B and 2). The resulting molecules are ligated with platform-specific adaptors for next-generation sequencing. Sequence reads are de-multiplexed using the barcode sequence and used to assemble long-range haplotypes that were present on the original template. In practice, we have applied this technology to perform single virion sequencing on the Hepatitis B virus, a DNA virus with a 3.2 kb genome. In general, we anticipate that this technology can be broadly applied to generate extended sequence reads and will be useful for long-range haplotype analysis on targeted genomic regions, or for improving *de novo* genome and transcriptome assemblies. A detailed description of our protocol is described in the following paragraphs.

[0046] There is described a method for generating extended sequence reads of long DNA molecules (>3 kb), in a sample. The method comprises the steps of: (i) assigning a specific barcode sequence to each template DNA molecule in a sample to obtain barcode-tagged molecules; (ii) amplifying the barcode-tagged molecules; (iii) fragmenting the amplified barcode-tagged molecules to obtain barcode-containing fragments; (iv) juxtaposing the barcode-containing fragments to random short segments of the original DNA template molecule during the process of generating a sequencing library to obtain demultiplexed reads; and (v) assembling the demultiplexed reads to obtain extended sequence reads for each DNA template molecule.

[0047] **a) Barcode Assignment.** In the first step, individual template molecules are assigned with a unique DNA barcode. In our example, two rounds of PCR amplification are performed using primers with template-specific sequence from opposite ends of the molecule (Figure 1A). This will generate uniquely tagged template molecules for preparing libraries and can be broadly applied for assigning barcodes to targeted genomic regions. Both primers contain a universal sequence on their 5'-ends and one of them contains a barcode, i.e., a string of 20 random nucleotides (encodes for $>10^{12}$ sequences).

To ensure that each template molecule is uniquely assigned, the template should be diluted to obtain a relatively small number of genomes ($<10^9$) compared to unique barcode sequences.

[0048] Subsequently, barcode-tagged molecules can be clonally amplified by PCR using universal primers and the PCR product can be used to prepare sequencing libraries. In other manifestations where the template sequence is unknown, the barcode can be assigned by ligation of double- or single-stranded DNA linkers carrying a random string of nucleotides flanked by universal sequences.

[0049] The use of unique barcodes to tag individual template molecules has been shown to greatly reduce the error rate of massively parallel sequencing. Using this strategy, mutations that pre-existed on the template and errors introduced during barcode assignment will be found in all daughter molecules.

[0050] In contrast, errors introduced in subsequent steps of library preparation, sequencing, or base-calling can be easily removed because they will only be present in a minority of daughter molecules (Figure 1A). Based on the published error rate of the DNA polymerase used in our protocol, this translates to one error in every 50 template sequences for template molecules that are 3 kb in size. Furthermore, by using barcodes as unique identifiers for individual genomes, sequences associated with each barcode can be assembled into a complete template sequence.

[0051] **b) Library Preparation.** The goal of library preparation is to tag overlapping fragments of each template molecule with its assigned barcode in order to obtain uniform sequence coverage. This concept is illustrated in Figure 1B and a detailed outline of the protocol is shown in Figure 2.

[0052] Firstly, clonally amplified barcode-tagged molecules are deleted from the barcode-distal end to achieve a broad size distribution of fragments ranging from ~300 bp to N bp, where N equals the length of the template molecule. Unidirectional deletion can be achieved by protecting the barcode-proximal end with nuclease-resistant nucleotides or a 3'-protruding overhang, and performing time-dependent digestion from the barcode-distal end using a 3' to 5' exonuclease (such as Exonuclease III), followed by treatment with an endonuclease (such as S1 Nuclease or Mung Bean Nuclease) to generate blunt-

ends.

[0053] Barcode-containing fragments are purified using streptavidin-coated beads, and these biotinylated fragments will be dissociated and subjected to end repair, such that both ends of the molecules are blunt and 5'-phosphorylated. The end-repaired molecules are circularized by intramolecular ligation using a DNA ligase (such as T4 DNA ligase). Uncircularized molecules will be removed by nuclease treatment (such as a combination of Exonuclease I and Lambda Exonuclease).

[0054] After circularization, different regions from the original template will be juxtaposed to its barcode. The circularized molecules will be used as template for random fragmentation and adaptor tagging using a transposome-based method, such as the Nextera XT kit (Illumina®).

[0055] Importantly, the primers used during PCR enrichment of the sequencing library will be designed such that the second sequencing read will be anchored by the barcode sequence. Thus, this PCR generates double-stranded DNA molecules that are "sequencing-ready". Finally, the PCR products are subjected to size selection before sequencing. A custom sequencing primer that anneals to the forward priming sequence is used for the second sequencing read.

[0056] There are several alternative approaches to generate a broad distribution of barcode-tagged molecules before circularization. One approach involves creating a nick at the barcode-distal end using a nicking endonuclease, nick translation towards the barcode-proximal end using DNA polymerase I, followed by treatment with endonuclease to generate a blunt end. Another approach involves performing random fragmentation using a mechanical method, such as using the Covaris instrument for focused-ultrasonication, or an enzymatic method, such as using the NEBNext dsDNA Fragmentase, followed by purification of barcode-containing fragments using streptavidin-coated paramagnetic beads.

[0057] An alternative, PCR-free approach to clonal amplification is contemplated, such as circularizing the barcoded template and performing rolling circle amplification using phi29 polymerase.

[0058] Barcodes can be assigned by linker ligation. Both linkers will contain universal sequences on their 5'-end to facilitate clonal amplification in the next step. The barcode-containing linker will also contain a unique universal sequence on its 3'-end for primer annealing during the PCR step at the end of the protocol.

[0059] Software packages for obtaining extended or extra-long sequence reads by reference-assisted assembly and for obtaining extended or extra-long sequence reads from template molecules of an unknown sequence are illustrated in Figures 3 and 4, respectively.

[0060] There is described hereinafter a system for obtaining extended sequence reads from template molecules of a DNA sequence. The system comprises (i) a quality filtering module for filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences; (ii) a barcode analysis module for identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes; (iii) a demultiplexing module for using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and (iv) an assembly module for assembling demultiplexed reads to obtain extended sequence reads for each template molecule (Figure 4). The template molecules are long, preferably >3 kb.

[0061] Where the DNA sequence is a known sequence, the system further comprises (i) a sequence alignment module for performing paired-end alignment to a reference sequence and removing disconcordant alignments; (ii) a demultiplexing module for using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs in place of the demultiplexing module shown in Figure 4; and (iii) a haplotyping module for obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule in place of the assembly module shown in Figure 4 (Figure 3).

[0062] There is also disclosed a computer-readable medium with an executable programme stored thereon, the programme comprising instructions for obtaining extended sequence reads from template molecules of a DNA sequence, wherein the programme

instructs a microprocessor to perform the following steps of (i) filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences; (ii) identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes; (iii) using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and (iv) assembling demultiplexed reads to obtain extended sequence reads for each template molecule.

[0063] Where the DNA sequence is a known sequence, the programme instructs the microprocessor to further perform the following steps of (i) performing paired-end alignment to a reference sequence and removing disconcordant alignments at the step of identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes; (ii) replacing the step of using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs described above with the step of using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs; and (iii) replacing the step of assembling demultiplexed reads described above with the step of obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule.

[0064] EXAMPLES

[0065] Hepatitis B virus (HBV), which contains a 3.2 kb dsDNA genome, was used as a template for methodology development and generating proof-of-concept data. The results presented below demonstrate the use of BAsE-Seq to obtain long (~3.2 kb) sequence reads from individual template molecules, thereby achieving single virion sequencing of HBV.

[0066] HBV DNA was isolated from a chronically infected patient, PCR-amplified to obtain full-length viral genomes, and cloned into a TOPO pCR2.1 vector (Life Technologies™). Sanger sequencing was performed across each clone to obtain full-length sequences, and two clones (Clone-1 and Clone-2) with 17 single nucleotide polymorphisms (SNPs) between them were used as input for barcode assignment. In the results presented hereafter, barcode-tagged whole-genome amplicons from 20,000

template molecules (HBV genomes) were used as input for library preparation using the BAsE-Seq protocol described above.

[0067] Summary statistics from a typical single virion sequencing experiment of HBV are shown in Table 1, and coverage data per template molecule are illustrated in Figures 5 and 6. In this library, 18,143,186 read-pairs were obtained from the MiSeq sequencer (Illumina®), from which 12,004,237 read-pairs contained the barcode in the expected orientation. After trimming for adaptor, barcode tag and universal sequences, and removing reads shorter than 15 bp, 7,336,915 pass-filter read-pairs were used for alignment to a HBV reference genome. From these read-pairs, 97% read-pairs aligned concordantly, and were distributed across 4,294 individual template molecules, 2,717 of which were identified as "high coverage" and were used for constructing long reads.

Table 1. Summary statistics from a BAsE-Seq run of HBV.

	MiSeq run (2 x 150 bp)
Genomes as input	20,000
Sequencing read-pairs	18,143,186
Barcode-associated read-pairs[1]	12,004,237 (66%)
Pass-filter read-pairs[2]	7,366,915
Concordantly aligned to HBV genome	7,151,142 (97% of pass-filter)
Unique barcodes observed[3]	4,294
High coverage HBV genomes[4]	2,717

[1] Read-pairs that contain the barcode in the expected orientation

[2] Read-pairs that are ≥15 bp after removal of adaptor and universal sequences

[3] Barcodes associated with at least 50 read-pairs

[4] ≥5 unique reads per base position across ≥85% of the genome

[0068] To test the sensitivity and accuracy of our methodology in generating long sequence reads, Clone-1 and Clone-2 were mixed at different ratios to generate a mixed template population where Clone-1 is present at approximately 1% or 10% frequency in the sample. BAsE-Seq was performed on each mixed-template pool. Firstly, barcodes were removed from each read-pair prior to alignment and the resulting data was treated as a "bulk" sequencing experiment to determine overall allele frequencies at the SNP positions. The minor allele frequencies in both libraries were very close to the mixing ratio—0.98%

for the "1% pool" (Lib_1:99) and 13.44% for the "10% pool" (Lib_1:9)—indicating that the mixed template pool was generated correctly and PCR bias was negligible (Table 2 and Figure 7). Subsequently, the "bulk" sequence data was de-multiplexed using barcode sequences and sequence reads from individual template molecules were analyzed to obtain ~3.2 kb reads. Using the long sequence reads, 17-SNP haplotypes were generated for each template molecule. In Lib_1:9, 240 molecules carried a Clone-1 haplotype and 1,639 molecules carried a Clone-2 haplotype, corresponding to a 12.77% minor haplotype frequency. In Lib_1:99, 20 molecules carried a Clone-1 haplotype and 1,912 molecules carried a Clone-2 haplotype, corresponding to a 1.04% minor haplotype frequency. Importantly, chimeric sequences where Clone-1 and Clone-2 SNPs were found on the same molecule were present at ≤0.1% frequency. Furthermore, the use of barcodes to correct for sequencing errors resulted in a very low error rate for BAsE-Seq, allowing for significant separation of true sequence variants from background noise in Lib_1:99 (Table 2 and Figure 8).

Table 2. Detection of low frequency haplotypes by BAsE-SEq

	Lib_1:99	**Lib_1:9**
Mixing ratio (Clone-1 vs. Clone-2)	1:99	10:90
Expected minor clone frequency[1]	0.98%	13.44%
Observed minor clone haplotypes (Clone-1/Clone-2)[2]	1.04% (20/1912)	12.77% (240/1639)
Chimeric haplotypes	0.10% (2/1912)	0.06% (1/1639)

[1] Based on average allele frequency of Clone-1 SNPs from "bulk" sequencing analysis

[2] Based on 17-SNP haplotypes observed in the data from individual template molecules

[0069] Those skilled in the art will appreciate that the invention described herein is susceptible to variations and modifications other than those specifically described. The invention includes all such variation and modifications. The invention also includes all of the steps, features, formulations and compounds referred to or indicated in the specification, individually or collectively and any and all combinations or any two or more of the steps or features.

[0070] Each document, reference, patent application or patent cited in this text, if any, is expressly incorporated herein in their entirety by reference, which means that it should

be read and considered by the reader as part of this text. That the document, reference, patent application or patent cited in this text is not repeated in this text is merely for reasons of conciseness.

[0071] Any manufacturer's instructions, descriptions, product specifications, and product sheets for any products mentioned herein or in any document incorporated by reference herein, are hereby incorporated herein by reference, and may be employed in the practice of the invention.

[0072] The present invention is not to be limited in scope by any of the specific embodiments described herein. These embodiments are intended for the purpose of exemplification only. Functionally equivalent products, formulations and methods are clearly within the scope of the invention as described herein.

[0073] The invention described herein may include one or more range of values (e.g. size, concentration etc). A range of values will be understood to include all values within the range, including the values defining the range, and values adjacent to the range which lead to the same or substantially the same outcome as the values immediately adjacent to that value which defines the boundary to the range.

[0074] Throughout this specification, unless the context requires otherwise, the word "comprise" or variations such as "comprises" or "comprising", will be understood to imply the inclusion of a stated integer or group of integers but not the exclusion of any other integer or group of integers. It is also noted that in this disclosure and particularly in the claims and/or paragraphs, terms such as "comprises", "comprised", "comprising" and the like can have the meaning attributed to it in U.S. Patent law; e.g., they can mean "includes", "included", "including", and the like; and that terms such as "consisting essentially of" and "consists essentially of" have the meaning ascribed to them in U.S. Patent law, e.g., they allow for elements not explicitly recited, but exclude elements that are found in the prior art or that affect a basic or novel characteristic of the invention.

[0075] Other definitions for selected terms used herein may be found within the detailed description of the invention and apply throughout. Unless otherwise defined, all other scientific and technical terms used herein have the same meaning as commonly understood to one of ordinary skill in the art to which the invention belongs.

[0076] While the invention has been described with reference to specific methods and embodiments, it will be appreciated that various modifications and changes may be made without departing from the invention.

[0077] Other embodiments may be included in the following statements.

1. A method for generating extended sequence reads of long DNA molecules in a sample, comprising the steps of:
   (i) assigning a specific barcode sequence to each template DNA molecule in a sample to obtain barcode-tagged molecules;
   (ii) amplifying the barcode-tagged molecules;
   (iii) fragmenting the amplified barcode-tagged molecules to obtain barcode-containing fragments;
   (iv) juxtaposing the barcode-containing fragments to random short segments of the original DNA template molecule during the process of generating a sequencing library to obtain demultiplexed reads; and
   (v) assembling the demultiplexed reads to obtain extended sequence reads for each DNA template molecule.

2. The method according to statement 1, further comprising the step of labelling the amplified barcode-tagged molecules with biotin.

3. The method according to statement 1 or 2, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the step of subjecting the amplified barcode-tagged molecules to unidirectional deletion from the barcode-distal end of the barcode-tagged molecules to obtain barcode-containing fragments.

4. The method according to statement 1 or 2, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the steps of:
   (i) creating a nick at the barcode-distal end of the amplified barcode-tagged molecules;
   (ii) performing a nick translation towards the barcode-proximal end; and
   (iii) treating with endonuclease the resulting molecules to generate blunt ends, to obtain barcode-containing fragments.

5. The method according to statement 1 or 2, wherein the step of fragmenting the amplified barcode-tagged molecules comprises the step of performing random fragmentation by a mechanical method or an enzymatic method to obtain barcode-containing fragments.

6. The method according to any preceding statements, wherein the barcode-containing fragments have lengths ranging from about 300 base pairs to N base pairs, wherein N equals to the length of the original DNA template molecule.

7. The method according to any one of statements 2 to 6, further comprising the step of purifying the barcode-containing fragments using streptavidin-coated paramagnetic beads.

8. The method of statement 7, wherein the step of purifying the barcode-containing fragments comprises dissociating the biotin-labelled molecules from the streptavidin-coated paramagnetic beads.

9. The method of statement 8, further comprising the step of circularizing the purified barcode-containing fragments by intramolecular ligation.

10. The method of any preceding statements, further comprising the step of ligating sequencing adaptors onto the ends of the barcode-containing fragments prior to the step of juxtaposing the barcode-containing fragments to random short segments of the original DNA template molecule.

11. The method of any preceding statements, wherein the specific barcode sequence is assigned by linker ligation to each template DNA molecule.

12. The method of any preceding statements, wherein the step of amplifying the barcode-tagged molecules is by circularizing the barcode-tagged molecules and performing rolling circle amplification.

13. The method of any preceding statements, wherein the extended sequences reads are compatible for sequencing on sequencing platforms.

14. A system for obtaining extended sequence reads from template molecules of a DNA sequence, comprising:
    (i) a quality filtering module for filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences;
    (ii) a barcode analysis module for identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;
    (iii) a demultiplexing module for using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and
    (iv) an assembly module for assembling demultiplexed reads to obtain extended sequence reads for each template molecule.

15. The system of statement 14 wherein the DNA sequence is a known sequence, further comprising:
    (i) a sequence alignment module for performing paired-end alignment to a reference sequence and removing disconcordant alignments;
    (ii) a demultiplexing module for using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs in place of the demultiplexing module of claim 14; and
    (iii) a haplotyping module for obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule in place of the assembly module of claim 14.

16. A computer-readable medium with an executable programme stored thereon, the programme comprising instructions for obtaining extended sequence reads from template molecules of a DNA sequence, wherein the programme instructs a microprocessor to perform the following steps:

(i) filtering raw paired-end sequence reads from a sequencer by removing read-pairs with low quality scores, removing read-pairs with missing barcode sequences and trimming platform-specific adaptor sequences;

(ii) identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;

(iii) using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs; and

(iv) assembling demultiplexed reads to obtain extended sequence reads for each template molecule.

17. The computer-readable medium of statement 16 wherein the DNA sequence is a known sequence, wherein the programme instructs the microprocessor to further perform the following steps:

(i) performing paired-end alignment to a reference sequence and removing disconcordant alignments at the step of identifying highly-represented barcodes and re-assigning sequences associated with poorly-represented barcodes;

(ii) replacing the step of using barcode sequences as identifiers to obtain reads associated with individual template molecules and removing duplicate read-pairs with the step of using barcode sequences as identifiers to obtain alignments to individual template molecules and removing duplicate read-pairs; and

(iii) replacing the step of assembling demultiplexed reads with the step of obtaining pileup of aligned reads at each position along the reference sequence, determining consensus base-call at each position and assembling base-calls to obtain extended sequence reads for each template molecule.

CLAIMS

Claim 1. A method comprising:

(i) assigning a specific barcode sequence to template DNA molecules in a sample, wherein the assigning comprises a polymerase chain reaction (PCR) amplification reaction using primers comprising universal sequences at the 5' ends of the primers, and wherein individual primers additionally comprise barcodes, thereby generating barcode-tagged molecules;

(ii) clonally amplifying the barcode-tagged molecules, thereby generating amplified barcode-tagged molecules;

(iii) randomly fragmenting the amplified barcode-tagged molecules using a mechanical method or an enzymatic method, thereby obtaining barcode-containing fragments with barcode-tagged ends and barcode-distal ends;

(iv) circularizing the barcode-containing fragments by intramolecular ligation, thereby juxtaposing the barcode-tagged ends of the barcode-containing fragments with the barcode-distal ends of the barcode-containing fragments and generating a sequencing library of overlapping fragments, wherein the sequencing library is compatible for sequencing on a massively parallel sequencing platform with a maximum read length of around 250 basepairs;

(v) obtaining demultiplexed reads from the sequencing library, wherein demultiplexed reads from the sequencing library comprise sequences of the barcode and the barcode-distal ends of the barcode-containing fragments; and

(vi) assembling the demultiplexed reads to obtain extended sequence reads for the template DNA molecules.

Claim 2. The method of claim 1, comprising protecting barcode-proximal ends of the amplified barcode-tagged molecules, thereby generating protected barcode-tagged molecules after step (ii).

Claim 3. The method of claim 1, further comprising labelling the amplified barcode-tagged molecules with biotin.

Claim 4. The method of claim 4, further comprising purifying biotin-labelled barcode-containing fragments using streptavidin-coated paramagnetic beads.

Claim 5. The method of claim 1, wherein the barcode-containing fragments comprise between about 300 base pairs to N base pairs, wherein N equals a length of one of the template DNA molecules.

Claim 6. The method of claim 1, wherein clonally amplifying the barcode-tagged molecules comprises circularizing the barcode-tagged molecules and performing rolling circle amplification.

Claim 7. The method of claim 6, wherein the rolling circle amplification is performed using a phi29 polymerase.

Claim 8. The method of claim 1, further comprising ligating sequencing adaptors to the barcode-containing fragments prior to the circularizing.

Claim 9. The method of claim 1, wherein the template DNA molecules comprise a population of mixed template DNA molecules.

Claim 10. The method of claim 1, wherein the sequencing library comprises double-stranded DNA circularized molecules.

Claim 11. The method of claim 1, wherein the template DNA molecules have a length of at least 3 kilobases (kb).

Claim 12. The method of claim 1, wherein the circularizing is performed by intramolecular ligation using a DNA ligase.

Claim 13. The method of claim 1, wherein:

(a) the mechanical method comprises focused ultrasound; or

(b) the enzymatic method comprises a dsDNA Fragmentase.

Claim 14. A method comprising:

(i) assigning a specific barcode sequence to template DNA molecules in a sample, wherein the assigning comprises a polymerase chain reaction (PCR) amplification reaction using primers comprising universal sequences at the 5' ends of the primers, and wherein individual primers additionally comprise barcodes, thereby generating barcode-tagged molecules;

(ii) clonally amplifying the barcode-tagged molecules, thereby generating amplified barcode-tagged molecules;

(iii) fragmenting the amplified barcode-tagged molecules comprising:

(a) creating nicks at the barcode-distal ends of the amplified barcode-tagged molecules,

(b) performing a nick translation towards the barcode-proximal ends of the amplified barcode-tagged molecules, and

(c) treating the resulting molecules with an endonuclease that generates blunt ends, thereby obtaining barcode-containing fragments;

(iv) circularizing the barcode-containing fragments by intramolecular ligation, thereby juxtaposing a barcode-tagged end of the barcode-containing fragments with a barcode-distal end of the barcode-containing fragments and generating a sequencing library of overlapping fragments, wherein the sequencing library is compatible for sequencing on a massively parallel sequencing platform with a maximum read length of around 250 basepairs;

(v) obtaining demultiplexed reads from the sequencing library, wherein demultiplexed reads from the sequencing library comprise sequences of the barcode and the barcode-distal ends of the barcode-containing fragments; and

(vi) assembling the demultiplexed reads to obtain extended sequence reads for the template DNA molecules.

Claim 15. The method of claim 14, comprising protecting barcode-proximal ends of the amplified barcode-tagged molecules, thereby generating protected barcode-tagged molecules after step (ii).

Claim 16. The method of claim 14, further comprising labelling the amplified barcode- tagged molecules with biotin.

Claim 17. The method of claim 16, further comprising purifying biotin-labelled barcode-containing fragments using streptavidin-coated paramagnetic beads.

Claim 18. The method of claim 14, wherein the barcode-containing fragments comprise between about 300 base pairs to N base pairs, wherein N equals a length of one of the template DNA molecules.

Claim 19. The method of claim 14, wherein clonally amplifying the barcode-tagged molecules comprises circularizing the barcode-tagged molecules and performing rolling circle amplification.

Claim 20. The method of claim 19, wherein the rolling circle amplification is performed using a phi29 polymerase.

Claim 21. The method of claim 14, further comprising ligating sequencing adaptors to the barcode-containing fragments prior to the circularizing.

Claim 22. The method of claim 14, wherein the template DNA molecules comprise a population of mixed template DNA molecules.

Claim 23. The method of claim 14, wherein the sequencing library comprises double-stranded DNA circularized molecules.

Claim 24. The method of claim 14, wherein the template DNA molecules have a length of at least 3 kilobases (kb).

Claim 25. The method of claim 14, wherein the circularizing is performed by intramolecular ligation using a DNA ligase.