illumina®

FastTrack Services
Long Reads Pipeline User Guide

FOR RESEARCH USE ONLY

ILLUMINA PROPRIETARY
Part # 15047621 Rev. A
November 2013

This document and its contents are proprietary to Illumina, Inc. and its affiliates ("Illumina"), and are intended solely for the contractual use of its customer in connection with the use of the product(s) described herein and for no other purpose. This document and its contents shall not be used or distributed for any other purpose and/or otherwise communicated, disclosed, or reproduced in any way whatsoever without the prior written consent of Illumina. Illumina does not convey any license under its patent, trademark, copyright, or common-law rights nor similar rights of any third parties by this document.

The instructions in this document must be strictly and explicitly followed by qualified and properly trained personnel in order to ensure the proper and safe use of the product(s) described herein. All of the contents of this document must be fully read and understood prior to using such product(s).

FAILURE TO COMPLETELY READ AND EXPLICITLY FOLLOW ALL OF THE INSTRUCTIONS CONTAINED HEREIN MAY RESULT IN DAMAGE TO THE PRODUCT(S), INJURY TO PERSONS, INCLUDING TO USERS OR OTHERS, AND DAMAGE TO OTHER PROPERTY.

ILLUMINA DOES NOT ASSUME ANY LIABILITY ARISING OUT OF THE IMPROPER USE OF THE PRODUCT(S) DESCRIBED HEREIN (INCLUDING PARTS THEREOF OR SOFTWARE) OR ANY USE OF SUCH PRODUCT(S) OUTSIDE THE SCOPE OF THE EXPRESS WRITTEN LICENSES OR PERMISSIONS GRANTED BY ILLUMINA IN CONNECTION WITH CUSTOMER'S ACQUISITION OF SUCH PRODUCT(S).

Read Before Using this Product

This Product, and its use and disposition, is subject to the following terms and conditions. If Purchaser does not agree to these terms and conditions then Purchaser is not authorized by Illumina to use this Product and Purchaser must not use this Product.

1 **Definitions**. "**Application Specific IP**" means Illumina owned or controlled intellectual property rights that pertain to this Product (and use thereof) only with regard to specific field(s) or specific application(s). Application Specific IP excludes all Illumina owned or controlled intellectual property that cover aspects or features of this Product (or use thereof) that are common to this Product in all possible applications and all possible fields of use (the "**Core IP**"). Application Specific IP and Core IP are separate, non-overlapping, subsets of all Illumina owned or controlled intellectual property. By way of non-limiting example, Illumina intellectual property rights for specific diagnostic methods, for specific forensic methods, or for specific nucleic acid biomarkers, sequences, or combinations of biomarkers or sequences are examples of Application Specific IP. "**Consumable(s)**" means Illumina branded reagents and consumable items that are intended by Illumina for use with, and are to be consumed through the use of, Hardware. "**Documentation**" means Illumina's user manual for this Product, including without limitation, package inserts, and any other documentation that accompany this Product or that are referenced by the Product or in the packaging for the Product in effect on the date of shipment from Illumina. Documentation includes this document. "**Hardware**" means Illumina branded instruments, accessories or peripherals. "**Illumina**" means Illumina, Inc. or an Illumina affiliate, as applicable. "**Product**" means the product that this document accompanies (e.g., Hardware, Consumables, or Software). "**Purchaser**" is the person or entity that rightfully and legally acquires this Product from Illumina or an Illumina authorized dealer. "**Software**" means Illumina branded software (e.g., Hardware operating software, data analysis software). All Software is licensed and not sold and may be subject to additional terms found in the Software's end user license agreement. "**Specifications**" means Illumina's written specifications for this Product in effect on the date that the Product ships from Illumina.

2 **Research Use Only Rights**. Subject to these terms and conditions and unless otherwise agreed upon in writing by an officer of Illumina, Purchaser is granted only a non-exclusive, non-transferable, personal, non-sublicensable right under Illumina's Core IP, in existence on the date that this Product ships from Illumina, solely to use this Product in Purchaser's facility for Purchaser's internal research purposes (which includes research services provided to third parties) and solely in accordance with this Product's Documentation, **but specifically excluding any use that** (a) would require rights or a license from Illumina to Application Specific IP, (b) is a re-use of a previously used Consumable, (c) is the disassembling, reverse-engineering, reverse-compiling, or reverse-assembling of this Product, (d) is the separation, extraction, or isolation of components of this Product or other unauthorized analysis of this Product, (e) gains access to or determines the methods of operation of this Product, (f) is the use of non-Illumina reagent/consumables with Illumina's Hardware (does not apply if the Specifications or Documentation state otherwise), or (g) is the transfer to a third-party of, or sub-licensing of, Software or any third-party software. All Software, whether provided separately, installed on, or embedded in a Product, is licensed to Purchaser and not sold. Except as expressly stated in this Section, no right or license under any of Illumina's intellectual property rights is or are granted expressly, by implication, or by estoppel.

Purchaser is solely responsible for determining whether Purchaser has all intellectual property rights that are necessary for Purchaser's intended uses of this Product, including without limitation, any rights from third parties or rights to Application Specific IP. Illumina makes no guarantee or warranty that purchaser's specific intended uses will not infringe the intellectual property rights of a third party or Application Specific IP.

3 **Regulatory**. This Product has not been approved, cleared, or licensed by the United States Food and Drug Administration or any other regulatory entity whether foreign or domestic for any specific intended use, whether research, commercial, diagnostic, or otherwise. This Product is labeled For Research Use Only. Purchaser must ensure it has any regulatory approvals that are necessary for Purchaser's intended uses of this Product.

4 **Unauthorized Uses**. Purchaser agrees: (a) to use each Consumable only one time, and (b) to use only Illumina consumables/reagents with Illumina Hardware. The limitations in (a)-(b) do not apply if the Documentation or Specifications for this Product state otherwise. Purchaser agrees not to, nor authorize any third party to, engage in any of the following activities: (i) disassemble, reverse-engineer, reverse-compile, or reverse-assemble the Product, (ii) separate, extract, or isolate components of this Product or subject this Product or components thereof to any analysis not expressly authorized in this Product's Documentation, (iii) gain access to or attempt to determine the methods of operation of this Product, or (iv) transfer to a third-party, or grant a sublicense, to any Software or any third-party software. Purchaser further agrees that the contents of and methods of operation of this Product are proprietary to Illumina and this Product contains or embodies trade secrets of Illumina. The conditions and restrictions found in these terms and conditions are bargained for conditions of sale and therefore control the sale of and use of this Product by Purchaser.

5 **Limited Liability. TO THE EXTENT PERMITTED BY LAW, IN NO EVENT SHALL ILLUMINA OR ITS SUPPLIERS BE LIABLE TO PURCHASER OR ANY THIRD PARTY FOR COSTS OF PROCUREMENT OF SUBSTITUTE PRODUCTS OR SERVICES, LOST PROFITS, DATA OR BUSINESS, OR FOR ANY INDIRECT, SPECIAL, INCIDENTAL, EXEMPLARY, CONSEQUENTIAL, OR PUNITIVE DAMAGES OF ANY KIND ARISING OUT OF OR IN CONNECTION WITH, WITHOUT LIMITATION, THE SALE OF THIS PRODUCT, ITS USE, ILLUMINA'S PERFORMANCE HEREUNDER OR ANY OF THESE TERMS AND CONDITIONS, HOWEVER ARISING OR CAUSED AND ON ANY THEORY OF LIABILITY (WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), STRICT LIABILITY OR OTHERWISE).**

6 **ILLUMINA'S TOTAL AND CUMULATIVE LIABILITY TO PURCHASER OR ANY THIRD PARTY ARISING OUT OF OR IN CONNECTION WITH THESE TERMS AND CONDITIONS, INCLUDING WITHOUT LIMITATION, THIS PRODUCT (INCLUDING USE THEREOF) AND ILLUMINA'S PERFORMANCE HEREUNDER, WHETHER IN CONTRACT, TORT (INCLUDING NEGLIGENCE), STRICT LIABILITY OR OTHERWISE, SHALL IN NO EVENT EXCEED THE AMOUNT PAID TO ILLUMINA FOR THIS PRODUCT.**

7 **Limitations on Illumina Provided Warranties. TO THE EXTENT PERMITTED BY LAW AND SUBJECT TO THE EXPRESS PRODUCT WARRANTY MADE HEREIN ILLUMINA MAKES NO (AND EXPRESSLY DISCLAIMS ALL) WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THIS PRODUCT, INCLUDING WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, NONINFRINGEMENT, OR ARISING FROM COURSE OF PERFORMANCE, DEALING, USAGE OR TRADE. WITHOUT LIMITING THE GENERALITY OF THE FOREGOING, ILLUMINA MAKES NO CLAIM, REPRESENTATION, OR WARRANTY OF ANY KIND AS TO THE UTILITY OF THIS PRODUCT FOR PURCHASER'S INTENDED USES.**

8 **Product Warranty**. All warranties are personal to the Purchaser and may not be transferred or assigned to a third-party, including an affiliate of Purchaser. All warranties are facility specific and do not transfer if the Product is moved to another facility of Purchaser, unless Illumina conducts such move.

a **Warranty for Consumables**. Illumina warrants that Consumables, other than custom Consumables, will conform to their Specifications until the later of (i) 3 months from the date of shipment from Illumina, and (ii) any expiration date or the end of the shelf-life pre-printed on such Consumable by Illumina, but in no event later than 12 months from the date of shipment. With respect to custom Consumables (i.e., Consumables made to specifications or designs made by Purchaser or provided to Illumina by, or on behalf of, Purchaser), Illumina only warrants that the custom Consumables will be made and tested in accordance with Illumina's standard manufacturing and quality control processes. Illumina makes no warranty that custom Consumables will work as intended by Purchaser or for Purchaser's intended uses.

b **Warranty for Hardware**. Illumina warrants that Hardware, other than Upgraded Components, will conform to its Specifications for a period of 12 months after its shipment date from Illumina unless the Hardware includes Illumina provided installation in which case the warranty period begins on the date of installation or 30 days after the date it was delivered, whichever occurs first ("Base Hardware Warranty"). "Upgraded Components" means Illumina provided components, modifications, or enhancements to Hardware that was previously acquired by Purchaser. Illumina warrants that Upgraded Components will conform to their Specifications for a period of 90 days from the date the Upgraded Components are installed. Upgraded Components do not extend the warranty for the Hardware unless the upgrade was conducted by Illumina at Illumina's facilities in which case the upgraded Hardware shipped to Purchaser comes with a Base Hardware Warranty.

c **Exclusions from Warranty Coverage**. The foregoing warranties do not apply to the extent a non-conformance is due to (i) abuse, misuse, neglect, negligence, accident, improper storage, or use contrary to the Documentation or Specifications, (ii) improper handling, installation, maintenance, or repair (other than if performed by Illumina's personnel), (iii) unauthorized alterations, (iv) Force Majeure events, or (v) use with a third party's good not provided by Illumina (unless the Product's Documentation or Specifications expressly state such third party's good is for use with the Product).

d **Procedure for Warranty Coverage**. In order to be eligible for repair or replacement under this warranty Purchaser must (i) promptly contact Illumina's support department to report the non-conformance, (ii) cooperate with Illumina in confirming or diagnosing the non-conformance, and (iii) return this Product, transportation charges prepaid to

Illumina following Illumina's instructions or, if agreed by Illumina and Purchaser, grant Illumina's authorized repair personnel access to this Product in order to confirm the non-conformance and make repairs.

e **Sole Remedy under Warranty**. Illumina will, at its option, repair or replace non-conforming Product that it confirms is covered by this warranty. Repaired or replaced Consumables come with a 30-day warranty. Hardware may be repaired or replaced with functionally equivalent, reconditioned, or new Hardware or components (if only a component of Hardware is non-conforming). If the Hardware is replaced in its entirety, the warranty period for the replacement is 90 days from the date of shipment or the remaining period on the original Hardware warranty, whichever is shorter. If only a component is being repaired or replaced, the warranty period for such component is 90 days from the date of shipment or the remaining period on the original Hardware warranty, whichever ends later. The preceding states Purchaser's sole remedy and Illumina's sole obligations under the warranty provided hereunder.

f **Third-Party Goods and Warranty**. Illumina has no warranty obligations with respect to any goods originating from a third party and supplied to Purchaser hereunder. Third-party goods are those that are labeled or branded with a third-party's name. The warranty for third-party goods, if any, is provided by the original manufacturer. Upon written request Illumina will attempt to pass through any such warranty to Purchaser.

9 **Indemnification**.

a **Infringement Indemnification by Illumina**. Subject to these terms and conditions, including without limitation, the Exclusions to Illumina's Indemnification Obligations (Section 9(b) below), the Conditions to Indemnification Obligations (Section 9(d) below), Illumina shall (i) defend, indemnify and hold harmless Purchaser against any third-party claim or action alleging that this Product when used for research use purposes, in accordance with these terms and conditions, and in accordance with this Product's Documentation and Specifications infringes the valid and enforceable intellectual property rights of a third party, and (ii) pay all settlements entered into, and all final judgments and costs (including reasonable attorneys' fees) awarded against Purchaser in connection with such infringement claim. If this Product or any part thereof, becomes, or in Illumina's opinion may become, the subject of an infringement claim, Illumina shall have the right, at its option, to (A) procure for Purchaser the right to continue using this Product, (B) modify or replace this Product with a substantially equivalent non-infringing substitute, or (C) require the return of this Product and terminate the rights, license, and any other permissions provided to Purchaser with respect this Product and refund to Purchaser the depreciated value (as shown in Purchaser's official records) of the returned Product at the time of such return; provided that, no refund will be given for used-up or expired Consumables. This Section states the entire liability of Illumina for any infringement of third party intellectual property rights.

b **Exclusions to Illumina Indemnification Obligations**. Illumina has no obligation to defend, indemnify or hold harmless Purchaser for any Illumina Infringement Claim to the extent such infringement arises from: (i) the use of this Product in any manner or for any purpose outside the scope of research use purposes, (ii) the use of this Product in any manner not in accordance with its Specifications, its Documentation, the rights expressly granted to Purchaser hereunder, or any breach by Purchaser of these terms and conditions, (iii) the use of this Product in combination with any other products, materials, or services not supplied by Illumina, (iv) the use of this Product to perform any assay or other process not supplied by Illumina, or (v) Illumina's compliance with specifications or instructions for this Product furnished by, or on behalf of, Purchaser (each of (i) – (v), is referred to as an "Excluded Claim").

c **Indemnification by Purchaser**. Purchaser shall defend, indemnify and hold harmless Illumina, its affiliates, their non-affiliate collaborators and development partners that contributed to the development of this Product, and their respective officers, directors, representatives and employees against any claims, liabilities, damages, fines, penalties, causes of action, and losses of any and every kind, including without limitation, personal injury or death claims, and infringement of a third party's intellectual property rights, resulting from, relating to, or arising out of (i) Purchaser's breach of any of these terms and conditions, (ii) Purchaser's use of this Product outside of the scope of research use purposes, (iii) any use of this Product not in accordance with this Product's Specifications or Documentation, or (iv) any Excluded Claim.

d **Conditions to Indemnification Obligations**. The parties' indemnification obligations are conditioned upon the party seeking indemnification (i) promptly notifying the other party in writing of such claim or action, (ii) giving the other party exclusive control and authority over the defense and settlement of such claim or action, (iii) not admitting infringement of any intellectual property right without prior written consent of the other party, (iv) not entering into any settlement or compromise of any such claim or action without the other party's prior written consent, and (v) providing reasonable assistance to the other party in the defense of the claim or action; provided that, the party reimburses the indemnified party for its reasonable out-of-pocket expenses incurred in providing such assistance.

e **Third-Party Goods and Indemnification**. Illumina has no indemnification obligations with respect to any goods originating from a third party and supplied to Purchaser. Third-party goods are those that are labeled or branded with a third-party's name. Purchaser's indemnification rights, if any, with respect to third party goods shall be pursuant to the original manufacturer's or licensor's indemnity. Upon written request Illumina will attempt to pass through such indemnity, if any, to Purchaser.

Revision History

Part #	Revision	Date	Description of Change
15047621	A	November 2013	Initial release.

Table of Contents

FastTrack Long Reads Sequencing Service

Overview

FastTrack Long Reads Sequencing Service through the Illumina Genome Network (IGN) delivers whole-genome sequencing using long read sample preparation technology. The Long Reads Sequencing Service is a cost-effective solution for genome finishing, metagenomics, and de novo sequencing.

The FastTrack Long Reads Informatics Pipeline consists of a suite of novel algorithms designed to assemble high quality synthetic long-read fragments using data generated from Illumina's Long Reads sequencing technology.

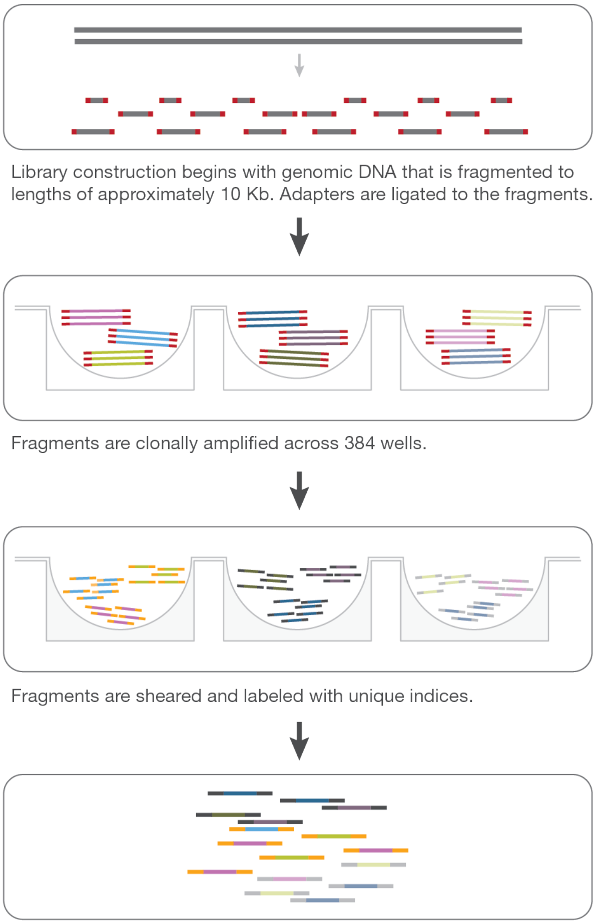
This user guide provides an overview of the sample preparation and the informatics pipeline included in the FastTrack Long Reads Sequencing Service, as well as a detailed description of the data provided in order to help you understand the Long Reads Informatics Data Package that you receive from Illumina.

Library Preparation

In the long-read library preparation, genomic DNA is initially sheared into 5-10 kb long fragments and diluted onto a 384-well plate. Each input DNA fragment is then ligated with PCR primers as well as an additional unique 8-base sequence, or end-marker sequence, which identifies the 5′ and 3′ ends of the molecule. The fragments in each well are clonally amplified, fragmented with Nextera technology and bar coded, to create a short-fragment library. The short reads fragments generated in all wells are finally pooled and sequenced on one HiSeq lane.

The relatively low number of fragments in each well facilitates the assembly process as there are fewer repetitive sequences in the input data to confound the assembly. In addition, the haploid nature of the input fragments eliminates the need to accommodate heterozygous variants and thus allows for more aggressive separation of repeat copies.

Figure 1 Sample Preparation for the Long Read Workflow



Library construction begins with genomic DNA that is fragmented to lengths of approximately 10 Kb. Adapters are ligated to the fragments.

Fragments are clonally amplified across 384 wells.

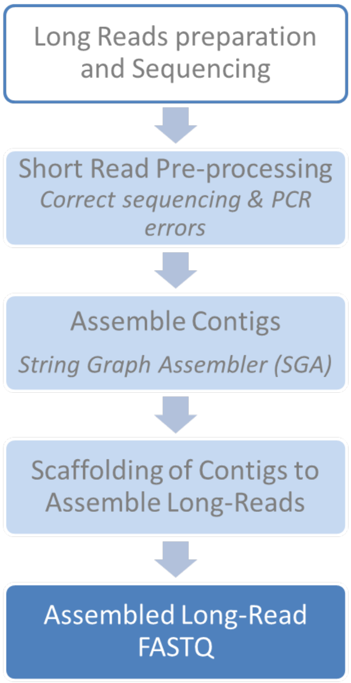
Fragments are sheared and labeled with unique indices.

Fragments are sequenced using Illumina technology. Short read sequences are assembled into long sequence fragments.

Long Reads Informatics Pipeline

The FastTrack Long Reads Informatics Pipeline begins by separating the sequence reads into the component 384 wells based on the barcode sequence. In the next stage, the reads in each individual well are pre-processed to correct sequencing and PCR errors. Next, a string graph is constructed using the String Graph Assembler (SGA) assembler[1]; the resulting graph is then cleaned by using the paired-end information from the short reads to produce an initial set of contigs. The contigs are further scaffolded together in the next step of the pipeline in order to resolve repeats and fill in gaps created due to low sequencing coverage. In the final stage, the scaffolds are examined for possible errors and misassemblies or where low-confidence regions are broken.

Figure 2 Overview of the FTS Long Reads Algorithm Workflow


Long Reads preparation and Sequencing
Short Read Pre-processing
Correct sequencing & PCR errors
Assemble Contigs
String Graph Assembler (SGA)
Scaffolding of Contigs to Assemble Long-Reads
Assembled Long-Read FASTQ

Analysis Deliverables

Data Files Delivery on Illumina Hard Drives

Illumina provides data for the long reads sequencing service on one or more hard drives. The hard drives are formatted with the NTFS file system and can optionally be encrypted using the open-source cross-platform TrueCrypt software (http://www.truecrypt.org) and the Advanced Encryption Standard (AES) algorithm (Federal Information Processing Standards Publication 197).

The data on the hard drive are organized in a folder structure with one top-level folder that is named by the barcode sample of the long fragment library.

This chapter details the files and folder structure for the Long Reads Sequencing deliverable. The files and folders generated for the Long Reads Sequencing results are all keyed off of the unique sample identifiers. In most cases, these unique identifiers are the barcodes associated with the samples in the lab (for example, LP600001-DNA_A01). They can also be a known sample IDs for reference samples (for example, HCC1187).

BaseSpace Delivery

The main outputs of the FastTrack long reads pipeline, the two FASTQ files, the scaffolds file, and report PDF will be delivered on both hard disk and via Illumina's genomics cloud computing environment, BaseSpace. Your project manager will be contacting you with further instructions on how you can access your data via BaseSpace.

Results Folder Structure

The files and folders generated for the long reads analysis pipeline results are all keyed off the unique sample identifiers. In most cases, these unique identifiers are the barcodes associated with the samples in the lab (for example, LP600001-DNA_A01) but can be a known sample id for reference samples (for example, HCC1187).

Under each long reads sample folder, you can find the following file structure that contains analysis results.

- [SampleBarcode]
  - LongRead_results – this folder contains all the output files resulting from the Long-Read sequencing run and analysis.
    - [LibraryName]_LongRead.fastq.gz
    - [LibraryName]_LongRead_500_1499nt.fastq.gz
    - [LibraryName]_Scaffolds.txt
    - [LibraryName]_LongReadsSummaryReport.pdf
    - [LibraryName]_ShortInsertSequencing.tar.gz

Long Reads Output File Details

[LibraryName]_LongRead.fastq.gz

FASTQ file containing the final assembled reads of 1500 bp or greater.

[LibraryName]_500-1499nt.fastq.gz

FASTQ file containing the final assembled reads of length 500–1499 bp. These reads are not used in the calculation of reported metrics but are made available to enable custom analysis by expert users.

[LibraryName]_Scaffolds.txt

A text file containing identifiers of long reads in the FastQ file that come from the same DNA fragment, with relative orientation and order preserved.

[LibraryName]_LongReadsSummaryReport.pdf

This compressed report contains an overview of the results for the sample. In the report you will find the following:

Metric	Section	Description
Number of Long Reads >= 1500nt	Assembly Metrics	Total number of assembled long reads >=1500bp
Total Bases Assembled in Long Reads >=1500bp	Assembly Metrics	Sum of bases in assembled long reads >=1500bp
Total Bases Assembled in All Long Reads	Assembly Metrics	Sum of bases in all assembled long reads
N50 of Assembled Long Reads >=1500 bp	Assembly Metrics	N50 value of the length of assembled long reads >=1500 bp

The report file also provides the following 2 plots:

- Yield of assembled sequence per read length bin. The sum of all assembled sequence for all long reads in a given read length bin is represented.
- Distribution of long reads with length 1500 or greater.

[LibraryName]_ShortInsertSequencing Folder

This folder contains the short read output from the long fragments library sequencing run. The output files are in FASTQ format and are demultiplexed by sample barcode, allowing a 1-base mismatch in the barcode sequence. End markers of the 5′-3′ sequence TACGCTTGCAT may be present in some short read sequences, indicating one end of a long fragment. Any sequence 5′ of the end marker, or 3′ of its reverse complement, is expected to be adapter rather than sample DNA. Note that this will not be true in the case where the sequence TACGCTTGCAT is present as a native part of your sample DNA.

Informatics Pipeline Details

Short Read Pre-Processing

Prior to the assembly of the long reads, the short reads in every well are pre-filtered to correct for errors which could lead to misassemblies. Reads that do not have a sufficient stretch of high-quality bases are filtered. Low-quality ends of remaining bases are trimmed (hard-clipped). Read pairs that appear to 'read through' one another, and thus potentially contain adapter sequence on the 3′ end(s) of one or both reads, are modified as follows. The first read is trimmed of bases that appear to extend beyond the second read, and the second read is discarded, resulting in an unpaired read that should have had any 3′ adapter sequence clipped off. If the trimmed reads in a pair are shorter than 30bp, the pair is discarded. If one read in a pair is shorter than 30bp, and the second read longer than 50bp, the longer read is kept. Adapter sequences are removed and the end-marker sequences identified and trimmed, and reads containing end-marker sequences are tagged for downstream use in the pipeline.

Assembly of Contigs

The assembly module consists of several steps: digital normalization, read error correction, graph construction, and clean-up using paired end reads. These steps are described in more detail in the following sections.

Digital Normalization

Due to bias introduced during PCR, the read coverage among input fragments in the sample can vary greatly. In order to normalize coverage variation across fragments (which improves the accuracy of the assembly as well as the computational performance of the algorithm), digital normalization methods outlined by Brown et al[2] are used. The digital normalization process smooths out highly biased sequence coverage by removing specific over-represented sequences. Coverage is normalized such that the highest coverage fragments are approximately 40x.

Error Correction

Following digital normalization, an error correction step is performed using an overlap-based method. The aim of this step is to correct PCR and sequencing artifacts which introduce false base substitutions or indels. At a high level, it operates as follows. An index of all k-mers of length 31 in the reads is constructed (the k-mer hash). For each read, k-mers in the read are compared to the index to find the set of reads which share the same k-mer. Matches to candidate overlapping reads are extended using semi-banded global alignment, and those which have a match length of at least 31 bases and share 95% identity, are retained. Multiple sequence alignment (MSA) of the set of overlapping reads is performed. Using both the base quality scores of the reads and the results of the MSA, a consensus sequence for the read is generated.

Graph Construction

The main assembly step is performed using the String Graph Assembler (SGA)[1], which is an overlap-based assembly method. In the first stage, SGA uses a k-mer overlap size of 31 to create an graph with reads as vertices and k-mer overlaps as edges.

After the construction of an initial graph, the next step of the algorithm is to clean the graph and remove spurious edges using several heuristics. The algorithm requires that paths in the graph are supported by paired-end reads. It checks for the existence of a path linking the two reads of a read pair within the expected insert size distribution (500 bp by default). Any edges in the graph which do not support read pairs are removed. In addition, tips and bubbles in the read graph which normally occur during *de novo* assembly are cleaned up using standard graph cleaning methods.

Scaffolding Contigs to Assemble Long Reads

The next stage in the pipeline is scaffolding, the goal of which is to use paired-end information to place and orient the contigs generated in the previous step and fill in gaps between contigs. The method employed in the long reads pipeline is based on the scaffolding method employed in the original SGA assembler, and the user is referred to Simpson et al[1] for further details.

In brief, scaffolding is accomplished by re-aligning the input short reads to the contigs using BWA aligner[3], and using the paired-end alignments to infer scaffold structure. The link between two contigs is made when 2 or more paired reads map such that read 1 from a read pair maps to one contig and read 2 from the same read pair maps to the other. The orientation of the contigs relative to one another is also inferred from the orientation of the read-pairs. In addition, the end-marker sequences are used to help guide and constrain the construction of our scaffold graph

Gap Filling

The next step of this module is to fill in scaffold gaps where possible in order to resolve repeats. In this step, we use the input short reads, making use of the FM index computed during the contig assembly. We begin by finding the highest scoring read which matches the end of one of the contigs, and continue to chain together reads iteratively. If a chain is found that overlaps another contig in the same scaffold, the consensus is retained and the gap filled with this sequence.

Assembly QC and Correction

The final stage of the analysis pipeline involves verification of the scaffolds and error correction. The short read data is again aligned against the scaffolds generated in the previous step usin BWA aligner[3]. Based on the alignments, the scaffolds are corrected for single-nucleotide errors and broken into smaller scaffolds should there be only partial alignment support. Quality scores for the final long reads are also estimated from the alignments.

Breaking Scaffolds

The short reads used during the Long Reads assembly are aligned to the scaffolds. The alignments are searched for read pairs in which one read aligns and the other one does not. Unaligned reads are re-aligned, and reads that are overlapping or running into scaffold gaps are counted and computed. In order to determine whether or not to break a scaffold gap, Illumina computes the following formula:

```
sqrt(0.3+(reads aligning to mid point of gap on fwd strand)*(0.3+
   (reads aligning to mid point of gap on rev strand)))/(total
   number of reads in gap)
```

If this ratio is smaller than 0.1, the gap is left as it is; if it is larger, the scaffold is broken at this gap. If there are only few reads or none, the scaffold for the region is left as it is.

Q-scores

From the alignments of short reads to the scaffolds, a pileup file is generated which provides the base quality scores of the aligned reads at each position in a scaffold. The quality score at each scaffold position is then estimated from the read base qualities as follows:

1. Remove Ns and indels from the pile-up.
2. If coverage > 5 and all nucleotides at this position agree and set Q-score to max of pileup.
3. If < 5% mismatches or > 3 matches, set Q-score to mean of pileup.
4. If all of the above steps fail, look at the most frequently occurring nucleotide in the pileup and the second most frequent one. Compute the posterior probability of most frequent base given the quality scores. This includes some correction factors from a PCR error rate model. Do the same for the second most frequent nucleotide. Choose the nucleotide with the highest posterior probability and compute the q-score from this probability

References

1 Simpson, JT. & Durbin, R. (2012) Efficient *de novo* assembly of large genomes using compressed data structures. Genome Research 22(3), 549-56.

2 A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. http://arxiv.org/abs/1203.4802

3 Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595.

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 1 Illumina General Contact Information

Illumina Website	www.illumina.com
Email	techsupport@illumina.com

Table 2 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Austria	0800.296575	Netherlands	0800.0223859
Belgium	0800.81102	Norway	800.16836
Denmark	80882346	Spain	900.812168
Finland	0800.918363	Sweden	020790181
France	0800.911850	Switzerland	0800.563118
Germany	0800.180.8994	United Kingdom	0800.917.0041
Ireland	1.800.812949	Other countries	+44.1799.534000

Safety Data Sheets

Safety data sheets (SDSs) are available on the Illumina website at www.illumina.com/msds.

Product Documentation

Product documentation in PDF is available for download from the Illumina website. Go to www.illumina.com/support, select a product, then click **Documentation & Literature**.

Illumina
Headquartered in San Diego, California, U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com