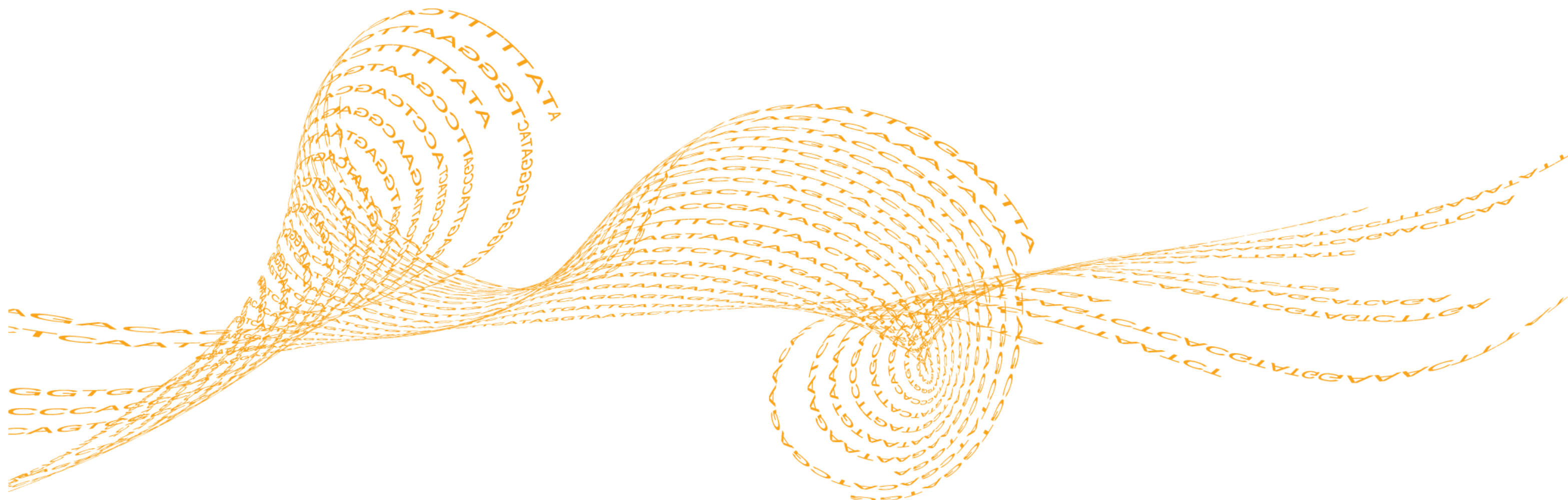
TruSeq Long-Read Assembly App

Revision History	3
Introduction	4
Running TruSeq Long-Read Assembly	6
TruSeq Long-Read Assembly Output	7
TruSeq Long-Read Assembly Methods	13
References	16
Technical Assistance	



ILLUMINA PROPRIETARY
15055850 Rev. B
January 2015

illumina®

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).

FOR RESEARCH USE ONLY

© 2015 Illumina, Inc. All rights reserved.

Illumina, **24sure**, **BaseSpace**, **BeadArray**, **BlueFish**, **BlueFuse**, **BlueGnome**, **cBot**, **CSPro**, **CytoChip**, **DesignStudio**, **Epicentre**, **GAIIx**, **Genetic Energy**, **Genome Analyzer**, **GenomeStudio**, **GoldenGate**, **HiScan**, **HiSeq**, **HiSeq X**, **Infinium**, **iScan**, **iSelect**, **MiSeq**, **NeoPrep**, **Nextera**, **NextBio**, **NextSeq**, **Powered by Illumina**, **SeqMonitor**, **SureMDA**, **TruGenome**, **TruSeq**, **TruSight**, **Understand Your Genome**, **UYG**, **VeraCode**, **verifi**, **VeriSeq**, the pumpkin orange color, and the streaming bases design are trademarks of Illumina, Inc. and/or its affiliate(s) in the U.S. and/or other countries. All other names, logos, and other trademarks are the property of their respective owners.

Revision History

Part #	Revision	Date	Description of Change
15055850	B	January 2015	Added support for combining samples. Updated log files.
15055850	A	June 2014	Initial Release.

Introduction

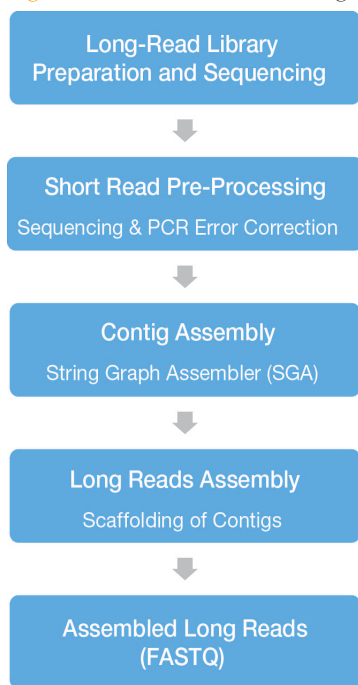
Illumina provides the TruSeq Synthetic Long-Read DNA Library Prep Kit, which generates libraries for long-read sequencing. The TruSeq Long-Read Assembly App consists of a suite of novel algorithms designed to assemble the high-quality synthetic long-read fragments.

This user guide provides an overview of the app, instructions on how to run it, and a detailed description of the analysis output.

TruSeq Long-Read Assembly App Overview

The input to the TruSeq Long-Read Assembly App app is a set of demultiplexed FASTQ files generated when TruSeq Synthetic Long-Read runs are streamed to BaseSpace. In the first stage, the reads from each individual well are pre-processed to correct sequencing and PCR errors. Next, a string graph is constructed using the String Graph Assembler (SGA)[1]. The resulting graph is then cleaned 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 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. Scaffolds are broken when low-confidence regions are identified.

Figure 1 Overview of the Long Reads Algorithm Workflow



Versions

The following module versions are used in the TruSeq Long-Read Assembly app:

- Tornado-v2.0
- BWA-0.6.2
- SGA-0.9.37
- SAMtools-0.1.18
- Khmer-0.2 (coverage normalization)

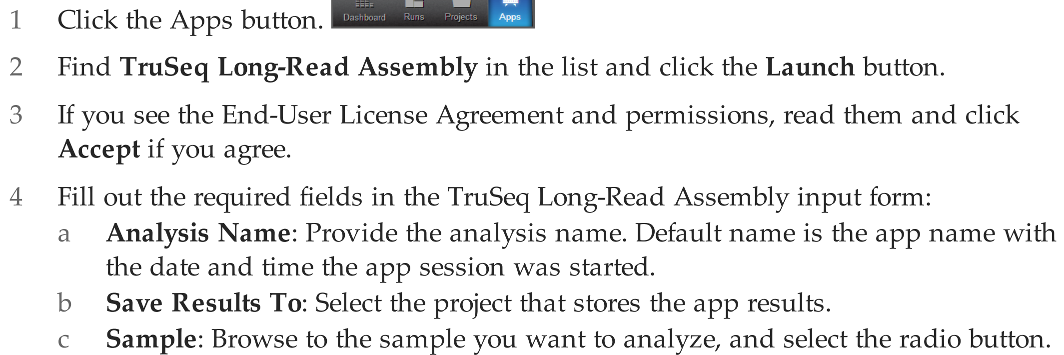
Current Limitations

Before running the TruSeq Long-Read Assembly app, be aware of the following limitations:

- Only samples generated with the TruSeq Synthetic Long-Read Library Prep Kit can be analyzed with this app.
- Needs paired-end data.
- The short read length lower limit is 2×100 bp. The app supports the longest read lengths supported by the sequencing chemistry.
- If you have less than 30 Gbp in sequence data, you could have insufficient data for optimal results. Perform an extra lane of sequencing with that library, and combine all lanes of sequencing for that library to ensure sufficient data.
- If you have more than 115 Gbp in sequence data, your analysis run can time out.
- The lower limit is 100 000 reads per barcode.
- The app restricts input to the first 1.5 million reads per barcode.
- Currently, you can only align and QC human long-reads. This function uses UCSC hg19 as the reference.

NOTE
If you must combine samples to reach the proper number of reads, see the *BaseSpace User Guide* for instructions.

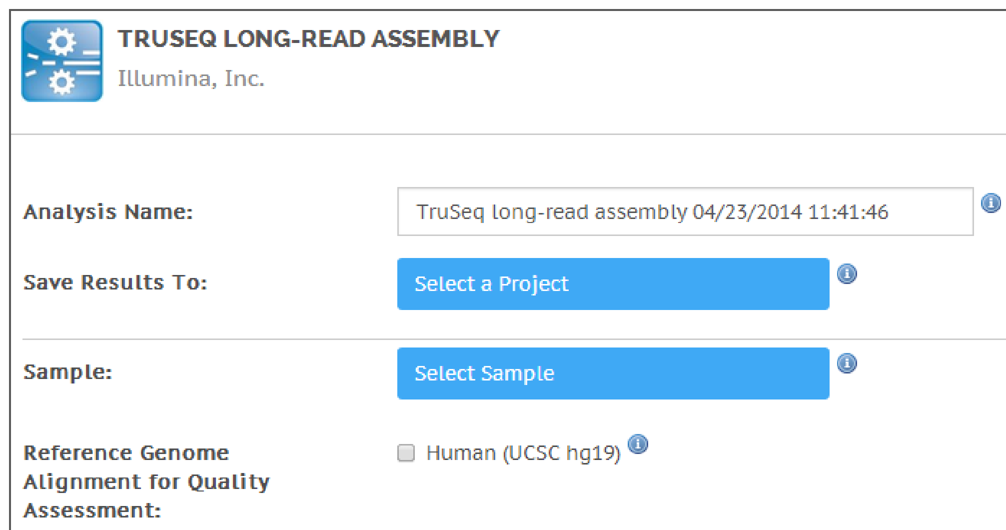
Running TruSeq Long-Read Assembly

1. Click the Apps button. 
2. Find **TruSeq Long-Read Assembly** in the list and click the **Launch** button.
3. If you see the End-User License Agreement and permissions, read them and click **Accept** if you agree.
4. Fill out the required fields in the TruSeq Long-Read Assembly input form:
   a. **Analysis Name**: Provide the analysis name. Default name is the app name with the date and time the app session was started.
   b. **Save Results To**: Select the project that stores the app results.
   c. **Sample**: Browse to the sample you want to analyze, and select the radio button.

NOTE
Indexing of the sample list causes a lag between the time the sample is created in BaseSpace and when the sample is available for selection in the apps.

   d. **Reference Genome Alignment for Quality Assessment**: Select the checkbox if you want to assess the quality of human synthetic long-reads by aligning them to the human reference genome. This option is only available for UCSC hg19.

Figure 2 TruSeq Long-Read Assembly Input Form



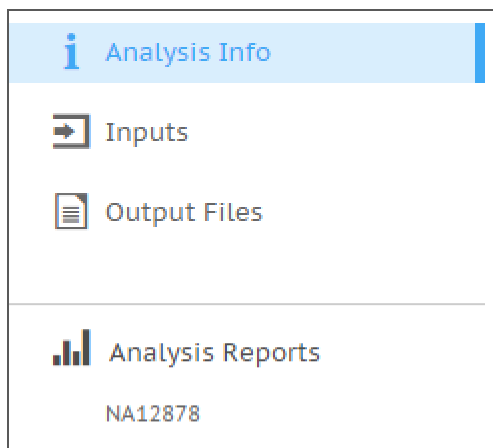
5. Click **Continue**.

The TruSeq Long-Read Assembly app now starts analyzing your sample. When completed, the status of the app session is automatically updated, and you receive an email.

TruSeq Long-Read Assembly Output

This chapter describes the output that the TruSeq Long-Read Assembly app produces. To go to the results, click the **Projects** button, then the project, then the analysis.

Figure 3 TruSeq Long-Read Assembly Output Navigation Bar



When the App Session is completed, you can access your output through the left navigation bar, which provides the following:

- **Analysis Reports**—An overview of the analysis results per sample.
  See *Analysis Reports* on page 7 for a description.
- **Analysis Info**—An overview of the app session settings.
  See *Analysis Info* on page 9 for a description.
- **Inputs**—An overview of input settings.
  See *Inputs* on page 10
- **Output Files**—Access to the output files, organized by sample and app session.
  The output files page contains links to a number of data files:
  - LongRead.fastq.gz—FASTQ file containing the final assembled reads of 1500 bp or greater. This file is the main output file of TruSeq Long-Read Assembly.
    For a description, see *FASTQ Files* on page 10.
  - LongRead_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.
  - Scaffolds.txt—Text file containing identifiers of long reads in the FASTQ file that come from the same DNA fragment, with relative orientation and order preserved.

  The output files contain two reports:
  - LongReadsSummaryReport.pdf—PDF report detailing sample information, sequencing metrics for phasing library and phasing summary statistics.
    See *Long Reads Summary Report* on page 11 for more information.
  - LibraryCharacteristics.csv—Comma-separated file describing the library characteristics.
    See *LibraryCharacteristics.csv* on page 11 for more information.

Analysis Reports

Located under Analysis Reports in the left navigation panel are links to Analysis Reports for each sample. These HTML reports contain an overview of the metrics of the long

fragments library and statistics from the assembly analysis. In addition, they provide links to the associated PDF Summary Report and associated FASTQ files.

The following statistics are reported:

Table 1 Sample Information

Metric	Description
Total Reads PF	Total number of passing filter short read pairs
Read 1 Length	Length of short Read 1
Read 2 Length	Length of short Read 2
Genome	Genome used for aligning and QC, if performed

Table 2 Sample Information

Metric	Description
Total Reads PF	Total number of passing filter short read pairs
Read 1 Length	Length of short Read 1
Read 2 Length	Length of short Read 2
Genome	Genome used for aligning and QC, if performed

Table 3 Well Metrics

Metric	Description
Reads PF	Total number of passing filter short read pairs The app does not process wells with less than 100 000 reads passing filter (PF) per well. For wells that are filtered out due to this threshold, Reads PF is set to 0.
Percent End Marker	Percentage of short reads containing end marker sequence tag
Number Unique Fragments Per Well (End Markers)	Number of unique long fragments per well, estimated by K-mer analysis of end marker containing sequence
Total Unique DNA Sequences Assembled (bp)	Sum of DNA bases assembled in long read before final QC step
Assembly N50 Length (pre-QC)	N50 Value of assembled long read length before final QC step

NOTE
Span90 means the least fold change to cover 90% of values in the given data set.

Table 4 Assembly Metrics

Metric	Description

Number of Long Reads Assembled	Total number of assembled long reads, grouped by length
Number of Bases in Assembled Long Reads	Sum of bases in assembled long reads, grouped by length
Long Reads Assembly N50 Length	N50 Value of long read assembled, grouped by length

NOTE
N50 is the length for which the sum of bases in the long reads of that length or longer is at least half the bases in the assembly. See also en.wikipedia.org/wiki/N50_statistic.

The report file also provides the following two plots:

- All Long Read Size Distribution. Plots the number of long reads by read length. The export button allows export of the data in a tab-separated file.
- End-Marked Long Read Size Distribution. Plots the number of end-marked long reads by read length. The export button allows export of the data in a tab-separated file.

Analysis Info

This app provides an overview of the analysis on the Analysis Info page.

A brief description of the metrics is below.

Table 5 Analysis Info

Row	Definition
Name	Name of the app session.
Application	App that generated this analysis.
Date Started	Date and time the app session started.
Date Completed	Date and time the app session completed.
Duration	Duration of analysis.
Session Type	The number of nodes used.
Size	Total size of all output files.
Status	Status of the app session.

Log Files

Clicking the **Log Files** link on the Analysis Info page provides access to the app log files. Additional log files are located in a folder in the Output Files section.

- **bsfs-{timestamp}.log**: Platform log file concerning FUSE drive mounting.
- **output-{timestamp}.log**: Platform log file that contains all the console output of the app.
- **uploader-{timestamp}.log**: File logging uploading.
- **spacedock-{timestamp}.log**: Shows console output from the SpaceDock and BaseSpace communication and input/output file staging.
- **spacedock-infrastructure-{timestamp}.log**: Log file used for debugging.

- **WorkflowLog.txt**: Pyflow logging output from the workflow.
- **WorkflowOut.txt**: Stdout output from the workflow.
- **WorkflowErr.txt**: Stderr output from the workflow.
- **config.txt**: Config file for running the workflow.

Inputs

The Inputs page provides an overview of the input samples and settings that were specified when the TruSeq Long-Read Assembly project was set up.

Output Files Overview

The TruSeq Long-Read Assembly app generates a number of output files. These files are discussed in this section. For more information on log files, see *Analysis Info* on page 9.

FASTQ Files on page 10

Long Reads Summary Report on page 11

LibraryCharacteristics.csv on page 11

FASTQ Files

TruSeq Long-Read Assembly App generates long read FASTQ files, which contain base call and quality information for reads passing filtering.

FASTQ files are saved compressed in the GNU zip format (an open source file compression program), indicated by the .gz file extension.

Format

Each entry in a FASTQ file consists of four lines:

- Sequence identifier:
  `@<read_and_barcode>:<length>`
- Sequence
- Quality score identifier line (consisting only of a +)
- Quality score

An example of a valid sequence identifier is as follows:

```
@Read_1-Barcode=BC002:length=1665
```

Wildcards

If the long-read assembler cannot determine the correct sequence at a given position, the assembler produces non-ACGT characters in the FASTQ files with reads ≥ 500 base pairs. In the regions with low coverage, the assembler determines the consensus sequence much less reliably and can introduce wildcards. This behavior is typical for alignment algorithms.

Examples of wildcard characters:

- `M` = `A` or `C` in the position
- `N` = base cannot be determined

A full list of ambiguity codes is available on www.bioinformatics.org/sms/iupac.html in the first table. The gap code is not used.

Long Reads Summary Report

This compressed report contains an overview of the results for the sample. This file is accessible through the Output File link. In the report you find the following:

Table 6 Long Reads Summary

Metric	Description
Number of Long Reads Assembled ≥ 1500bp	Total number of assembled long reads ≥ 1500 bases
Number of Bases in Assembled Long Reads ≥ 1500bp	Sum of bases in assembled long reads ≥ 1500 bases
Number of Bases in Assembled Long Reads ≥ 500bp	Sum of bases in all assembled long reads ≥ 500 bases

The report file also provides the following two 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 bases or greater.

LibraryCharacteristics.csv

The LibraryCharacteristics.csv report is accessible through the Output File link. This topic details the metrics reported in this file.

NOTE
Not all metrics are provided if you did not select the alignment option when setting up the analysis.

Table 7 Assembly Metrics

Metric	Definition
Number of Long Reads Assembled (≥1500bp)	Total number of long read assembled with length ≥ 1500 bases
Number of Bases in Assembled Long Reads (≥1500bp)	Sum of DNA bases delivered in long reads with length ≥ 1500 bases
Long Reads Assembly N50 Length (≥1500bp)	N50 value of long read assembled with length ≥ 1500 bases
Number of Long Reads Assembled (500-1499bp)	Total number of long read assembled with length ≥ 500 and ≤ 1499 bases
Number of Bases in Assembled Long Reads (500-1499bp)	Sum of DNA bases delivered in long reads with length ≥ 500 and ≤ 1499 bases
Long Reads Assembly N50 Length (500-1499bp)	N50 value of long read assembled with length between ≥ 500 and ≤ 1499 bases

NOTE
N50 is the length for which the sum of bases in the long reads of that length or longer is at least half the bases in the assembly. See also en.wikipedia.org/wiki/N50_statistic.

Table 8 Well Metrics—Summary Statistics

Metric	Definition
n	Number of bar code/well with valid value for analysis
sum	Summation of all wells. For the following stats, sum is calculated: Reads PF, Number Unique Fragments Per Well (End Markers), Total Unique DNA (bp) Aligned, Number Unique Fragments Per Well (Alignment). All other stats display N/A.
avg	Average
med	Median
std	Standard deviation
span50	Least fold change to cover 50% of values in the given data set
span90	Least fold change to cover 90% of values in the given data set

Table 9 Well Metrics - Details

Metric	Definition
Reads PF	Total number of passing filter short read pairs The app does not process wells with less than 100 000 reads passing filter (PF) per well. For wells that are filtered out due to this threshold, Reads PF is set to 0.
Percent End Marker	Percentage of short reads containing end marker sequence tag
Number Unique Fragments Per Well (End Markers)	Number of unique long fragments per well, estimated by K-mer analysis of end marker containing sequence
Total Unique DNA Sequences Assembled (bp)	Sum of DNA bases assembled in long read before final QC step
Assembly N50 Length (pre-QC)	N50 value of assembled long read length before final QC step
Percent Reads Aligned	Percentage of short reads aligned against the reference genome
Mismatch Rate Read 1	Percentage of mismatch of short Read 1 against the reference genome
Mismatch Rate Read 2	Percentage of mismatch of short Read 2 against the reference genome
Median Library Size	Median insert size of short paired-end library calculated by aligning to the reference genome + adapter length
Number Unique Fragments Per Well (Alignment)	Number of unique long fragments estimated by aligning to the reference genome
Total Unique DNA (bp) Aligned	Sum of DNA bases from all unique long fragments identified by aligning to the reference genome
Long Fragment Length Average	Length average of long fragments identified by aligning to the reference genome
Long Fragment Length Span90	Length span90 of long fragments identified by aligning to the reference genome

TruSeq Long-Read Assembly Methods

The TruSeq Long-Read Assembly App uses the following steps to assemble the long reads:

1 Pre-processes the reads from each individual well to correct sequencing and PCR errors.
2 Produces an initial set of contigs using a string graph.
3 Scaffolds contigs together to resolve repeats.
4 Fills in gaps created due to low sequencing coverage.
5 Examines the scaffolds for possible errors and misassemblies. Scaffolds are broken when low-confidence regions are identified.

These steps are explained in the following topics.

Short Read Pre-Processing

Before the assembly of the long reads, the short reads in every well are pre-filtered to correct for errors that 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 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. This step results in an unpaired read that has any 3’ adapter sequence clipped off.
- If the trimmed reads in a pair are shorter than 30 bp, the pair is discarded.
- If one read in a pair is shorter than 30 bp, and the second read longer than 50 bp, 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.

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. Coverage variation is normalized across fragments using digital normalization[2]. This normalization improves the accuracy of the assembly as well as the computational performance of the algorithm. 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 40×.

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. It operates as follows:

1. Constructs an index of all k-mers of length 31 in the reads (the k-mer hash).
2. Compares k-mers in the read to the index for each read to find the set of reads that share the k-mer.
3. Extends matches to candidate overlapping reads using semi-banded global alignment, and retains those reads that have a match length of at least 31 bases and share 95% identity.
4. Performs multiple sequence alignment (MSA) of the set of overlapping reads.
5. Generates a consensus sequence for the read using both the base quality scores of the reads and the results of the MSA.

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 a graph with reads as vertices and k-mer overlaps as edges.
- Then SGA cleans up the graph and removes spurious edges using several heuristics.
- Paths in the graph require paired-end read support. The algorithm 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). SGA removes any edges in the graph that do not support read pairs.
- In addition, SGA cleans up tips and bubbles in the read graph, which normally occur during *de novo* assembly, using standard graph cleaning methods.

Scaffolding Contigs to Assemble Long Reads

The next stage is scaffolding, which uses paired-end information to place and orient the contigs generated in the previous step, and fill in gaps between contigs. The method employed is based on the scaffolding method in the original SGA assembler[1].

Scaffolding is accomplished by realigning the input short reads to the contigs using the BWA aligner[3], and using the paired-end alignments to infer scaffold structure. The link between contigs is made when 2 or more paired reads map where Read 1 maps to one contig and Read 2 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 help guide and constrain the construction of our scaffold graph.

Gapfilling

The next step is to fill in scaffold gaps where possible to resolve repeats. In this step, the algorithm uses the input short reads and the FM index computed during the contig assembly. The algorithm begins by finding the highest scoring read that 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 is verification of the scaffolds and error correction. The short read data are again aligned against the scaffolds generated in the previous step using BWA aligner[3]. Based on the alignments, the scaffolds are corrected for single-nucleotide errors and if there is only partial alignment support, broken into smaller scaffolds. The algorithm also calculates quality scores for the final long reads 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 realigned, and reads that are overlapping or running into scaffold gaps are counted and computed. To determine whether 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 a few reads or none, the scaffold for the region is left as it is.

Q-scores

The algorithm generates a pileup file from the alignments of short reads to the scaffolds. This file 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, set Q-score to maximum 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 calculation 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. 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 10 Illumina General Contact Information

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

Table 11 Illumina Customer Support Telephone Numbers

Region	Contact Number	Region	Contact Number
North America	1.800.809.4566	Italy	800.874909
Australia	1.800.775.688	Netherlands	0800.0223859
Austria	0800.296575	New Zealand	0800.451.650
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 support.illumina.com/sds.html.

Product Documentation

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

Illumina
San Diego, California 92122 U.S.A.
+1.800.809.ILMN (4566)
+1.858.202.4566 (outside North America)
techsupport@illumina.com
www.illumina.com