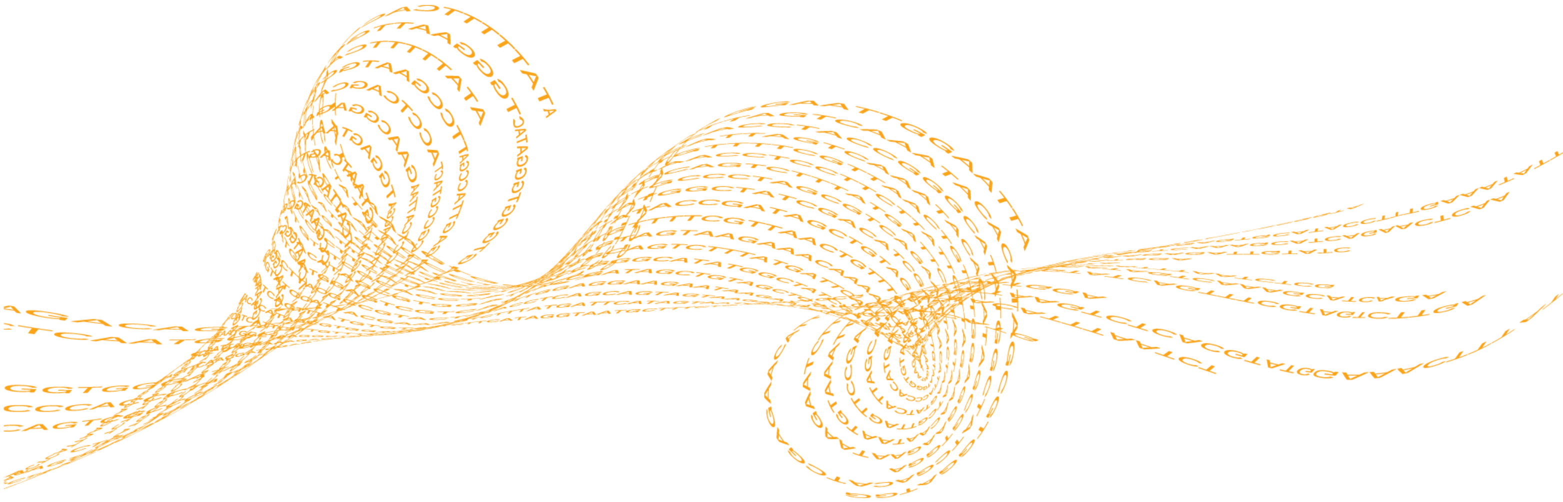
TruSeq Phasing Analysis App



ILLUMINA PROPRIETARY
15055852 Rev. B
December 2014

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

Revision History

Part #	Revision	Date	Description of Change
15055852	B	December 2014	Updated log files.
15055852	A	June 2014	Initial Release.

Introduction

Phasing analysis complements whole-genome sequencing (WGS) by distinguishing between variants originating from separate chromosomes. This analysis enables the generation of haplotype fragments for allele-specific analysis.

Illumina provides the TruSeq® Synthetic Long-Read DNA Library Prep Kit, which generates libraries for phased sequencing. The TruSeq Phasing Analysis App then provides phasing analysis by combining sequencing information from two sources:

- A long-read library generated with the TruSeq Synthetic Long-Read DNA Library Prep Kit, which provides the phasing information.
- A VCF file containing the variant calls from a conventional 30× WGS project of the same sample.

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

TruSeq Phasing Analysis Informatics Overview

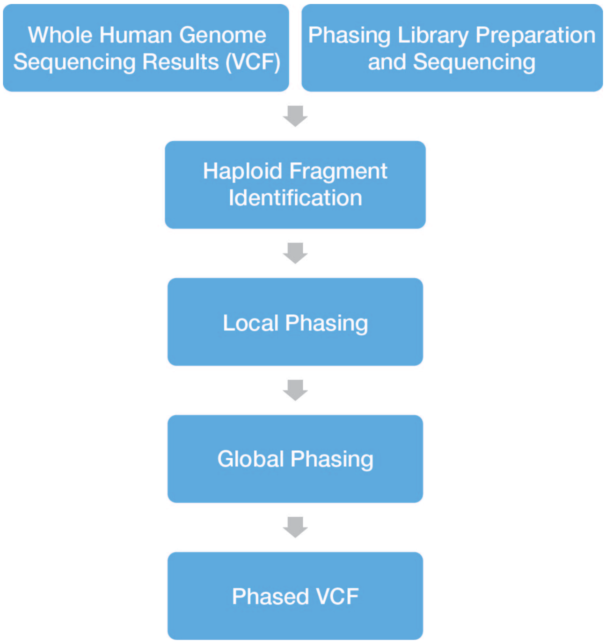
The TruSeq Phasing Analysis App consists of a suite of novel algorithms designed to provide whole-genome phasing information using the synthetic long-read sample preparation technology from Illumina. The TruSeq Phasing Analysis App uses the variants discovered from a previous deep whole-genome sequencing (WGS) run and the base calls from a long fragments library as input.

The first stage of the phasing algorithm identifies contiguous segments of haploid fragments, or clouds, in each well. In the next stage, local phasing, clouds with overlapping heterozygous SNP sites are joined, resulting in blocks that are normally around 5–10× larger in size than the input clouds. At the final stage of the analysis, global phasing, the local blocks that have been computed are phased using the 1000 Genomes phased reference panels to form long haplotype contigs.

The TruSeq Phasing Analysis App reports haplotype blocks across the genome and confidence scores for the phasing in a phased VCF file.

See *TruSeq Phasing Analysis Methods* on page 16 for a more detailed description.

Figure 1 Overview of TruSeq Phasing Analysis Algorithm Workflow


Whole Human Genome Sequencing Results (VCF)
Phasing Library Preparation and Sequencing
Haploid Fragment Identification
Local Phasing
Global Phasing
Phased VCF

Versions

The following module versions are used in the TruSeq Phasing Analysis app:

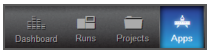
- Prism v2.2:
  - BEDTools-v2.17.0
  - BWA-0.7.5a
  - SAMtools-0.1.18
  - Tabix-0.2.5
  - VCFtools-0.1.10

Current Limitations

Before running the TruSeq Phasing Analysis app, be aware of the following limitations:

- Only reads generated using the TruSeq Synthetic Long-Read Library Prep Kit can be analyzed with this app.
- The short read length lower limit is 2×100 bp. The app supports the longest read lengths supported by the sequencing chemistry.
- If you use a VCF file from a whole-genome sequencing analysis done outside of BaseSpace, the file must be from a whole-genome sequencing run with > 30× coverage for optimal results. See *Input VCF Requirements* on page 7 for a description.
- Currently, only diploid human genomes are supported.
- Cancer samples are currently not supported.
- If you have less than 30 Gbp in sequence data, you have insufficient data for optimal results. The recommended upper limit is 115 Gbp.
- If there are < 1000 heterozygous SNVs in a single chromosome the TruSeq Phasing Analysis app exits.
- The app restricts input to the first 1.5 million reads per barcode.

Running TruSeq Phasing Analysis

1 Click the Apps button. 

2 Find **TruSeq Phasing Analysis** 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 Phasing Analysis 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.

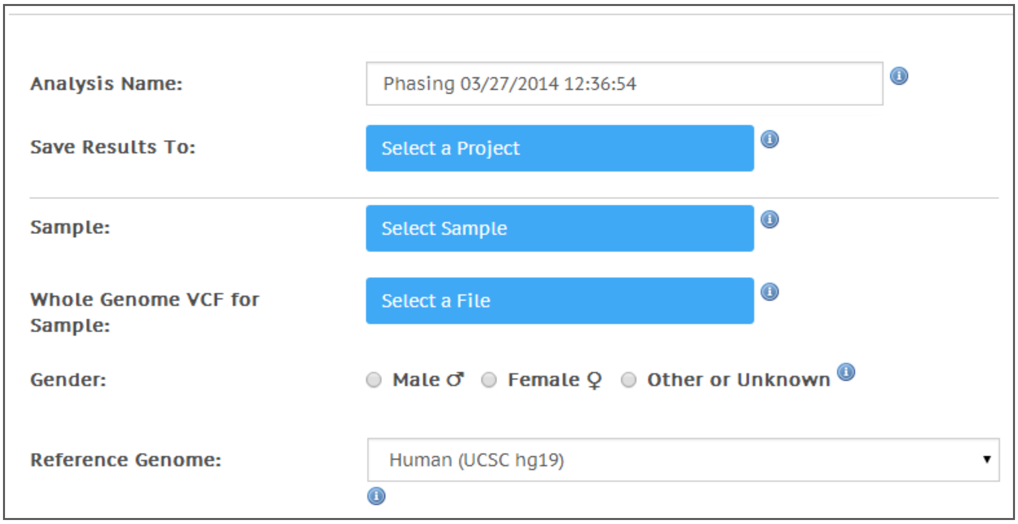
d **Whole Genome VCF for Sample**: Browse to and select the VCF file of your whole-genome sequencing run for the selected sample.
If you performed the whole-genome sequencing analysis outside of BaseSpace, upload the VCF file using the BaseSpace Import Analysis button. The VCF file must have the proper format; see *Input VCF Requirements* on page 7.

e **Gender**: Select the gender of the sample. These choices result in the following behavior:

- **Female**: Phase entire X chromosome (pseudoautosomal regions and non-pseudoautosomal regions).
- **Male**: Phase pseudoautosomal regions if they exist.
- **Unknown**: only phase chromosomes 1 through 22.

f **Reference Genome**: Select the reference genome. Currently, you can only use Human (UCSC hg19).

Figure 2 TruSeq Phasing Analysis Input Form



Analysis Name:	Phasing 03/27/2014 12:36:54
Save Results To:	Select a Project
Sample:	Select Sample
Whole Genome VCF for Sample:	Select a File
Gender:	○ Male ♂ ○ Female ♀ ○ Other or Unknown
Reference Genome:	Human (UCSC hg19) ▾

5 Click **Continue**.

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

Input VCF Requirements

If you use a variant call format (VCF) file from a whole-genome sequencing analysis done outside of BaseSpace, the input VCF file must adhere to VCF version 4.1. The file also has the following requirements:

- ***VCF file header and columns*** require the following characteristics:
  - Columns are in the following order, and separated from each other by tabs:
    `#CHROM POS ID REF ALT QUAL FILTER INFO`

    If these requirements are not met, the VCF Upload app does not upload the file. See also support.basespace.illumina.com.
- ***VCF file entries*** require the following characteristics:
  - The chromosome field (`CHROM`) begins with the string `chr` followed by the chromosome number `1-22`, `X`, `Y`, or `M`.
  - Annotations for the reference bases in the ref column match the UCSC hg19 reference; see genome.ucsc.edu/cgi-bin/hgTables.
  - Multi-sample VCF files are not allowed for analysis. If you do have a multi-sample VCF file, filter the file to include only a single sample of interest with the VCFtools package (vcftools.sourceforge.net) or another tool.
  - 16,000 or more passing variants, else there are insufficient variants to perform analysis.
  - If there are < 1000 heterozygous SNVs in a single chromosome the TruSeq Phasing Analysis app exits.

  If these requirements are not met, the TruSeq Phasing Analysis app does not proceed.
- ***Variants*** in the VCF file require the following characteristics; if not, the TruSeq Phasing Analysis app skips the variant for phasing analysis. The analysis proceeds with all variants that adhere to the rules.
  - Entries are labeled as `1-22`, `X`, `Y`, or `M` only in the `CHROM` column. For example, `chr1` is valid whereas `chr6_cox_hap2` is not.
  - Variants in the `ref` and `alt` columns are 50 bases long or less.
  - Variants in the `ref` and `alt` columns only contain ACGT characters.
  - Only bi-allelic variants are allowed. Haploid calls and multi-allelic entries are discarded.
  - Variants have a `PASS` label in the filter column.
  - Variants have information in the `GT` field.

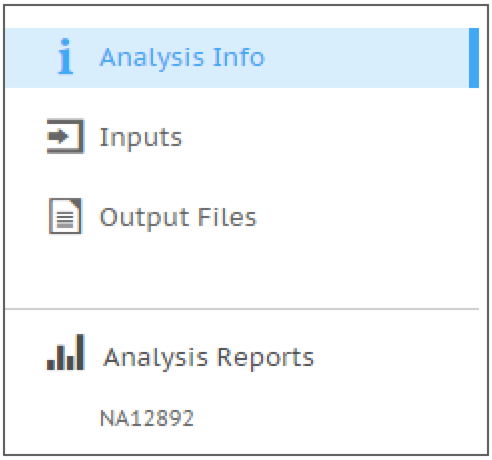
When your VCF file has the proper format, you can upload it to BaseSpace. For more information, see the Import Analysis section in the *BaseSpace User Guide*.

NOTE
If the VCF file passes the variant filtering criteria but has less than 1 million SNPs and indels after filtering, the TruSeq Phasing Analysis app issues a warning. VCF files with less than 1 million passing variants are rare for a human whole-genome sequencing run with depth of coverage of at least 30×. Low numbers of variants indicate problems with the sample, sequencing, or analysis.

TruSeq Phasing Analysis Output

This chapter describes the output that the TruSeq Phasing Analysis app produces. To go to the results, click the **Projects** button, then select the project, then the analysis.

Figure 3 TruSeq Phasing Analysis 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 8 for a description.
- **Analysis Info**—An overview of the app session settings.
  See *Analysis Info* on page 10 for a description.
- **Inputs**—An overview of input settings.
  See *Inputs Overview* on page 11
- **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:
  - Phased.vcf.gz—Variant file in VCF 4.1 format with markup to indicate phased variants and haplotype blocks.
    See *Phased.vcf.gz* on page 15 for more information.
  - Unphased.vcf.gz—Variants file in VCF4.1 format listing the variants of your sample before phasing.
    See *Unphased.vcf.gz* on page 15 for more information.

  The output files contain two reports:
  - PhasingSummaryReport.pdf—PDF report detailing sample information, sequencing metrics for phasing library and phasing summary statistics.
    See *PhasingSummaryReport.pdf* on page 11 for more information.
  - LibraryCharacteristics.csv—Comma-separated file describing the library characteristics.
    See *LibraryCharacteristics.csv* on page 13 for more information.

  In addition, there are files that the app uses to generate other plots and reports: Frac_phased_raw, Mintl_raw, and Stats.xml.
    See *Output Files Overview* on page 11 for more information.

Analysis Reports

Under Analysis Reports in the left navigation panel, you find links to Analysis Reports for each sample. The HTML reports contain an overview of the metrics of the long fragments library and statistics from the phasing analysis. In addition, they provide links to the associated PDF Summary Report, and the phased VCF file.

The following statistics are reported:

Table 1 Well Metrics

Metric	Description
Reads PF	Total number of short read pairs that pass filter.
Cloud N50 Length	N50 value of long fragments/clouds identified across all barcodes by aligning to the hg19 reference genome. The N50 is the length for which the collection of all blocks or contigs of that length, or longer, contains half of the total bases included in blocks/contigs.
Cloud N50 Yield	N50 yield in bases of long fragments/clouds identified across all barcodes by aligning to the hg19 reference genome.

Table 2 Phasing Metrics

Metric	Description
Number of All Variants	Total number of indels and single nucleotide polymorphisms (SNVs) used in the phasing analysis.
Number of SNVs	Total number of SNVs used in the phasing analysis.
Number of Het SNVs	Total number of heterozygous SNVs using in the phasing analysis.
Percent Inconsistency (ICF)	Percentage of variants reflecting the positions where the alleles detected in the long fragment library do not match the expected alleles from the input WGS VCF file.
Cloud N50 Length	N50 value of long fragments/clouds identified across all barcodes by aligning to the hg19 reference genome.
Cloud Mean Coverage	Genome coverage depth by long fragment/cloud identified by aligning to the reference genome.
Percent Phased Locally	The overall percentage of heterozygous variants included in locally phased blocks, and derived by parsing the PS value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.
Local Block N50 Length	N50 value of locally phased blocks generated without imputation from the second step of the analysis, and derived by parsing the Local Phase Set (LPS) value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.
Percent Phased Globally	The overall percentage of heterozygous variants included in globally phased blocks, and derived by parsing the PS value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.
Global Block N50 Length	N50 value of globally phased blocks generated using imputation from the third step of the analysis, and derived by parsing the Phase Set (PS) value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.

Table 3 Gene Phasing Statistics

Metric	Description
100% Variants Phased	Number of autosomal genes on chromosomes 1–22 that have 100% of their variants with EL value ≥ 0.95 phased.
>70% Variants phased:	Number of autosomal genes on chromosomes 1–22 that have at least 70% of their variants with EL value ≥ 0.95 phased.

The report file also provides two plots illustrating gene phasing metrics:

- **Phasing completeness by gene**—This graph provides the distribution of the fraction of variants phased per gene, over all autosomal RefSeq genes. This metric considers all variants phased across the gene, regardless of the number of haplotype blocks that span the gene.
- **End-to-End Phasing Confidence**—This graph provides the distribution of the minimum Transition Likelihood (TL) value within each autosomal gene in RefSeq. Phased blocks break when the TL falls below a selected accuracy threshold. Therefore, the minimum TL value per gene indicates our confidence that the gene is phased in one block, end-to-end, at a given accuracy threshold. Usually, most genes are phased in one block, indicated by the peak observed at a TL value of 1.

Analysis Info

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

A brief description of the metrics is below.

Table 4 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.

- **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.
- **TimingStats.txt**: Contains timing statistics about the algorithm of the app.

- **uploader-{timestamp}.log**: File logging uploading.
- **UsageLog.txt**: Shows system resources logging information.
- **Warnings.txt**: Log file with warnings, which do not stop analysis.
- **WorkflowLog.txt**: Workflow standard output (contains details about workflow steps, command line calls with parameters, timing, and progress).
- **WorkflowOut.txt**: Shows the raw console output.

Inputs Overview

The TruSeq Phasing Analysis app provides an overview of the input samples and settings that were specified when setting up the TruSeq Phasing Analysis run.

Table 5 Inputs Overview

Metrics	Description
Analysis Name	Analysis name. Default name is the app name with the date and time the app session was started.
Input	Sample that is analyzed.
Gender	The gender of the sample.
Whole Genome VCF	The whole-genome VCF of the sample that is analyzed.
Reference Genome	The reference genome.
Save Results To	The project that stores the app results.

Output Files Overview

The TruSeq Phasing Analysis app generates the following main output files:

- *PhasingSummaryReport.pdf* on page 11
- *LibraryCharacteristics.csv* on page 13
- *Phased.vcf.gz* on page 15
- *Unphased.vcf.gz* on page 15

In addition, there are a number of files that the app uses to generate the report pages:

- Frac_phased_raw.txt: raw data file that is used to generate the *Phasing completeness by gene* plot.
- Mintl_raw.txt: raw data file that is used to plot the *End-to-end phasing confidence* plot.
- Stats.xml: xml file that is used to generate the PDF report.

PhasingSummaryReport.pdf

When you click the Output File link, you have access to the PhasingSummaryReport.pdf file. This report contains an overview of the metrics of the long fragments library and statistics from the phasing analysis. The following statistics are reported:

Sequencing Metrics

Table 6 Sequencing Metrics

Metric	Description
Total Reads (Pass Filter)	Total number of short read pairs that pass filter.
Percent Reads Mapped	Percentage of short reads across all barcodes that align to the hg19 reference.
Cloud N50 Length	N50 value of long fragments/clouds identified across all barcodes by aligning to the hg19 reference genome. The N50 is the length for which the collection of all blocks or contigs of that length, or longer, contains half of the total bases included in blocks/contigs.

Phasing Statistics

Table 7 Phasing Statistics

Metric	Description
Local Block N50 Length	N50 value of locally phased blocks generated without imputation from the second step of the analysis, and derived by parsing the Local Phase Set (LPS) value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.
Global Block N50 Length	N50 value of globally phased blocks generated using imputation from the third step of the analysis, and derived by parsing the Phase Set (PS) value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.
Percent Phased Globally	The overall percentage of heterozygous variants included in globally phased blocks, and derived by parsing the PS value. The value is reported separately for haplotype blocks considering SNVs alone and both SNVs and indels.

Gene Phasing Statistics

For all autosomal genes in RefSeq, we report the fraction of phased variants with EL value ≥ 0.95 that intersect the coordinates of the RefSeq genes.

Table 8 Gene Phasing Statistics

Metric	Description
100% Variants Phased	Number of autosomal genes on chromosomes 1–22 that have 100% of their variants with EL value ≥ 0.95 phased.
>70% Variants phased:	Number of autosomal genes on chromosomes 1–22 that have at least 70% of their variants with EL value ≥ 0.95 phased.

The report file also provides two plots illustrating gene phasing metrics:

- **Phasing completeness by gene**—This graph provides the distribution of the fraction of variants phased per gene, over all autosomal RefSeq genes. This metric considers all variants phased across the gene, regardless of the number of haplotype blocks that span the gene.
- **End-to-End Phasing Confidence**—This graph provides the distribution of the minimum Transition Likelihood (TL) value within each autosomal gene in RefSeq.

Phased blocks break when the TL falls below a selected accuracy threshold. Therefore, the minimum TL value per gene indicates our confidence that the gene is phased in one block, end-to-end, at a given accuracy threshold. Usually, most genes are phased in one block, indicated by the peak observed at a TL value of 1.

LibraryCharacteristics.csv

The LibraryCharacteristics.csv file provides metrics that help determine the quality of your synthetic long-read library. This report is accessible through the Output File link. This topic details the metrics reported in this file.

Table 9 Phasing Metrics

Metric	Definition
Number of All Variants	Total number of indels and single nucleotide polymorphisms (SNVs) used in the phasing analysis.
Number of SNVs	Total number of SNVs used in the phasing analysis.
Number of All Het Variants	Total number of heterozygous indels and SNVs used in phasing analysis.
Number of Het SNVs	Total number of heterozygous SNVs using in the phasing analysis.
Percent Inconsistency (ICF)	Percentage of variants reflecting the positions where the alleles detected in the long fragment library do not match the expected alleles from the input WGS VCF file.
Cloud N50 (bp)	N50 value of long fragments/clouds identified across all barcodes by aligning to the hg19 reference genome. The N50 is the length for which the collection of all blocks or contigs of that length, or longer, contains half of the total bases included in blocks/contigs.
Cloud Fold Coverage of Genome	Genome coverage depth by long fragment/cloud identified by aligning to the reference genome.
Percentage of SNVs Phased Locally	The overall percentage of heterozygous SNVs included in locally phased blocks, and derived by parsing the LPS value.
Local SNV Block N50 Length (bp)	N50 value of locally phased blocks generated without imputation from the second step of the TruSeq Phasing Analysis, and derived by parsing the Local Phase Set (LPS) value. The value is reported for haplotype blocks considering only SNVs.
Percentage of SNVs Phased Globally	The overall percentage of heterozygous SNVs included in globally phased blocks, and derived by parsing the PS value.
Global SNV Block N50 Length (bp)	N50 value of globally phased blocks generated using imputation from the third step of the TruSeq Phasing Analysis, and derived by parsing the Phase Set (PS) value. The value is reported for haplotype blocks considering only SNVs.

The following tables refer to metrics that are reported per barcode and help troubleshoot or QC library preparation.

Table 10 Well Metrics—Summary Statistics

Metric	Definition
n	Number of wells (barcodes) 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), Cloud Yield, 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 11 Well Metrics—Details

Metric	Definition
Reads PF	Number of short read pairs that pass the sequencing quality filter
Percent Reads Aligned	Percentage of short read aligned against the reference genome
Percent End Marker	Percentage of short reads containing end marker sequence tag
Number Unique Fragments Per Well (End Markers)	Number of unique long fragments estimated by K-mer analysis of end marker containing sequence
Mismatch Rate Read 1	Percentage of bases mismatched on short Read 1 against the reference genome
Mismatch Rate Read 2	Percentage of bases mismatched on 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)	Sum of DNA bases from all unique long fragments identified by aligning to the reference genome
Cloud Yield	Number of base-pairs in unique long fragments estimated by aligning to the reference genome
Cloud N50 Length (bp)	Length N50 of long fragments identified by aligning to the reference genome
Cloud Length Average (bp)	Average length of long fragments identified by aligning to the reference genome
Cloud Length Span90(bp)	Length Span90 of long fragments identified by aligning to the reference genome

Phased.vcf.gz

The principal results of the TruSeq Phasing Analysis are returned in the phased VCF file. This file recapitulates most of the information in the input VCF file, but the phased VCF file differs from the input VCF in the following way:

- The presence of additional markup providing information about the phase of variants, our confidence in the phase assignments, and on sets of phased heterozygous variants. Phased variants in the phased VCF file (*phased.vcf.gz) contain a modification in the genotype (GT) field and also have five additional INFO fields. These fields are described in the table in this topic. Heterozygous variants that could not be phased are output in the final VCF file unchanged.
- The variants excluded from analysis as described in *Input VCF Requirements* on page 7

Phased variants in the phased VCF file (*phased.vcf.gz) contain a modification in the genotype (GT) field and also have five additional INFO fields, which are described in the following table.

Table 12 Phased Variants

Metric	Description
Genotype (GT)	Phased variants (heterozygous and homozygous) contain a modified GT field, using a pipe symbol (|) instead of forward slash in accordance with VCF 4.1 specifications. Variants whose genotype contains a / in the phased VCF are unphased.
Emission Likelihood (EL)	An indication of the degree of confidence that the phasing of one locus is correct relative to its neighbors. Although evaluation of quantitative correctness is still ongoing, the value can be interpreted as a likelihood. Values range from 0.5 (a random guess) to 1 (complete confidence). See *Phase Scoring* on page 18 for a more detailed description.
Transition Likelihood (TL)	The confidence in the phasing of variants between the current (heterozygous) variant locus and the previous heterozygous locus. Although evaluation of quantitative correctness is still ongoing, the value can be interpreted as a likelihood. Values range from 0.5 (random guess) to 1 (complete confidence). A more detailed description is provided in the section *Phase Scoring* on page 18.
Phase Set (PS)	Used in the TruSeq Phasing Analysis App to define the heterozygous variants that belong to a globally phased block set with TL > 0.95. The integer value represents the coordinate of the first locus in a given phase set. PS values are unique within a chromosome locus. Therefore, variants with the same PS value on different chromosomes are not phased relative to each other.
Local Phase Set (LPS)	Interpretation is analogous to that of PS, except that the phasing indicated by LPS annotation is based solely on the blocks determined by local phasing. In general, confidence in the relative phasing of variants that share the LPS is higher than the confidence for variants that share a PS value but not an LPS value.
Inconsistency Flag (ICF)	An indication of positions where the alleles detected in the long fragment library do not match the expected alleles from the input WGS VCF file.

Unphased.vcf.gz

The original VCF file contains a subset of variants from the input VCF that pass the filtering criteria. No additional phasing information is provided in this file.

TruSeq Phasing Analysis Methods

The TruSeq Phasing Analysis App works in three steps:

- ***Identification of Haploid Fragments***: The TruSeq Phasing Analysis App separates the sequence reads into the component 384 wells based on the barcode sequence and aligns the reads to the human reference sequence. The algorithm then detects haploid fragments of DNA, or *clouds*, in each well, and records the overlapping heterozygous variants called in the input WGS VCF file. Clouds that indicate an overlap of fragments from both maternal and paternal chromosomes are removed.
- ***Local Phasing***: The resulting fragments (up to ~10 kb long) are pooled together and longer haplotypes reconstructed by chaining together fragments that share heterozygous SNPs. The resulting blocks, referred to as locally phased blocks, are on average 5–10× longer than the individual clouds and provide highly accurate haplotype blocks derived entirely from the data.
- ***Global Phasing***: In the final step of the algorithm, the locally phased blocks are phased relative to one another to reconstruct longer haplotypes by statistical imputation using the 1000 Genomes phased reference panels. Final block sizes are normally increased another 5–10× relative to locally phased blocks. Phased blocks and confidence scores are output in the phased VCF file.

This chapter describes these three steps, and explains the phase scoring scheme that the TruSeq Phasing Analysis app uses. See also the explanation in the following PDF: www.nature.com/nbt/journal/v32/n3/extref/nbt.2833-S1.pdf.[1]

Identification of Haploid Fragments

The first stage of the phasing algorithm involves the identification of contiguous segments of haploid fragments, or clouds, in each well. The reads from each well are aligned to the human reference genome using the BWA aligner[2] and clustered into groups separated from each other by at least 2 Kbp. Reads that cluster according to this criterion are likely to have originated from the same fragment. The heterozygous variants from the WGS VCF that intersect the long fragments are retained as input to the subsequent steps of the phasing pipeline. The following fragments are removed from consideration:

- Fragments with insufficient sequencing read coverage
- Fragments with many variant-covering base calls with low Q-scores
- Fragments with evidence of two different alleles in the same fragment
- Fragments with inferred length in the top three percentiles

The result of this stage is a set of *clouds* from every well. Each cloud corresponds to an interval on the reference genome and the base calls of reads that overlap heterozygous variant calls from the input VCF. A cloud provides a tentative phasing of the set of variants that map to the original long fragment.

Reference Genome Definition

The phasing pipeline currently uses hg19 from UCSC as a reference genome (genome.ucsc.edu). The chromosome naming scheme follows the UCSC conventions of chr1-22, chrX, chrY, chrM. The pseudoautosomal region (PAR) of the Y chromosome is masked out with N's. The result of this is that any mappings occurring in the PAR region map to the X chromosome. Currently, only the assembled chromosomes and mitochondria are used in the reference; none of the non-mapped contigs are included, nor are alternate haplotypes.

The hg19 PAR regions are defined as follows:

Table 13 hg19 PAR Regions

Name	Chr	Start	Stop
PAR #1	X	60,001	2,699,520
PAR #2	X	154,931,044	155,260,560
PAR #1	Y	10,001	2,649,520
PAR #2	Y	59,034,050	59,363,566

Local Phasing

In the next stage of the pipeline, the clouds are combined across wells. Clouds with overlapping heterozygous SNP sites are joined, resulting in blocks that are normally around 5–10× larger in size than the input clouds. Each of these blocks is analyzed to infer the phasing of the variants it covers.

The local phasing algorithm finds the haplotypes and cloud-to-haplotype assignments that best fit the data (the maximum likelihood answer), using an algorithm based on Hidden Markov Models (HMM). The algorithm works iteratively, by finding and scoring the best partial solution for positions 1 through k for each possible choice of phasing and cloud-to-haplotype assignments at position k. The partial solution is based on the best partial solutions at positions 1 through k-1.

Phasing and cloud assignments at position k are scored by considering the quality scores of the alleles in the clouds at k and also requiring that cloud assignments are 'consistent'.

For a cloud assignment to be consistent, we require that fragments that span positions k and k-1 must be assigned to the same haplotype. Ignoring the cloud assignments for simplicity, the best possible phasing of positions 1 through k such that position k is phased 0|1 is found (chromosome A | chromosome B). The best possible phasing of positions 1 through k such that position k is phased 1|0 is also is found.

The best phasing of 1..k with phasing 0|1 at position k is obtained by comparing extensions of the following:

- The best solution for 1..k-1 with 0|1 phasing at k-1
- The best solution for 1..k-1 with 1|0 phasing at k-1.

Thus, each step, is built on the optimal partial solutions found at the previous step. When analysis has finished with the final position N, the local phasing algorithm compares the best solution with 0|1 phasing at N against the best solution with 1|0 phasing at N. The algorithm then determines the optimal phasing. The optimal phasing of the final position also aids in determining the optimal phasing of all positions.

Global Phasing

At the next stage of the pipeline, the local blocks that have been computed are phased relative to each other using a statistical algorithm to form long haplotype contigs.

The global phasing algorithm is based on a hidden Markov model (HMM) which extends statistical phasing algorithms such as IMPUTE2[3] and SHAPE-IT[4]. In brief, the idea is to view the locally phased blocks as an *imperfect mosaic* from a reference panel of pre-phased haplotypes from the 1000 Genomes project[5]. The algorithm assigns the most likely phase to each locally phased block. This assignment is based on the best matches of the local blocks to the individuals in the reference panel and patterns of linkage disequilibrium inferred statistically from the panel.

In addition to finding the optimal phase, the confidence scores (Emission Likelihood (EL) and Transition Likelihood (TL)) are derived. These scores represent the likelihood of a switch or single-site error between the current SNP and the previously phased SNP. Each heterozygous variant that is given a TL score < 0.95 causes the previous phase set, if any, to end and typically starts a new phase set. An exception to this rule occurs when breaking a block at a variant with TL < 0.95 results in a singleton block. In other words, if the next heterozygous locus also has TL < 0.95 or if the current heterozygous locus is the last one of the chromosomes. In this case, the preceding phase set is ended but no Phase Set is output for the singleton variant.

The scores can also be used to break the inferred haplotypes into blocks at a given level of accuracy. For applications requiring highly accurate haplotype blocks, haplotype contigs can be broken at a higher threshold, resulting in shorter but more accurate blocks. Alternatively, a more-aggressive reconstruction can be obtained by breaking blocks at a lower threshold.

Phase Scoring

Two scores communicate the confidence in phasing given to each phased heterozygous variant, specified as the Emission Likelihood (EL) and Transition Likelihood (TL) values in the phased VCF. The scores indicate two different modes of possible error in phasing. The EL score addresses the possibility of a *single-site error* at a locus. In other words, this score addresses the possibility that the inferred phasing of one locus is wrong relative to its neighbors. The TL score addresses the possibility of a *switch error*. In other words, the relative phase of variants before a given locus is wrong compared to the phase of variants at and after the locus.

The following examples illustrate these two types of errors.

Accurate Haplotypes

Figure 4 Haplotype alleles arbitrarily labeled 0 & 1

```
0000000000000000000000000 haplotype 1
1111111111111111111111111 haplotype 2
```

Single-Site Errors

Figure 5 Indicated by *; Probability represented by Emission Likelihood

```
0000001000000000000100000 haplotype 1
1111110111111111111011111 haplotype 2
      *            *
```

Switch Error

Figure 6 Indicated by |; Probability represented by Transition Likelihood

```
0000000000000011111111111 haplotype 1
1111111111111100000000000 haplotype 2
              |
```

Intended Use of Phase Scoring

The reason for separating the Emission Likelihood and Transition Likelihood scores is that doing so permits the algorithm to assign scores to individual loci, rather than all pairs of loci. The scores represent the fact that a given locus can be weakly phased but confidence in the phasing between variants separated by that locus can remain high. This situation might arise if a locus has low coverage and/or poor quality base calls while clouds spanning the variant also connect other variants with better coverage. In this case, you can have low single-site confidence, and thus a low EL score, while a high TL signals that confidence in phasing spanning these loci is still good.

These scores can help answer the following questions:

- How confident can we be in the reported relative phasing of two adjacent sites? (Two sites are ‘adjacent’ for this purpose if they are on the same chromosome and there are no intervening loci for which phasing is reported.)
- How confident can we be in the reported relative phasing of two distal sites?

The first question is addressed through calculating the combined probability (PQ1) of a switch error between the two (TL value of the second site) and a single-site error at either locus (EL values of both sites). Thus, for loci i and i+1, the relative phasing confidence can be estimated as follows:

$$P_{Q1} = TL_{i+1} * EL_i * EL_{i+1}$$

The second question is addressed through calculating the combined probability (PQ2) of switch errors anywhere between the two loci (TL value at the two loci and between) and no single-site error at either locus (EL value at the two sites). Thus, for loci i and i+n, the relative phasing confidence can be estimated as follows:

$$P_{Q2} = TL_i * TL_{i+1} * TL_{i+2} *\ldots* TL_{i+n} * EL_i * EL_{i+n}$$

Using Phase Scoring to Extract Globally Phase Blocks

In the provided phased VCF files, the default Transition Likelihood (TL) cutoff of TL ≥ 0.95 used for the determination of Phase Set (PS) annotations is a reasonable starting point for defining haplotype blocks: all variants on a given chromosome sharing the PS value.

NOTE
The Emission Likelihood (EL) score is not taken into account in determining PS. Therefore, a locus with a low EL score generally receives a PS annotation even though it is not confidently phased relative to the other members of the phase set. To determine whether two variants can be confidently phased relative to one another:

- Check whether the two loci both have PS annotations with the same value
- Check whether both loci have EL scores ≥ 0.95.

Some applications require greater confidence or, alternatively, more aggressive (less confident) phasing. To extract phased blocks using alternative probability cutoffs, you can use the TL and EL scores together to parse Globally Phased Blocks from the VCF file. Consider a set of loci on the same chromosome to be phased relative to one another provided the following:

- All SNPs have TL scores at or above an alternative cutoff except for the first locus in the set. This alternative cutoff is relatively high for greater confidence and relatively lower for more aggressive phasing.
- EL scores are at or above an alternative cutoff (relatively high for greater confidence and more aggressive phasing) for all loci in the set.

References

1 Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., Snyder, M. Whole-genome haplotyping using long reads and statistical methods. Nat. Biotechnol. 32, 261–266 (2014).

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

3 Howie, B.N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5, e1000529 (2009).

4 Delaneau, O., Zagury, J. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Meth 10, 5–6 (2012).

5 McVean, G.A. et al. An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56–65 (2012).

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 14 Illumina General Contact Information

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

Table 15 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