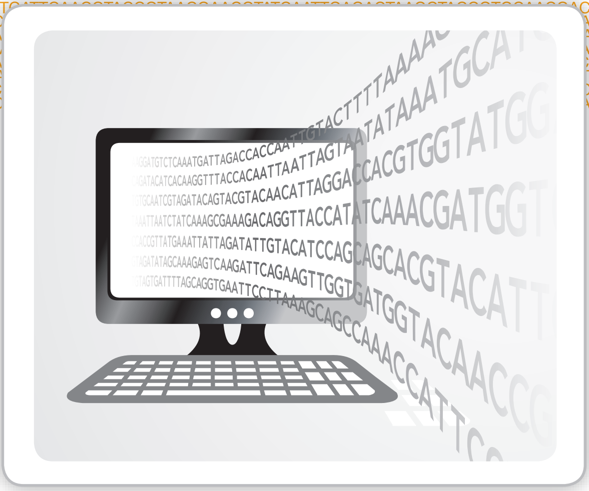
illumina®

FastTrack Phasing Analysis Services Guide



For Research Use Only. Not for use in diagnostic procedures.

ILLUMINA PROPRIETARY

Document # 15047619 v01
September 2015

Customize a short end-to-end workflow guide with the Custom Protocol Selector
support.illumina.com/custom-protocol-selector.html

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).

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

Document	Date	Description of Change
Document # 15047619 v01	September 2015	Corrected Phased VCF, Unphased VCF, and Phasing Summary Report file names in the Analysis Deliverables chapter.
Part # 15047619 Rev. B	June 2015	Revised documentation to reflect changes in version 1.1.0 of the Illumina FastTrack Phasing Analysis Service pipeline.
Part # 15047619 Rev. A	November 2013	Initial Release.

Table of Contents

Chapter 1

Getting Started

Phasing Analysis Service

The Phasing Analysis Service leverages TruSeq® Synthetic Long-Read technology to complement whole-genome sequencing. Phased sequencing captures the unique content of homologous chromosomes, including mutations that can differ across chromosome copies. Phase information distinguishes between variants originating from separate chromosomes, enabling the generation of haplotype fragments for allele-specific analysis.

The FastTrack Phasing Informatics Pipeline uses algorithms designed to provide whole-genome phasing information. TruSeq Synthetic Long-Read uses the variants discovered from a Whole-Genome Sequencing (WGS) run and the base calls from a TruSeq Synthetic Long-Read Library Prep as inputs. The FastTrack Phasing Informatics Pipeline then reports haplotype blocks across the genome and phasing confidence scores in a phased VCF file.

The purpose of this document is to help you understand the Phasing Analysis Service data package you receive from Illumina. This document provides an overview of the FastTrack Phasing Informatics Pipeline 1.1.0. deliverables and the Prism v2.2 phasing algorithm.

Data Delivery

Illumina FastTrack Services currently provides data delivery through the following choices.

Illumina Hard Drive Data Delivery

Illumina FTS ships data on 1 or more hard drives. The hard drives are formatted with the NTFS file system and can optionally be encrypted.

The data on the hard drive are organized in a folder structure with 1 top-level folder per sample or analysis.

Illumina Cloud Data Delivery

Illumina FTS uploads data to a cloud container. Illumina currently supports uploads to the Amazon S3 service. Upload data are organized per upload batch by date with an Illumina_FTS prefix. For example, a sample in a batch uploaded on February 1, 2014 would be found in the container with the prefix Illumina_FTS/20140201/SAMPLE_BARCODE. Contact your FastTrack Services project manager to enable cloud delivery.

Chapter 2

Analysis Deliverables

Overview

This section details the files and folder structure for the Phasing Analysis Service results. The files and folders are named based on the unique sample identifiers. Usually, these unique identifiers are the barcodes associated with the samples in the lab, but can be a known sample ID for reference samples.

Result Folder Structure

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

- 📁 [SampleBarcode]/
  - 📁 Phased_results—Contains all the output files resulting from the phasing library sequencing run and analysis.
    - 🗎 [SampleName]_Phased.vcf.gz—Variant file in VCF 4.1 format with markup to indicate phased variants and haplotype blocks.
    - 🗎 [SampleName]_Unphased.vcf.gz—Variants file in VCF 4.1 format listing the variants of your sample before phasing.
    - 🗎 [SampleName]_PhasingSummaryReport.pdf—PDF report detailing sample information, sequencing metrics for phasing library, and phasing summary statistics.
    - 🗎 md5sum.txt—Checksum file to confirm file consistency.
    - 📁 [SampleName]_ShortInsertSequencing.tar.gz—Contains the original short read data for the phasing library.
  - 📁 WGS_results—Folder containing the results of the whole-genome sequencing pipeline. Refer to the folder structure described in the *Whole-Genome Sequencing Services User Guide, document # 15040892*.

Result Files

Phased VCF File

The [SampleName]_Phased.vcf.gz file contains the principal results of the FastTrack Phasing Informatics Pipeline. This file includes most of the information in the input VCF file, which is generated from the FastTrack Services Whole-Genome Sequencing Informatics Pipeline. For more information, see the *Whole-Genome Sequencing Services User Guide, document # 15040892*.

The phased VCF file differs from the input VCF in the following ways:

- Contains additional markup about the phase of variants, our confidence in the phase assignments, and on sets of phased heterozygous variants. Heterozygous variants that could not be phased are unchanged in the final VCF file.
- Does not include the variants excluded from analysis, which are based on the following criteria:
  - Indels > 50 bp in length
  - Nonpassing variants
  - Haploid
  - Nonvariant
  - Variants with nonstandard VCF formatting

The following table describes fields in the phased VCF file.

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 slash (/) in the phased VCF are unphased.
Emission Likelihood (EL)*	An indication of the degree of confidence that the phasing of 1 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).
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).
Phase Set (PS)	Used in the FastTrack Phasing Informatics Pipeline 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.

*For more information, see the Phase Scoring *Overview* on page 17.

Unphased VCF File

The [SampleName]_Unphased.vcf.gz file contains a subset of variants from the input VCF that pass the filtering criteria. No additional phasing information is provided in this file. For more information, see the *Whole-Genome Sequencing Services User Guide, part # 15040892*.

Phasing Summary Report

The [SampleName]_PhasingSummaryReport.pdf file contains an overview of the metrics of the long fragments library and statistics from the phasing analysis.

Input VCF Snapshot

This section describes details related to the input whole-genome sequencing VCF file.

Metric	Description
Reference Genome	Version of human reference genome used in the analysis.
Number of Variants	Total number of indels and single nucleotide variants (SNVs) used in the phasing analysis.
Number of heterozygous variants	Total number of heterozygous indels and SNVs used in the phasing analysis.
Number of heterozygous SNPs	Total number of heterozygous SNVs used in the phasing analysis.

Sequencing Metrics

This section describes details related to the sequencing of reads from a TruSeq Synthetic Long-Read Library Prep.

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

This section describes details related to the phasing of variants.

Metric	Description
Local N50	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 N50	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	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.

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 2 plots illustrating gene phasing metrics:

- **Phasing Completeness by Gene**—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**—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 1 block, end-to-end, at a given accuracy threshold. Usually, most genes are phased in 1 block, indicated by the peak observed at a TL value of 1.

Short Read Output Folder

The [SampleName]_ShortInsertSequencing.tar.gz folder contains the short read output from the long fragments library sequencing run. The output files are in FASTQ format and are demultiplexed using the sample barcode, allowing a 1-base mismatch in the barcode sequence. The presence of end markers of the 5′–3′ sequence TACGCTTGCAT in short read sequences indicate 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, except where the sequence TACGCTTGCAT is a native part of your sample DNA.

Data Integrity

The md5sum.txt file is provided to check the integrity of the sample files and folders. Immediately after sample quality check, the md5sums, or compact digital fingerprint, for every file in the directory tree are generated. If media failures compromise data integrity, you can use the md5sum tool to find the inconsistencies. Use the tool to compare the hash from the provided md5sum file to the hash generated from the downloaded file.

On a Unix system, you can use the following commands to perform an md5sum check, assuming the utility is installed:

```
% cd [Sample_Barcode]
% md5sum -c md5sum.txt
```

The check verifies every file in ~30–45 minutes. Any errors are listed in the output.

In Windows, there are various command line and GUI tools available to perform an md5sum check. The Cygwin tools provide a utility identical to Linux.

Analysis Overview

Chapter 3

Overview

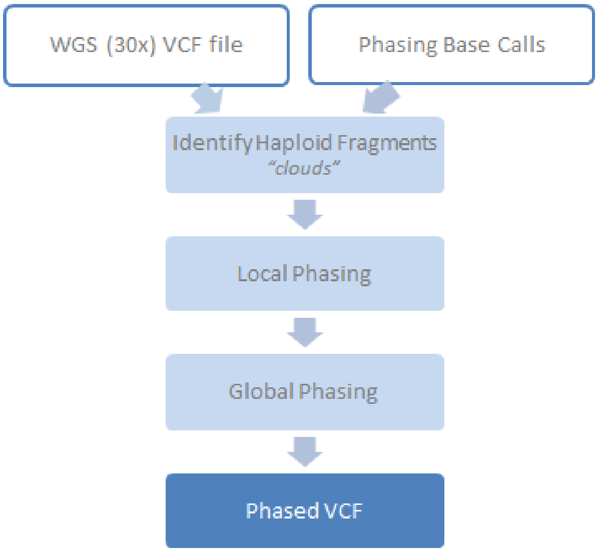
The FastTrack Phasing Informatics Pipeline works in 3 steps.

1 **Identification of Haploid Fragments**— The FastTrack Phasing Informatics Pipeline 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.

2 **Local Phasing**—The resulting fragments (up to ~10 kb long) are pooled together. Fragments that share heterozygous SNPs are chained together to reconstruct longer haplotypes. On average, these locally phased blocks are 5–10× longer than the individual clouds and provide highly accurate haplotype blocks derived entirely from the data.

3 **Global Phasing**—Using the 1000 Genomes phased reference panels for statistical imputation, the locally phased blocks are phased relative to each other to reconstruct longer haplotypes. Final block sizes are increased another 5–10× relative to locally phased blocks. Phased blocks and confidence scores are output in the phased VCF file.

Figure 1 FastTrack Phasing Informatics Algorithm Workflow



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[1] 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. Fragments with the following criteria are removed from consideration:

- Insufficient sequencing read coverage
- Many variant-covering base calls with low Q-scores
- Evidence of 2 different alleles in the same fragment
- Inferred length in the top 3 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. Therefore, 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 nonmapped contigs are included, nor are alternate haplotypes.

Table 1 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 are assigned to the same haplotype. Ignoring the cloud assignments for simplicity, we find the best possible phasing of positions 1 through k such that position k is phased 0|1 or 1|0 (chromosome A | chromosome B).

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

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

Using a statistical algorithm, the local blocks that have been computed are phased relative to each other 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[2] and SHAPE-IT[3]. The locally phased blocks are viewed as an *imperfect mosaic* from a reference panel of prephased haplotypes from the 1000 Genomes project[4]. 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 (the next heterozygous locus also has TL < 0.95 or the current heterozygous locus is the last 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.

References

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

2 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).

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

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

Phase Scoring

Chapter 4

Overview

The Emission Likelihood (EL) and Transition Likelihood (TL) values in the phased VCF file provide the degree of confidence in phasing for each phased heterozygous variant. The scores indicate 2 different modes of possible error in phasing.

- **Single-site error**—The EL score addresses the possibility of a single-site error at a locus, which means that the inferred phasing of one locus is wrong relative to its neighbors.
- **Switch error**—The TL score addresses the possibility of a switch error, which means that 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 the accurate haplotype and the 2 types of errors.

Table 2 Accurate Haploytype and Error Type Examples

Description	Example

Accurate Haplotype—Haplotypes arbitrarily labeled 0 and 1	`00000000000000000000000000 haplotype 1`
	`11111111111111111111111111 haplotype 2`

Single-site Error—Indicated by *	`00000010000000000000100000 haplotype 1`
	`11111101111111111111011111 haplotype 2`
	`      *             *`

Switch Error—Indicated by |	`00000000000000111111111111 haplotype 1`
	`11111111111111000000000000 haplotype 2`
	`              |`

Intended Use of Phase Scoring

The Emission Likelihood and Transition Likelihood scores allow the algorithm to assign scores to individual loci, rather than all pairs of loci.

For example, a given locus can be weakly phased but confidence in the phasing between variants separated by that locus can remain high. This situation can arise if a locus has low coverage and/or poor quality base calls, but the clouds spanning the variant connect other variants with better coverage. In this case, you can have a low single-site confidence (a low EL score) and a high confidence in phasing spanning these loci (a high TL score)

These scores can help answer the following questions.

Table 3 Phase Scoring Questions and Answers

Question	Answer
How confident can we be in the reported relative phasing of 2 adjacent sites? If 2 sites are on the same chromosome and there are no intervening loci for which phasing is reported, these 2 sites are 'adjacent'.	Calculate the combined probability (PQ1) of a switch error between the 2 (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: $P_{Q1} = TL_{i+1} * EL_{i} * EL_{i+1}$
How confident can we be in the reported relative phasing of 2 distal sites?	Calculate the combined probability (PQ2) of switch errors anywhere between the 2 loci (TL value at the 2 loci and between) and no single-site error at either locus (EL value at the 2 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 Phased Blocks

In the provided phased VCF files, the default Transition Likelihood (TL) cutoff of TL ≥ 0.95 is used to determine Phase Set (PS) annotations. The PS annotations are used as a starting point to define haplotype blocks, which are all variants on a given chromosome with the same PS value.

A locus with a low EL score usually receives a PS annotation even though it is not confidently phased relative to the other members of the phase set. Before phasing 2 variants relative to each other, make sure that the 2 loci both have PS annotations with the same value and make sure that both loci have EL scores ≥ 0.95.

Some applications require greater confidence or 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.

If a set of loci on the same chromosome meet the following criteria, then they are phased relative to each another.

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

Technical Assistance

For technical assistance, contact Illumina Technical Support.

Table 4 Illumina General Contact Information

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

Table 5 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 (SDSs)—Available on the Illumina website at support.illumina.com/sds.html.

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

🏭
Illumina
5200 Illumina Way
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