TruSeq® Synthetic Long-Read DNA Library Prep Kit for Whole Human Genome Phasing

Accurate genome phasing identifies co-inherited alleles and haplotype information, delivering a comprehensive view of the complex human genome.

Highlights

- **Detailed Insight into Genomic Variation**
  Phase information enables more comprehensive haplotyping and allele-specific analysis compared to traditional methods
- **Inclusive Library Preparation**
  Master-mixed reagents and primed 384-well plates increase protocol efficiency by minimizing manual pipetting steps
- **Simple Phasing Analysis**
  Push-button software simplifies genome-wide phasing
- **Highly Accurate Data**
  Method combines accuracy of short reads with long-read information to deliver greater phased coverage of the genome

Figure 1: TruSeq Synthetic Long-Read DNA Library Prep and Barcode Kits

The TruSeq Synthetic Long-Read DNA Library Prep and Barcode Kits are designed for preparing DNA libraries for whole-genome sequencing and phasing. The TruSeq Phasing Analysis App synthetically constructs long fragments from shorter sequencing reads to identify haplotype information.

Introduction

Conventionally, whole-genome sequencing generates a single consensus sequence without identifying the chromosome from which variants originate. Current short-read sequencing technology can identify single nucleotide polymorphisms (SNPs) and insertions and deletions (indels) in a given genome, but cannot determine the chromosome of origin. Recent findings suggest that haplotype or phase information—the unique content of the two homologous chromosomes in humans—can help researchers understand the relationship between DNA sequence and phenotype[1]. If two potentially deleterious variants occur in a gene, knowing how the alleles affect the function of one or both copies of the gene can help to identify genetic causes of disease.

Illumina phasing technology uses long, synthetic DNA sequences to make whole human genome phasing simpler, more accurate, and more cost-effective than conventional methods[2]. The TruSeq Synthetic Long-Read DNA Library Prep Kit and accompanying TruSeq Synthetic Long-Read DNA Barcode Kit (Figure 1) are designed to prepare DNA libraries for whole-genome sequencing and subsequent phasing analysis. The BaseSpace® TruSeq Phasing Analysis App constructs long fragments from the highly accurate shorter reads and assigns haplotype information to homologous chromosomes. TruSeq Synthetic Long-Read technology enables researchers to phase *de novo* mutations and identify co-inherited alleles in a population, providing greater insight into the human genome.

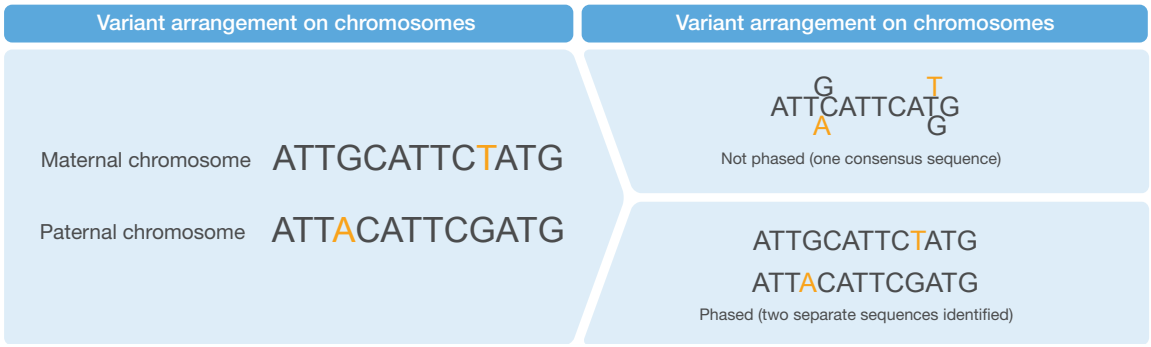
Greater Insight into Human Genomes

Phasing provides haplotype information, enabling allele-specific and variant linkage analyses (Figure 2). These inheritance patterns can inform studies of complex trait susceptibility, which are often influenced by interactions among multiple genes and alleles. Phasing can also provide valuable information for genetic disease research, as disruptions to alleles in *cis* or *trans* positions on a chromosome or chromosomes can cause some genetic disorders[3].

By leveraging the high accuracy of shorter reads to assemble long sequences into haplotype fragments, TruSeq Synthetic Long-Read technology eliminates traditional reliance on incomplete phasing methods, such as trio studies and statistical inference. Trio studies involve sequencing the maternal and paternal genomes and therefore cannot phase *de novo* mutations in the offspring. Because the TruSeq method requires only one genome instead of three, it uses a fraction of the sequencing resources used for trio studies. Statistical inference can be compromised if a database is not often updated, if the genome under study does not reflect statistical patterns, or if the genome is not represented adequately. In contrast, the TruSeq method analyzes the genome of interest directly by complementing sequence data with statistical imputation. This approach delivers long haplotype fragments for genomes from well-represented populations.

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA

Figure 2: Phasing Identifies Haplotype Information



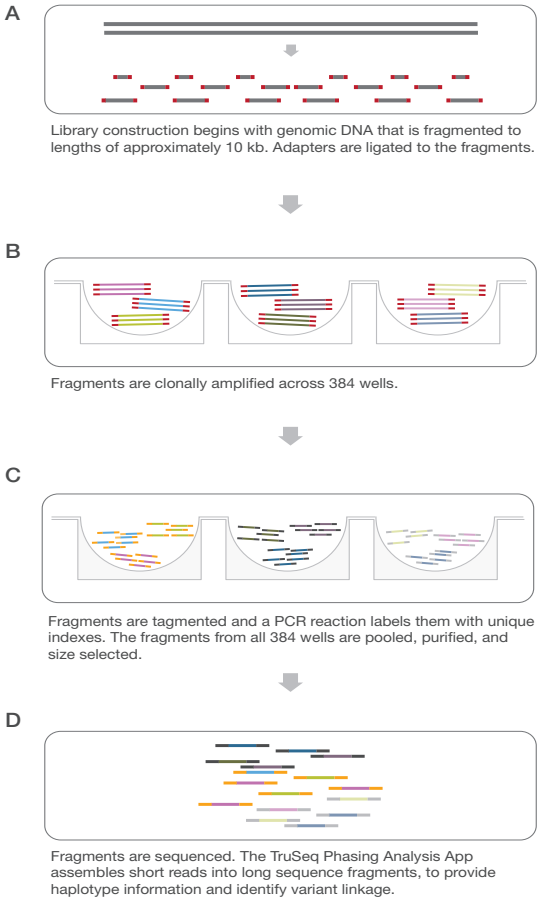
Genome phasing distinguishes between variants on homologous chromosomes, enabling researchers to investigate linkage patterns and allele-specific expression.

Unique Library Preparation Chemistry

TruSeq Synthetic Long-Read DNA Library Prep and Barcode Kits combine TruSeq and Nextera® chemistries with synthetic long-read technology to prepare DNA libraries for sequencing and phasing (Figure 3). The TruSeq Synthetic Long-Read DNA Barcode Kit includes 384 indexes for labeling the samples in each well. After sequencing, these indexes are used to construct synthetic long reads accurately. It also includes master-mixed reagents and pre-plated index primers in 384-well plates. This design minimizes manual pipetting steps so researchers can perform barcoding by PCR through simple centrifugation. The optional TruSeq Synthetic Long-Read DNA Accessory Kit contains alignment rings for arranging and centrifuging the plates.

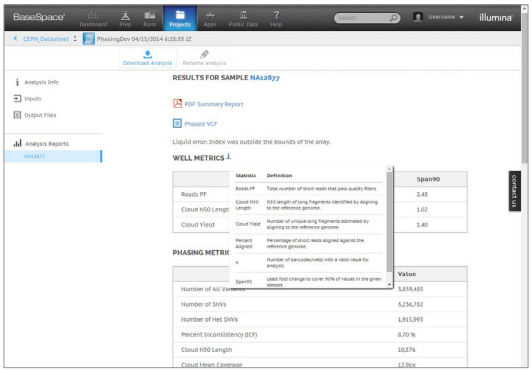
Library preparation begins by fragmenting 500 ng of genomic DNA to approximately 10 kb. Adapters are then ligated and 10 kb fragments are size selected (Figure 3A). A single Nextera "tagmentation" reaction simultaneously fragments and tags the DNA in each well (Figure 3B). PCR amplification then adds unique indexes to the samples. Fragments from all 384 wells are pooled, purified, and size selected (Figure 3C). Library preparation can be completed in 3 days, with 6 hours of hands-on time. The libraries are then sequenced (Figure 3D). After sequencing, the TruSeq Phasing Analysis App generates long synthetic fragments from the short reads and performs genome phasing, identifying haplotype information and variant linkage.

Figure 3: TruSeq Synthetic Long-Read DNA Library Preparation Workflow



The TruSeq Synthetic Long-Read DNA Library Prep Kit prepares DNA for sequencing and phasing. The TruSeq Phasing Analysis App identifies variant linkage and chromosome-specific variation for a comprehensive view of the human genome.

Figure 4: Simple Phasing Analysis



Featuring an intuitive interface, the TruSeq Phasing Analysis App simplifies genome phasing for any biological researcher, regardless of bioinformatics expertise. Results are displayed in simple tabular and graphical formats.

Simple Phasing Analysis

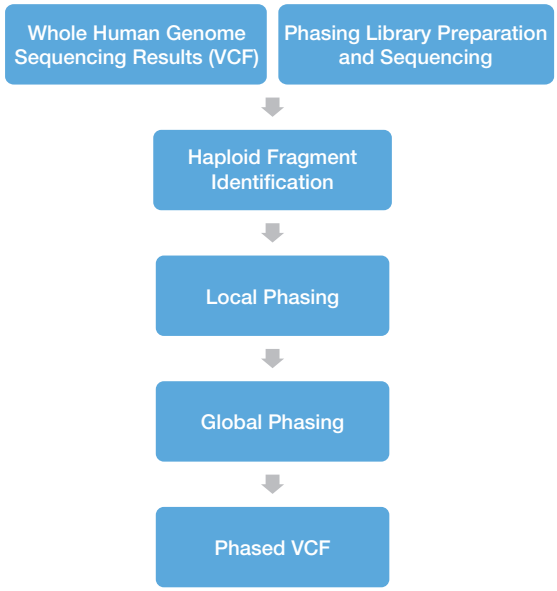
In addition to library preparation, the TruSeq phasing method includes push-button analysis in the BaseSpace environment. Designed for use with the TruSeq Synthetic Long-Read DNA Library Prep and Barcode Kits, the TruSeq Phasing Analysis App[4] combines whole-genome sequencing results with long-read data to generate haplotype contigs. The intuitive user interface (Figure 4) simplifies data analysis. Researchers can phase entire genomes simply by uploading the whole-genome sequencing results in variant call format (VCF) and launching the app.

First, the app uses shorter reads associated with the indexes from the TruSeq Synthetic Long-Read DNA Barcode Kit to identify contiguous haploid fragments from each well. Next, it performs local phasing by joining fragments with overlapping heterozygous SNPs together. Finally, during the global phasing process, it leverages statistical imputation to assemble the haplotype fragments into phased, genome-wide sequences (Figure 5). The app delivers haplotype contigs and confidence scores in a phased VCF file. This industry-standard format is compatible with various downstream tools for further analysis and visualization. The TruSeq Phasing Analysis App enables one-click genome phasing without requiring complicated statistical analyses.

Highly Accurate Data

The TruSeq method couples unique library preparation chemistry and statistical analysis to phase ≥ 96% of SNPs. The TruSeq Phasing Analysis App phases ≥ 70% variants within ≥ 94% of all genes[5] (Table 1) and achieves > 99% accuracy (Figure 6). By combining high-quality data with a novel algorithm, TruSeq Synthetic Long-Read technology provides more comprehensive haplotype information than solely experimental or statistical approaches.

Figure 5: Phasing Analysis Workflow



The TruSeq Phasing Analysis App enables genome-wide phasing without requiring complicated analyses or statistical calculations. Researchers simply upload the VCF file and select the appropriate sequencing run in BaseSpace. The app assembles haplotype fragments and phases genes and variants across the genome, delivering a phased VCF file.

Table 1: TruSeq Phasing Analysis App Results

Parameter	Individual NA12877	Individual NA12878	Individual NA12891	Individual NA12892
SNPs				
Global N50 (bp)	846,345	581,317	700,672	597,340
Phased (Local)	92.2%	90.7%	92.5%	90.7%
Phased (Global)	98.7%	98.2%	98.5%	98.1%
SNPs and Indels				
Global N50 (bp)	860,793	587,396	712,926	606,037
Phased (Local)	86.3%	85.4%	87.9%	85.5%
Phased (Global)	92.8%	92.8%	93.8%	92.8%
Gene Phasing (SNPs and Indels)				
100% Variants Phased	45%	44%	47%	44%
70% Variants Phased	95%	95%	95%	94%

Individuals NA12877, NA12878, and NA12891 form a related trio. Individual NA12892 is unrelated. DNA libraries were prepared using the TruSeq Synthetic Long-Read DNA Library Prep Kit and analysis was performed with the TruSeq Phasing Analysis App. The output from the TruSeq Phasing Analysis App shows the percentage of variants phased and the N50 length of the haplotype fragments, demonstrating that the app phases most variants and genes in the human genome.

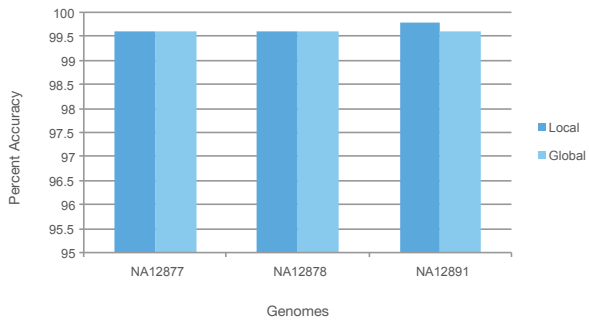
Summary

The TruSeq Synthetic Long-Read DNA Library Prep and Barcode Kits are designed for preparing libraries for whole human genome sequencing and phasing. By constructing long fragments from shorter sequencing reads, this method provides more comprehensive and accurate phasing compared to conventional trio studies or statistical inference. Simple analysis in BaseSpace enables researchers to phase an entire genome with minimal hands-on time. Together, the TruSeq Synthetic Long-Read DNA Library Prep Kit, Barcode Kit, and TruSeq Phasing Analysis App deliver a simple, accurate, and cost-effective whole-genome phasing solution.

References

1. Tewhey R, Bansal V, Torkamani A, Topol EJ, Schork NJ (2011) The importance of phase information for human genomics. Nat Rev Genet 12: 215–23.
2. Kuleshov V, Xie D, Chen R, Pushkarev D, Ma Z, et al. (2014) Whole-genome haplotyping using long reads and statistical methods. Nat Biotechnol 32: 261–6.
3. Kleinjan DJ, Coutino P (2009) *Cis*-ruption mechanisms: disruption of *cis*-regulatory control as a cause of human genetic disease. Brief Funct Genomic Proteomic 8: 317–32.
4. BaseSpace (basespace.illumina.com) Accessed 04 June 2014.
5. Public Data - BaseSpace (basespace.illumina.com/datacentral) Accessed 04 June 2014.
6. Platinum Genomes | Illumina (www.illumina.com/platinumgenomes) Accessed 04 June 2014.
7. Complete Genomics > > LFR (www.completegenomics.com/LFR) Accessed 04 June 2014.
8. Browning SR, Browning BL (2011) Haplotype phasing: existing methods and new developments. Nat Rev Genet 12: 703–14.

Figure 6: High Phasing Accuracy



Phasing accuracy is based on the local haplotype edit rate and the global switch error rate. The TruSeq Phasing Analysis App achieves > 99% accuracy when compared to reference data from the Platinum Genomes Project[6] (NA12877 and NA12878) and Complete Genomics[7] (NA12891) using a method adapted from Browning, SR et al. (2011)[8]. These results were generated using high-quality DNA from a CEPH family.

Ordering Information

Product	Catalog No.
TruSeq Synthetic Long-Read DNA Library Prep Kit (4 samples)	FC-126-1001
TruSeq Synthetic Long-Read DNA Barcode Kit (1 sample)	FC-126-1002
TruSeq Synthetic Long-Read DNA Barcode Kit (4 samples)	FC-126-1003
TruSeq Synthetic Long-Read DNA Accessory Kit	FC-126-1004

Illumina • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

© 2014 Illumina, Inc. All rights reserved.
Illumina, BaseSpace, Nextera, TruSeq, 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.
Pub. No. 770-2014-010 Current as of 09 June 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA