Application Note

Low-Pass Whole-Genome Sequencing Enabled by Scalable Library Preparation Offers a Competitive Alternative to Microarray-Based Genotyping

- The plexWell™ LP 384 Library Preparation Kit and open source GLIMPSE pipeline offer an accessible, robust, and cost-effective pipeline for high-confidence genotyping based on low-pass WGS and imputation.
- 10 M read pairs per human genome (<1X coverage) is sufficient for precise and accurate genotyping, and offer significantly more information than microarrays.

The plexWell LP 384 Library Preparation Kit requires less than one hour of hands-on time to generate two 48-library pools. The streamlined workflow and up to 2,304 barcode combinations facilitate implementation in high-throughput genotyping imputation pipelines.

Introduction

Advances in massively parallel sequencing and bioinformatics have enabled the routine use of whole-genome sequencing (WGS) in both research and clinical settings. While most applications require coverage depths in the range of 20X – 100X, low-pass or low-coverage (<1X) WGS is emerging as a powerful technique for detecting genome-wide genetic variation. Raw NGS coverage depths of human genomic DNA as low as 0.4X have been shown to enable imputation (statistical inference of unobserved genotypes from the haplotypes/genotypes of a characterized reference) of common single-nucleotide polymorphisms (SNPs) with accuracy comparable to that of widely used commercial micro-arrays.[1]

High-throughput low-pass WGS offers a wide range of benefits, including low data requirements, unbiased analysis when working with species or populations that are underrepresented in genetic databases, and the opportunity to perform off-target analysis. As such, low-pass WGS is suited for a variety of applications including genome-wide associations studies (GWAS), animal and plant breeding, cell bank profiling, and patient testing for disease risk or treatment outcome assessments.

Broad-based use of low-pass WGS as an alternative to microarray-based technology is highly dependent on the ability to prepare large, normalized pools of genomic DNA libraries for multiplexed sequencing in a single lane. seqWell's plexWell™ library preparation technology employs a unique, sequential transposase-based strategy to fragment and simultaneously tag genomic DNA with barcoded Illumina® adapters.[2] The cost-effective, streamlined, and highly scalable workflow generates pools of auto-normalized libraries that yield uniform read distributions without careful normalization of input DNA or individual sequencing-ready libraries.

In this study, we demonstrate the utility of the plexWell LP 384 Library Preparation Kit for the preparation of genomic DNA libraries for accurate, routine, high-throughput, low-pass WGS-based genotyping.

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

Materials and Methods

Genomic DNA – Twenty-five individual human genomic DNA (hgDNA) preparations were obtained from the Coriell Institute for Medical Research. The set included the well-characterized CEPH/Utah pedigree 1463 HapMap reference, NA12878,[3] as well as a panel of 24 uncharacterized genomes (LP01 – LP24) from males and females of different ethnicity and age, compiled to represent genetic diversity in the human population.

DNA quantification – Where applicable, input DNA and/or sequencing-ready libraries or library pools were quantified using an Infinite® F200 PRO microplate reader (Tecan) and Quant-IT™ PicoGreen dsDNA Assay Kit (ThermoFisher Scientific).

Library construction with the plexWell™ LP 384 Library Preparation Kit – A total of 48 libraries were prepared in the first six columns of a 96-well plate. Eight of the 25 genomes (LP17 – LP23 and NA12878) were processed in triplicate and seven (LP01 – LP07) in duplicate. Only one library was prepared from each of the remaining ten genomes (LP08 – LP16, LP24).

Of each hgDNA preparation, 6 µL was used in the Sample Barcoding reaction (first library preparation step, see Figure 1). This translated to a DNA input range across the 48 libraries of 7.97 – 12.94 ng (median: 10.17 ng). Libraries were prepared with the plexWell LP 384 Library Preparation Kit according to the standard protocol,[4] with eight cycles of library amplification.

Sequencing – Sequencing was outsourced to a service provider. The pool of 48 auto-normalized libraries was loaded in one lane of an S4 flow cell, for paired-end (2 x 150 bp) sequencing using an Illumina® NovaSeq™ 6000 System and S4 Reagent Kit v1.5.

Sequencing data analysis – Sequencing data were demultiplexed, and fastq files generated with bcl2fastq. Paired end reads for each sample were aligned to the GRCh38 human reference genome using BWA MEM. Library QC metrics and sequencing statistics such as library insert size, library complexity, and genome coverage were calculated using standard tools from the Picard suite.

Microarray analysis – Four of the eight uncharacterized genomes (LP17, LP19, LP22, and LP23) were submitted to a service provider for microarray analysis. Samples were analyzed with the Infinium® Omni5-4 v1.2 and Infinium Global Screening Array-24 v3.0 Kits. These two bead-based microarrays support genotyping of up to ~4.3 million (M) and ~0.65 M single-nucleotide polymorphisms (SNPs), respectively; some of which map to chromosomes X and Y.

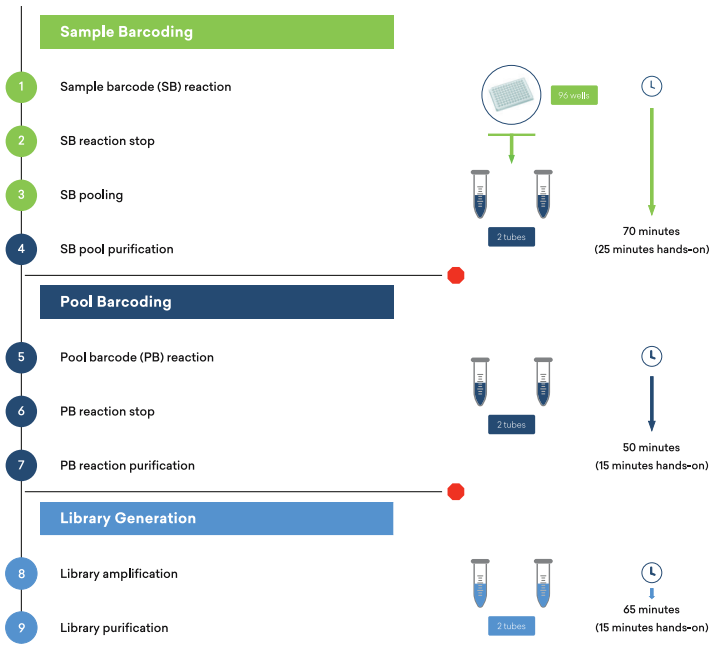

Figure 1. Overview of the plexWell LP 384 library preparation workflow. The Sample Barcoding and Pool Barcoding steps are designed to fragment DNA, add adapters (including sequencing barcodes), and normalize libraries via sequential tagging with a transposase. A limited number of PCR cycles (eight) are performed in the Library Generation step to fill gaps, and produce a sufficient amount of each of the two 48-library pools for sequencing. Red hexagons designate safe stopping points. Integrated normalization eliminates laborious dilution of input DNA. Pooling of libraries in the first (Sample Barcoding) step significantly streamlines library construction by allowing for single-tube bead purifications. This simplifies the preparation of library pools for sequencing, facilitates automation, and supports higher throughput.

Imputation – Imputation was performed with the open source GLIMPSE pipeline (GLIMPSE_Phase algorithm, Version 1.1.1)[5] using default settings. This pipeline includes steps to genotype, impute, and phase variants using reference data from the 1000 Genomes Project.[6] Sequencing data for the three NA12878 replicates were individually downsampled to 10 M, 20 M, 30 M, 40 M, or 50 M random reads pairs prior to variant calling and imputation. Variants (SNPs only) were called for a total of 57,420,428 positions on chromosomes

1 to 22 identified as potential polymorphic sites in the human genome (based on reference data from the 1,000 Genomes Project). Variant calls were compared across replicates, and to 3,180,406 of the 3,485,898 known heterozygous/homozygous Chr 1 – Chr 22 SNPs in the Genome in a Bottle (GIAB) Consortium HG001 reference genome.[7]

The four genomes from the "human diversity" panel that were submitted for microarray analysis (LP17, LP19, LP22, and LP23) were also genotyped using the GLIMPSE pipeline. For this analysis, sequencing data were downsampled to 10 M random read pairs per replicate (n=3).

Results and Discussion

The plexWell™ LP 384 Kit produces consistent and reproducible library and sequencing metrics

Library and sequencing metrics for the pool of 48 libraries (comprised of replicates of NA12878 and the 24 genomes from the "human diversity" panel), as well as the seven individual genomes from the panel that were processed in triplicate, are given in Table 1. Median library insert sizes in the range of 300 – 375 bp were obtained for all libraries. These cluster efficiently on Illumina® sequencers, thereby maximizing the unique data generated from 2 x 150 bp paired-end sequencing.

Read counts were consistent and reproducible across the entire pool and between replicate libraries for individual genomes, with an average read count coefficient of variation (CV) of 17.7% across the data set. Sequencing data quality was high across all samples, with ≥99.3% of reads aligning to the GRCh38 reference genome.

The sequencing yield from one lane of a NovaSeq™ 6000 S4 flow cell exceeded 20 M read pairs for each of the 48 individual libraries. Data were downsampled to 10 M or 20 M read pairs per library for coverage analysis, and consistently yielded mean coverage depths of 0.7X and 1.3X, respectively. This translated to approximately 48% or 68% of all genomic positions covered at a minimum depth of 1X.

plexWell technology supports precise and accurate genotyping from <1X coverage of a well-characterized human genome

To determine the precision and accuracy of SNP genotyping with the plexWell LP 384-GLIMPSE workflow, NA12878 imputation results were compared across replicates (Figure 2) and to the GIAB HG001 truth data set (Figure 3).

Table 1. Select library and sequencing metrics for the 48-library pool and individual human genomes processed in triplicate.

					10 M read pairs/sample			20 M read pairs/sample	
Sample/ Genome	Mean insert size (bp)	Median insert size (bp)	Read count CV (%)	% reads aligned	Duplication rate (%)	Mean coverage depth (X)	% genome covered at ≥1X	Mean coverage depth (X)	% genome covered at ≥1X
Pool of 48	322 ±29	346 ±29	20.1	99.5	6.8	0.70	48.2	1.32	68.4
LP17	321 ±3	342 ±4	18.9	99.4	7.9	0.70	47.9	1.28	67.9
LP18	321 ±3	343 ±3	8.1	99.4	7.1	0.70	47.7	1.29	67.9
LP19	322 ±4	345 ±3	13.2	99.7	6.3	0.71	48.4	1.33	68.7
LP20	325 ±4	348 ±5	16.6	99.6	6.9	0.71	48.4	1.32	68.6
LP21	324 ±2	346 ±2	18.3	99.6	6.4	0.71	48.2	1.33	68.5
LP22	324 ±3	346 ±4	17.7	99.6	6.9	0.70	48.1	1.31	68.2
LP23	335 ±14	313 ±12	41.4	99.3	7.8	0.69	47.5	1.27	67.5
NA12878	317 ±2	338 ±2	4.9	99.5	5.9	0.71	48.8	1.35	69.3

All values (except for the read count coefficient of variance, CV) represent the average for the individual libraries comprising each sample, i.e., is the average of 48 individual measurements/data points for the pool of 48 libraries, and the average of three individual measurements/data points for each genome processed in triplicate. Genomes highlighted in green were genotyped with the GLIMPSE imputation pipeline and a commercial microarray platform.

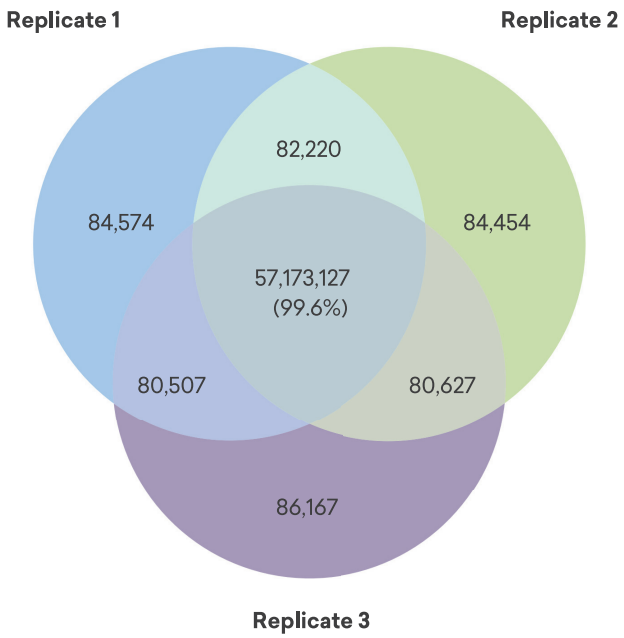

Figure 2. Concordance between genotyping results for NA12878 replicates. Unweighted Venn diagrams represent SNP calls for 57,420,428 potentially polymorphic positions on chromosomes 1 – 22, generated from 10 M random read pairs per replicate.

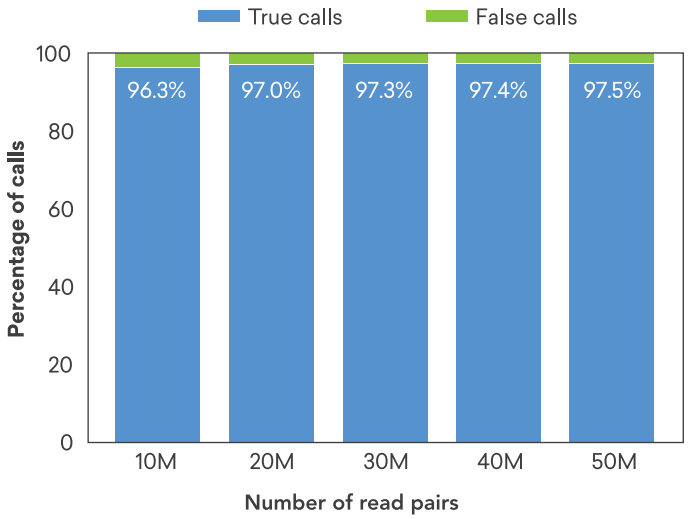

Figure 3. Accuracy of low-pass WGS genotyping for NA12878 replicates from increasing amounts of sequencing data. Data correspond to the 3,180,406 confirmed heterozygous/homozygous SNPs in the GIAB HG001 truth data set for which a genotype was obtained with the plexWell LP 384-GLIMPSE workflow. Blue bars represent the average percentage of SNP calls (from three replicates) confirmed to be true calls, whereas false calls are designated in green. Error bars are not visible at the scale of the graph.

Of the 57,420,428 potentially polymorphic Chr 1 – Chr 22 positions, variant calls obtained from 10 M read pairs were concordant for 57,173,127 positions (99.6%) across the three NA12878 replicates (Figure 2). This confirmed that the plexWell™ LP 384-GLIMPSE workflow supports precise (reproducible) genotyping from a mean coverage depth <1X, or the amount of data obtainable when 96 human genomes are sequenced in one lane of a NovaSeq™ S4 flow cell.

Low-pass WGS yielded NA12878 genotyping results for 3,180,406 of the 3,485,898 known polymorphic (non-reference) positions in the GIAB HG001 reference genome. A very high proportion of imputed variant calls (>96%) were identical to those in truth data set (Figure 3), confirming that highly accurate genotyping can be achieved from a minimal amount of WGS data (10 M read pairs or <1X genome coverage). As shown in the figure, a significant (2- to 5-fold) increase in sequencing depth (and cost) had a marginal impact on genotyping accuracy.

Low-pass WGS enables precise and accurate genotyping from uncharacterized genomes

In real-world settings genotyping is typically performed on uncharacterized genomes. Results generated with the plexWell LP 384-GLIMPSE workflow for four genomes from the "human diversity" panel are given in Table 2 on the next page. As for NA12878, a very high degree of reproducibility (99% concordant calls) was observed across replicates.

To provide more context for the results obtained with the plexWell LP 384-GLIMPSE workflow, the same four genomes were analyzed using industry-leading SNP microarray platform. As shown in Figure 4, low-pass WGS yielded genotyping data for an order of magnitude more positions on Chr 1 – Chr 22 (57 M) as compared to the high-density microarray (3.2 M positions), and two orders of magnitude more data than the low-density array (0.50 M positions). Low-pass WGS genotyping results were compared to those obtained with the high-density microarray (Figure 5). Concordance between genotype calls (for approximately 3 M positions called by both methods) approached 100%, confirming that the accuracy of low-pass WGS-based genotyping is comparable to that of microarray technology.

Table 2. Genotyping results for four genomes from the "human diversity" panel, obtained from 10 M read pairs per replicate.

		Concordant calls				Unique calls		
Genome	**Total variant calls**	**All 3 replicates**	**Replicates 1 and 2**	**Replicates 1 and 3**	**Replicates 2 and 3**	**Replicate 1**	**Replicate 2**	**Replicate 3**
LP17	57,758,232	57,088,552 (98.8%)	108,838	108,124	108,986	114,914	114,052	114,766
LP19	57,708,315	57,138,275 (99.0%)	91,312	96,197	88,910	94,644	101,931	97,046
LP22	57,700,432	57,143,828 (99.0%)	90,313	94,539	88,344	91,748	97,943	93,717
LP23	57,739,177	57,105,854 (98.9%)	91,413	96,499	122,487	126,662	100,674	95,588
Average	57,726,539	57,119,127 (99.0%)	95,469	98,840	102,182	106,992	103,650	100,279

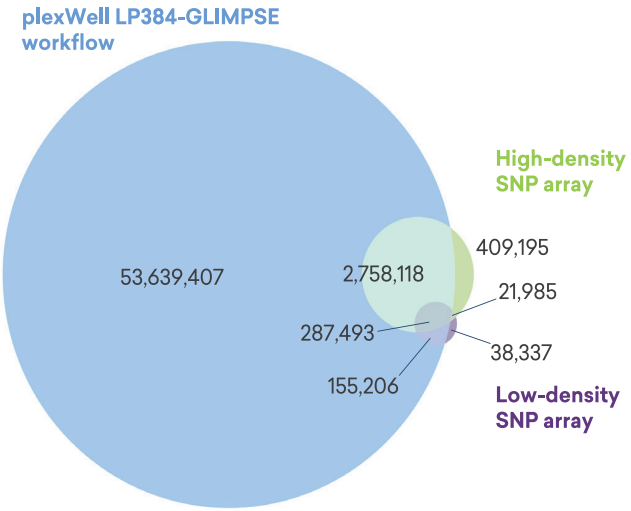

Figure 4. Data yields from low-pass WGS-based genotyping (10 M read pairs) and the two commercial microarrays used in this study. Weighted Venn diagrams represent SNP calls for the number of positions on chromosomes 1 to 22 for which genotyping results were obtained with each method. The number of positions genotyped by microarray (3.2 M and 0.50 M for the high- and low-density arrays, respectively) exclude markers that do not map to Chr 1 – Chr 22 on GRCh38, as well as redundant positions.

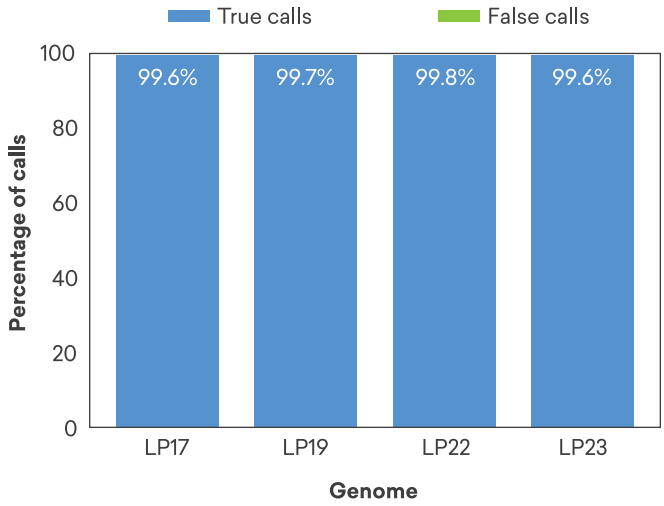

Figure 5. Comparison of variant calls generated with the plexWell™ LP 384-GLIMPSE workflow (10 M read pairs) vs. the high-density SNP array for four uncharacterized genomes from the "human diversity" panel. Data correspond to the 3,045,611 positions called with both methods. For the sake of the comparison, microarray results are regarded as the reference ("truth"). Blue bars represent the average proportion of "true" SNP calls (for three replicates) from imputation False calls are designated in green. Neither these, nor error bars, are visible at the scale of the graph.

Contact us at sales@seqwell.com to learn how to implement the plexWell LP 384 Kit in your low-pass WGS genotyping workflow

Conclusions

plexWell™ library preparation technology supports truly multiplexed and highly scalable construction of library pools for low-pass WGS. Together with the open source GLIMPSE pipeline, this technology offers an accessible, robust, and sample-type agnostic alternative to established microarray technology for genotype imputation.

Specific benefits of the plexWell LP 384-GLIMPSE workflow include:

- A unique, streamlined library preparation workflow that enables the preparation of two auto-normalized 48-library pools (for sequencing in a single lane of a NovaSeq™ S4 flow cell) in approximately three hours, with less than one hour of hands-on time.
- High precision and accuracy from a minimal amount of sequencing data (10 M read pairs or <1X coverage).
- High-confidence genotyping data for significantly more positions than those that can be interrogated with SNP microarrays, including positions that are not specifically targeted by predefined microarray content.
- Plate-based reagents, flexible kit configurations, and up to 2,304 sequencing barcodes that support different batch sizes and facilitate implementation in a wide range of laboratory settings.

Robust performance and ease-of-use make the plexWell LP 384-GLIMPSE pipeline ideally suited for high-throughput, low-pass WGS-based genotyping in both research and commercial settings.

References

1. Wasik K, et al. BMC Genomics 2021; 22:197. doi: 10.1186/s12864-021-07508-2
2. https://seqwell.com/technology/
3. Coriell Institute for Medical Research, NA12878. https://www.coriell.org/0/Sections/Search/Sample_Detail.aspx?Ref=NA12878&Product=DNA
4. User Guide: plexWell™ LP 384 Library Preparation Kit for Illumina® Sequencing Platforms. https://seqwell.com/wp-content/uploads/2021/06/plexWell_LP384_Library_Preparation_Kit_User_Guide_v20210609.pdf
5. Rubinacci S, et al. Nat Genet 2021, 53:120-126. doi: 10.1038/s41588-020-00756-0.
6. https://www.internationalgenome.org/
7. Genome in a bottle—a human DNA standard. Nat Biotechnol 2015, 33:675. doi: 10.1038/nbt0715-675a.

seqWell Inc.
66 Cherry Hill Drive
Beverly, MA 01915

+1 (855) SEQWELL (737-9355)
sales@seqwell.com

Learn more about plexWell kits at seqwell.com/products/

Unless otherwise stated in references, presented data on file.
For Research Use Only. Not for use in diagnostic procedures.
All product names and trademarks are the property of their respective owners.
© 2022 seqWell Inc. All rights reserved.