Application Note

Low-Pass Whole-Genome Sequencing on the Element AVITI™ System Enables Cost-Effective Genotyping

- **10 million read pairs, 1X coverage, per human genome analyzed with GLIMPSE demonstrate consistent and accurate genotyping.**
- **purePlex™ library prep and the AVITI System enable flexible, scalable, and efficient low-pass sequencing and genotyping, including kinship analysis.**
- **Lower sequencing costs coupled with high quality data and low duplication rates make low-pass sequencing accessible on a benchtop system.**

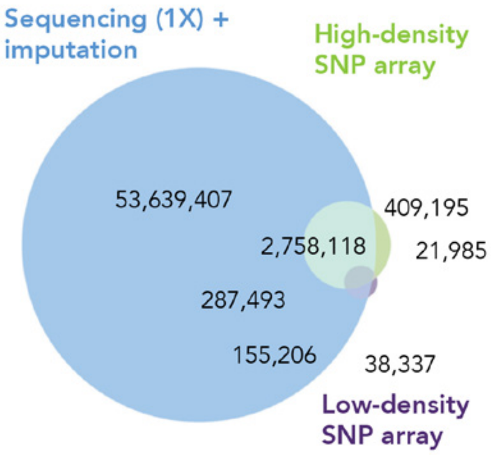

Figure 1. Weighted Venn diagrams represent SNP calls for the number of positions on chromosomes 1-22 for which genotyping results were obtained from WGS-based genotyping (10 million read pairs) and two commercial microarrays.[3]

Introduction

Advances in next-generation sequencing (NGS) and bioinformatics have empowered routine use of whole-genome sequencing (WGS) in research and clinical settings. Although certain applications require > 20x coverage, low-pass or low-coverage (0.4–4x) sequencing with imputation — the statistical inference of unobserved genotypes from the haplotypes/genotypes of a characterized reference — is emerging as a powerful technique for detecting genome-wide genetic variation. Raw coverage depths of 0.5–1x have comparable accuracy to microarrays[1] and at ≥ 1x coverage can outperform high-density microarrays in rare variant burden tests.[2]

We previously demonstrated that low-coverage WGS with imputation is a cost-effective alternative to microarrays that allows genotyping at orders of magnitude more positions (Figure 1).[3] The number of samples and sequencing platform heavily influence the price of low-coverage sequencing and imputation. The lowest prices are feasible only when fitting hundreds or thousands of samples onto the highest output flow cell, rendering this approach inaccessible to most researchers. This application note demonstrates that the purePlex DNA Library Prep Kit paired with the AVITI System improves the performance of mid-throughput, low-pass WGS at a competitive price point (Table 1).

Table 1. Price per gigabase (Gb) and 1x human genome sequencing on benchtop and high-output sequencers.

Sequencer	$ per Gb	$ per 1X Human Genome
Illumina NextSeq 2000	$17	$54
Illumina NovaSeq 6000	$5-$15	$15-45
Element AVITI Sytem	$5	$15

Prices reflect list price by vendor as of February 2023.

Materials and Methods

Genomic DNA and DNA quantification – Eight 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, the Ashkenazi cohort, and the Han Chinese cohort (Table 2). These samples were quantified with an Infinite® F200 PRO microplate reader (Tecan) and Quant-IT™ PicoGreen dsDNA Assay Kit (ThermoFisher Scientific) before library prep.

Table 2. Summary of hgDNA samples assessed in this study.

Coriell ID	NIST ID	Ethnicity	Relationship	Sample ID
NA12878	HG001	Utah/Mormon	Daughter	A0
NA12891	N/A	Utah/Mormon	Father	A1
NA12892	N/A	Utah/Mormon	Mother	A2
NA24385	HG002	Ashkenazi	Son	B0
NA24149	HG003	Ashkenazi	Father	B1
NA24631	HG005	Chinese	Son	C0
NA24694	HG006	Chinese	Father	C1
NA24695	HG007	Chinese	Mother	C2

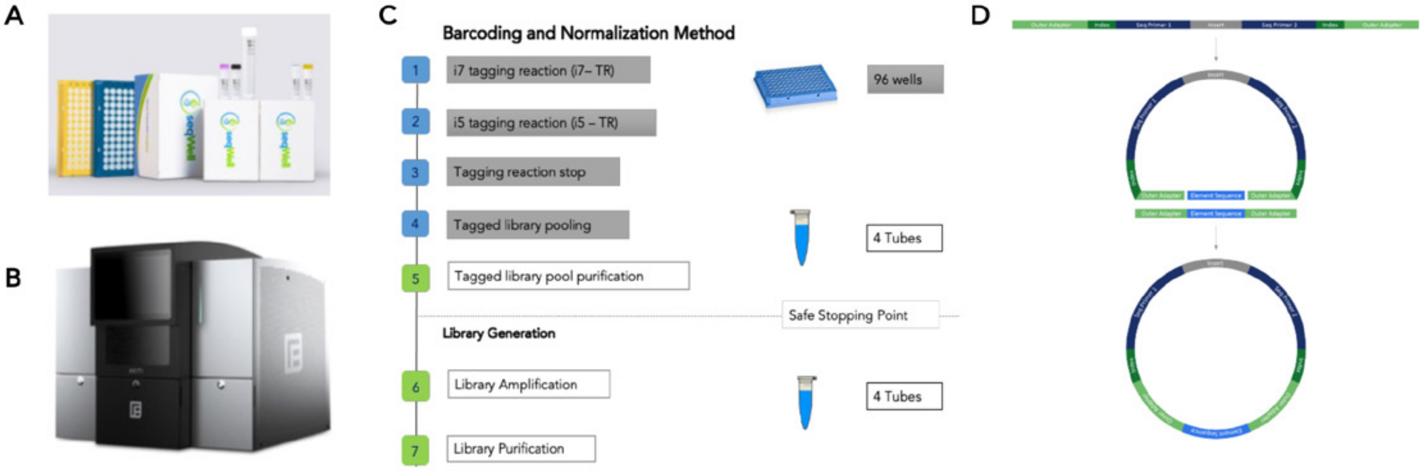

Figure 2. The workflow starts with the purePlex DNA Library Preparation Kit (A) and culminates with sequencing on the AVITI System (B). Steps for the purePlex DNA Library Prep Kit and Element Adept Library Compatibility Workflow are respectively depicted in (C) and (D).

Library construction with the purePlex™ DNA Library Preparation Kit – The purePlex DNA Library Preparation Kit processed the eight hgDNA samples in triplicate, generating a normalized 24-plex library pool. Samples were processed according to the manufacturer protocol[4] (Figure 2C). The input ranged from 45.3 to 50.1 ng DNA with a median of 47.6 ng. A Qubit 1x dsDNA High Sensitivity (HS) Assay Kit and Agilent TapeStation 2200 HSD5000 quantified and sized the library pool.

Library conversion using the Adept™ Workflow – The Adept Workflow further prepared the library for sequencing on the AVITI System.[5] The workflow anneals 0.5 pmol input linear library to splint oligos, adding the Element surface primers. A ligation reaction then circularizes the library and a digestion reaction removes any leftover splint oligos or linear library material without amplification (Figure 2D). A final bead cleanup removes any remaining small materials, salts, and enzymes.

Sequencing and data analysis – A 2 x 150 bp run sequenced the purePlex libraries to a depth of ≥ 20 million read pairs per sample. Bases2Fastq demultiplexed the sequencing data and generated FASTQ files. Paired-end reads for each sample were aligned to the GRCh38 human reference genome using BWA MEM. Library quality control (QC) metrics and sequencing statistics such as insert size, library complexity, and genome coverage were calculated using standard tools from the Picard suite.

Imputation and trait analysis – The open-source GLIMPSE pipeline (GLIMPSE Phase algorithm, Version 1.1.1)[6] performed imputation using default settings. Leveraging reference data from the 1000 Genomes Project, the pipeline includes steps to genotype, impute, and phase variants.[7] Sequencing data for the three replicates from eight DNA samples were individually downsampled to 10 million random reads pairs before variant calling and imputation. Variants (SNPs only) were called with a total of 929,854 positions on chromosome 22 identified as potential polymorphic sites in the human genome based on reference data from the 1000 Genomes Project. Variant calls were compared across replicates and to ~45,000 known heterozygous/homozygous chromosome 22 SNPs in the Genome in a Bottle (GIAB) Consortium[8] HG001, HG002, HG003, HG005, HG006, and HG007 reference genomes.

PLINK software uses hidden Markov model (HMM) to infer identity-by-descent (IBD).[9] Cotterman coefficients of relatedness k0, k1 and k2 are estimated using PLINK (Version 1.9) with genome option, which are represented using Z0, Z1 and Z2. Twenty-eight human genome pairs are formed by any possible combination of any two out of the eight. Z0, Z1 and Z2 are estimated using variant call files generated by GLIMPSE on chr22.

Results and Discussion

The purePlex DNA Library Prep Kit uses sequential transposase steps to incorporate full-length Illumina P7 and P5 adapter sequences with indexes. Each kit contains 96 unique i7 and i5 indexes that enable processing of up to 96 samples. Additional index sets achieve multiplex levels > 96. Briefly, i7 tagging reactions are applied to 5–50 ng of DNA input. After this initial indexing reaction, a novel normalization reagent and i5 tagging reagent are added to each sample for the second tagging reaction. Samples are pooled volumetrically, purified, and converted into libraries to complete the 2.5-hour workflow, which includes 45 minutes of hands-on time. Built-in auto-normalization obviates the need to normalize sample input. Pooling samples during the workflow results in normalized pools of 24 samples, greatly reducing the QC burden before further multiplexing for sequencing.

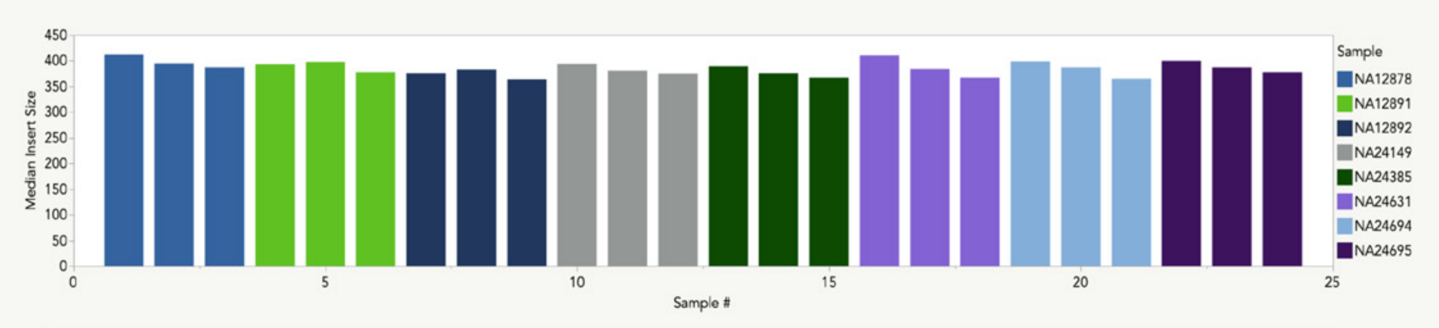

Figure 3. Insert size uniformity of a 24-plex purePlex multiplexed library that includes eight individual hgDNA samples. A 350–400 bp insert size is optimal for generating unique data from a 2 x 150 run and increasing the usable data per dollar of sequencing.

The Adept Workflow preserves the insert size and indexing of the purePlex library pool. Median insert sizes > 300 nt (Table 3, Figure 3) coupled with the low duplication rate of the AVITI System and 300 Gb output are well-suited to multiplexing 96 samples per flow cell for 1x coverage with a 2 x 150 read length. For this study, higher sequencing depth per sample demonstrated library and sequencing performance (Table 3).

Table 3. Summary of 24-plex purePlex library performance on the AVITI System. Average duplication rate and genome coverage after samples were down-sampled to 20 million read pairs.

Reads Demultiplexed	≥Q30	Mean Quality	Average Median Insert Size	Reads Aligned	Duplication Rate	%Genome covered at ≥1X	Mean Coverage Depth (X)
92.4%	92.4%	40.2	385nt	99.6%	3.45%	72.8%	1.77

Precise and accurate genotyping from 1x coverage of well-characterized human genomes

The precision and accuracy of SNP calls were determined for chromosome 22. Prior analysis of a genetically diverse group of human samples sequenced at a low depth indicated that summary statistics of chromosome 22 statistically agreed with the imputation results from all autosomes. The concordance for each triplicate was calculated as > 91% for seven of the eight samples. Concordance for the eighth sample remained high (89.3%). Although chromosome 22 contains 929,854 positions, only ~45,000 positions are known heterozygous/homozygous chromosome 22 SNPs in the Genome in a Bottle (GIAB) Consortium. Calls that are negative for a SNP in both the reference and imputed calls were omitted from the counts presented in Figure 4. Including these counts raises the concordance obscuring differences. When these calls are included for NA12878 the concordance increases from 91.9% to 98.7%.

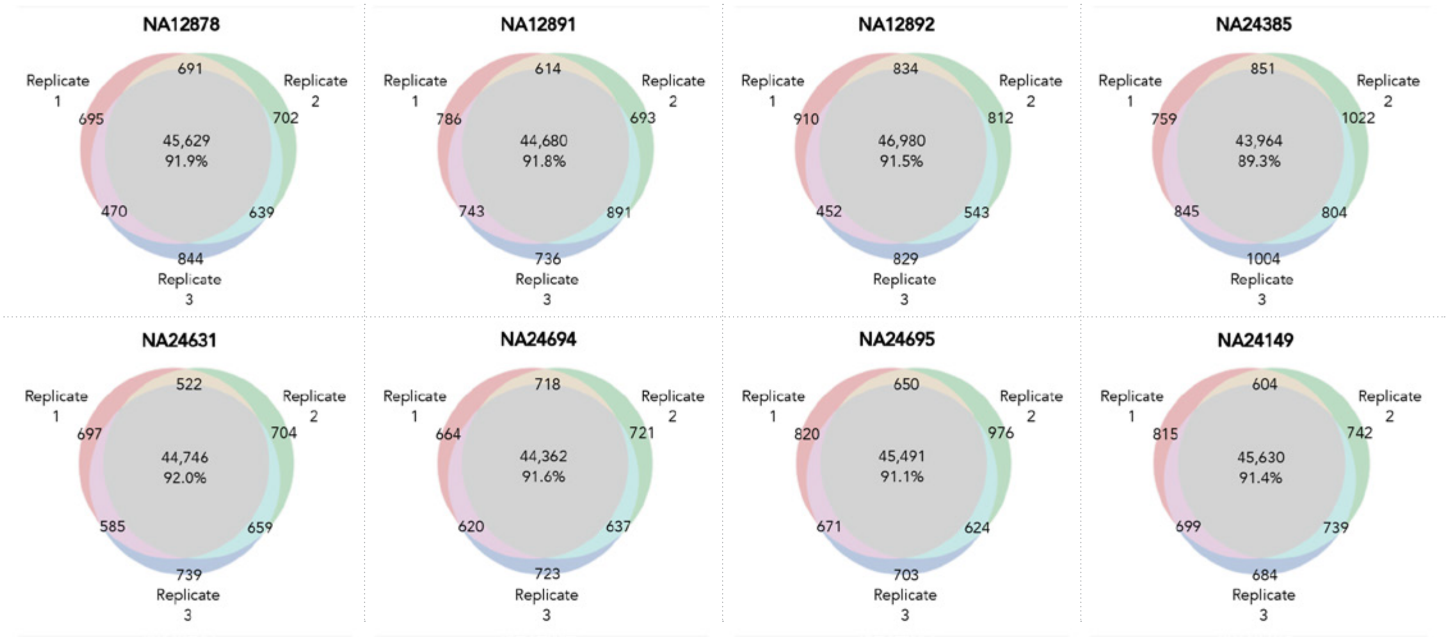

Figure 4. High chromosome 22 SNP concordance for three replicates of eight individual human samples at 1x coverage. Each triplicate set for the CEPH/Utah cohort, the Ashkenazi cohort, and the Han Chinese cohort were down-sampled to 1x coverage and analyzed using GLIMPSE imputation pipeline. Venn diagrams represent SNP calls for 929,854 potentially polymorphic positions on chromosome 22.

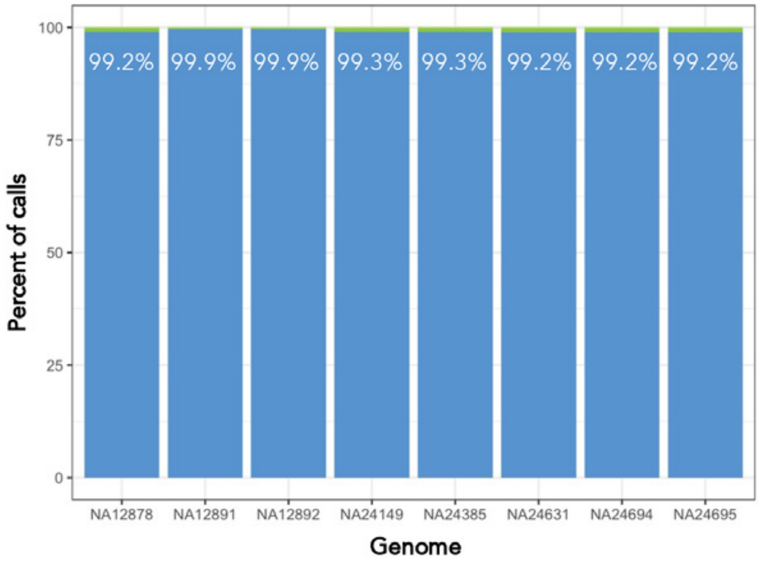

Figure 5. Accuracy of low-pass WGS genotyping for eight individual human genomic samples with purePlex library prep on an AVITI System. Data correspond to the confirmed heterozygous/homozygous chromosome 22 SNPs in the GIAB Consortium HG001, HG002, HG003, HG005, HG006, and HG007 truth data set. Blue bars represent the average percentage of SNP calls from three replicates confirmed as true calls. False calls are designated in green. Error bars are not visible at this scale.

In addition to reproducibility, data were compared to reference sequencing data for each of the eight samples. Low-pass WGS yielded genotyping results for an average of 924,240 of the 929,854 known polymorphic (non-reference) positions in the GIAB HG001, HG002, HG003, HG005, HG006, and HG007 reference genomes. A high proportion of imputed variant calls (> 99%) were identical to those in the truth data set (Figure 5), confirming that a minimal amount of WGS data—only 10 million read pairs or < 1x genome coverage—can achieve highly accurate genotyping.

Construction of relatedness matrices using genotyping-by-sequencing data

A common application of low-pass WGS is the analysis of genetic relationships. This powerful approach to genetic mapping of disease is based on comparing the genetic marker profiles of affected relatives. Kinship assess the probability that a set of genes was derived from a single ancestral gene and the probability that they are identical-by-descent (IBD).[10,11] If one IBD allele or alleles are present in the other IBD allele, two individuals are considered related.[10,11] In first-degree relationships, the probability of parent-child pairs sharing 0, 1 and 2 of the IBD alleles (Z0, Z1, and Z2) are expected to be close to 0, 1 and 0, respectively.[10] Figure 6 demonstrates successful imputation of relatedness between each set of the CEPH/Utah cohort, the Ashkenazi Jewish cohort, and the Han Chinese cohort. See Table 2 for information regarding relation to each family.

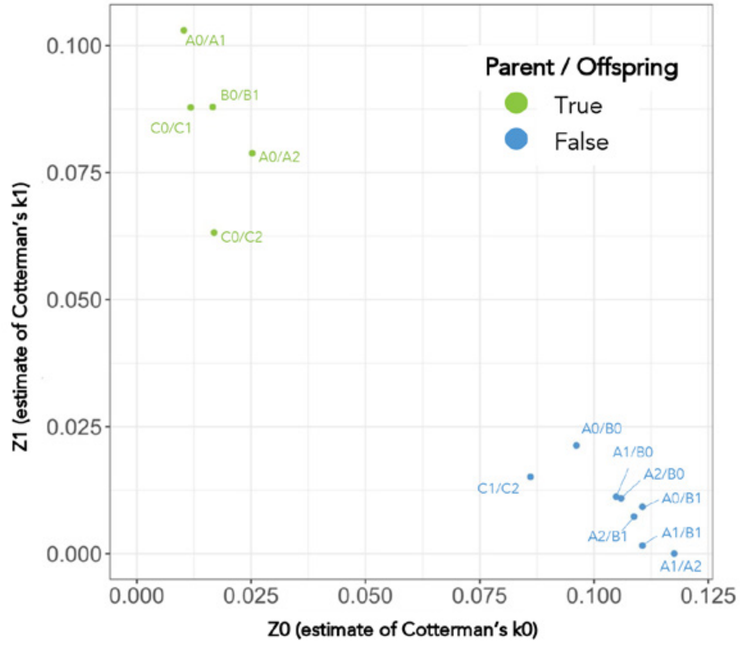

Figure 6. Relatedness coefficient from GLMPSE impute variant call format (VCF) of chromosome 22. Human genome pairs are plotted using Z0 and Z1, respectively. Z0 and Z1 scores can be estimated from VCF with PLINK genome analysis. A green dot represents the high probability of parent-offspring pairs sharing 1 of the IBD alleles. A blue dot, designates low probability.

Summary

The purePlex DNA Library Prep Kit supports truly multiplexed and highly scalable construction of library pools for low-pass WGS. Together with the Element AVITI system, the NGS technology offers an accessible and robust alternative to microarrays for genotyping. More specifically, the NGS technology includes the following benefits:

- A novel and streamlined workflow with unique dual indexes (UDIs) that enables the preparation of four auto-normalized 24-library pools in fewer than three hours with only 45 minutes of hands-on time.
- A benchtop sequencer featuring unprecedented performance backed by a low duplication rate, high accuracy, low cost and flexibility.
- High precision and accuracy from a minimal amount of sequencing data.
- Plate-based reagents, flexible kit configuration, and scalable UDI multiplexing that support various batch sizes and facilitate implementation in a wide range of labs.

Robust performance and ease-of-use make the purePlex Library Prep Kit and AVITI System workflow ideally suited for mid-throughput low-pass WGS applications that assess human and similarly-sized genomes. The benchtop system delivers NGS at the same price point as the highest-throughput sequencers.

References

1. Wasik K, et al. BMC Genomics 2021; 22:197. doi: 10.1186/s12864-021- 07508-2
2. Rubinacci S, et al. Nat Genet 2021, 53:120-126. doi: 10.1038/s41588-020-00756-0
3. Application Note: Low-Pass Whole-Genome Sequencing Enabled by Scalable Library Preparation Offers a Competitive Alternative to Microarray-Based Genotyping. Download here
4. User Guide: purePlexTM DNA Library Preparation Kit for Illumina® Sequencing Platforms. Download here
5. User Guide: Element Adept™ Library Compatibility Workflow. https://www.elementbiosciences.com/products/adept
6. Rubinacci S, et al. Nat Genet 2021, 53:120-126. doi: 10.1038/s41588- 020-00756-0
7. https://www.internationalgenome.org/
8. Genome in a bottle — a human DNA standard. Nat Biotechnol 2015, 33:675. doi: 10.1038/nbt0715-675a
9. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81, 559–575 (2007). https://www.cog-genomics.org/plink/
10. Arab, M.M., Marrano, A., Abdollahi-Arpanahi, R. et al. Genome-wide patterns of population structure and association mapping of nut-related traits in Persian walnut populations from Iran using the Axiom J. regia 700K SNP array. Sci Rep 9, 6376 (2019). https://doi.org/10.1038/s41598-019-42940-1
11. Dodds, K.G., McEwan, J.C., Brauning, R. et al. Construction of relatedness matrices using genotyping-by-sequencing data. BMC Genomics 16, 1047 (2015). https://doi.org/10.1186/s12864-015-2252-3

Learn more about purePlex™ DNA Library Preparation Kit and AVITI™ System at
seqwell.com/products/pureplex-dna-library-prep-kit/
elementbiosciences.com/products/aviti

Unless otherwise stated in references, presented data on file.
For Research Use Only. Not for use in diagnostic procedures.

seqWell

Element Biosciences