Automated Low-Pass Whole-Genome Sequencing to Scale and Accelerate Genotype Imputation

Sabina Gude[1], John Palys[1], Ken Tenan[1], Michelle Rahardja[1], Yanyan Liu[1], Nicole Madamba[2], Mary Barter[3], Joe Brown[3], Qingchang Meng[3], Samuel Widmayer[3], and Dan Gatti[3]

[1]seqWell, Inc., Beverly, MA [2]Revvity, Waltham, MA [3]The Jackson Laboratory, Bar Harbor, ME

Introduction

Genotyping arrays, which sample a set of ~106 known polymorphisms, in conjunction with imputation software, have been traditionally used in genome-wide association studies, HLA typing, animal genotyping, and other applications. As the cost of NGS sequencing declines, low-pass whole-genome sequencing has become an attractive and cost-effective alternative to arrays, with similar concordance outcomes at sequencing depths as low as 0.1x. The time required for traditional NGS library preparation can constrain large genotype-imputation projects for which hundreds, thousands, or more samples may be required. Intelligently designed and rapid sequencing chemistries coupled with workflow automation provide a path towards efficient library preparation. seqWell's purePlex™ DNA Library Preparation Kit features a streamlined workflow such that a user can prepare Illumina sequencing libraries for 96 samples in under three hours. The kit auto-normalizes DNA inputs over an order of magnitude, which mitigates the need for extensive sample normalization, and it displays reduced insertion bias compared to other transposase-based methods. Additionally, the purePlex workflow is readily automatable. Here we present depth-of-sequencing outcomes deriving from an implementation of the purePlex workflow on a Revvity Sciclone™ G3 NGSx workstation. A total of sixteen 12-plex libraries were prepared from hgDNA (Coriell NA12878) using purePlex library preparation kit, with 8 libraries prepared manually and the other 8 automated on the Sciclone workstation. The 16 libraries were sequenced on an Illumina MiSeq Nano at 2 x 150 bp to assess read output uniformity across the whole 96-well plate from manual and automated preparation. A subset of the libraries from each method was used for deeper sequencing on an Illumina NovaSeq X at 2 x 150 bp. Sequencing data were individually down-sampled to various depth of coverage prior to variant calling and imputation using the open-source GLIMPSE2 pipeline. Our results show despite the order-of-magnitude variation in sample concentration, sequencing outcomes were uniform using purePlex Library Preparation Kit automated on the Revvity Sciclone G3 NGSx workstation.

Automated purePlex Library Prep Enables Highly Scalable Low-Coverage Whole-Genome Sequencing

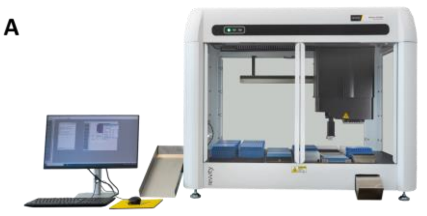
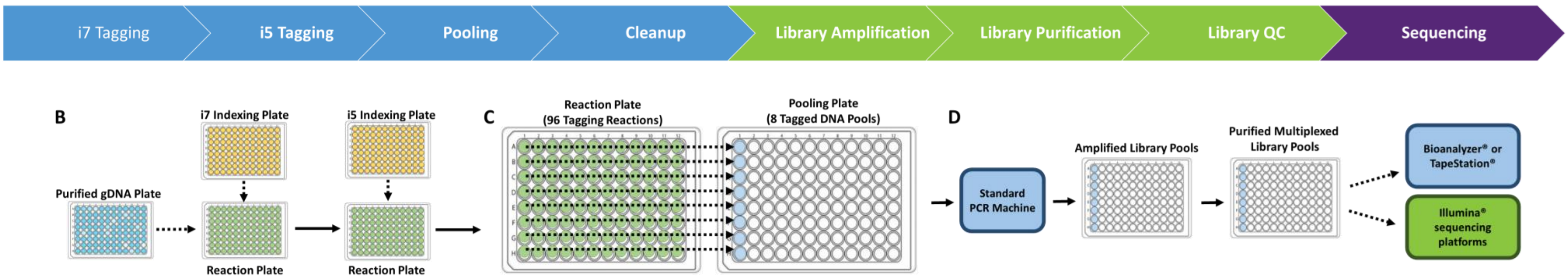


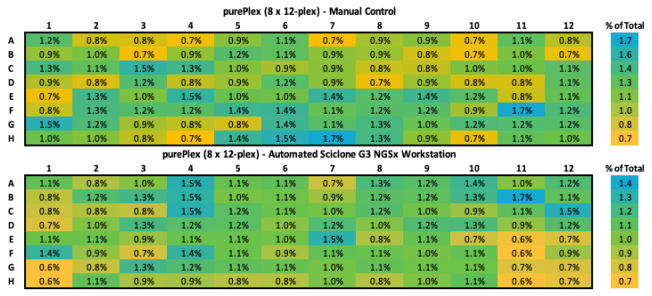
Figure 1. Automated purePlex DNA Library Preparation Workflow using Revvity's Sciclone G3 NGSx Workstation **(A)**. In the Sciclone system purePlex DNA Library Preparation protocol, samples are tagged with unique dual indexes (UDIs) in the first steps of the library prep **(B)** via sequential Tn5 transposition with full-length adapters. Following tagging, samples are pooled for purification and amplification. In the automated workflow, 8 µl of tagged DNA from each well in a row was pooled together **(C),** resulting in 8 pools each containing 12 tagged samples. All subsequent purification, amplification, and QC steps are performed on pooled libraries **(D)**. **This reduced tip usage over 80% compared to other library prep methods, significantly reducing costs in addition to reduced labor costs with the succinct three hour turn around time.**



Methods

- Eight 12-plex libraries were prepared from hgDNA (Coriell NA12878) using purePlex library preparation kit on the Sciclone G3 NGSx workstation, with four of the 8 libraries prepared using a fixed input of 10 ng DNA (row A to D) and the other 4 libraries using six distinct dilutions from 5 to 50 ng (row E to H). The same sample layout was used to prepare eight 12-plex libraries using purePlex manually.
- The 16 libraries were sequenced on an Illumina MiSeq Nano at 2 x 150 bp to assess read output uniformity across the whole 96-well plate for each method of preparation.
- A subset of the libraries made using variable inputs from manual and automated prep were used for deeper sequencing on an Illumina NovaSeq X at 2 x 150 bp.
- Sequencing data were individually down-sampled to 1M, 2M, 5M, 8M, 10M, and 15M random paired-end reads (0.1X, 0.2X, 0.5X, 0.8X, 1X, and 1.5X coverage, respectively) before variant calling and imputation.
- The open-source GLIMPSE pipeline (v2.0.0) performed imputation using default settings.

Benchmarking Automated and Manual Workflow



purePlex (8 x 12-plex) - Manual Control

	1	2	3	4	5	6	7	8	9	10	11	12	% of Total
A	1.2%	0.8%	0.8%	0.7%	0.9%	1.1%	0.7%	0.9%	0.9%	0.7%	1.1%	0.8%	1.7
B	0.9%	1.0%	0.7%	0.9%	1.2%	1.1%	0.9%	0.9%	0.8%	0.7%	1.0%	0.7%	1.6
C	1.3%	1.1%	1.5%	1.3%	1.0%	0.9%	0.9%	0.8%	0.8%	1.0%	1.0%	1.1%	1.4
D	0.9%	0.8%	1.2%	0.8%	0.9%	1.2%	0.9%	0.7%	0.9%	0.8%	0.8%	1.1%	1.3
E	0.7%	1.3%	1.0%	1.5%	1.0%	1.0%	1.4%	1.2%	1.4%	1.2%	0.8%	1.1%	1.1
F	0.8%	1.3%	1.2%	1.2%	1.4%	1.4%	1.1%	1.2%	1.2%	0.9%	1.7%	1.2%	1.0
G	1.5%	1.2%	0.9%	0.8%	0.8%	1.4%	1.1%	1.3%	1.0%	1.2%	1.2%	1.2%	0.8
H	1.0%	1.0%	0.8%	0.7%	1.4%	1.5%	1.7%	1.3%	0.9%	0.7%	1.1%	1.0%	0.7

purePlex (8 x 12-plex) - Automated Sciclone G3 NGSx Workstation

	1	2	3	4	5	6	7	8	9	10	11	12	% of Total
A	1.1%	0.8%	1.0%	1.5%	1.1%	1.1%	0.7%	1.3%	1.2%	1.4%	1.0%	1.2%	1.4
B	0.8%	1.2%	1.3%	1.5%	1.0%	1.1%	0.9%	1.2%	1.2%	1.3%	1.7%	1.1%	1.3
C	0.8%	0.8%	0.8%	1.5%	1.2%	1.1%	1.0%	1.2%	1.0%	0.9%	1.1%	1.5%	1.2
D	0.7%	1.0%	1.3%	1.2%	1.2%	1.0%	1.2%	1.0%	1.2%	1.3%	0.9%	1.2%	1.1
E	1.1%	1.1%	0.9%	1.1%	1.1%	1.0%	1.5%	0.8%	1.1%	0.7%	0.6%	0.7%	1.0
F	1.4%	0.9%	0.7%	1.4%	1.1%	0.9%	1.1%	1.1%	1.0%	1.1%	0.6%	0.9%	0.9
G	0.6%	0.8%	1.3%	1.2%	1.1%	1.1%	1.0%	1.1%	1.1%	1.1%	0.7%	0.7%	0.8
H	0.6%	1.1%	0.9%	0.9%	0.8%	0.8%	1.0%	0.8%	1.0%	1.1%	0.6%	0.7%	0.7

Figure 2. Uniformity of read output of the eight 12-plex libraries of NA12878 across fixed (row A-D) and variable DNA mass inputs of 5 to 50 ng (row E-H) for both manual (top) and automated (bottom) preparation of purePlex. Color gradient maps display the proportion of the sequencing run's total capacity that each sample occupies. The read output CV of 20% and 23% for manual and automated, respectively, across 96 samples with varying DNA input indicates the effectiveness of purePlex ability to auto-normalizes DNA inputs over an order of magnitude, which mitigates the need for extensive sample normalization and the efficiency of the Sciclone G3 NGSx Workstation on automating purePlex workflow, which reduces tip usage over 80% compared to other library prep methods and significantly reducing costs and labor.

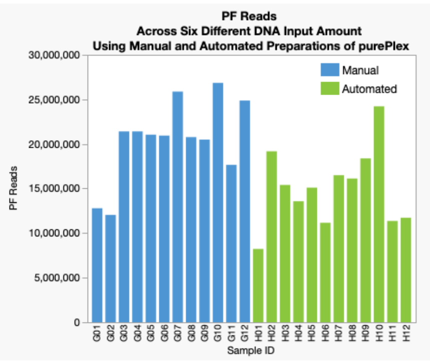


Figure 3. Comparison of read output of the two 12-plex libraries of NA12878 across six distinct dilutions from 5 to 50 ng DNA input prepared manually and automated. Blue bars represent the passed filter reads of each sample prepared manually using purePlex library preparation kit. Passed filter reads from the automated workflow on the Sciclone G3 NGSx Workstation are designated in green. The CV of read outputs across 6 DNA inputs (5, 10, 20, 25, 40, and 50 ng) for both manual and automated preparation are comparable at 22% and 29%, respectively.

Genotyping Imputation from Various Coverage Depth

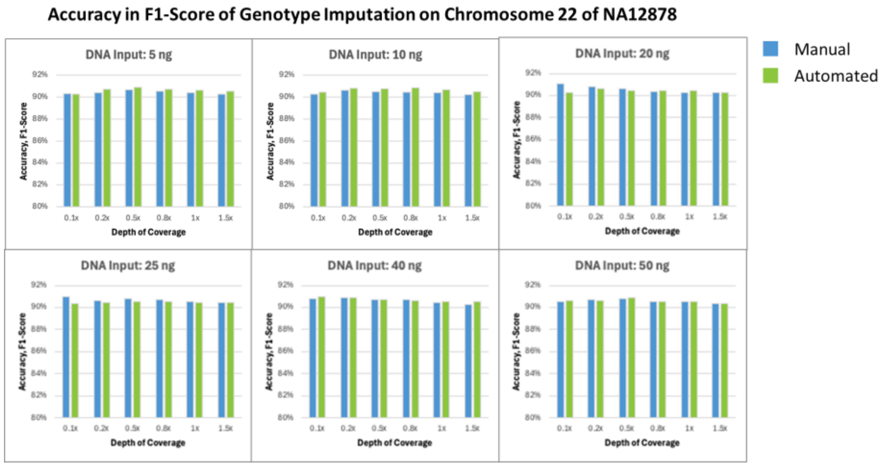


Figure 4. Accuracy in F1-score of genotype imputation for chromosome 22 in NA12878 prepared using purePlex library preparation kit manually and automated on Sciclone G3 NGSx Workstation at various depths of coverage across six different DNA inputs. The F1-score reflects the overall accuracy of the imputation process. A higher F1-score indicates better imputation performance, with a perfect score of 1 signifying perfect prediction of missing values. The bar graphs demonstrate that the proportions of correct imputed genotype remain high despite various DNA inputs used at various coverage depth, showing the equivalency of manual and automated preparation of purePlex on the Sciclone G3 NGSx Workstation.

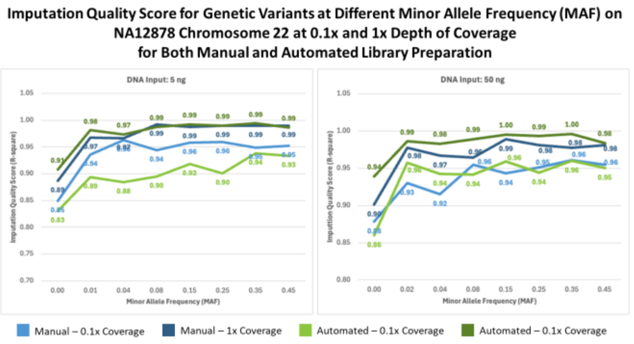


Figure 5. Imputation quality score for all genetic variants at different minor allele frequency (MAF) for chromosome 22 in NA12878 at 0.1x and 1x depth of coverage across low (5 ng) and high (50 ng) DNA inputs comparing manual and automated workflow of purePlex on the Sciclone G3 NGSx Workstation. The imputation quality score is an estimate of imputation quality on a scale of 0 to 1, where 1 indicates that a genotype has been imputed with high certainty.

Summary and Conclusions

- The purePlex Library Preparation Kit enables highly multiplexed and scalable library pool construction for low-pass WGS by alleviating the burden of individual sample normalization prior to sequencing and its ability to be automated on Revvity's Sciclone G3 NGSx Workstation (Figure 1, Figure 2).
- A high proportion of imputed calls (F1-score of 0.91) were identical to those in the truth data set (Figure 4) despite various DNA input used at various coverage depth. Increasing depth of coverage from 0.1x to 1x improved the imputation quality score for genetic variant MAF (Figure 5), though not statistically significant (p-value <0.05), suggesting high certainty of imputation can be obtained with minimal sequencing data.