One-step Library Preparation Method Supporting Skim-seq, Low Pass Sequencing Approach for Genotype Imputation

Michelle Rahardja, Stella Huang, Yanyan Liu, Jack Leonard, Rebecca Feeley, Jenna Couture, Joe Mellor

seqWell, Inc. Beverly, MA USA

Introduction

Recent advancements in next-generation sequencing (NGS) and bioinformatics have enabled whole-genome sequencing (WGS) to become a routine tool in both research and clinical applications. Low-coverage WGS, also known as skim-sequencing or "low-pass" sequencing, combined with imputation is a highly effective and cost-conscious alternative approach to microarray-based genotyping. Here, we demonstrate as a proof of concept, the use of the ExpressPlex™ 2.0 Custom High Strength Formulation, a one-step library preparation method, for skim-seq approach to genotype imputation on human samples. The ExpressPlex 2.0 Custom High Strength Formulation was used to process two individual human genomic DNA at four different total mass inputs, generating a normalized 8-plex library pool. A 2 x 150 bp run on a NextSeq 2000 P3 sequenced the ExpressPlex libraries to ≥ 20 million paired-end reads per sample. Paired-end reads for each sample were aligned to the GRCh38 human reference genome, the precision and accuracy of SNP calls were determined for chromosome 22. The open-source GLIMPSE pipeline performed imputation using default settings. Our results show the utility of ExpressPlex 2.0 Custom High Strength Formulation for routine low-pass WGS applications, where we characterize multiplexing uniformity and genotype imputation accuracy on a collection of reference samples.

ExpressPlex Library Preparation Kit Workflow

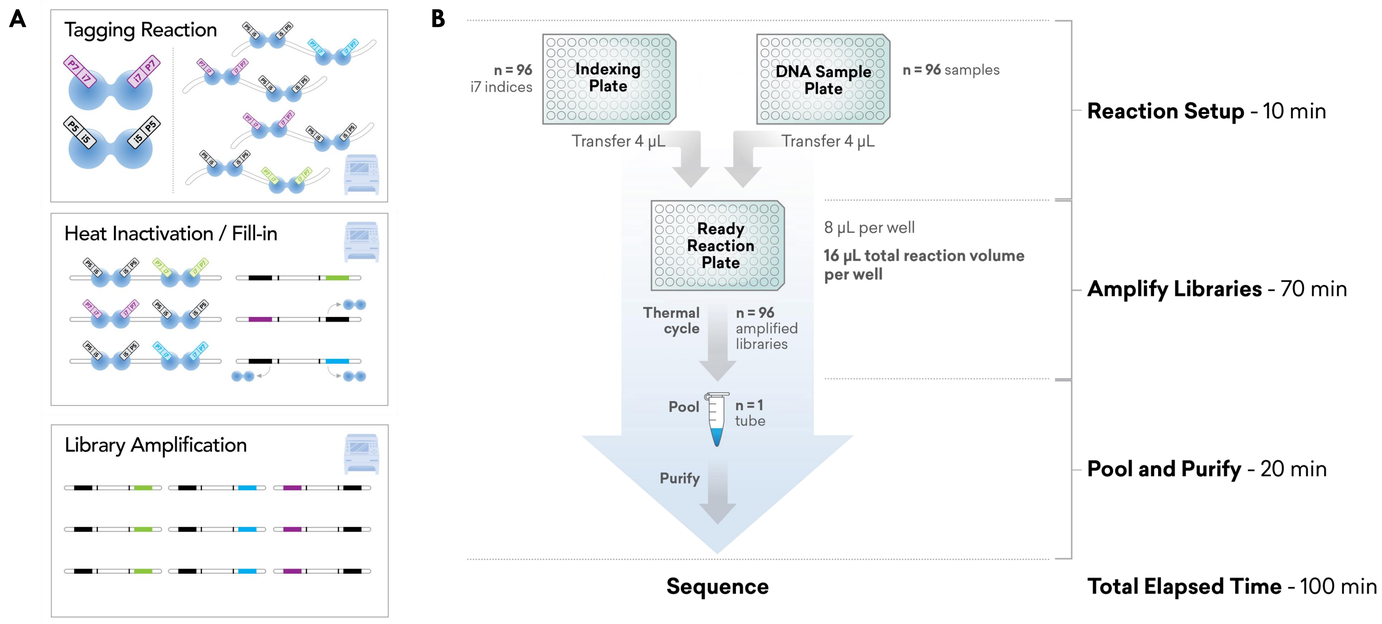

Figure 1. ExpressPlex 2.0 uses seqWell's high performance TnX™ transposase that was specifically engineered for NGS library preparation. (A) ExpressPlex 2.0 library preparation kits utilize a proprietary mixture of enzymes to tag input DNA with indexed adapters and amplify libraries all in a single reaction. Different full-length i7 indexed adapters tag the 96 DNA samples and barcoded libraries are amplified in separate wells, making for a highly efficient, one-step multiplexed library prep workflow. (B) Using the ExpressPlex 2.0 (96-well) kit, a 96-plex library can be prepared for library QC and sequencing in under 120 minutes, with less than 30 minutes of hands-on time. The ExpressPlex 2.0 Custom High Strength Formulation uses the same workflow as our standard ExpressPlex 2.0 Library Preparation Kit.

Methods

- Two individual human genomic DNA from the reference cohorts of CEPH/Utah and Han Chinese cohorts (Table 1) varying in mass input from 10 – 100 ng were processed using the ExpressPlex 2.0 Custom High Strength Formulation.
- The 8-plex ExpressPlex library was sequenced on the NextSeq 2000 (P3 300). Sequencing data were individually down-sampled to 1M, 2M, 5M, 10M, and 20M random paired-end reads (0.1X, 0.2X, 0.5X, 1X, and 2X coverage, respectively) before variant calling and imputation.
- The open-source GLIMPSE pipeline (v2.0.0) performed imputation using default settings.

Table 1. Summary of hgDNA samples assessed in the study.

Coriell ID	NIST ID	Reference Cohort	Sample ID	DNA Input Amount (ng)
NA12878	HG001	CEPH/ Utah	C01	10
			C02	20
			C03	50
			C04	100
NA24631	HG005	Chinese	C05	10
			C06	20
			C07	50
			C08	100

Genotyping Imputation from Various Depth of Coverage

Figure 2. Accuracy of low-pass WGS genotyping for two individual human genomic samples with ExpressPlex 2.0 library prep on a NextSeq System at various depth of coverage across four different DNA inputs.

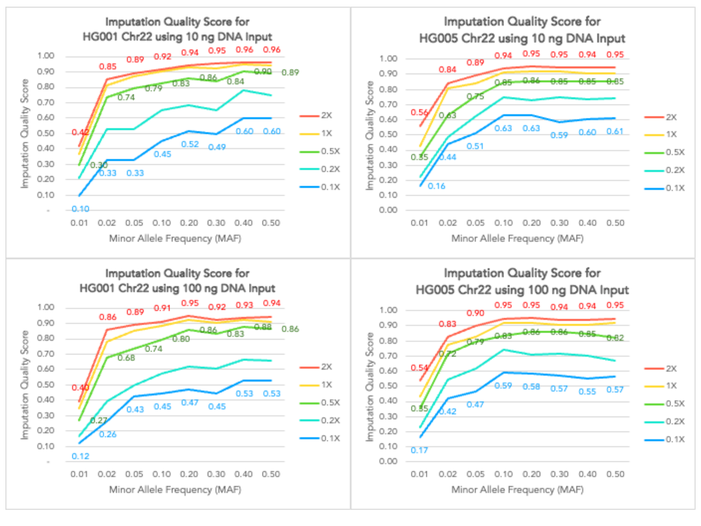

Figure 3. Imputation quality score for all genetic variants at different minor allele frequency (MAF) for chromosome 22 in the GIAB Consortium HG001 and HG005 at various depth of coverage across low (10 ng) and high (100 ng) DNA inputs. The imputation quality score is an estimate of imputation quality on a scale of 0 to 1, where 1 indicates that a genotype has been imputed with high certainty.

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. Figure 2 and Table 2 show the proportion of correct imputed genotype remains high despite various DNA input used at various coverage depth.

Table 2. Summary of genotype imputation accuracy for chromosome 22 in HG001 and HG005.

Number of Paired End Reads	Coverage Depth	Imputation Accuracy
1M	0.1X	98.3%
2M	0.2X	98.6%
5M	0.5X	98.9%
10M	1X	98.9%
20M	2X	99.0%

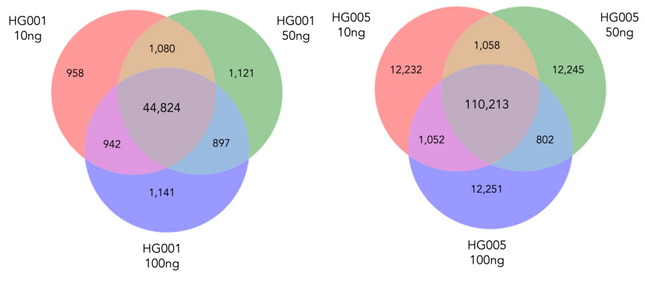

Figure 4. Unweighted venn diagrams represent high chromosome 22 SNP concordance for three replicates at different DNA input (10 ng, 50ng, and 100ng) of two individual human samples HG001 (left) and HG005 (right) at 1X coverage.

Sequencing Metrics

Table 3. Sequencing metrics summary of 8-plex ExpressPlex 2.0 library performance on the NextSeq 2000 after samples were down-sampled to 10M paired-end reads (1X coverage).

Reads Aligned	Average Mean Read Length	Average Median Insert Size	Average Duplication Rate	Mean Coverage Depth (X)
99.8%	140.3	301 nt	4.8%	1.31

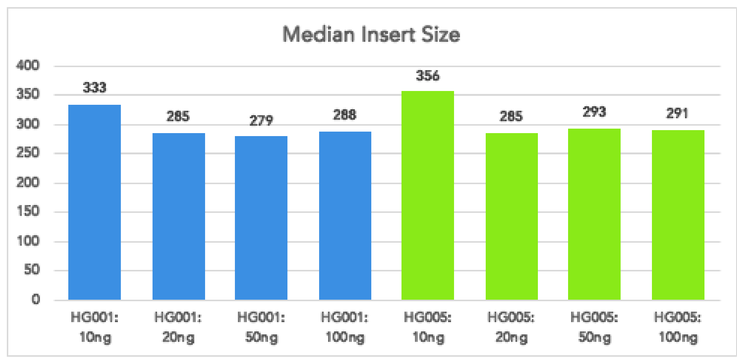

Figure 5. Insert size uniformity of an 8-plex ExpressPlex multiplexed library across two individual human genomic DNA HG001 and HG005 at various DNA input: 10, 20, 50, and 100 ng.

Summary and Conclusions

- The ExpressPlex 2.0 Custom High Strength Formulation supports truly multiplexed and highly scalable construction of library pools for low-pass WGS through its one-step workflow (Figure 1).
- A high proportion of imputed calls (>98%) were identical to those in the truth data set (Figure 2, Table 2) despite various DNA input used at various coverage depth, confirming that high accuracy of genotyping can be achieved from a minimal sequencing data.
- Increasing depth of coverage from 0.1X to 1X significantly improve the imputation quality score for genetic variant MAF (Figure 3).