Low-pass whole genome sequencing as a cost-effective and improved alternative to genotyping arrays

Peng Zhang, Hua Ling, Justin Paschall, Beth Marosy, Jessica Gearhart, Kimberly Doheny

Center for Inherited Disease Research (CIDR), Johns Hopkins Genomics, Institute of Genetic Medicine, The Johns Hopkins School of Medicine

Contact: Peng Zhang (pzhang13@jhmi.edu), Kimberly Doheny (kdoheny@jhmi.edu)

PB3067

Introduction

We tested using low-pass whole genome sequencing (lpWGS) as a cost-effective, and potentially improved alternative for SNP arrays in genome-wide association studies (GWAS).

We designed and analyzed an experiment with 92 samples to evaluate aspects of practical implementation including: library preparation methodology, impact across ancestry groups, lpWGS coverage levels, DNA source, imputation algorithm, imputation reference and compute resources.

The validation 'truth' dataset was created by deep sequencing 92 samples with diverse ancestry background (down sampled to the same coverage of 28X). The lpWGS 'test' dataset ("lps") was generated using a cost-effective and high-throughput library prep (plexWell LP384 from seqWell) for the same set of 92 samples to target 1X coverage, then imputed to the 1000 Genomes (1KG) deep sequencing reference panel (NYGC) using GLIMPSE.

The SNP array 'test' dataset ("GSA") was in silico array data derived by extracting the Illumina Global Screening Array (GSA, 654K variants) genotypes from the deep WGS validation dataset, and then performing imputation with the 1KG reference panel, and the more recent TOPMed reference panel to mimic current standard practice for GWAS studies.

Library Prep & Data Generation

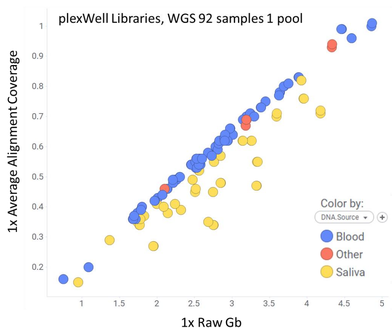
Validation ('truth') deep WGS dataset: PCR-free library was created with 500-750ng of genomic DNA, sheared , and processed using the Kapa Hyper Prep kit (Roche) for End-Repair, A-Tailing and Ligation of IDT (Integrated DNA Technologies) unique dual-indexed adapters according to the Kapa protocol. The Illumina NovaSeq 6000 platform was used to generate raw sequencing data (2x150). Coverage normalized to 28x, alignment and variant calling was performed using the DRAGEN 3.7.5 pipeline on the Illumina BaseSpace cloud platform.

lpsWGS dataset: Libraries were generated using 10ng of genomic DNA. Samples underwent a transposase-based library prep using the plexWell LP384 kit (SeqWell) which attaches a unique sample (i7) barcode directly into the input DNA. After pooling (92->1), samples undergo a second barcoding step which inserts the secondary (i5) unique pool barcode into each sample within the pool. Dual barcoded, pooled libraries were amplified using 8 cycles followed by a bead-based size selection. Sequencing reads and alignment were done the same as the deep WGS with a targeted coverage of 1x.

GLIMPSE imputation: We used BCFtools (v1.9) to compute the Genotype Likelihoods (GLs) for all bi-allelic sites in the reference panel using the lpWGS reads in CRAM format. GLIMPSE (v1.1.1) then used those GLs as input to impute the genotypes for all the SNVs from to the 1KG deep sequencing reference panel (NYGC).

Traditional Library Prep with Enzymatic Fragmentation		PlexWell LP384	
Steps	***Sample wells***	***Steps***	***Sample wells***
Frag/Atail	768	Tagmentation/Stop	768
Ligation	768	Pool/Clean Up	32
Clean Up	768	Barcode/Stop	32
Amplification	768	Clean Up	32
Clean up	768	Amplification	32
QC	768	Clean Up	32
Pool	32	QC	32

Table 1. Workflow comparison. The plexWell LP384 prep reduces the number of sample wells from 768 to 32 after the first step, where traditional library prep continues with 768 samples until the last step of the process.



The plexWell LP384 prep workflow saves time and cost:

Time/Labor saving: 1 day versus 3 days, automation required for processing 768 samples only for 1st step versus all steps.

Cost saving: reduced library prep costs by 44% when collapsing down to 32 sample wells early without the need to purchase additional equipment specifically for lp384 kit.

Fig 1. The plexWell chemistry uses a reagent limiting initial transposition step which normalizes the sample input into pooling. Here we show the 1x average alignment coverage (y-axis) vs the 1x Raw Gb (x-axis) for each of the 92 samples processed into 1 pool. Input sample DNA was derived from both blood and saliva, as saliva is a common DNA source used in array studies.

Genetic Ancestry Determination: We determined five genetic ancestry groups by performing principal component analysis (PCA) on the 92 samples (GSA) along with an external set of 1201 unrelated HapMap 3 samples genotyped on the Illumina 1M array. Results: 9 African, 6 Admix_CEU_African, 69 European, 7 Admix_CEU_Asian and 1 Asian.

Contamination Detection

To determine if we could detect sample mixtures, we derived in silico mixtures from the lps datasets. We selected six sample pairs to mix, with the same or similar genetic ancestry and similar level of coverage (within 0.1x). Sample mixtures were detectable at a 5-10% range (dependent on coverage), similar to detection limits for arrays.

Table 2. Contamination results from HaploCheck, which examines mitochondrial DNA and returns a contamination metric. verifyBamID did not return results for lps as it was designed for higher sequencing depths.

0.5X			**1X**		
Mixture	Contamination prediction percent	MT avg cov	Mixture	Contamination prediction percent	MT avg cov
50%	42.90%	67	50%	48.10%	138
30%	29.40%	68	30%	32.40%	137
15%	16.50%	71	15%	17.20%	138
10%	10.30%	69	10%	12.10%	138
7%	None	NA	5%	6.70%	135
5%	None	NA	2%	None	NA
2%	None	NA			

lpWGS on Illumina DRAGEN v4.0

Recently Illumina DRAGEN Server v4 incorporated GLIMPSE with accelerated running speed and a single command-line instruction that allows imputation of lpWGS. A high quality IRPv1 reference panel curated from the NYGC 1000G dataset is also provided. Thus allows an integrated pipeline from processing raw reads to running lpWGS imputation all from one place.

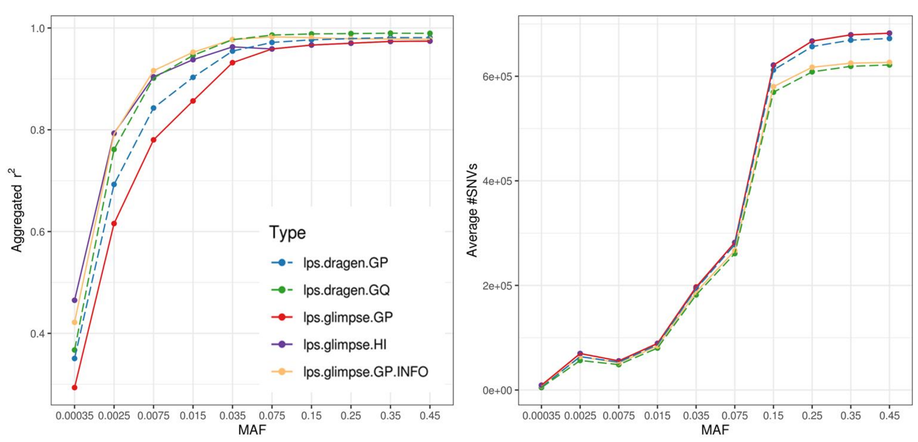


Fig 2. r2 concordance and number of SNVs for filtered lpWGS GLIMPSE imputation run on local server (lps.glimpse) versus on DRAGEN v4 (lps.dragen).
Shown are the HI, GP, and GP & INFO filters for the local run (solid line), and GQ (≥20) and GP filters for the DRAGEN runs (dashed line).
GQ of 10 is equivalent of max(GP) of 0.90, and GQ of 20 is similar as the GP&INFO for the local run.

Analyses & Results

Table 3. Sensitivity (%) and other metrics (%) from different filters compared to validation (average #SNVs: 4,043,555) dataset.

Filter.Type	#SNVs (m)	Sensitivity	FPR	FNR	NTPR
lps.glimpse.all	61,715,567 (100)	89.7	0.140	0.157	96.8
lps.glimpse.GP	61,294,115 (99.32)	85.0	0.055	0.070	98.6
lps.glimpse.INFO	59,218,836 (95.95)	82.5	0.116	0.067	97.0
lps.glimpse.GP.INFO	59,051,021 (95.68)	79.3	0.050	0.035	98.7
lps.glimpse.HI	59,791,338 (96.88)	85.2	0.083	0.054	98.2
GSA.all.TOPMed	292,323,483 (100)	84.1	0.092	0.022	92.3
GSA.R2.TOPMed	16,018,114 (5.48)	81.3	1.632	0.301	92.2
GSA.all.1KG	43,779,550 (100)	83.3	1.47	0.384	81.9
GSA.R2.1KG	13,377,178 (30.56)	79.1	4.36	0.909	82.1

Results shown are average across 92 samples and all autosomal chromosomes (chr1-22). Filter.type: for lps, we applied GP≥0.9, INFO≥0.3, HI was defined as INFO≥0.3 for MAF<0.05 and GP≥0.9 for MAF≥0.05. For GSA, R2≥0.3. Sensitivity here is the non-reference sensitivity. **Note:** Validation data have no reference homozygote (0/0)

Sensitivity = (e+f+h+i)/ (b+e+f+h+i+k)
FPR = FalsePosRate = (d+h)/m
FNR = FalseNegRate = (b+f)/m
NTPR = Non-referenceTruePositiveRate = (e+i)/(d+e+f+h+i)
m is the #SNVs in test datasets (a+b+d+e+f+h+i)

		Validation WGS(28X)		
		NA	0/1	1/1
Lps or GSA	0/0	a	b	
	0/1	d	e	f
	1/1		h	i
	No call	*NA*	k	

Fig 3. Concordance (r-squared Pearson correlation) and number of SNVs compared from different filters compared to the validation dataset.

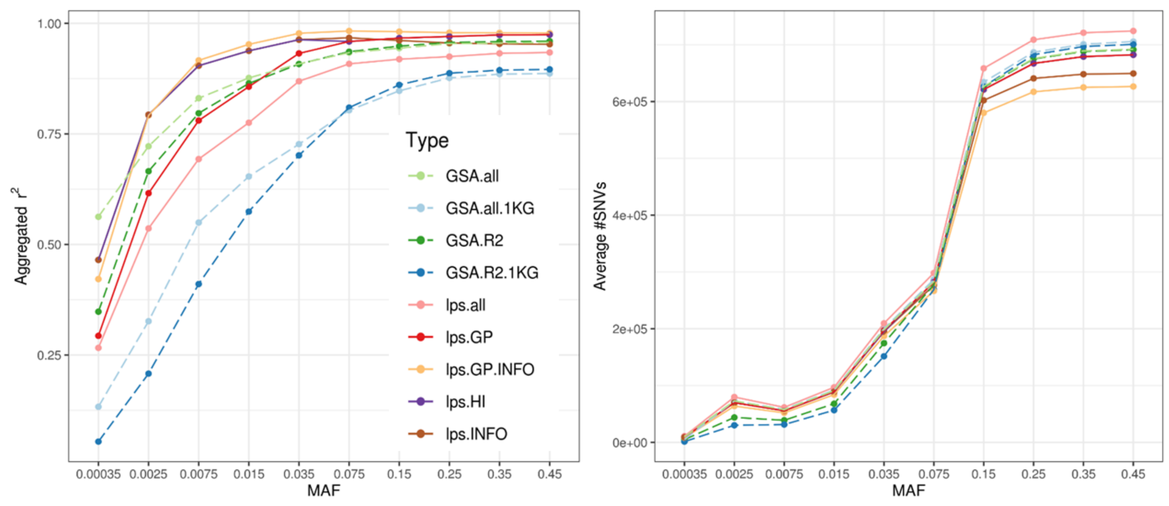
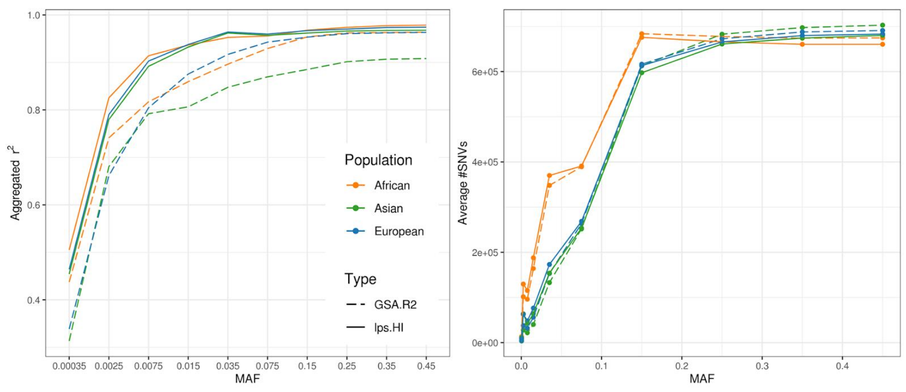


Fig 4. Concordance and # of SNVs compared by genetically defined ancestry.
lpWGS has higher concordance is less affected by genetic ancestry than array.



Ethnicity check: **Fig 5.** showed comparable PCA results from lpWGS as those from GSA.
Population refers to self-reported race.

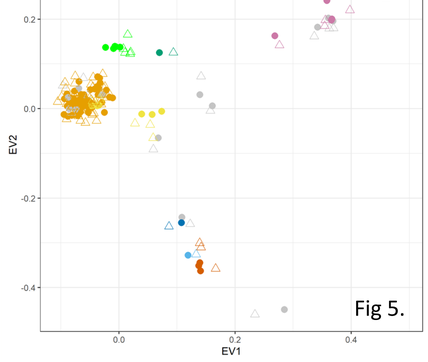


Fig 5.

Relatedness check: **Fig 6**. showed similar Identical By Descent (IBD) results by KING using GSA as results from lpWGS (**Fig 7**).

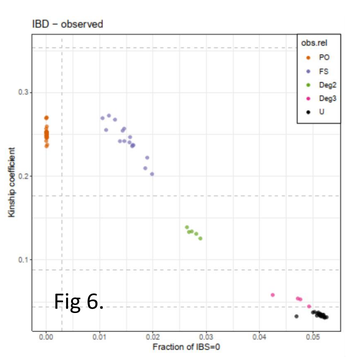


Fig 6.

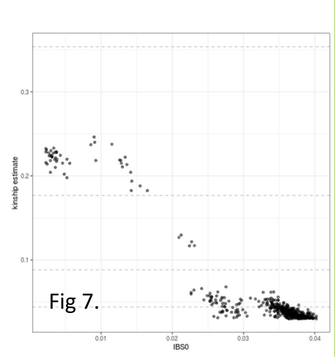


Fig 7.

Summary & Discussions

- We demonstrated the feasibility of using lpWGS as a promising alternative for SNP array with currently available resources.
- The plexWell LP384 library prep method provides the dramatically increased throughput and reduced costs required to replace array for extremely large studies.
- Using results from deep coverage WGS (28X) as the validation dataset, we show that lpWGS (1X) can achieve higher sensitivity and concordance compared to array based imputation.
- The gain in sensitivity and concordance of lpWGS is more prominent for non-European ancestry or less common variants (e.g. MAF < 5%).
- Using lpWGS data, we can detect sample mixtures, and perform analysis for ethnicity (PCA) and relatedness (IBD) with similar results as using array data.
- The recent incorporation of lpWGS to illumina DRAGEN v4 greatly simplified the pipeline with accelerated speed and a high quality reference panel, making the lpWGS more accessible as an alternative to array.

1. GLIMPSE: Rubinacci Simone et al. (2021) *Nature Genetics*, l53:120-126; 2. HaploCheck: Weissensteiner H, et al. (2021) *Genome Research*, 31: 309-316;
3. GWAStools: Gogarten SM et al. (2012) *Bioinformatics*, 28(24), 3329-3331; 4. Summary metrics FPR etc: Kishikawa T et al. (2019) *Scientific Report*, 9:1784