A Unified Sequencing Workflow for Population-Scale Genomic Studies

B Marosy[1], J Gearhart[1], M Kokosinski[1], J Paschall[1], P Zhang[1], H Ling[1], M Mawhinney[1], C Oncagco[1], L Aker[1], B Craig[1],K Hetrick[1] and K Doheny[1]
[1]Johns Hopkins University, Department of Genetic Medicine, Baltimore, MD, USA

Introduction:

Recent advances in next generation sequencing chemistry and instrumentation have significantly reduced per-GB costs, making low pass whole genome sequencing (lpWGS) followed by imputation (AKA genotyping by sequencing) a viable alternative to array genotyping for genome-wide association studies. At the Center for Inherited Disease Research (CIDR), we aim to provide the best quality genetic services at the lowest cost possible, by continuous re-evaluation and optimization of protocols. In a large pilot study (n=5,600), we evaluated library preparation protocols for lpWGS to develop a cost-effective, streamlined method competitive with array genotyping. In addition, while deep whole genome sequencing (WGS) remains the gold standard for comprehensive genetic data, its cost limits scalability. Combining lower coverage whole exome sequencing (WES) with lpWGS (1-2x) followed by imputation, efficient genome sequencing (EGS), offers comparable association results to WGS data at a lower cost (7-fold reduction in reagent cost) [1,2]. We further evaluated the lpWGS library prep methods for applicability to small capture panels, enabling a unified workflow at the bench, covering all current population scale sequencing services.

Methods: Workflow

Traditional LP w/Enzymatic Fragmentation			96 Prep - lpWG			FlexPrep - lpWG		
Steps	1 Plate Batch	8 Plate Batch	Steps	1 Plate Batch	8 Plate Batch	Steps	1 Plate Batch	8 Plate Batch
Frag/ER-ATail	96	768	Primer Extension *(Sample Barcoding)*	96	768	Frag/ER-ATail	96	768
Ligation *(Sample Barcoding)*	96	768	Pooling	1	8	Ligation *(Sample Barcoding)*	96	768
Clean Up	96	768	Capture/Wash	1	8	Pooling	8	64
Amplification	96	768	Primer Extension	1	8	Bead Clean Up	8	64
Clean Up	96	768	Amplification *(Pool Barcoding)*	1	8	Amplification *(Pool Barcoding)*	8	64
QC	96	768	Size Selection Bead Clean up	1	8	Bead Clean Up	8	64
Pool	1	8	QC	1	8	QC	8	64

Table 1a: Workflow comparison between traditional library prep, 96-prep and FlexPrep. In both 96-prep and FlexPrep, samples are collapsed/pooled earlier in the workflow providing a scalable method which reduces overall costs without the need to purchase additional equipment specifically for higher throughput. By reducing the sample footprint within each plate early in the process up to a seven-fold reduction in cost savings for QC reagents and consumables is achieved.

FlexPrep - EGS		
Steps	1 Plate Batch	8 Plate Batch
Frag/ER-ATail	96	768
Ligation *(Sample Barcoding)*	96	768
Pooling	8	64
Bead Clean Up	8	64
Amplification *(Pool Barcoding)*	8	64
Bead Clean Up	8	64
QC	8	64
Pooling & QC	1	8
Hyb	1	8
Capture/Wash	1	8
Amplification (8 cycles)	1	8
QC	1	8

Library Prep

Enrichment

Portion retained for lpWGS

Exome library re-combined with lpWG at specific ratio

Table 1b: Workflow using Twist FlexPrep as library prep for Efficient Genome Sequencing. Post library prep, 12 x 8-plexes are pooled to a final 96 plex pool. A portion is retained as the lpWG library, while the rest undergoes exome enrichment. After enrichment, both portions are recombined to form a final 96 sample pool containing exome-enriched and lpWG molecules per sample. Multiplexing 96 samples in one pool for enrichment reduces costs and scales throughput. For small captures, the same workflow is applied without retaining any lpWG material.

Methods/Results: Initial pilot using 96-prep and adoption of FlexPrep for lpWG

Initially 2,112 samples of the lpWG pilot study were processed with the Twist 96-prep library kit using 50ng of gDNA as input according to manufacturer protocol. Sample barcodes were incorporated during the initial step using a random primer extension, followed by pooling which collapses 96 samples down from into a single well. The 96-plex pool undergoes a streptavidin bead capture to isolate the extended product. The pool is amplified with indexed primers incorporating a pool barcode based on each 96-plex. (Table 1a)

The Twist FlexPrep library prep kit was then adopted into practice and used to process 1524 samples of the pilot study using 50-100ng of gDNA as input according to the manufacturer protocol. Samples underwent enzymatic fragmentation, end-repair and A-Tailing in one step, followed directly by ligation with incorporation of short normalizing adapters containing inline barcodes to differentiate between samples and generate even amounts of library post ligation. Samples are pooled from 96 individual wells into 8 x 12-plex pools and undergo amplification with dual indexed primers to differentiate between each 12 plex pool. (Table 1a)

Sequencing was performed on the Illumina NovaSeqXPlus platform. Samples are demultiplexed in a two-step process. Using Illumina BCLconvert, the BCL files are first converted to FASTQ, with each file containing a 12-plex pool. Fgbio is then used to further demultiplex the 12plex FASTQs by each inline barcode to generate individual sample level FASTQs. For imputation, the pipeline runs DRAGEN germline variant calling (v4.3.6f) followed by imputation on the Illumina Connected Analytics (ICA). The imputation reference panel is the Human Genome Diversity Project (HGDP)+1kGP version 2. Concordance was determined against the Illumina Global Screening Array (GSA). (Table 2)

Table 2: Compares sequencing QC pipeline metrics between samples process using 96-prep and FlexPrep for the lpWG pilot. Due to the smaller insert size obtained with FlexPrep, sequencing was performed as 2x100. An increase in uniformity was observed with the FlexPrep samples, which improved the call rates by decreasing the missing rate. Concordance to Illumina GSA array was 97.5% and 98.5% between 96 prep and Flex Prep methods.

lpWG Sequencing QC Pipeline Metrics		
Library prep	96 Prep - lpWG	FlexPrep - lpWG
Sample number	2112	1524
DNA input (ng)	50	50-100
Insert size (bp)	380	263
Read Length	2x150	2x100
Mean coverage	1.69	1.39
Pct at 1x	52.45	60.46
Pct dups	39.7	33.49
Sequencing Yield (raw Gb)	9.6	7.56
Imputation Call Rate	99.1536	99.4731
F_Missing Q10	0.008464	0.005269
Uniformity	52.25	60.46
Concordance	97.5116 (n=2095)	98.5597 (n=1494)

Methods/Results: Utility of FlexPrep for small capture panels

The Twist FlexPrep library prep and enrichment kit was used to process 96 HapMap samples using 100ng of gDNA as input according to the manufacturer protocol. In 2 separate experiments, 10ug of the 96 plex library was captured with a small custom panel (0.5Mb) and a mitochondrial capture (16kb). (Table 1b)

Sequencing was performed on the Illumina NovaSeqXPlus platform. Samples are demultiplexed in a two-step process. Using Illumina BCLconvert, the BCL files are first converted to FASTQ, with each file containing a 12-plex pool. Fgbio is then used to further demultiplex the 12plex FASTQs by each inline barcode to generate individual sample level FASTQs. Sample level FASTQ were aligned with BWA mem and variant calling with GATK/Haplotype Caller. (Table 3)

Table 3: Summary of the sequencing qc metrics produced from processing 96 HapMap samples enriched with a small capture (0.5Mb) and mitochondrial panel (16kb) in 2 independent experiments. Samples were pooled down to 1-96plex for both enrichments. While over sequenced, we obtained on average 195x coverage from 0.4 raw Gb yield per sample with 96% of bases covered at 20x for the small capture. The mito panel reached 98% on target at 5000x and 88% on target at 10000x coverage. This provides feasibility data for increased multiplexed enrichments and the flexibility of utilizing this method for multiple applications.

Metric	Small capture	Mito
Library prep	FlexPrep	FlexPrep
Capture product	Custom panel	Mito panel
Capture size (Mb)	0.524	0.016
Sample number	96	96
DNA Input (ng)	100	100
Insert size (bp)	242	211
Read Length	2x150	2x150
Mean coverage	195	16313
Sequencing yield (raw Gb)	0.388	4.2
Pct Duplication	8.37	15.6
Pct on Target at 10x	97.07	na
Pct on Target at 20x	96.73	na
Pct on Target at 10000x mito only	na	88.28
Pct on Target at 5000x mito only	na	98.47

Methods/Results: Utility of FlexPrep for Efficient Genome Sequencing (EGS)

The Twist FlexPrep library prep and enrichment kit was used to process 96 HapMap samples using 100ng of gDNA as input according to the manufacturer protocol, except for retaining a small portion of the initial library as reserve for lpWG sequencing downstream. Enrichment was performed on the 96-plex pool in one reaction with 10ug input of pooled library and the CIDR Custom Exome panel (35.1Mb). Post Enrichment, the exome library was combined with the lpWG portion at a ratio 64:36 (lpWG:Exome). (Table 1b)

Sequencing was performed on the Illumina NovaSeqXPlus platform. Samples are demultiplexed in a two-step process. Using Illumina BCLconvert, the BCL files are first converted to FASTQ, with each file containing a 12-plex pool. Fgbio is then used to further demultiplex the 12plex FASTQs by each inline barcode to generate individual sample level FASTQs. The imputation method of lpWG here is the same as lpWG only stated in previous initial pilot data set. Non-Reference Concordance to HGDP+1kGP was performed on 87/96 samples using the lpWG data. The exome portion of the EGS sequencing was aligned with BWA mem and variant calling with GATK/Haplotype Caller. Concordance to Illumina GSA array was performed on 53/96 samples. (Table 4)

EGS Library Prep, Catpure and Sequencing Metrics			
Library prep	FlexPrep - EGS	PCR cycles PostHyb	8
Sample number	96 (HapMaps)	Insert size (bp)	230
Capture product	CIDR custom exome	Read length	2x150
Capture size	35.1 Mb	Sequencing yield (raw Gb)	8
DNA input (ng)	100	Pct duplication	7.7
Library yield after pooling 96plex (ug)	22	TiTv dbSNP 129	3.06
PCR cycles PreHyb	6	Pct dbSNP138 SNV	98.8005
Hyb input (ug)	10	Pct dbSNP138 INDEL	92.881
Combined Ratio lpWG:Exome = 64:36			
lpWG		**Exome**	
Mean coverage	1.2	Mean coverage	32.5
Pct at 1x	66.86	Pct at 10x	96.76
Imputation call rate	99.7332	Pct at 20x	82.5
Uniformity	66.86	Pct selection*	40.8
Non-ref concordance SNP	97.5 (n=87)	Fold 80	1.45
Non-ref concordance Indel	61.6 (n=87)	AT dropout	4.0
*Pct Selection takes into account all reads (lpWG & exome)		GC dropout	1.0
		Estimated library size (Mil)	172
		Concordance (n=53)	99.3428
		Sensitivity 2 het (n=53)	98.7277

Table 4: Summary of the library prep, capture and sequencing metrics produced from the EGS sequencing. From a single library, both lpWG and Exome libraries were generated and combined at a ratio of 64:36 (lpWG to Exome). LpWG results achieved on average 1.2x coverage with 99.7% imputation call rate. Non-reference SNP concordance to HGDP+1kGP data was 97.5%. WES results achieved on average 33x coverage from 8 raw Gb yield, with 97% and 83% of bases covered at 10x and 20x, respectively. Concordance to Illumina GSA array was 99.3% and sensitivity to heterozygous SNPs was 98.7%.

Discussion/Conclusion:

- Evolving library prep methods that provide improvements to uniformity will help to increase completeness of coverage and reduce missing rates.
- Sample level indexing combined with pool indexing provides cost reductions by reducing the sample footprint early in the library prep process, minimizing labor and automation costs, qc reagents and consumables.
- Enrichment methods and reagents that allow for increased sample multiplexing up to 96 plex, further reduce costs.
- One workflow solution offers more flexible and cost-effective applications (low pass whole genome sequencing, efficient genome sequencing or small capture sequencing) for population scale genome studies.

References:
[1]Gaynor et al, Nature Genetics 2024 - https://doi.org/10.1038/s41588-024-01930-4
[2]DeFelice et al BioRxiv 2024 - https://doi.org/10.1101/2024.04.03.587209