Development of a Low Input FFPE workflow for Whole Exome Sequencing

B Marosy, B Craig, K Hetrick, PD Witmer, H Ling, S Griffith, B Myers, KF Doheny
Center for Inherited Disease Research (CIDR), IGM, JHU-SOM, Baltimore, MD

Introduction

The Center for Inherited Disease Research (CIDR) provides high quality next-generation sequencing (NGS), genotyping and statistical genetics consultation to investigators working to discover genes that contribute to disease. Historically, sample requirements for a variety of targeted selection and NGS platforms have been a limiting factor for investigators. Formalin Fixed Paraffin Embedded (FFPE) samples are a useful source of DNA for NGS studies but present additional challenges in the lab due to age of the sample, damage to the DNA from fixation resulting in changes to the nucleotide sequence and fragmentation of the DNA. Evaluation and optimization of a variety of conditions and commercial kits has enabled us to establish a workflow for low input and compromised samples for library prep using 50ng of DNA to generate high quality exome enriched sequencing data.

Results

Comparison of library kits for low input

50ng of HapMap DNA (NA12891 & NA12892) was used as input into library prep using the Kapa Hyper Prep (Fig 1a) and Swift 2S (Fig 1b) library kits, processed according to the manufacturer's protocol[1,2]. Eight experimental samples (50ng DNA input) were processed using the Kapa Hyper Prep kit. Additional data generated from these two HapMap samples was used from a previously processed project using CIDRs current production protocol (Fig 1d)[3,4].

Figure 1: Library Prep Workflows for 1a) Kapa Hyper Prep Kit; 1b)Swift2S Library prep; 1c)Swift1S; 1d)CIDR's standard workflow using a modified Agilent XT workflow.

Figure 1a – Kapa Hyper Prep workflow

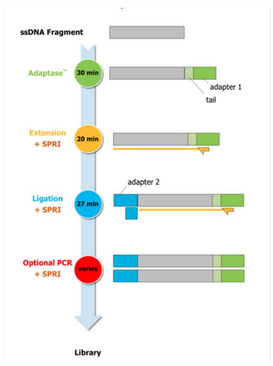

Figure 1c – Swift 1S workflow

Figure 1d – CIDR's standard workflow

Figure 1b – Swift2S workflow

Table 1 compares QC metrics generated during library prep, capture and sequencing. Overall QC metrics are comparable between the three methods. Insert size for the Kapa Hyper prep method was lower than Swift 1S and CIDR's standard method. Percent duplication was similar between methods when comparing HapMap samples. Percent selection metrics were based on data prior to downsampling.

Table 1. Low Input - Library Prep, Capture and Sequencing Metrics

	Standard	LowInput		
Experiment	CIDR	KapaHyperPrep HapMap	Swift2S HapMap	KapaHyperPrep Experimental
Library and Capture Metrics				
Input DNA [ng]	1000	50	50	50
Yield at PCR Amp [ng]	964	1513	635	1114
Capture	v5+UTR	v5+UTR	v5+UTR	v4
% Selection	**74.26**	**80.13**	**79.42**	**85.42**
Sequencing Data Quality				
Concordance	100	99.8388	99.8177	99.8148
Sensitivity (Het)	95.4471	94.7577	96.6945	96.7783
TiTv (ALL_Ratio)	**3.0988**	**3.1473**	**3.1012**	**3.0305**
Sequencing Data Output				
Raw GB	5.089	5.175	5.176	7.709
Mean Target Coverage	36	39	40	90
% On Target @ 10x	**92.66**	**92.28**	**94.63**	**98.11**
% Zero Bases	0.23	0.41	0.24	0.18
Mean Insert Size	**299**	**233**	**270**	**227**
% Dups	**2.32**	**2.41**	**2.38**	**6.28**

Comparison of library kits for degraded/small fragment size

Due to degradation, small fragment size presents a challenge for generating libraries using traditional workflows. The Swift Accel-NGS™ 1S DNA Library Kit uses single stranded ligation to prepare libraries (Fig 1c). The Kapa Hyper prep kit combines End Repair/A-Tailing into one step without a clean up prior to ligation (Fig 1a). Two HapMap samples (NA12891 & NA12892) underwent 40min of shearing using the Covaris® E210 to generate fragment sizes ~100bp in size. 500ng of DNA was used as input into library prep and processed according to the manufacturer's protocol[5]. An experimental sample, extracted from a 38yr old tissue slide (500ng DNA input) was processed using the Kapa Hyper prep kit. No shearing was performed on this sample since initial QC showed that the sample was highly fragmented ~100bp (data not shown).

Degraded/small fragment size – Cont'd

Table 2 compares QC metrics generated during library prep, capture and sequencing. Kapa has a higher sensitivity than Swift1S. Duplication rate was lower with the Swift 1S libraries. The tissue sample has a much lower TiTv ratio and higher duplication rate.

Table 2. Small Fragment Size – Library Prep, Capture and Sequencing Metrics

	Degraded		
Experiment	Degraded HapMap Kapa	Degraded HapMap Swift1S	Tissue Sample Kapa
Input DNA [ng]	500	500	500
Shear Time	40min	40min	0 time
Yield at PCR Amp [ng]	524	711	607
Capture	v5+UTR	v5+UTR	v4
% Selection	**81.83**	**66.63**	**48.65**
Sequencing Data Quality			
Concordance	99.7489	99.5694	na
Sensitivity (Het)	**86.3392**	**78.9622**	na
TiTv (ALL_Ratio)	**3.1265**	**3.1312**	**2.0063**
Sequencing Data Output			
Raw GB	5.612	5.231	4.998
Mean Target Coverage	48	27	23
% On Target @ 10x	**87.5**	**80.7**	**73.88**
% Zero Bases	1	1.16	0.71
Mean Insert Size	**97**	**145**	**100**
% Dups	**7.68**	**4.97**	**23.68**

Further analysis of the tissue sample revealed a large amount of degradation (Fig 2a and 2b) which impacted the quality of the data. This was anticipated due to the age and condition of the sample. Figure 2a shows an increase in the number of novel SNPs compared to other non-degraded samples, as well as a decrease in the percent of SNVs found in dbSNP138. Figure 2b shows an increase in the oxidative error rate from this sample indicating poor quality.

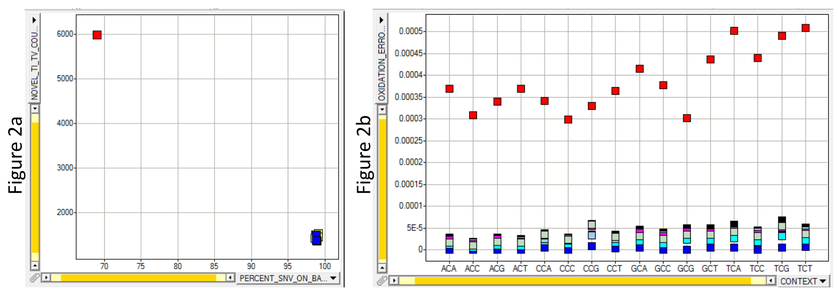

Figure 2a – Percent SNV in SNP138 and counts of novel SNV metrics were generated using CIDRSeqSuite v6.0. Oxidation metrics were generated using Picard v1.129 CollectOXOGmetrics. Small fragment size sample is highlighted in red, all other samples are non-degraded or HapMap.

Comparison of DNA repair methods of FFPE

Four FFPE samples of varying age and quality, plus 1 HapMap sample underwent DNA repair (NEB USER enzyme, NEB PreCR repair mix, NEBNext FFPE DNA Repair mix and no repair) prior to library prep. 50ng of DNA was used as input into DNA Repair. Samples were then sheared and libraries generated using the Kapa Hyper Prep library kit according to the manufacturer's protocol. Figure 3 plots the Percent SNV in SNP138 vs count of novel SNVs showing the difference between samples undergoing repair vs no repair (black circle/squares).

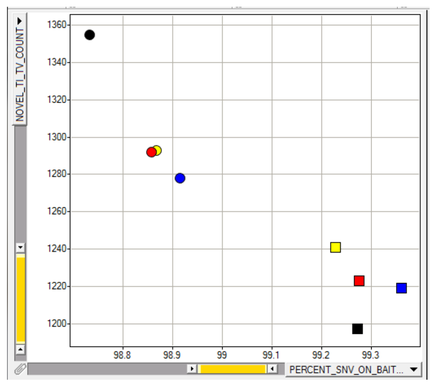

Figure 3. Percent SNV in SNP138 and counts of novel SNV metrics were generated using CIDRSeqSuite v6.0. Two out of the 5 samples are shown (FFPE-2 – circles; HapMap Control – squares. Black = no repair, Red = NEBNExt, Blue = PreCR, Yellow = USER)

Figure 4 shows higher oxidative error rates in the unrepaired samples compared to repaired samples. Overall NEB PreCR shows the lowest error rate across samples tested.

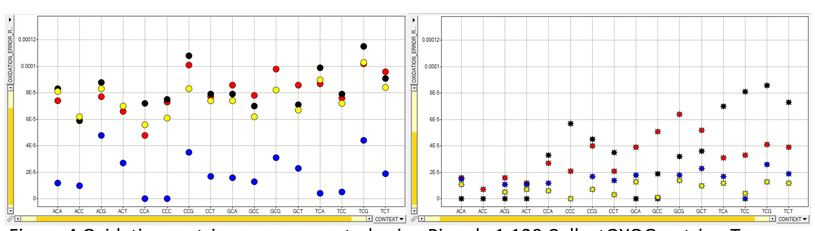

Figure 4 Oxidation metrics were generated using Picard v1.129 CollectOXOGmetrics. Two out of the 5 samples are shown (FFPE-2 left; HapMap Control right; Black = no repair, Red = NEBNExt, Blue = PreCR, Yellow = USER).

Capture/Sequencing/Data Analysis – Methods

All samples were hybridized according to the Agilent SureSelect™ XT protocol using either the Agilent Human All Exon V4, V5 or V5+UTR capture. Modifications to the hyb protocol included exchanging the universal blocker for base/base blocking to accommodate the indexed adapters used during ligation. Libraries were then clustered and sequenced on the Illumina® HiSeq™ 2500 platform using on-board clustering and Rapid Run SBS chemistry, 2x100. Sequencing data for the Low Input experiments was downsampled (Picard; DownsampleSam) to the experiment with the lowest raw data yield and the intersection of the BED files between the capture products was used to ensure equal comparison between methods. Subsequent data analysis was performed using CIDRSeqSuite v6.0 (http://www.cidr.jhmi.edu/next_gen_seq_serv/sequencing_qc_pipelines_workflow.pdf). Modifications to the pipeline for the degraded sample included trimming raw FASTQ files prior to alignment using CutAdapt (http://code.google.com/p/cutadapt/) to remove the default Illumina adapter sequences.

References:

1-Kapa Hyper Prep Kit Illumina platforms Technical Data Sheet – KR0961-v1.14; 2-Swift Accel-NGS 2S DNA Library Kit for the Illumina platform instruction manual – cat. no. DL-ILM2S-12/48; ver. 04291444; 3-SureSelectXT Target Enrichment System for Illumina Paired-End Sequencing Library; SureSelectXT Target Enrichment for Illumina Multiplexed Sequencing Protocol v1.1.1, January 2011 p/n G7530-90000; 4-Fisher *et al*; A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. *Genome Biology* 2011, 12:R1; 5-Swift Accel-NGS 1S DNA Library Kit for the Illumina platform Instruction manual – cat. no. DL-ILM1S-12/48; ver. 12011444

FFPE Pilot - Quantitation and Quality Assessment of FFPE samples

20 FFPE samples of varying age and quality were processed according to the Low Input/FFPE workflow depicted in Figure 5. Sample concentrations based on picogreen were provided by the investigator. To ensure accurate quantitation all samples were amplified by qPCR using the Kapa Human Genomic DNA Quantification and QC kit according to the manufacturer's protocol. Figure 6a compares the concentrations of each sample generated by nanodrop, picogreen or Kapa qPCR. A 50ng input was calculated using the Kapa qPCR values. Figure 6b plots the observed total DNA for each sample after shearing based on BioAnalyzer.

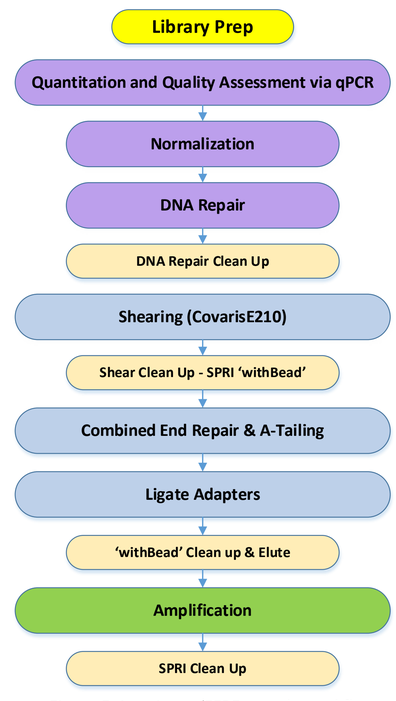

Figure 5. Low Input/FFPE sample workflow

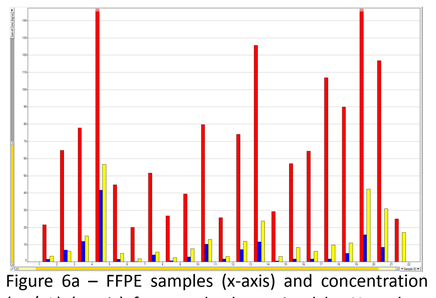

Figure 6a – FFPE samples (x-axis) and concentration (ng/uL) (y-axis) for sample determined by Nanodrop (red), Picogreen (blue) and Kapa qPCR (yellow).

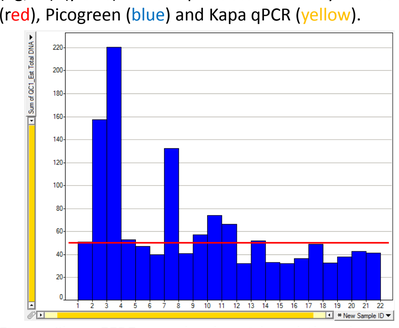

Figure 6b – FFPE samples (x-axis) and the observed total DNA input (y-axis) after DNA repair and Covaris shearing determined by the Agilent BioAnalyzer. Red line indicates expected DNA input at 50ng.

FFPE Pilot – Sequencing Data

Table 3 compares QC metrics generated during library prep, capture and sequencing. Sample metrics are compiled according to age of sample. Trends in percent duplication, mean insert size and yield at PCR follow age of sample. Data quality (TiTv, Percent SNV OnBait SNP138 & Count SNV OnTarget) across samples are consistent.

Table 3. FFPE Pilot – Library Prep, Capture and Sequencing Metrics

Experiment	FFPE Samples - ≥20years (n=8)	FFPE Samples - ≤10 years (n=12)	HapMap Control (n=1)
Year of Sample	1995-1996	2002-2005	2015
Library and Capture Metrics			
Input DNA [ng]	50	50	50
Yield at PCR Amp [ng]	**183.6**	**416.2**	**801.2**
Capture	v5	v5	v5
% Selection	83.78	84.29	84.29
Sequencing Data Quality			
Concordance	na	na	99.8476
Sensitivity (Het)	na	na	99.2203
TiTv (ALL_Ratio)	3.0281	3.0452	3.0305
Percent SNV OnBait SNP138	98.66	98.45	98.91
Count SNV On Target	25070	26024	25474
Sequencing Data Output			
Raw GB	9.3792	8.5336	8.2456
Mean Target Coverage	74	79	92
% On Target @ 10x	**96.73**	**97.24**	**98.21**
% Zero Bases	0.65	0.62	0.69
Mean Insert Size	**173**	**213**	**223**
% Dups	**36.27**	**21.55**	**7.42**

FFPE Pilot – Q-ratio, PCR yield and sequencing yield

The Kapa Quantification and QC kit provides a q-ratio that can be used to estimate the quality of DNA samples. We compared the q-ratio to the yield at PCR and sequencing depth to determine appropriate cut-offs that would flag lower quality samples which may require additional sequencing or DNA input (Figures 7a,b & c).

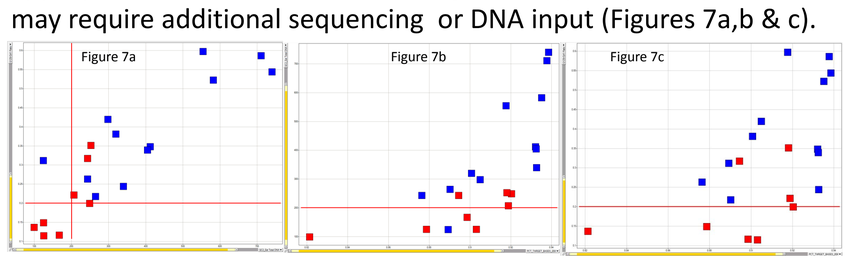

Figure 7a,b & c. Fig 7a plots the PCR yield prior to hybridization (x-axis) vs the Q-ratio (y-axis). Q-ratio correlates with PCR yield. Samples ≥20 yrs (red squares) tend to be of lower quality and have lower yields compared to samples ≤10 yrs (blue squares). Fig 7b & c plots % of targeted bases @20x vs PCR yield or Q-ratio, respectively. Q-ratio is helpful to flag samples (red lines) which may be of lower quality and require additional sequencing.

Discussion/Conclusions

- Which library prep method to use may depend on a number of variables (cost differences, streamlined reagents, benefits from specific chemistry). With any platform it is important to validate overall performance and quality within a laboratory.
- qPCR prior to library prep improves quantitation accuracy and allows for the assessment of sample quality as an indicator of downstream sequencing requirements.
- Repair of DNA reduces the rate of age related artifacts and improves data quality. Lower quality correlates with age.
- Small fragment sized DNA can be used to generate libraries for NGS. Lower quality samples will impact data quality. Additional methods for repairing small fragmented DNA need to be explored.