Preliminary study for whole genome sequencing from dried blood spots using high throughput linked-read sequencing compared to short read sequencing

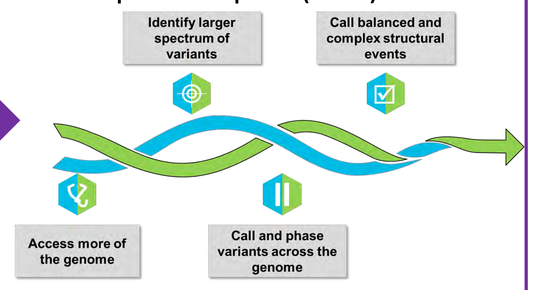
Zeqiang Ma[1], Gerard Irzyk[1], C. Alexander Valencia[1], Sarah Garcia, Edward Szekeres Jr[1], Zdenek Markovic[1], Yang Wang[1], Alice Tanner[1], Christin Collins[1], Madhuri Hegde[1,3]
[1]PerkinElmer Genomics, [2]10x Genomics, [3]Emory University

ABSTRACT

Whole genome sequencing (WGS) is becoming increasingly attractive as an alternative to standard clinical testing approaches, due to its uniform coverage, ability to detect all types of variants, and decreasing cost. Standard sequencing methodologies involve the use of short reads, however, there are disadvantages to this approach particularly for the detection of structural variants, phasing and variants in difficult regions of the genome (homologous areas). DNA extraction from dried blood spot samples (DBS) is challenging because they only yield several hundred nanograms of DNA. One approach to overcoming these limitations is to utilize high throughput Linked-Read whole genome sequencing (lrWGS) by 10X Genomics. lrWGS utilizes haplotype-level dilution of high molecular weight DNA molecules, typically of >10 Kb length, into >1 million barcoded partitions to create a novel data type referred to as 'Linked-Reads'. These Linked-Reads enable high-resolution genome analysis with minimal DNA input (~1 ng), which is an advantage when working with DBS. We applied this technology to the analysis of genomic DNA obtained from 3 DBS samples and one Coriell control sample, NA12878. Extraction of the DBS was conducted using the Chemagic 360 platform and yielded DNA compatible with the 10X Genomics technology, allowing for genome-wide phasing analysis. Samples were prepared using 10X Chromium technology and compared to samples prepared with the Illumina PCR-free library kit (standard). All samples were sequenced on the NovaSeq 6000 using 2 X 150 paired-end reads. The average coverages obtained were as follows (Standard:Chromium): DBS1 39X: 46X, DBS2 40X:50X, DBS3 31X:32X and NA12878 31X:29X. Sequencing data generated from the Illumina PCR-free libraries was analyzed using the Edico Genome's DRAGEN Platform and sequencing data generated from Chromium library samples were analyzed using the latest versions of the 10X Genomics Long Ranger pipeline and Loupe visualization tools. Libraries prepared with Chromium technology yielded a greater number of SNPs compared to the standard library method. Specifically, the SNP pairwise comparisons of Standard:Chromium were DBS1 3,848,077: 3,893,116 (Δ of 45,039 SNPs), DBS2 3,867,808: 3,927,551 (Δ of 59,328 SNPs), DBS3 3,842,844: 3,871,172 (Δ of 28,328 SNPs) and NA12878 3,838,443: 3,856,038 (Δ of 17,595 SNPs). In addition, the Chromium technology and Long Ranger analysis identified short deletions (range 50 bp – 30 Kb) in all samples analyzed: 3,619 (DBS1), 3,779 (DBS2), 4,302 (DBS3) and 4,028 (NA12878). Similarly, Chromium technology and the Long Ranger pipeline detected structural variants (>30 Kb) in all samples analyzed: 16 (DBS1), 14 (DBS2), 20 (DBS3) and 19 (NA12878). In summary, we demonstrate that complex, high-quality Linked-Read sequencing can be generated from dried blood spots. Although comparable SNP calling was observed in all samples, the Linked-Reads sequencing provided a noticeable increase of SNPs over the standard library results. In addition, samples prepared by Chromium technology and Long Ranger analysis provided information on deletions, structural variations, overall greater coverage of the genome and trinucleotide repeat detection in the heterozygous/homozygous normal range, with phasing. Continued improvement on the algorithm may be applied to larger trinucleotide repeats. These results indicate the compatibility of 10X Chromium technology with DNA extracted from DBS for the generation of more comprehensive high quality genomic sequence data sets.

INTRODUCTION

- **Challenge 1**: NGS using short reads has disadvantages:
  - Detection of structural variants
  - Unable to do phasing
  - Detection of variants in difficult regions of the genome (homologous areas)
- **Challenge 2:** DNA extraction from dried blood spot samples (DBS) is low (nanograms of DNA)
- **Approach:** To utilize high throughput Linked-Read whole genome sequencing (lrWGS) by 10X Genomics

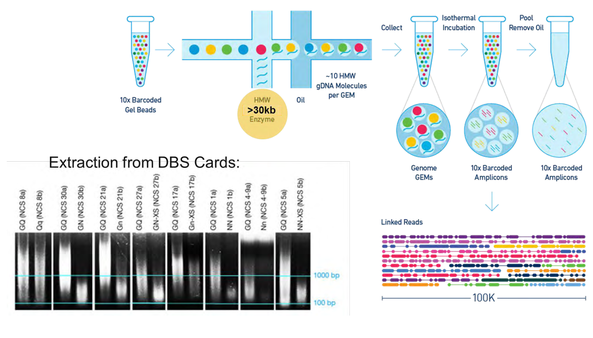


METHODS

PerkinElmer / 10x Genomics™ DBS Workflow

Sample to Answer

Length Requirements for Linked-Reads vs. Blood Spot Length



- Showing the labeling of "HMW" to contrast against DBS extractions

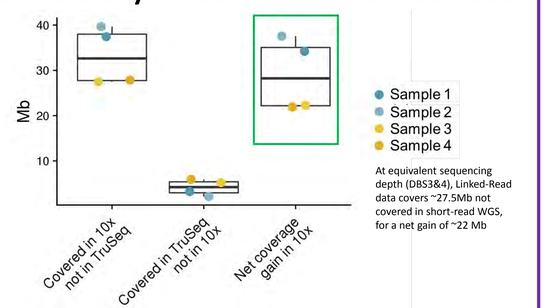
RESULTS

Dried Blood Spot Sequencing Metrics

Sample	Assay	Mean Depth (duplicates removed)	Median Depth (duplicates removed)
DBS1 (male)	Standard (2ug input)	39x	37x
	Chromium (1.25ng input)	46x	36x
DBS2 (male)	Standard (2ug input)	40x	38x
	Chromium (1.25ng input)	50x	42x
DBS3 (female)	Standard (2ug input)	31x *	30x
	Chromium (1.25ng input)	32x	27x
DBS4 (female)	Standard (2ug input)	31x *	30x
	Chromium (1.25ng input)	29x	23x

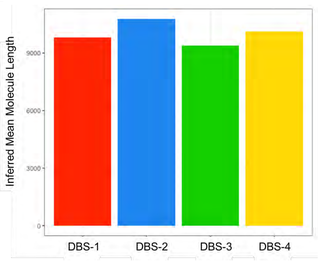
All sequenced on NovaSeq, run through 10x Pipeline
*Downsampled from total available sequence for matched depth comparisons

Access More of your Genome with Linked-Reads

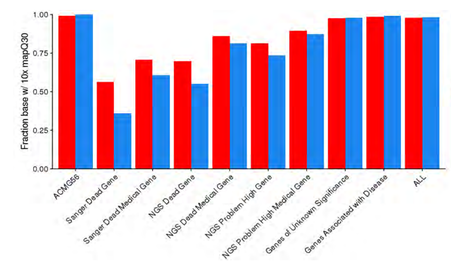


At equivalent sequencing depth (DBS3&4), Linked-Read data covers ~27.5Mb not covered in short-read WGS, for a net gain of ~22 Mb

Mean Length from Dried Blood Spot Extractions: 10kb



Linked-Read Data Improves Per Gene Coverage



Fraction of bases in various gene sets with >10x MAPQ >= 30 coverage

Variant Calling Performance

SVs calculated using barcode-dependent SV calling algorithms

Sample	Assay	Length	Mean Depth (dedup ed)	# SNPs	#SNPs gained by Linked-Reads	# short deletions (50bp-30kb)	# structural variants (>30kb)
DBS1 (male)	Standard	n/a	39x	3,848,077	n/a	n/a	n/a
	Chromium	9.8kb	46x	3,893,116	45,039	3,619	16
DBS2 (male)	Standard	n/a	40x	3,867,808	n/a	n/a	n/a
	Chromium	10.7kb	50x	3,927,551	59,743	3,779	14
DBS3 (female)	Standard	n/a	31x *	3,842,844	n/a	n/a	n/a
	Chromium	9.4kb	32x	3,871,172	28,328	4,302	20
DBS4 (female)	Standard	n/a	31x *	3,838,443	n/a	n/a	n/a
	Chromium	10.1kb	29x	3,856,038	17,595	4,028	19

*Downsampled to ~30x to match mean depth of Chromium runs

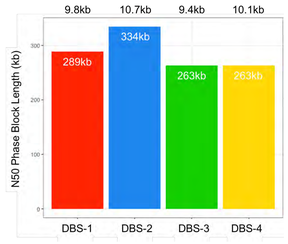
Triplet Repeat Disorders- Calling with HipSTR

	DBS1_LR	DBS1_SR	DBS2_LR	DBS2_SR	DBS3_LR	DBS3_SR	DBS4_LR	DBS4_SR
AR	✓	✓	✓	✓		✓	✓	✓
ATXN1	✓	✓		✓	✓	✓	✓	✓
ATXN2	✓	✖	✓	✓	✓	✓	✓	✓
ATXN3	✓	✓	✓	✓	✓	✓	✓	✓
ATXN7	✓	✓	✓	✓	✖	✓	✓	✓
CACNA1A	✓	✓	✓	✓		✓	✓	✓
DMPK	✓	✓	✓	✓		✓	✓	✓
ATN1	✓	✓	✓	✓	✓	✓	✓	✓
HTT	✓	✓	✓	✓	✓	✓	✓	✓
PABPN1	✓	✓	✓	✓	✓	✓	✖	✓
FXP2B	✓	✓	✓	✓	✓	✓	✓	✓

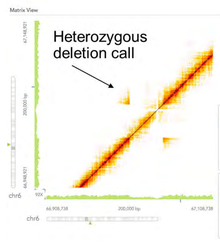
Of the 26 HipSTR-called sites with genotypes in both LR and SR, 4 are discrepant

✓: Called by HipSTR
43/44 sites in short read (SR) data
26/44 sites in Linked-Read (LR) data
✓: Called by manual review of phased alignments
An additional 11 LR sites could be confidently called (algorithmic improvements needed)
✖: Single haplotype detected

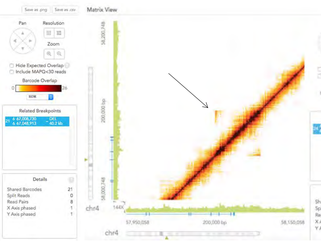
Chromium Genome Phasing Performance



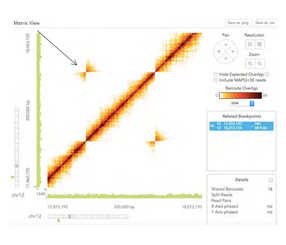
Heterozygous Deletion



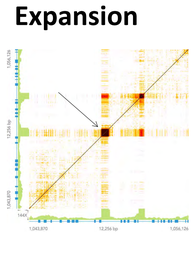
Duplication



Balanced Inversion



ABCA7 VNTR Expansion



DISCUSSION/CONCLUSIONS

- Complex, high quality Linked-Read sequencing can be generated from dried blood spots
  - Phase blocks 100s of kbs long
- Comparison to standard short read
  - Comparable SNP calling between Linked-Reads and Short Reads
  - Linked Reads provide advantages in:
    - Coverage over a larger portion of the genome
    - Access to repetitive regions
    - Structural variant calling
    - Resolution of normal length triplet repeat alleles
- Extraction optimization
- Sequencing of additional samples
- Optimization of software to take full advantage of Linked-Read data type