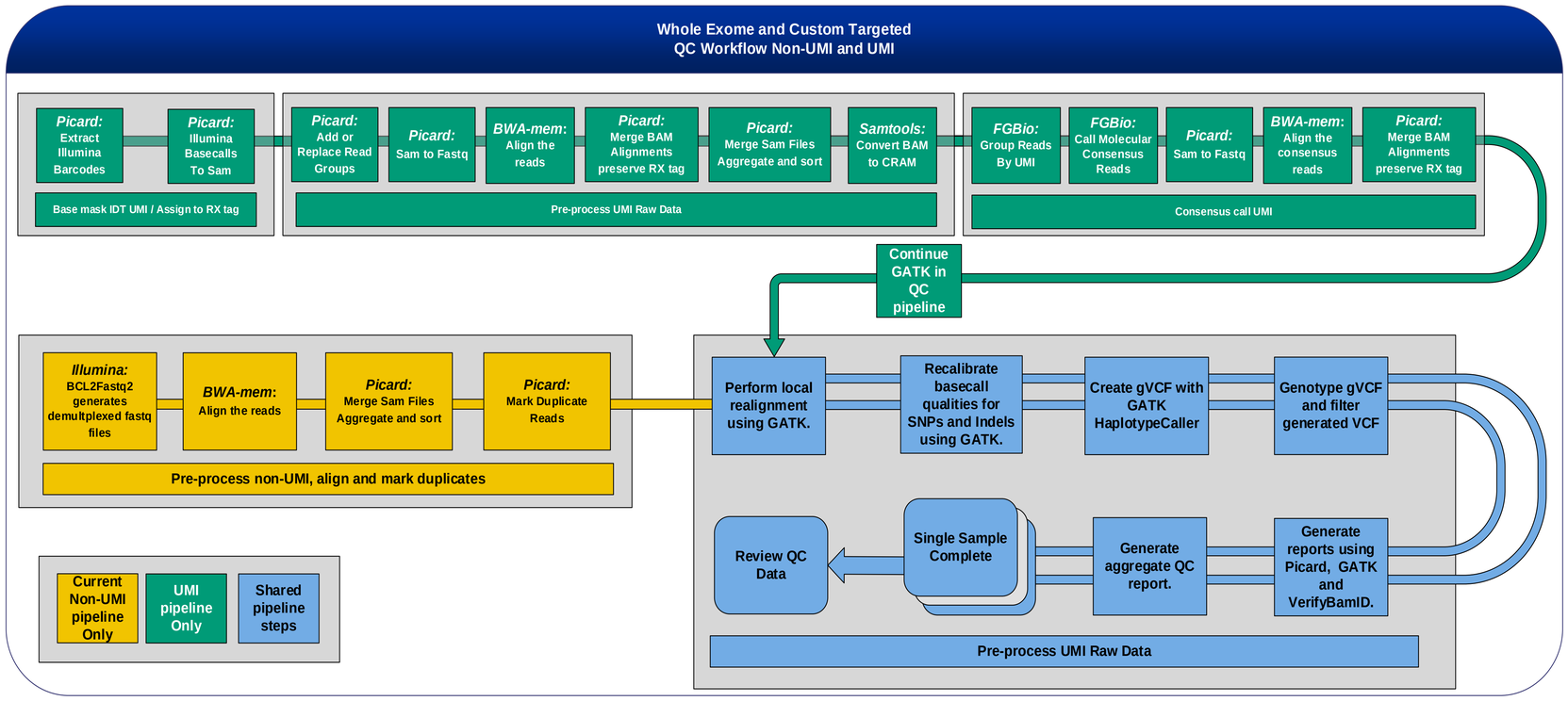
Evaluation of UMI tagged whole exome sequencing data comparing conventional best practice to consensus calling methods

B. Craig[1], B. Marosy[1], A. Robinson[1], M. Mawhinney[1], K. Doheny[1]

[1] Department of Genetic Medicine/Center for Inherited Disease Research (CIDR), Johns Hopkins Genomics (JHG), Johns Hopkins University School of Medicine, Baltimore, MD, USA

Introduction

CIDR is continually seeking new informatics and laboratory methods in order to improve its workflow, lower costs and generate high quality data. Unique molecular identifiers (UMIs) are short random sequences incorporated into libraries prior to amplification aimed at decreasing variant calls produced by PCR errors. Here we describe development of analysis pipelines to incorporate and consensus call UMI aware reads to evaluate effect on sample and variant quality metrics. The current single sample QC non UMI aware analysis pipeline (CIDR), uses BCL2Fastq2 to de-multiplex samples, and runs GATK's BQSR and HaplotypeCaller with filters tuned for depth, strand bias, and mapping quality. The development UMI pipeline (UMI) uses Picard to de-multiplex samples and incorporate the UMI tag, group, consensus call and filter the UMI families, and runs GATK's BQSR and HaplotypeCaller with and without CIDR hard filters. Target enriched libraries (2.9Mb) were constructed from tumor and normal sample types with unique dual-indexed UMIs, sequenced on NovaSeq and analyzed with CIDR and UMI pipelines.



Design

Sample type and quality selection
- Tumor: *FFPE tissue and Fresh Frozen*
- Normal: *Blood spot, Buccal, Buffy Coat, Cell Line, Fresh Frozen, Lymphocytes, Saliva and Whole Blood*
- Preliminary analysis of 1342 pairs

Library preparation
- Kapa Hyper Prep reagents
- IDT Dual index UMI adapters

Target enrichment
- IDT xGen target enrichment panel
- Custom designed content 2.9 Mb

Cluster and Sequencing
- Illumina NovaSeq 6000
- 46bp paired end template reads
- 8bp sample barcode + 9bp UMI I7 read
- 8bp sample barcode I5 index read

Analysis
- Single sample QC analyses scripted for local high performance computing (HPC) cluster based on IDT and GATK best practices (QC workflow diagram)

QC Pipeline Comparison results

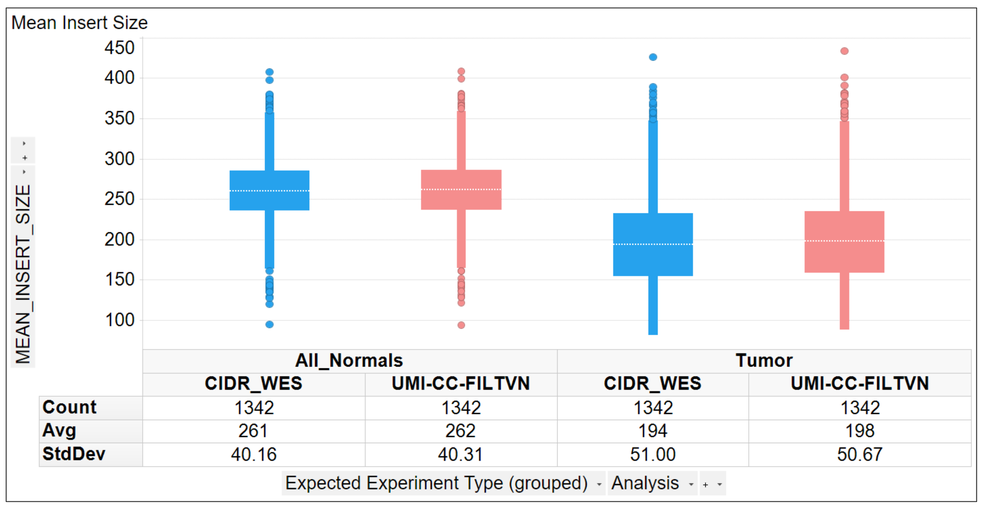


Figure 1: Box plot Average of Mean Insert Sizes grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

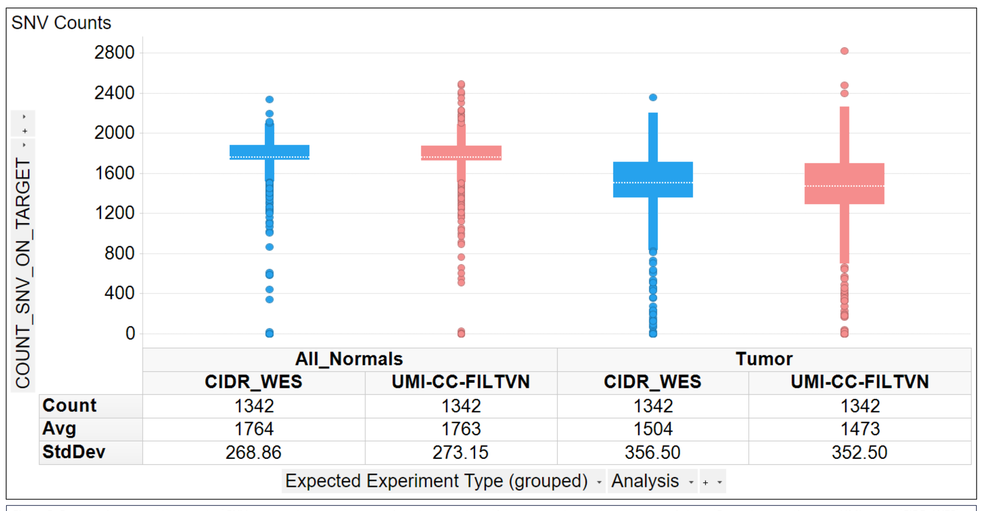


Figure 2: Box plot Average of the count of SNVs on target grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

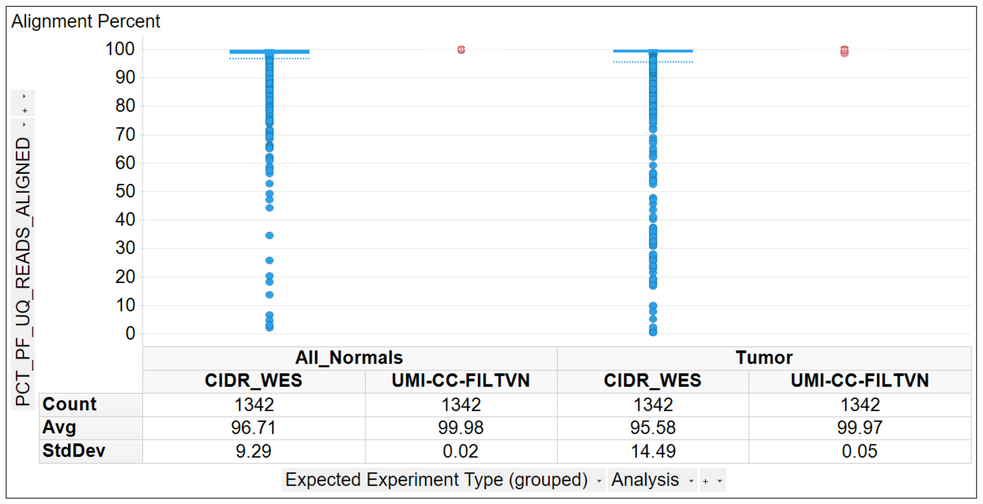


Figure 3: Box plot Average of the percent of reads aligned grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

Figure 4: Box plot Average of the percent of C to T variant calls grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

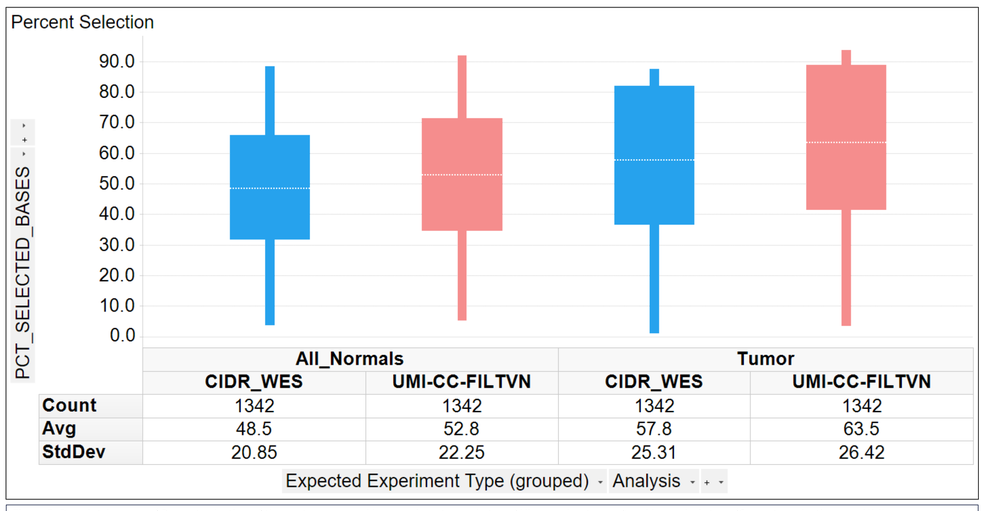


Figure 5: Box plot Average of Percent selection of the capture grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

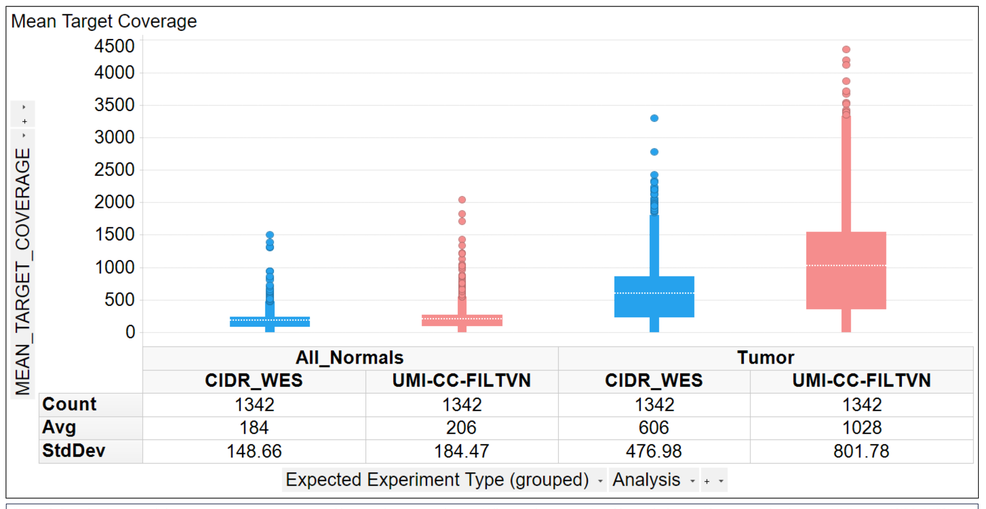


Figure 6: Box plot Average of the mean target coverage grouped by experiment sample type (All normal or tumors) and by analysis type (Current production non-UMI aware (CIDR_WES) and development UMI with production hard filters(UMI-CC-FILTVN) and colored by analysis type)

Discussion

- **Incorporation and consensus calling UMIs**
  - Does not negatively affect the average insert size estimation (Figure 1)
  - Marginally decreases the average count of SNVs on target and percent C to T variant calls for tumor sample types (Figures 2 and 4)
  - Shows ~4% increase in the average alignment quality for both sample types (Figure 3)
  - Slightly increases the average percent selection of the capture (~5% increase for both tumor and normal) (Figure 5)
  - Increases the mean target coverage by considering more unique reads vs traditional start and stop duplicate marking techniques (Figure 6)
  - Estimates lower duplicate percentages using Picard's experimental UmiAwareMarkDuplicatesWithMateCigar compared with the current start / stop MarkDuplicates algorithm
    - Normals: 21% vs 18%
    - Tumors: 46% vs 27%
    - non-UMI and UMI aware respectively
- Future directions
  - Run the consensus called read data with a somatic caller (e.g., VarDict or Mutect 2) to determine if UMI use increases the sensitivity of the variant calls.

References

UMI Demultiplex and Consensus Analysis Resources
- https://www.idtdna.com/pages/products/next-generation-sequencing/adapters/xgen-dual-index-umi-adapters-tech-access
- http://fulcrumgenomics.github.io/fgbio/

CIDR Non-UMI aware analysis resources
- https://github.com/Kurt-Hetrick/CIDR_WES

Shared pipeline resources
- https://gatk.broadinstitute.org/hc/en-us
- https://broadinstitute.github.io/picard/

Presentation Information

Advances in Genome Biology and Technology (AGBT) General Meeting 2020
Poster # 910