Understanding low allele variant detection in heterogeneous samples, required read coverage and the utility of unique molecular indices (UMIs)

BRENDAN BLUMENSTIEL[1], MARK FLEHARTY[1], MATTHEW DEFELICE[1], LISA GREEN[1], JONNA GRIMSBY[1], YOSSI FARJOUN[1], NIALL LENNON[1] & STACEY GABRIEL[1]

[1]Broad Institute of MIT and Harvard, 320 Charles St. Cambridge, MA 02142

Abstract

As interest in tumor genomic profiling and circulating cell-free DNA sequencing continues to evolve, the ability to detect low allele frequency and rare genomic alterations in heterogeneous samples is becoming increasingly important. With clinically relevant variants often presenting with allele frequencies lower than 1%, deep sequencing coverage and highly sensitive and specific methods are needed for efficient screening and discovery of actionable alterations. Here we describe methods used for detecting low allele fraction events across a range of variant frequencies from 0.5-5%. This is accomplished by creating a serial spike-in curve using two ethnically different germline DNAs and observing minor allele frequencies for known variant points. In addition to understanding raw sequencing depth requirements, we test the utility of incorporating unique molecular indices (UMIs) into the sequencing read structure to enable a second level of marking duplicate reads and calculating true unique molecular library complexity and coverage. Performance with cfDNA was also observed

Design

Genomic DNA from a single Yoruban sample (NA19238) was spiked into a single CEPH sample (NA12144) at 5%, 1% and .2% (NEB 0.1%)

Targeted by two different enrichment methods

- Illumina Rapid Capture of Kapa Hyper UMI libraries. Panel size = ~3MB target territory
- New England BioLabs NEBNext Direct. Panel size = ~40KB target territory

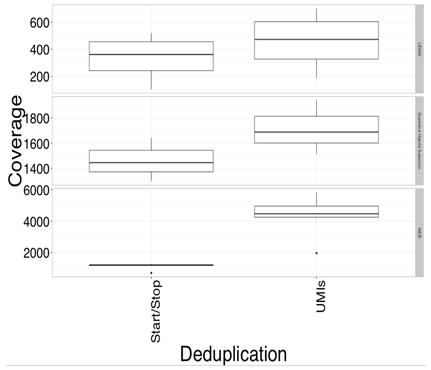
Sequenced to deep coverage 5,000-30,000X raw target coverage on Illumina 2500

Unique Molecular Indices increase coverage

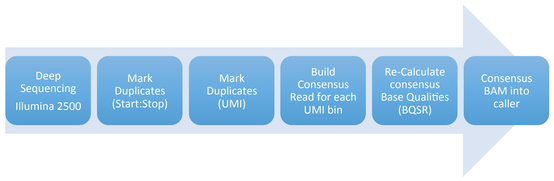
- Reads are binned and de-duplicated by 12-bp UMI code
- Modest increase in coverage observed for germline Hyb Selection due to broad fragment size distribution
- Moderate increase observed for cfDNA due to tighter fragment size distribution around 160bp
- Significant increase in coverage observed for NEBDirect Hyb Selection due to fixed fragment start position and random sheared end

	Raw	start/stop	UMI	% inrease w/ UMI
CfDNA Hyb Sel	11928	336	458	**36%**
Germline Hyb Sel	29190	1462	1714	**17%**
NEB	5546	1131	4394	**289%**

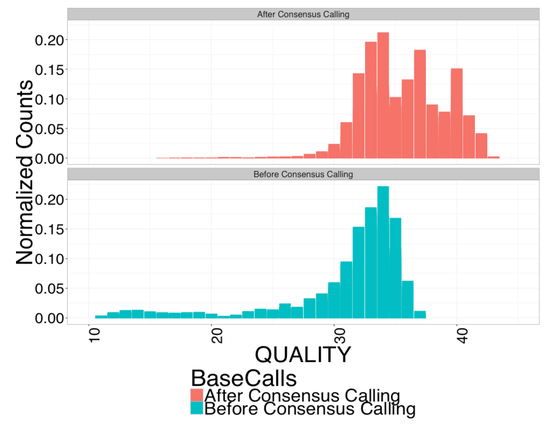
Sequencing and de-duplicated coverage (Mean Target Coverage)



Consensus Synthetic Reads increase base quality scores

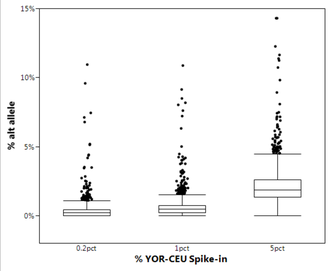


- Post marking duplicates by UMI code, multiple reads per UMI bin are compiled to generate a consensus read
- Consensus reads generated using the CallMolecularConsensusReads tool provided by fgBio
- Quality scores estimated empirically using GATK BQSR both before and after consensus reads were called
- Consensus reads improve Q Scores by ~5 Phred scales



Hybrid Selection

Observed minor allele frequency Hybrid Selection



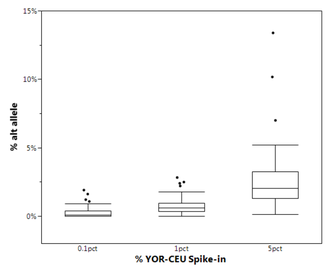
UMI Library Construction Workflow

- Random sheared library
- End Repair and A-tailing (Kapa Hyper)
- Ligate 12bp UMI-containing adapter
- Illumina Rapid Capture Broad Custom Panel (3Mb)

Final Hyb Sel UMI dual indexed sequencing construct

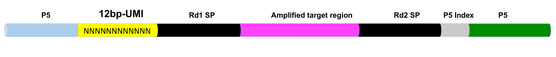
NEBNext Direct

Observed minor allele frequency NEBNext Direct



UMI Library Construction Workflow

- Random sheared library
- Direct capture with biotin probe
- 3' blunting and A-tailing
- Hairpin adapter ligation
- 5' probe extension and blunting
- Ligation of 12bp UMI containing adapter
- Adapter cleavage and PCR amplification

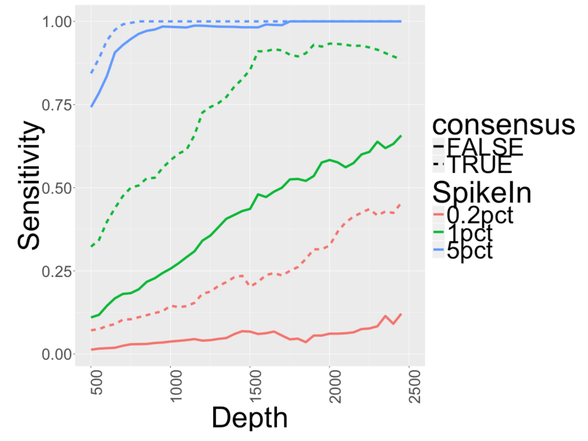


Final NEB UMI single indexed sequencing construct

Consensus reads increase sensitivity

Consensus reads increase call sensitivity using Mutect II

- Variant detection was performed using Mutect II on consensus and non-consensus data sets
- Analysis performed for germline hybrid selection spike-in libraries
- Increase in sensitivity observed across full range of allele frequencies (0.2-5% slike-in) when using consensus
- No filters or modifications applied for Mutect calling



Conclusions

- Incorporation of Unique Molecular Indices (UMIs) increases usable coverage for germline Hyb Selection, cfDNA Hyb Selection and NEBNext Direct targeting methods
- Building consensus reads from UMI de-duplicated data increases base quality scores
- Increased base quality scores increase low allele fraction variant calling sensitivity using Mutect II

Acknowledgments

Data used in this poster was generated at the Broad Institute, for more information please visit: http://genomics.broadinstitute.org/