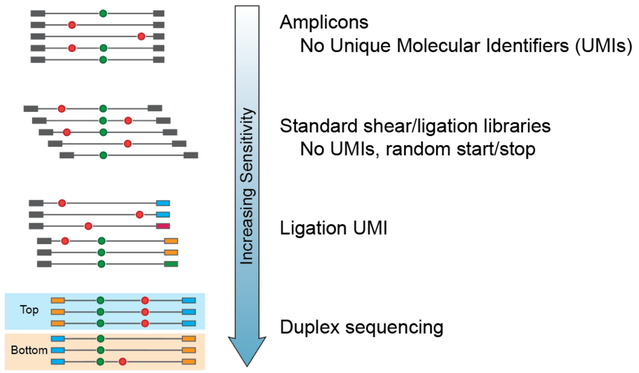
Subclonal mutation profiling – characterization of targeted deep sequencing with molecular tagging

Kristina Giorda[1], Joshua Xu[2], Guangchun Chen[3], Madelyn Light[1], Jiashi Wang[1], Kevin Lai[1], Binsheng Gong[2], Meeiyueh Liu[4], Isaac Meek[5], Natalia Novoradovskaya[6], Don Johann[4], Quan-Zhen Li[3], Leming Shi[7], Weida Tong[2], and Mirna Jarosz[1]

[1] Integrated DNA Technologies, Redwood City, CA; [2] National Center for Toxicological Research, Food and Drug Administration, Jefferson, AR; [3] University of Texas Southwestern Medical Center, Dallas, TX; [4] University of Arkansas for Medical Sciences, Little Rock, AR; [5] New England BioLabs Inc., Ipswich, MA; [6] Agilent Technologies, La Jolla, CA; [7] Fudan University, Shanghai, China

* Corresponding author: kgiorda@idtdna.com

Introduction



- Incorporating molecular barcodes during library preparation enables error correction and greater sensitivity for low frequency variants, which is critical for precision medicine in oncology.
- Molecularly barcoded ligation libraries were used in this pilot study to support the US FDA-led Sequencing Quality Control Phase 2 (SEQC2) consortium project.

Consensus analysis

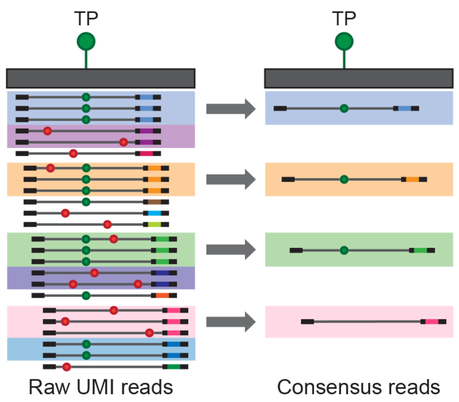


Figure 1. Consensus analysis. Reads that map to the same region of the genome and share the same UMI are used to build consensus reads. In this schematic, a minimum of 3 reads is used to build consensus reads, which removes false positives (red) but retains true positives (TP, green).

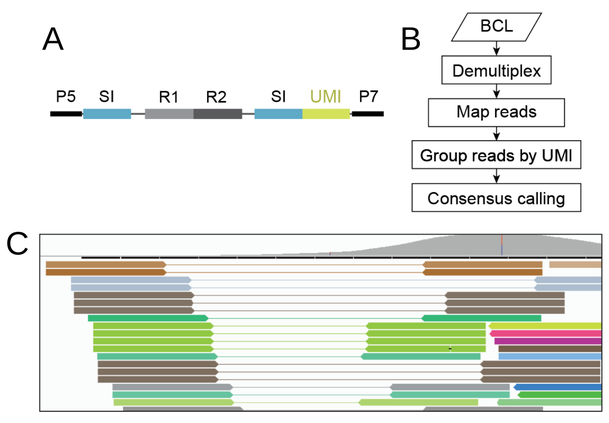


Figure 2. Library design and analysis workflow. The **(A)** structure of the ligation UMI library and **(B)** schematic of the bioinformatics pipeline used for consensus calling are shown. **(C)** An IGV screenshot shows reads color-coded by the UMI tag.

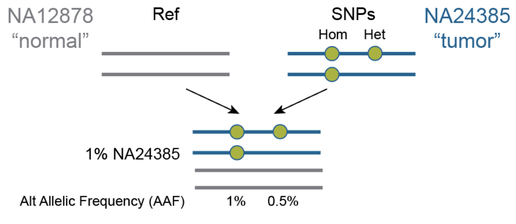
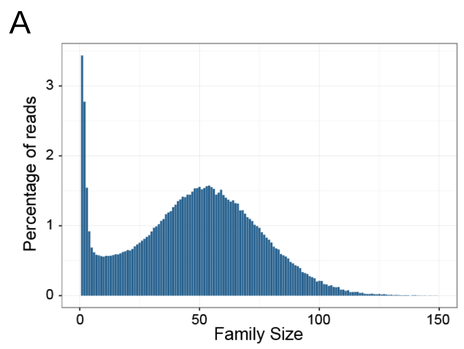
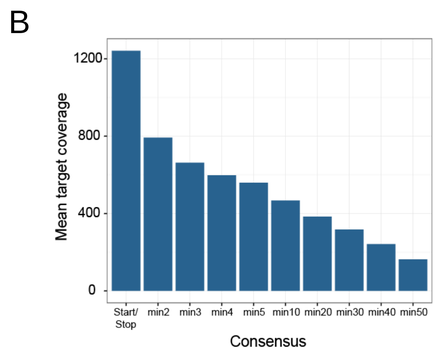
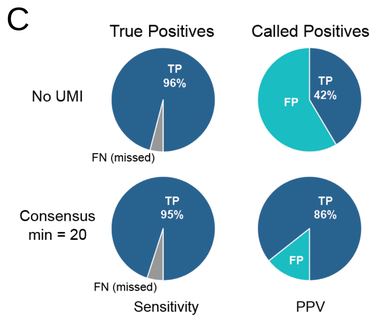


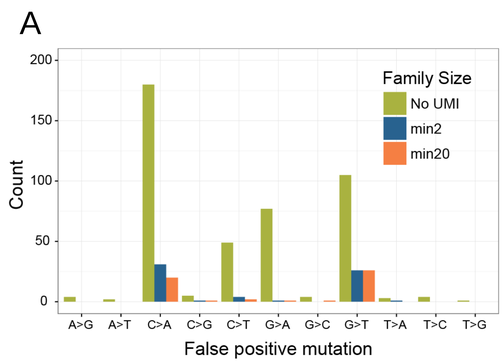
Figure 3. Rare variant detection model system. The Genome in a Bottle samples NA12878 and NA24385 were blended together to create a mock somatic sample.

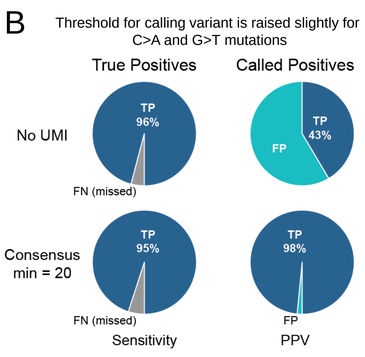
Consensus analysis enables variant detection at 0.5% allele frequency











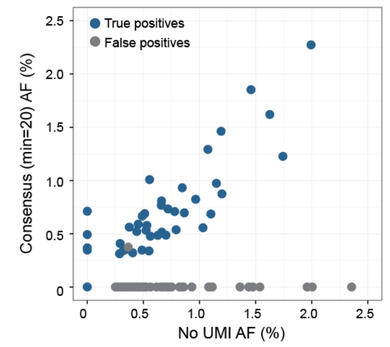
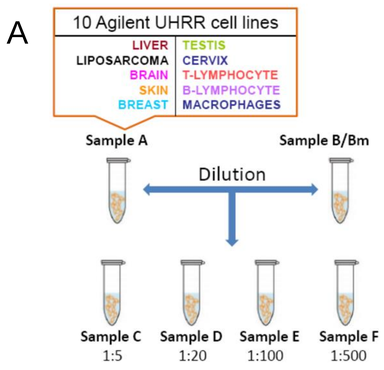
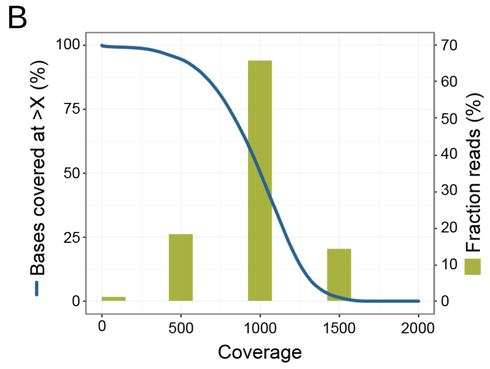


Figure 4. Consensus reads increase accuracy. Libraries were made with 1% NA24385 using 10 ng input and captured with a set of custom xGen® Lockdown® Probes covering 288 common SNPs over a target area of ~35 kb. Variant calling was done with VarDict using a 0.25% threshold. The **(A)** family size distribution and **(B)** mean target coverage are shown. **(C)** Using a minimum of 20 reads to build consensus molecules increased the positive predictive value (PPV) for variants down to 0.5% allelic frequency (AF).

Figure 5. oxoG errors are corrected with consensus reads. (A) The reference imbalance of C verses G is removed with consensus reads. **(B)** Increasing the threshold for C>A and G>T mutations improves the PPV for consensus reads. **(C)** Consensus reads reduce false positives for low frequency variants.

SEQC2 pilot study: low-frequency variant detection





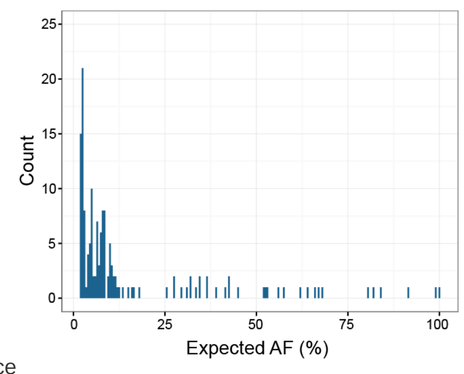
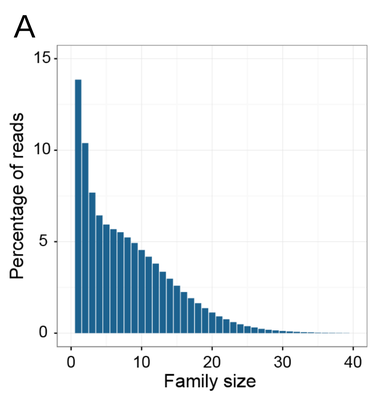
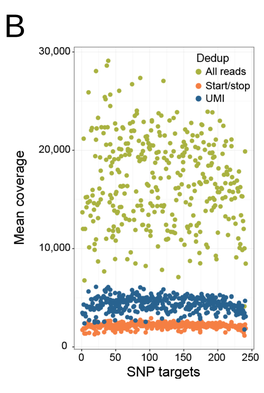


Figure 6. SEQC2 pilot study. (A) gDNA control samples, made from Universal Human Reference RNA (UHRR) cell lines and the Agilent male control (provided by Agilent Technologies), were tested individually and as mixtures. Libraries were constructed using the NEBNext® Ultra™ II DNA Library Prep Kit (New England BioLabs) and captured with a custom NSCLC-focused xGen® Lockdown® panel designed for targeted deep sequencing. **(B)** The by-base coverage of the 111 kb target region was highly uniform (subsampled to 1M reads).

Figure 7. Distribution of expected variants. As a first pass, libraries constructed from the individual cell lines (Fig. 6A) were captured with the xGen® Pan–Cancer Panel, and a 20% threshold was used for variant calling. The samples were mixed *in silico* using tools from Fulcrum Genomics to generate our ground truth for Sample A.





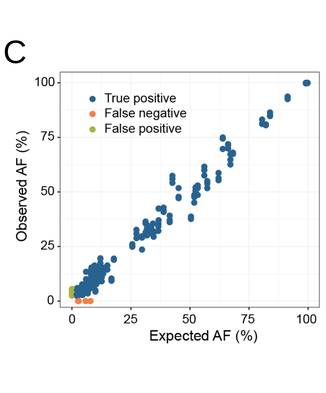


Figure 8. UHRR cell line mixture analysis. Libraries were constructed with 25 ng of DNA from a UHRR cancer cell line mixture (Sample A) and captured with the NSCLC panel. The **(A)** family size distribution and **(B)** mean target coverage are shown using all reads (blue), reads de-duplicated by start and stop (red), or de-duplicated by UMI (green). **(C)** Using a minimum of 3 reads to build consensus reads had high sensitivity of 98% and PPV of 94%, and the expected and observed allelic frequencies were well correlated.

Conclusions

- Without molecular barcoding, it is difficult to distinguish true and false positives at frequencies below ~1%
- Using UMIs to build consensus reads dramatically increases variant calling accuracy
- Ligation UMIs enable error correction of mutation artifacts

References

- Cancer Genome Atlas Research Network. (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature, 511:543–550.
- Imielinski M, Berger AH, et al. (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell, 150:1107–1120.
- Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature, 489:519–525.
- Costello M, Pugh TJ, et al. (2013) Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. Nucleic Acids Res, 41:e67.