**Supplementary Material**

**Quantifying sequencing error and effective sequencing depth of liquid biopsy NGS with UMI error correction**

Malene Støchkel Frank1,2\*, Janina Fuß3,\*, Tim Alexander Steiert3, Greta Streleckiene4, Julie Gehl1,2, Michael Forster3

1Department of Clinical Oncology and Palliative Care, Zealand University Hospital, Denmark

2Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

3Institute of Clinical Molecular Biology, Christian-Albrechts-University Kiel, 24105 Kiel, Germany

4Institute for Digestive Research, Lithuanian University of Health Sciences, Kaunas, Lithuania

\*shared first authors

Correspondence: Michael Forster, m.forster@ikmb.uni-kiel.de

|  |  |  |
| --- | --- | --- |
| **Supplementary Table 1: Bioinformatic tools, versions, and settings that were used** | | |
| **Tool** | **Version** | **Settings used** |
| bwa | v0.7.12 | mem -M -R |
| samtools | v0.1.18 | uhSt |
| UMI Error Correction Local App | 1.0.0.1 | minSupportingReads 2 (default) 1 |
| fastp | v0.20.1 | -q 32 -u 0 |
| GenSearchNGS | v1.6.973 | Min frequency = 0, min coverage = 0, min allele occurence = 0, ignore X bases from read borders = 0, max homopolymer length = 10, min alignment quality (phred) = 20, min base quality (phred) = 20 |
| R-command: plot(density(list)) 2 | R version 3.5.3 | gaussian kernel (default) |
| R-command: boxplot(list) 2 | R version 3.5.3 | - |
| 1 This setting uses a minimum of 2 sequences for the UMI-based error correction by collapsing reads to an error-corrected consensus read.  2 list is the list of allele frequencies of the variants that were previously exported from GenSearchNGS | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Supplementary Table 2: Allelic frequency of PIK3CA p.E545K mutation for non-UMI libraries vs UMI libraries and bioinformatic non-filtering vs filtering, in wild-type standard HD776 (0.00%) and mutated standard HD779 (0.13%)** | | | | | |
| **Type** | **Library** | **Filtering1** | **Depth2** | **AF3** |
| non-UMI | TruSeq 776-1 | - | 107478 | 0.02% |
| dedup | 41127 | 0.02% |
| phred > 31 | 6274 | 0.00% |
| phred > 31, dedup | 679 | 0.00% |
| TruSeq 776-2 | - | 80521 | 0.01% |
| dedup | 25609 | 0.00% |
| phred > 31 | 3749 | 0.00% |
| phred > 31, dedup | 628 | 0.00% |
| TruSeq 776-Norm | - | 90163 | 0.01% |
| dedup | 34945 | 0.01% |
| TruSeq 779-1 | - | 76134 | 0.44% |
| dedup | 23300 | 0.10% |
| phred > 31 | 1990 | 0.00% 4 |
| phred > 31, dedup | 555 | 0.00% 4 |
| TruSeq 779-2 | - | 105402 | 0.40% |
| dedup | 37131 | 0.43% |
| phred > 31 | 5635 | 0.12% |
| phred > 31, dedup | 726 | 0.14% |
| TruSeq 779-Norm | - | 128423 | 0.40% |
| dedup | 46339 | 0.35% |
| UMI | Oncology 776 | - | 83687 | 0.02% |
| dedup | 31141 | 0.03% |
| consensus | 863 | 0.00% |
| Oncology 779 | - | 112187 | 0.22% |
| dedup | 43624 | 0.27% |
| consensus | 932 | 0.11% |
| 1 Filtering: dedup considered only deduplicated reads (i.e. reads after removal of redundant reads), phred > 31 considered only read-pairs with base quality > 31 for all bases in the 2x150bp read-pair, consensus filtering was performed with the Illumina UMI app which collapsed duplicate reads with identical UMIs to a consensus read.  2 Depth: number of reads at the mutation, but counting only reads where the base at the mutation has phred > 19.  3 AF: Allele frequency of the mutation (HD779) or of the PCR or sequencing error (HD776).  4 The stringent filtering led to a drop-out of sequences with the PIK3CA mutation. | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Supplementary Table 3: Deduplication (removal of redundant sequences) increases the allele frequency of signal noise (PCR or sequencing errors) and reduces the effective sequencing depth** | | | | | |
| **Type** | **Library** | **Filtering1** | **Mean depth2** | **AF3 95% Quantile** | **AF3 Max** |
| non-UMI | TruSeq 776-1 | - | 55472 | 0.02% | 0.04% |
| **dedup** | **18026** | **0.04%** | **0.13%** |
| phred > 31 | 2160 | 0.17% | 0.49% |
| phred > 31, dedup | 461 | 0.25% | 0.69% |
| TruSeq 776-2 | - | 52573 | 0.02% | 0.04% |
| **dedup** | **13667** | **0.04%** | **0.09%** |
| phred > 31 | 1416 | 0.17% | 0.39% |
| phred > 31, dedup | 412 | 0.32% | 0.69% |
| TruSeq 776-Norm | - | 51340 | 0.02% | 0.03% |
| **dedup** | **17559** | **0.03%** | **0.06%** |
| TruSeq 779-1 | - | 47552 | 0.02% | 0.04% |
| **dedup** | **11711** | **0.04%** | **0.12%** |
| phred > 31 | 789 | 0.26% | 0.53% |
| phred > 31, dedup | 332 | 0.00% | 0.75% |
| TruSeq 779-2 | - | 57491 | 0.02% | 0.03% |
| **dedup** | **17618** | **0.04%** | **0.11%** |
| phred > 31 | 1948 | 0.16% | 0.70% |
| phred > 31, dedup | 494 | 0.23% | 0.87% |
| TruSeq 779-Norm | - | 61746 | 0.02% | 0.04% |
| **dedup** | **23525** | **0.03%** | **0.05%** |
| UMI | Oncology 776 | - | 36786 | 0.05% | 0.12% |
| **dedup** | **13612** | **0.06%** | **0.16%** |
| consensus | 602 | 0.00% | 0.30% |
| Oncology 779 | - | 48924 | 0.05% | 0.13% |
| **dedup** | **21080** | **0.06%** | **0.12%** |
| consensus | 670 | 0.00% | 0.30% |
| 1 Filtering: dedup considered only deduplicated reads (i.e. reads after removal of redundant reads),  phred > 31 considered only read-pairs with base quality > 31 for all bases in the 2x150bp read-pair, consensus filtering with the Illumina UMI app collapsed duplicate reads with identical UMIs to a consensus read.  2 Mean depth: considered the entire coding region of PIK3CA but counted only bases with phred > 19.  3 AF: Allele frequency of the PCR or sequencing errors, computed at the 604 genomic coordinates that had at least 500X depth after filtering in every library. The AF values in this table have been corrected using the known PIK3CA allele frequency of 0.13% vs the sequenced AF in the HD779 DNA. | | | | | |



**Supplementary Figure 1. Ratio of forward-to-reverse reads is no useful quality criterion for UMI-libraries after error-correction bioinformatics.** The density plots show the position balance of forward reads to reverse reads at each genomic position, from targeted sequencing of the *PIK3CA* coding regions. The position balance was computed with GenSearchNGS, with 1 being optimal and 0 being worst. (A) Position balance in the four non-UMI-libraries (TruSeq 776-1, 776-2, 779-1, 779-2). The majority of genomic positions have a near-optimal read balance. (B) Position balance in the two UMI-libraries (Oncology 776, 779). Nearly all positions are covered by error-corrected reads oriented in just one genomic direction. The conventional quality check of forward/reverse balance is therefore no longer useful after UMI-based error-correction has been performed.