BEST PRACTICES FOR BIOINFORMATICS PIPELINES FOR MOLECULAR-BARCODED TARGETED SEQUENCING

Marcel Caraciolo[b], Murilo Cervato[b] , George Carvalho[a], Wilder Galvão[a], Rodrigo Alexandre[b], João B Oliveira[a]

[a] Genomika Diagnósticos, Recife-PE | [b] Hospital Israelita Albert Einstein, São Paulo-SP

INTRODUCTION AND MOTIVATION

In cancer research the detection of mutations is critical, for tumor samples and blood samples mutations may be present in very low fractions of DNA molecules. By using molecular barcoding technology, more than reduce the impact of enrichment, the sequencing errors can be eliminated by tagging each input molecule with an unique molecular identifier (UMI) [1,2]. In contrast to sample barcoding, molecular barcoding assigns a unique sequence not just to all the molecules from a certain sample, but to all molecules being amplified and sequenced. Recent works on this approach show outstanding performance in targeted high-throughput sequencing, being the most promising approach for the accurate identification of rare variants in complex DNA samples, and has application in several areas such as detecting DNA mutations at very low allele fractions with high accuracy for cancer samples and reducing sequencing artifacts occurrences. However, at the sample preparation, the residual PCR errors might be introduced at first PCR cycles and during UMI tag attachment, which decrease the accuracy of variant calling.

In order to perform the variant detection on those input data, a different approach is required for bioinformatics pipelines that handles the caveats of UMI-based analysis. In this poster we present the best practices and strategies for handling the UMI-tagged data, by showing the steps and related software tools to the audience when building the variant calling pipeline.

BIOINFORMATICS PIPELINE WITH UNIQUE DEDUPLICATION USING UMIS

1. **Align reads**
2. **Group read pairs to designed probes based on read start-stop postition**
3. **For each probe: group reads with identical molecular barcode sequence**
4. **Consolidate read informatio to one read per molecule (remove PCR duplicates)**

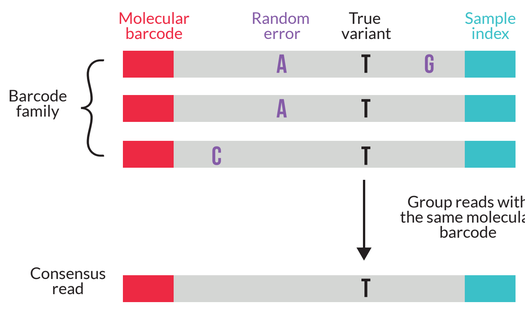

Figure 3: Molecular Barcode Analysis Step-By-Step.

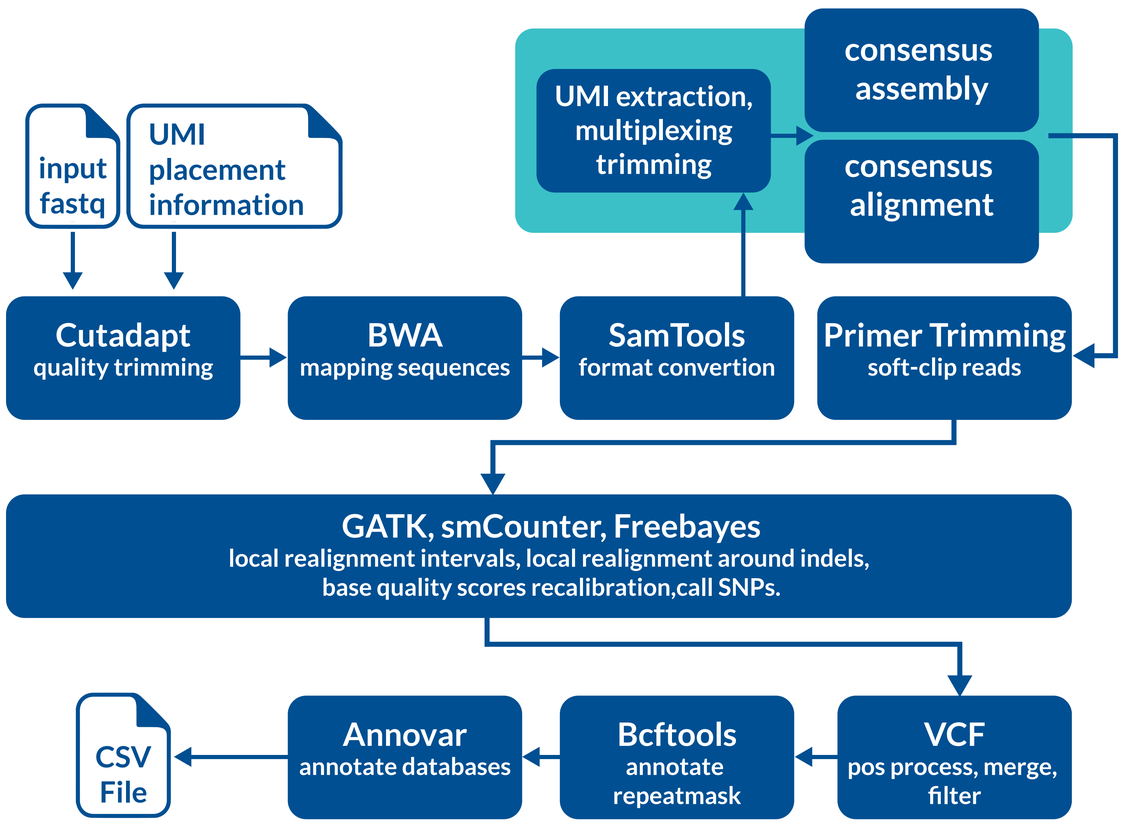

Figure 4: Variant Calling Pipeline

REFERENCES

[1] Kivioja T, Vaharautio A, Karlsson K, Bonke M, Enge M, Linnarsson S, et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Methods 2012, 9:72–74

[2] Peng Q, Vijaya Satya R, Lewis M, Randad P, Wang Y. Reducing amplification artifacts in high multiplex amplicon sequencing by using molecular barcodes. BMC Genomics. 2015;16(1):589. doi:10.1186/s12864-015-1806-8.

[3] Xu, Chang, et al. "Detecting very low allele fraction variants using targeted DNA sequencing and a novel molecular barcode-aware variant caller." BMC genomics 18.1 (2017): 5.

MOLECULAR BARCODE OVERVIEW

Molecular Barcode (UMI) is a tag barcode of random nucleotides incorporated into the original DNA molecules before amplification to preserve their uniqueness. It enables digital sequencing and identification of unique progenitor DNA fragments (de-duplication).

Figure 1: Biases and errors from PCR amplification or sequencing steps can be detected.

IMPROVED ACCURACY ON MOLECULAR QUANTIFICATION WITH UMIS

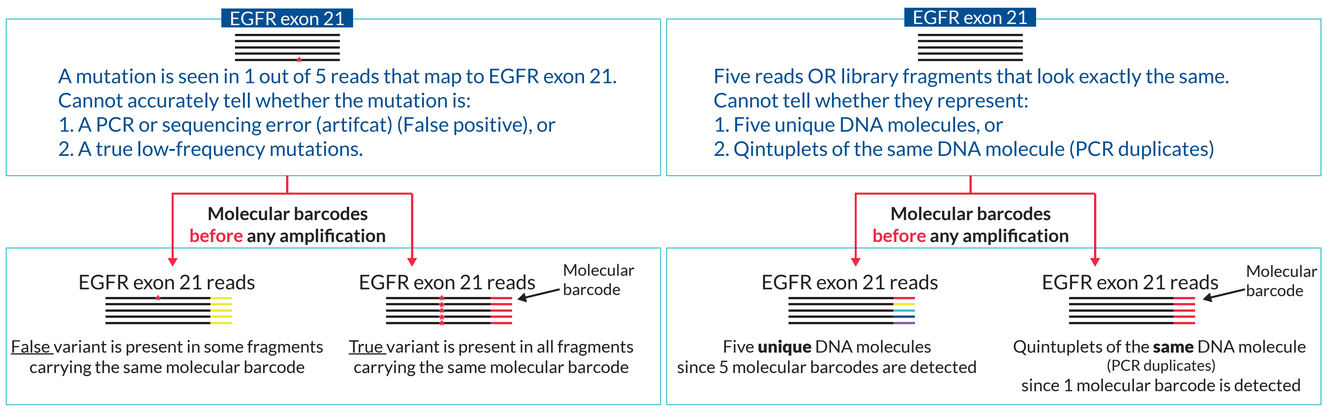

Figure 2: Conventional targeted DNA sequencing based on total reads x single original molecules quantification.

OPEN-SOURCE BIOINFORMATICS TOOLS SUGGESTED

Tool	Step	Link for Further information
fgbio	For variant calling, fgbio collapses sequencing duplicates for each UMI into a single consensus read prior to running re-alignment and variant calling.	http://bit.ly/2fUtqyd
samtools	For variant calling, samtools performs operations on BAM files such as sorting, header tags addition, alignment extractions.	http://bit.ly/1aNGfNY
picard	For variant calling, picard has several tools to perform operations on BAM files. In this scenario is used the MarkDuplicates command with BARCODE_TAG parameter.	http://bit.ly/2fWZ1zm
smCounter[3]	A versatile UMI-aware variant caller to detect both somatic and germline SNVs and indels.	http://bit.ly/2xHrKje

NEXT STEPS

A comprehensive benchmark is undergoing at our lab, in order to evaluate the accuracy and detection rate for somatic and germline variants with different allele frequencies. The goal is to assess the accuracy of UMI-tagged data processing and ultra-rare variant calling software using this proposed pipeline. Published results using the smCounter variant caller has demonstrated very good sensitivity and specificity in detecting 1% SNVs and indels within targeted coding regions. We still need to evaluate another callers that can be used in the pipeline and evaluate mixed experiment settings such as sample input, DNA quality, sequencing depth, sequencing platform and other sequencing factors.

CONCLUSION

In summary, we have been working on an NGS bioinformatics pipeline for somatic and germline variant calling using the molecular barcoding step into high multiplex PCR amplicon sequencing. The benefits of this approach is the promising results of in reducing low level sequencing artifacts, which would otherwise plague the detection of SNVs at very low fractions and the decreased error rate, with an increased accuracy for variant-calling, specially for low-input DNA.

bioinfo@genomika.com.br | genomika.com.br

Recife, PE / São Paulo, SP | Brazil

A-meeting 2017