Primerclip™: A tool for trimming primer sequences

IMPORTANCE OF SOFT-CLIPPING PRIMER SEQUENCES FOR xGEN™ PREDESIGNED AMPLICON PANELS

Our **xGen™ Predesigned Amplicon Panels** leverage a short overlapping (tiled) amplicon design (**Figure 1**) to provide continuous coverage of regions of interest in a single-tube format, yet maintain compatibility with short fragments of cell-free DNA (cfDNA) and damaged/degraded formalin-fixed paraffin embedded (FFPE) samples.

Due to the 2 x 151 read length and short amplicon design, synthetic primer sequences introduced during PCR research will be encountered at both the beginning and end of reads. These artificial sequences must be clipped before variant calling occurs. To accomplish this, we designed the Primerclip tool for soft clipping primer bases after alignment. Primerclip bioinformatically clips the 5′ and 3′ primer bases, eliminating the risk of variants being called from these synthetic sequences. In addition to speed, Primerclip also has the advantage of improving variant calling at the ends of alignments which may be otherwise compromised due to edge effect. Variants present at the edges/ends of amplicons will have a better chance of getting called if the primer bases are present during the alignment and then trimmed right before variant calling.

Note: Soft clipping by Primerclip is only performed at the ends of reads and not when the primer sequence is present internally (within a read). Because amplicons are tiled contiguously across the target region, coverage of the trimmed primer sequence is provided by sequence from an adjacent amplicon.

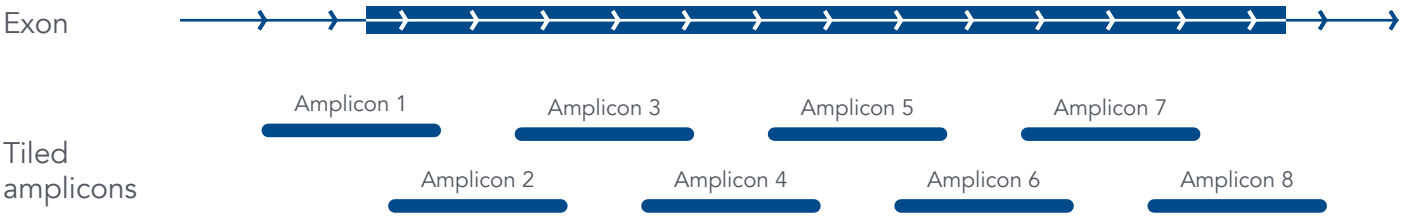

Figure 1. xGen™ Predesigned Amplicon Panels leverage overlapping amplicons to allow for contiguous regions of coverage in a single-tube format. This illustration represents eight overlapping amplicons across the region of interest to provide continuous coverage of the entire exon in a single-tube format.

PIPELINE WORKFLOW

Trimming primer sequences is mandatory for reliable data output and variant calling from xGen Amplicon Panels. A suggested data analysis pipeline for **xGen Predesigned Panels** includes 1) adapter trimming, 2) alignment, and 3) post-alignment primer soft-clipping using the panel-specific Masterfile and the Primerclip pre-compiled binary file supplied by IDT. Following soft clipping, target enrichment evaluation and variant calling are performed using the appropriate amplicon panel's target BED file. See **Figure 2** for an example data analysis workflow of the xGen Amplicon Panels.

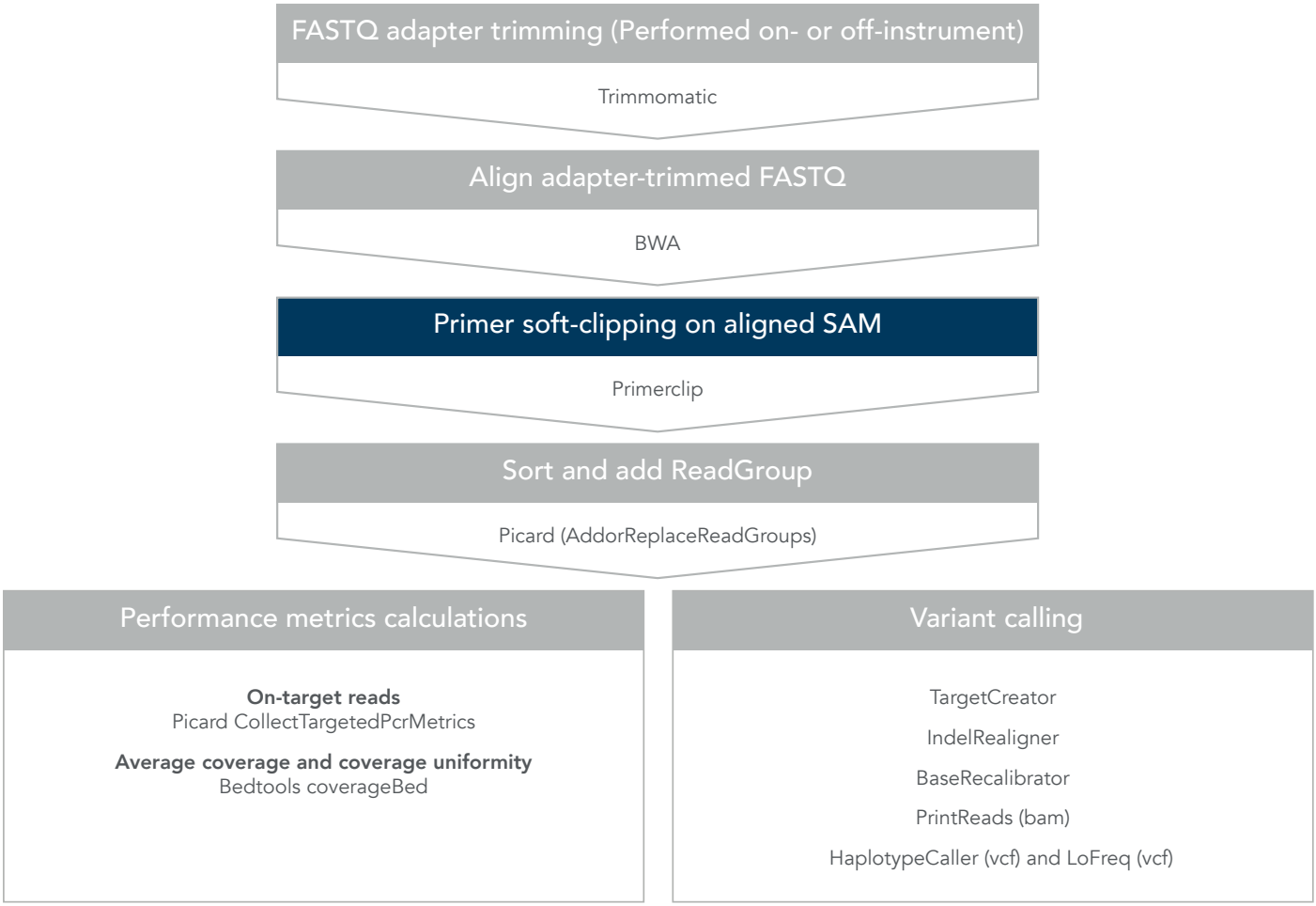

Figure 2. Suggested bioinformatics pipeline for xGen Amplicon Panels. Post-alignment soft clipping of primer sequences using the Primerclip tool (highlighted in dark blue) is a mandatory step when performing analysis on sequencing results from xGen Amplicon Panels. The other steps shown may be performed as listed, or a similar data analysis pipeline may be followed.

ADAPTER TRIMMING

Primerclip is designed to be used after Illumina® adapter trimming with a program such as Trimmomatic [1]. Adapter trimming can be done on- or off- instrument. For on-instrument use, ensure that adapter-trimming setting is enabled while setting up the sequencing run. Alternatively, this step can be performed bioinformatically before data analysis.

ALIGNMENT

After adapter sequences have been trimmed, use an alignment tool such as BWA [2] to align sequenced reads to the reference genome.

PRIMER TRIMMING

After performing the alignment step, use Primerclip to trim primer sequences unique to the xGen Amplicon Panels. To execute Primerclip, use the panel-specific Masterfile provided by IDT, and your SAM alignment file as input to the tool. The tool will output a primer-clipped SAM file, which can then be converted to BAM and used for regular downstream analysis. See Primerclip's README document for comprehensive instructions to install and use this tool.

Go here to download Primerclip. The Primerclip program is compatible with xGen Predesigned Amplicon Panels. Panel-specific BED files and Masterfiles are available via applicationsupport@idtdna.com.

After primer trimming is complete, use the Picard tool, AddOrReplaceReadGroups, (Broad Institute) to sort and add ReadGroups to the SAM file. The processed SAM file can then be converted to BAM for downstream analysis.

PERFORMANCE METRICS

To assess data quality, calculate the percent of both on-target and coverage uniformity with the following equations.

On-target

The on-target metric is a measure of the proportion of the total aligned reads sequenced from an NGS library that are aligned to the intended target regions on the reference genome. On-target percent (OT%) is calculated as the ratio of "total bases (B) that map/align to the on-target (OT) region" to the "total (Tot) aligned bases."

$$OT\% = \frac{B^{OT}}{B^{Tot}} \times 100.0$$

Where,
B^{OT} = number of aligned bases which coincide with or are adjacent to a target base
B^{Tot} = total number of aligned bases

Coverage uniformity

Coverage uniformity percent (CU%) is calculated as the ratio of "number of target bases that have coverage at or above 20% (0.2μ)" and "total number of target bases".

$$CU\% = \frac{B^{0.2\mu}}{B^{OT}} \times 100.0$$

Where,
$B^{0.2\mu}$ = total number of bases with coverage ≥ 20% of mean
B^{OT} = total number of target bases

Important: While the use of the Picard tool, MarkDuplicates, is a common quality control step to identify low-complexity libraries, MarkDuplicates cannot be used on data derived from PCR-based target enrichment methods such as our xGen Amplicon Panels. Since these targeted panels contain high numbers of identical library fragments (particularly regarding alignment start position), MarkDuplicates cannot appropriately analyze xGen Amplicon libraries.

CONSIDERING PANELS WITH THE xGEN SAMPLE ID AMPLICON PANEL AS A SPIKE-IN

Important: Consider the following information to achieve the most reliable sequencing results from libraries prepared from xGen Oncology & Inherited Disease Amplicon Panels:

- **On-target percentage (OT%):** Use the combined BED file (available by emailing applicatonsupport@idtdna.com) to calculate overall OT% for the entire panel. Calculating OT% for the entire panel based on a BED file specific to main panel content only or xGen Sample ID Amplicon Panel only will result in incorrect assessment of OT%, as the reads from the excluded panel will be incorrectly reported as off-target.
- **Coverage uniformity percentage (CU%):** Use the BED file specific to main panel content or xGen Sample ID Amplicon Panel to calculate CU% for each subset of the data. By design, the xGen Sample ID targets in the combined panel have lower coverage than main panel targets because they are specific to germline variants; therefore, analyzing CU% across the combined panel will inaccurately reflect Sample ID targets as low coverage dropouts.

For other xGen Amplicon Panels

Use the panel-specific BED or Masterfile file supplied with each panel to calculate the OT% and CU% on the sequenced reads.

VARIANT CALLING TOOLS

There are various publicly available tools to call variants in your samples, such as:

- LoFreq (Genome Institute of Singapore) [3]
- LoFreq-somatic
- Vardict [4]
- Varscan [5-7]
- FreeBayes [8]
- Strelka [9]
- Mutect [10]
- HaplotypeCaller (GATK, Broad Institute)

REFERENCES

1. Bolger AM, Lohse M, Usadel B. **Trimmomatic: a flexible trimmer for Illumina sequence data**. Bioinformatics. 2014;30(15):2114-2120.
2. Li H, Durbin R. **Fast and accurate short read alignment with Burrows-Wheeler transform**. Bioinformatics. 2009;25(14):1754-1760.
3. Wilm A, Aw PP, Bertrand D, et al. **LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets**. Nucleic Acids Res. 2012;40(22):11189-11201.
4. Lai Z, Markovets A, Ahdesmaki M, et al. **VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research**. Nucleic Acids Res. 2016;44(11):e108.
5. Koboldt DC, Larson DE, Wilson RK. **Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection**. Curr Protoc Bioinformatics. 2013;44:15 14 11-17.
6. Koboldt DC, Zhang Q, Larson DE, et al. **VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing**. Genome Res. 2012;22(3):568-576.
7. Koboldt DC, Chen K, Wylie T, et al. **VarScan: variant detection in massively parallel sequencing of individual and pooled samples**. Bioinformatics. 2009;25(17):2283-2285.
8. Garrison E, Marth, G. . **Haplotype-based variant detection from short-read sequencing**. arXiv. 2012;1207:3907v3902 [q-bio.GN].
9. Kim S, Scheffler K, Halpern AL, et al. **Strelka2: fast and accurate calling of germline and somatic variants**. Nat Methods. 2018;15(8):591-594.
10. Cibulskis K, Lawrence MS, Carter SL, et al. **Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples**. Nature biotechnology. 2013;31(3):213-219.

Primerclip: A Tool for Trimming Primer Sequences

Technical support: applicationsupport@idtdna.com.

Doc ID: RUO22-0695_001 03/22