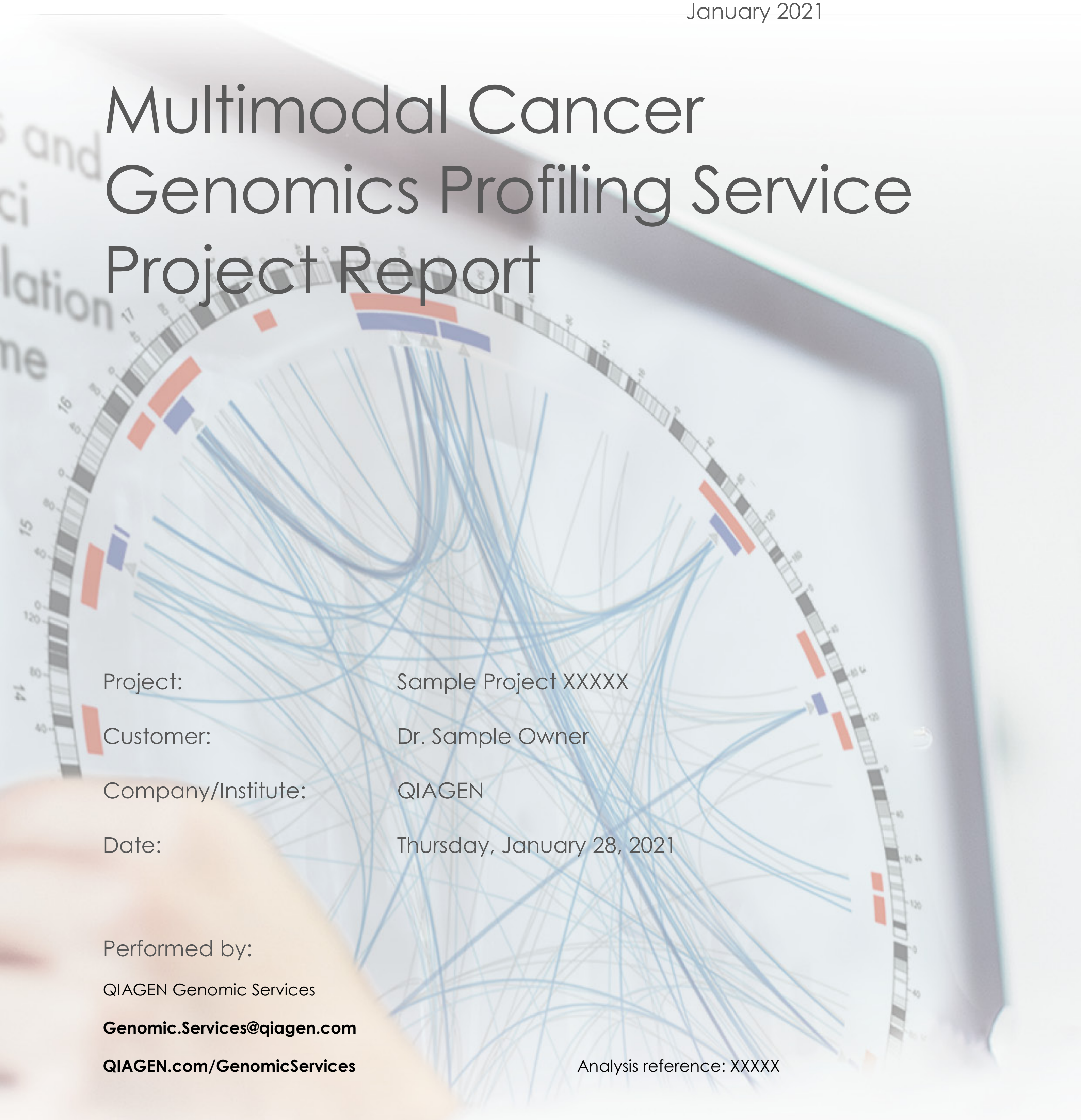
January 2021

Multimodal Cancer Genomics Profiling Service Project Report



Project:	Sample Project XXXXX
Customer:	Dr. Sample Owner
Company/Institute:	QIAGEN
Date:	Thursday, January 28, 2021

Performed by:

QIAGEN Genomic Services

Genomic.Services@qiagen.com

QIAGEN.com/GenomicServices

Analysis reference: XXXXX

Sample to Insight

Contents

Project Summary

Sample Overview and Metadata

Sample ID	Sample name	Condition
99999-001	Sample 1	Infected
99999-002	Sample 2	Infected
99999-003	Sample 3	Healthy control

Main findings, Conclusions, Next steps

Dear Customer,

We have now finalized the Next Generation Sequencing (NGS) analysis for the samples you have submitted to QIAGEN Genomic Services.

QIAseq® Multimodal Panel NGS libraries were successfully prepared, quantified, and sequenced for all your samples. The collected reads were subjected to quality control, alignment, and downstream analysis.

The principal structure of your result data is summarized in this report. In addition, the report contains details on the technical background.

For more information about QIAGEN's products for validation and functional analysis of your RNA or DNA of interest, please go to our **Next-Generation Sequencing Content World** at QIAGEN.com.

If you have any questions, please contact your local QIAGEN representative or our Genomic Services lab scientists at **Genomic.Services@qiagen.com**.

Kind regards,

QIAGEN Genomic Services

Data Package

Overview

All analyses were performed using CLC Genomics Workbench (version 20.0.4) and CLC Genomics Server (version 20.0.4).

Human genome version: GRCh38

Gene annotation: NCBI Homo sapiens Updated Annotation Release 109.20190607

Panel: QIAseq Multimodal Human Pan Cancer-Panel (UHZ-5000Z)

Variant interpretation: QIAGEN's professional variant interpretation service

Content	Description
Fastq Quality control	**[1]_Quality Control** QC report and supplementary QC report (per sample)
Fastq adapter- and quality-trimming	**[2]_Trimming** Trimming report (per sample)
Mapping Statistics	**[3]_Mapping** **DNA** Combined report (per sample) UMI report (per sample) UMI groups report (per sample) Coverage report (per sample) Per-region statistics (per sample) **RNA** UMI report (per sample) RNA-seq report (per sample) RNA target QC report (per sample) Fusion report (per sample) Gene expression matrix
Variants	**[4]_Variants** For each comparison: 1. Excel table and VCF file for unfiltered DNA variants (per sample) 2. Excel table and VCF file for filtered DNA variants (per sample) 3. Excel table for fusion gene (per sample) 4. VCF file for combined DNA and RNA variants (per sample) 5. Variant interpretation pdf report (per sample)

Technical Background

DNA and RNA reads quality control

The QC reports were generated by the "QC for Sequencing Reads" tool from CLC Genomics Workbench and can be found in **[1] Quality Control** to assess and visualize statistics on sequence read lengths and base coverages, nucleotide contributions and base ambiguities, quality scores as emitted by the base-caller, and over-represented sequences and hints that suggest contamination events. The information from the reports describe if the sequencing of the libraries were performed in high quality. Good sequencing metrics is defined by having average PHRED scores > 30.

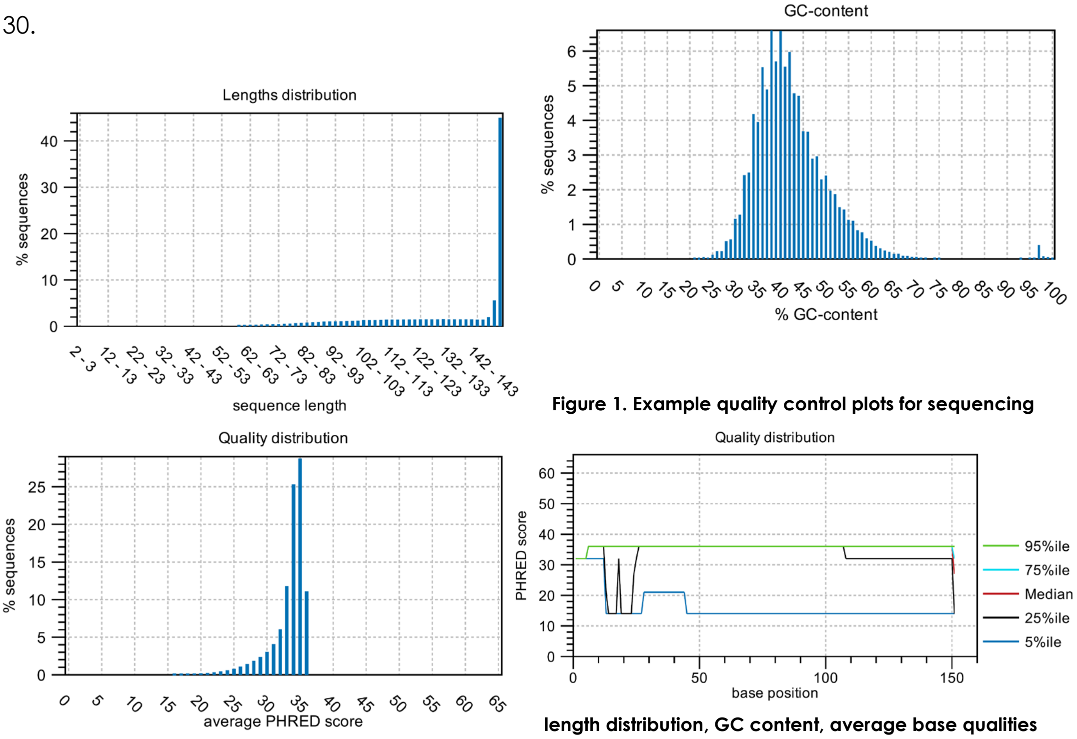


Figure 1. Example quality control plots for sequencing length distribution, GC content, average base qualities for reads, and quality distributions over the read length (from top left to bottom right panels).

DNA and RNA reads preprocessing: Removal and annotation with UMI

Unique Molecular Indices (UMI) located at the beginning (DNA: 12 nt, RNA: 10 nt) and the linker sequences were removed from read 2 sequences, and each read pairs were annotated with the UMI information. The UMI information is important to identify true variants from false-positive variants introduced by sequencing or PCR errors (Figure 2).

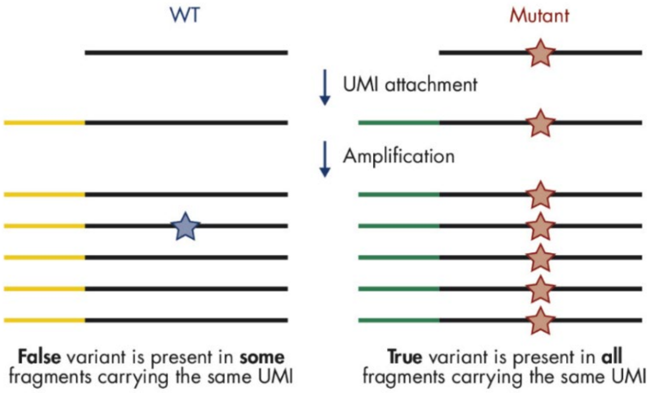

WT
Mutant
UMI attachment
Amplification
False variant is present in some fragments carrying the same UMI
True variant is present in all fragments carrying the same UMI

Figure 2. UMI reads remove false-positive variants.

DNA and RNA reads preprocessing: Adapter and quality trimming

After removal of UMIs and linker sequences, adapter and quality trimming was done by the “Trim Reads” tool from CLC Genomics Workbench. While adapters are often removed directly by the sequencer, part of the adapter region may be included in the sequenced reads. Such adapter artefacts were removed by identifying read-through adapter sequences, whereby the 3’ end of one read includes the reverse complement of the adapter from the other read.

Furthermore, reads were trimmed based on quality scores and ambiguous nucleotides (e.g., due to stretches of Ns). A maximum of 2 ambiguous nucleotides were allowed in a read. The trimming reports can be found in **[2]_Trimming**. For multimodal panels, the read length after trimming is usually shorter than before trimming because some of the single-primer extended amplicons can be short depending on the fragmentation of the sample and read-through adapters would be detected in the 2 x 150 paired-end sequencing of these short fragments. Another possible reason if read lengths are significantly shorter than before trimming is if the sequencing qualities are low at the end of the reads, in which case should be reflected in the QC report from **[1]_Quality Control**.

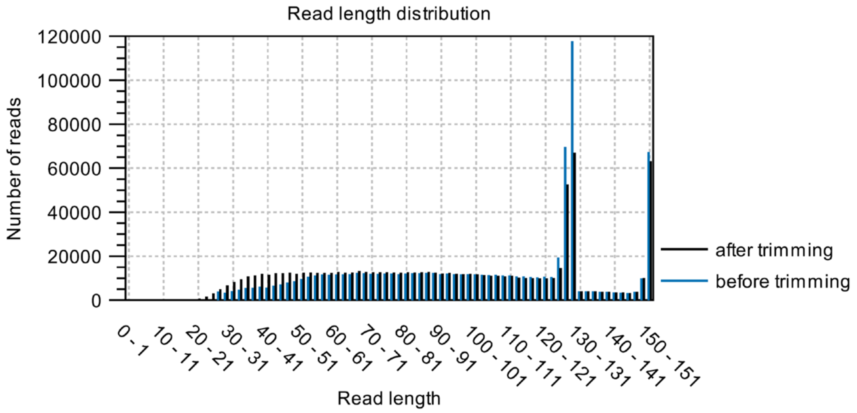


Figure 3. Example read length distribution before and after trimming.

DNA reads mapping and UMI grouping

After adapter trimming, the DNA variant calling analysis was performed using the "Perform QIAseq Multimodal Analysis" workflow from CLC Genomics Workbench. In this workflow, reads are first mapped to the reference genome with mapping statistics summarized in the combined report in [3]_Mapping/DNA. A QC file describing the read coverage statistics for each targeted region can also be found in a per-region spreadsheet in the same folder. After mapping to the genome, reads with the same UMI were grouped together, and the result for UMI groupings can be viewed in the UMI_report in [3]_Mapping/DNA. For UMI groups with UMIs that differ by one mismatched base, these groups are merged, and for each merged UMI group, a consensus sequence is then calculated by evaluating bases from all reads at each position at a time. At each position along the read, a consensus base is only reported when at least 50% of the reads have the same base identity (A, C, G, or T). The summary for this step can be found in UMI group report in [3]_Mapping/DNA. In this report, you can find information of the UMI groups such as the number of UMI groups found, and how many reads are in each UMI group. An example figure of number of reads in a UMI group is shown below (Figure 4) A median of ~3 reads per group is usually optimal for balancing sensitivity and sequencing depth.

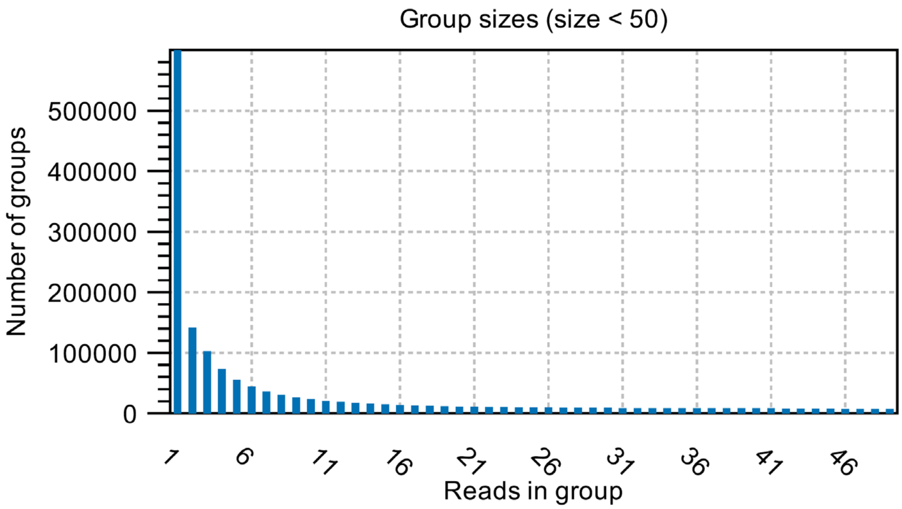


Figure 4. Example histogram of UMI group sizes.

DNA variant callings

Variant callings can be sensitive to alignment artefacts that will lead to incorrect reporting of variants at genome loci with insertions and deletions, such as homopolymer regions. To improve variant callings, a local realignment step is performed using the CLC Genomic Workbench. Variants were then identified from the refined read alignments with at least 0.5% variant frequency. You can find the list of variants in the VCF file and spreadsheet named as unfiltered_variants under [4]_Variants. In addition, a filtered list of variants can also be found in the same folder. The variants are filtered by 2 criteria:

1. **Filter based on quality criteria**: Average Quality (quality of the sequenced bases that carry the variant), QUAL (significance of the variant), and Read Direction Test Probability (relative presence of the variant in the reads from different directions that cover the variant position).
2. **Remove homopolymer error type variants**: Errors of the indel type that occur in homopolymer regions. These regions are known to be harder to sequence than non-homopolymeric regions. Note that the definition of homopolymeric regions differ between the pipelines due to differences in sequencing technology. Homopolymer is defined as regions with the same base repeating at least 3 times (AAA, CCC, TTT, or GGG).

The filtered variant tables contain the following columns describing the details of each variant:

- Chromosome: The name of the reference sequence on which the variant is located.

- Region: The region on the reference sequence at which the variant is located. The region may be either a 'single position', a 'region', or a 'between position region'.
- Type: The type of variant. This can be SNV (single-nucleotide variant), MNV (multi-nucleotide variant), insertion, deletion, or replacement.
- Reference: The reference sequence at the position of the variant.
- Allele: The identity of the variant.
- Reference allele: Describes whether the variant is identical to the reference.
- Count: The number of 'countable' fragments supporting the allele. The 'countable' fragments are those that are used by the variant caller when calling the variant.
- **Coverage**: The fragment coverage at this position.
- **Frequency**: Variant frequency calculated by 'Count' divided by 'Coverage'.
- **Average quality**: The average base quality score of the bases supporting a variant.
- **Read count**: The number of 'countable' reads supporting the allele. Note that each read in an overlapping pair contribute 1.
- **Read coverage**: The read coverage at this position.
- **# Unique start positions**: The number of unique start positions for 'countable' fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same start position, you could suspect that it is a result of an amplification error.
- **# Unique end positions**: The number of unique end positions for 'countable' fragments that support the variant. This value can be important to look at in cases with low coverage. If all reads supporting the variant have the same end position, you could suspect that it is a result of an amplification error.
- **BaseQRankSum**: The BaseQRankSum column contains an evaluation of the quality scores in the reads that has a called variant compared with the quality scores of the reference allele. Variants for which no corresponding reference allele is called does not have a BaseQRankSum value. Likewise, no values are calculated for reference alleles. The score is a Z score, so a value of 2.0 means that the observed qualities for the variant two standard deviations below the qualities for the reference allele. The scoring is performed using a Mann-Whitney U for comparing the two sets of quality scores from the reference allele and the variant.
- **Read position test probability**: The test probability for the test of whether the distribution of the read positions variant in the variant carrying reads is different from that of all the reads covering the variant position.
- **Read direction test probability**: The test probability for the test of whether the distribution among forward and reverse reads of the variant carrying reads is different from that of all the reads covering the variant position.
- **Homopolymer**: The column contains "Yes" if the variant is likely to be a homopolymer error and "No" if not.

- **Amino acid change**: If the reference sequence of the mapping is annotated with ORF or CDS annotations, the variant caller will also report whether the variant is synonymous or non-synonymous. If the variant changes the amino acid in the protein translation, the new amino acid will be reported.

RNA Reads Mapping and UMI grouping

After adapter trimming, the RNA fusion discovery analysis was performed using the "Perform QIAseq Multimodal Analysis" workflow from CLC Genomics Workbench. The UMI from the trimmed RNA reads were removed from the reads. The reads were grouped into UMI groups and a consensus read (UMI read) was calculated from each UMI group. A report describing the summary statistics for the UMI reads can be found under **[3]_Mapping/RNA**.

The UMI reads were then sequentially mapped to the annotated transcriptome reference and genome reference, wherein a gene expression matrix is computed from. The gene expression matrix and the RNA panel QC information can be found under **[3]_Mapping/RNA.**

The genomic loci of the mapped UMI reads with at least 20 nucleotides on the 5' or 3' ends unaligned to the genome reference were then annotated. The unaligned segments on these partially-mapped UMI reads were extracted and remapped to the genome to annotate fusion break points and create artificial fusion chromosomes, as illustrated in Figure 5. All RNA reads were then remapped to the genome reference and the artificial chromosomes to quantify fusion RNAs. With recount of fusion crossing supporting reads at each fusion site, a binomial model is used to calculate p-values and z-scores for each fusion event by considering the read coverage information at the upstream and downstream of the fusion site. A fusion gene report can be found in **[3]_Mapping/RNA**, and a table describing statistics of the fusion site can be found in **[4]_Variants**.

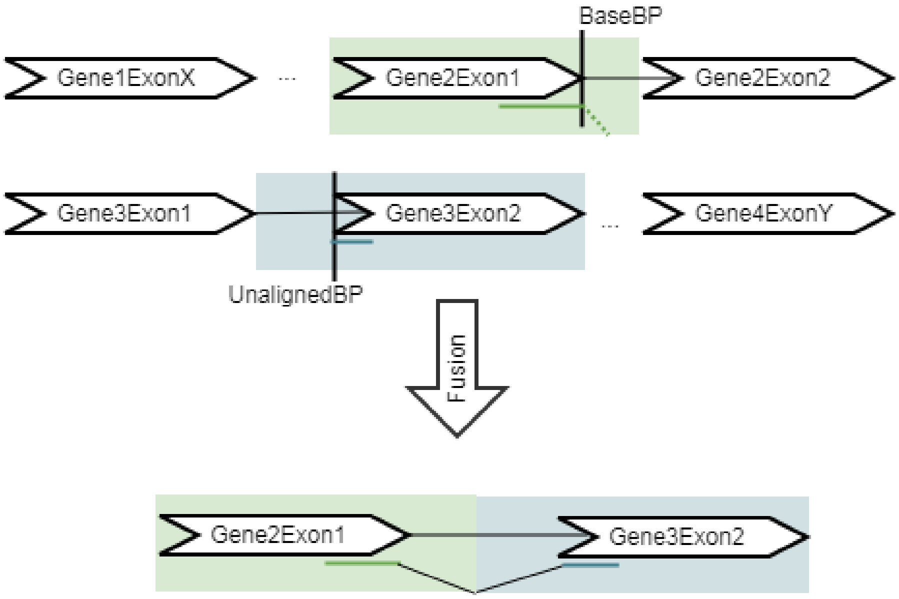


Figure 5. Example histogram of UMI group sizes.

Variant interpretation report

Confidently detected somatic variants are vetted for biological and clinical relevance based on the AMP/ASCO/CAP and ACMG/AMP guidelines for variant interpretation. Expert variant and disease specific summaries are provided for biologically or clinically relevant variants. No variant specific interpretive comments are provided for variants of uncertain significance.

The variants presented on the final report summary page (page 1) according to AMP/ASCO/CAP guidelines are for research purposes only. **Tier 1** contains variants of strong clinical significance: variants that are associated with therapies approved within the specific indication or with level A or B prognostic or diagnostic significance. Also shown on the first page are therapies with evidence of resistance, as well as if clinical trials are matched to that variant. **Tier 2** contains variants of potential clinical significance. Variants with biological significance (pathogenic/likely pathogenic) without associated targeted therapies or trials and without significant prognostic or diagnostic relevance are listed as well on the first page to indicate a potential role as likely drivers of cancer. Please note that this information is intended to support your research studies and should never be used for clinical purposes.

Strong clinical significance	Tier 1A	- Biomarker predicts response or resistance to an FDA or EMA approved therapy, according to drug label or professional guidelines for this diagnosis. - Biomarker included in professional guidelines is prognostic or diagnostic for this diagnosis.
	Tier 1B	- Biomarker predicts response or resistance to a therapy for this diagnosis based on well-powered studies. - Biomarker is prognostic or diagnostic for this diagnosis based on well-powered studies.
Potential clinical significance	Tier 2C	- Biomarker is associated with response or resistance to an FDA or EMA approved therapy, according to drug label or professional guidelines but only for different diagnosis. - Biomarker is an inclusion criterion for an active clinical trial. - Biomarker is prognostic or diagnostic based on multiple small studies.
	Tier 2D	- Biomarker shows plausible response or resistance based on case or preclinical studies. - Biomarker may assist in disease diagnosis or prognosis based on small studies.
Uncertain clinical significance	Tier 3	- Biomarker has uncertain clinical significance and not known to be likely benign or benign.

The variant details section of the report contains for all biological or clinically significant variants a highly structured interpretation summary with subheadings and increasing levels of detail. The high-level summary indicates the type of alteration, the role of variants in this gene in cancer, and association to any therapies that target this gene. The next sections describes the molecular function of the specific variant on the protein, along with all the supporting references, followed by a section on the incidence of alterations in this gene in this cancer type and the role of this gene in the development of cancer in general, or this cancer type. The clinical evidence section summarizes the diagnostic or prognostic significance for this mutation, variant specific, or alteration type specific evidence on drug sensitivity or resistance including a summary of clinical trial outcome studies, if applicable

Variants of uncertain significance will be listed in a tabular format. Benign/likely benign variants will not be reported.

Material and Methods

DNA and RNA Isolation

An amount of 25 mg tissue was used for RNA/DNA isolation using the AllPrep® DNA/RNA Mini Kit after manufacturer`s instructions. The RNA and DNA were eluted with 50 µl RNase-free water 100 µl of Buffer EB, respectively.

DNA and RNA QC

The DNA and RNA concentration were quantified by fluorescence-based method, and the integrity was determined by electrophoresis.

Multimodal sequencing library preparation

The library preparation was done using the QIAseq Multimodal Panel.

Briefly, 250 ng RNA and 40 ng DNA are used as starting material. While the RNA is heat fragmented, the DNA is enzymatically fragmented, end repaired, and A-tailed. Specific to RNA, synthetic polyadenylation is performed to create a binding site for subsequent reverse transcription. Specific to DNA, UMI-containing adapters are ligated at the 3` ends of the molecules. Specific to RNA, reverse transcription and template switching are performed, including the introduction of UMI. For combined target enrichment, ligated DNA molecules and reverse-transcribed/template-switched cDNA molecules are subject to 8 cycles of targeted PCR using a single primer extension (SPE) approach. A Universal PCR is ultimately carried out separately on DNA and RNA libraries to both optimally amplify each library as well as add the second UDI. The libraries size distribution was validated on an automated capillary electrophoresis system. High quality libraries are pooled based in equimolar concentrations based on the Bioanalyzer® automated electrophoresis system (Agilent Technologies). The library pool(s) were quantified using qPCR, and optimal concentration of the library pool used to generate the clusters on the surface of a flow cell before sequencing on a NextSeq® instrument (2 x 149 pb) according to the manufacturer instructions (Illumina Inc.).

Data Analysis

Data analysis was done using QIAGEN CLC Genomics Workbench (version 20.0.4) and QIAGEN CLC Genomics Server (version 20.0.4). The sequencing data were processed with default parameters using “Perform QIAseq Multimodal Analysis” workflow.

For DNA reads, bases at the UMI positions are removed and each read was annotated with the UMI information. Read-through adapters sequences and low quality bases were trimmed from the UMI-annotated DNA reads, and the high quality trimmed reads were mapped to the hg38 human genome reference. After mapping to the genome, the reads were deduplicated using the UMI information and read mapping coordinates, such that the sequencing and PCR errors were corrected. Reads that were likely ligation artifacts were then filtered out and insertion and deletion variants were detected from the filtered read mappings as a guidance for local realignments. The gene specific DNA primer sequences were then trimmed from the locally realigned read mappings and variants with >1% frequencies were reported.

For RNA reads, bases at the UMI positions are removed and each read was annotated with the UMI information. Sequencing adapters, low quality bases, and homopolymers from the 3’ ends of the reads were then trimmed from the UMI-annotated reads. The trimmed reads were then grouped by the UMI information and merged to generate high quality reads, which were then aligned to the hg38 genome using “RNA-seq Analysis” tool from CLC. Reads that were likely ligation artifacts were then filtered out and potential fusion events with at least 4 supporting reads were reported.

Variant interpretation and results of the interpretation will be provided in a final pdf report.

The QIAGEN Genomics Profiling Service is intended exclusively for research use only (RUO). This service is not for use in diagnostic procedures.

For up-to-date licensing information and product-specific disclaimers, see the respective QIAGEN kit handbook or user manual. QIAGEN kit handbooks and user manuals are available at **www.qiagen.com** or can be requested from QIAGEN Technical Services or your local distributor.