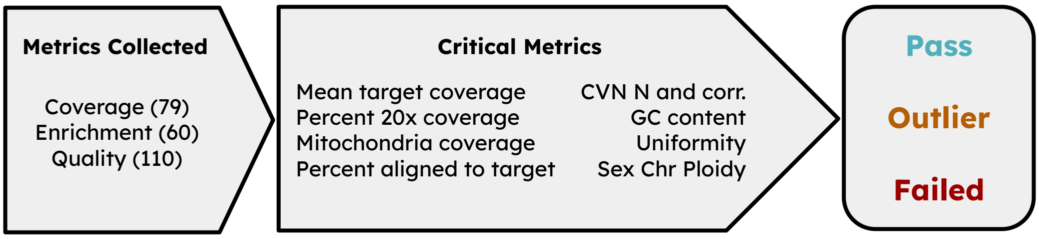
Streamlined QC for Clinical Sequencing

Reut Fluss, Michal Naftali, Dina Marek-Yagel, Rotem Greenberg, Shay Ben-Shachar and Ofer Isakov
The Clalit Genomic Center, Clalit Research Institute

Background: Advanced genetic sequencing, specifically Next Generation Sequencing (NGS), is used for clinical diagnosis and genomic research. Rigorous bioinformatic quality control (QC) is critical to ensure the accuracy and reliability of sequencing outputs and data analysis

Methods: The Clalit Genomic Center has developed an automated process for comprehensive quality control. Quality metrics are collected for each sequencing run at every stage of the process. Sequencing quality metrics include base calling and target DNA enrichment quality. Alignment quality is based on coverage data across target regions. Variant calling quality is assessed for both single nucleotide variants (SNVs) and copy number variations (CNVs). To evaluate the reliability of Copy Number Variants (CNVs) detection, coverage correlation is calculated between each sample and all other samples sequenced in the same run. Additionally, a comparison is made between the reported and estimated sex of the subject based on sequencing results.



Metrics Collected

Coverage (79)
Enrichment (60)
Quality (110)

Critical Metrics

Mean target coverage
Percent 20x coverage
Mitochondria coverage
Percent aligned to target

CVN N and corr.
GC content
Uniformity
Sex Chr Ploidy

Pass

Outlier

Failed

Results: Overall, 249 different metrics corresponding to coverage (79), enrichment (60) and quality (110) are collected. Nine critical metrics are used to defined outlier and failed samples. These metrics include: Mean target coverage, percent 20x coverage, mitochondria coverage, percent aligned to target, CVN number and correlation, GC dropout rate, uniformity and ploidy match. After completing the quality control process, an automated summary is generated, including aggregated metrics, and samples flagged due to outlier or failed metrics.

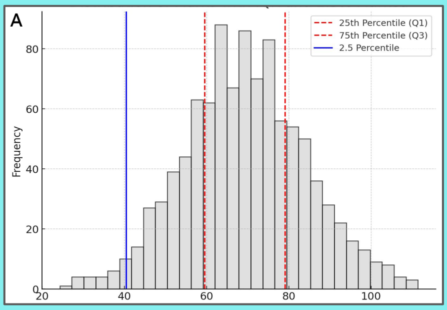
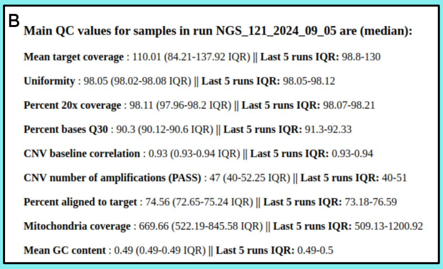


Figure A: For each critical quality metric, outlier thresholds are set dynamically by collecting its values across samples from the last five sequencing runs. Samples with a metric value above the 97.5th or below the 2.5th percentile are reported as outliers. **Figure B**: A screenshot from the QC report sent to the lab and bioinformatics teams including aggregated mean and interquartile ranges for each critical metric.



Main QC values for samples in run NGS_121_2024_09_05 are (median):

Mean target coverage : 110.01 (84.21-137.92 IQR) || **Last 5 runs IQR:** 98.8-130

Uniformity : 98.05 (98.02-98.08 IQR) || **Last 5 runs IQR:** 98.05-98.12

Percent 20x coverage : 98.11 (97.96-98.2 IQR) || **Last 5 runs IQR:** 98.07-98.21

Percent bases Q30 : 90.3 (90.12-90.6 IQR) || **Last 5 runs IQR:** 91.3-92.33

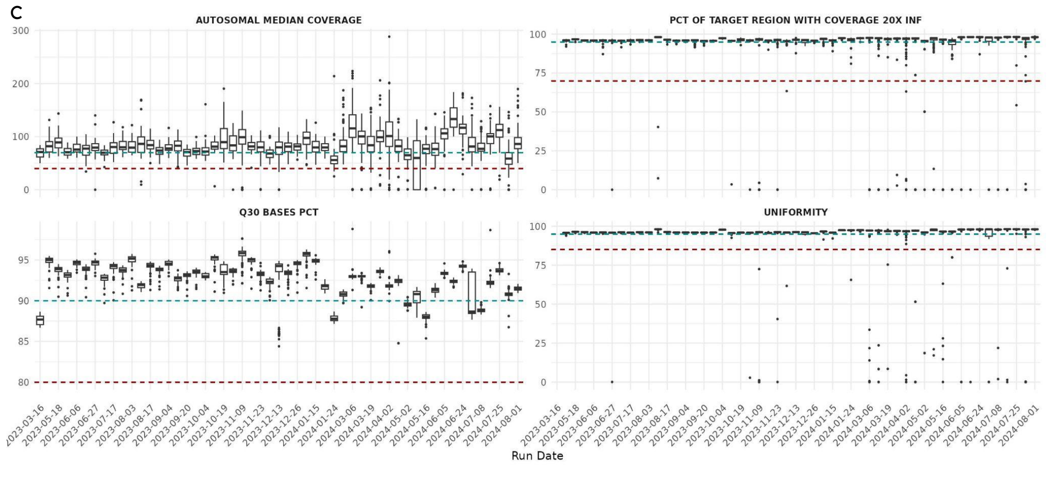
CNV baseline correlation : 0.93 (0.93-0.94 IQR) || **Last 5 runs IQR:** 0.93-0.94

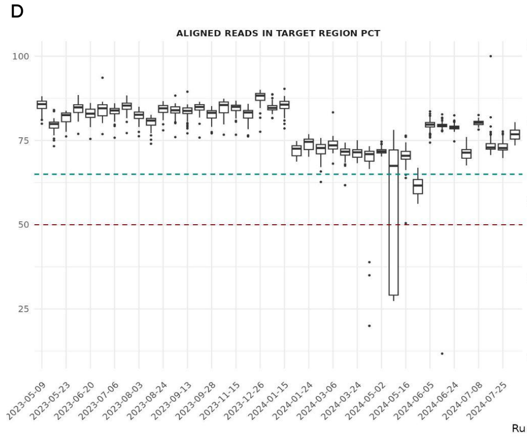
CNV number of amplifications (PASS) : 47 (40-52.25 IQR) || **Last 5 runs IQR:** 40-51

Percent aligned to target : 74.56 (72.65-75.24 IQR) || **Last 5 runs IQR:** 73.18-76.59

Mitochondria coverage : 669.66 (522.19-845.58 IQR) || **Last 5 runs IQR:** 509.13-1200.92

Mean GC content : 0.49 (0.49-0.49 IQR) || **Last 5 runs IQR:** 0.49-0.5





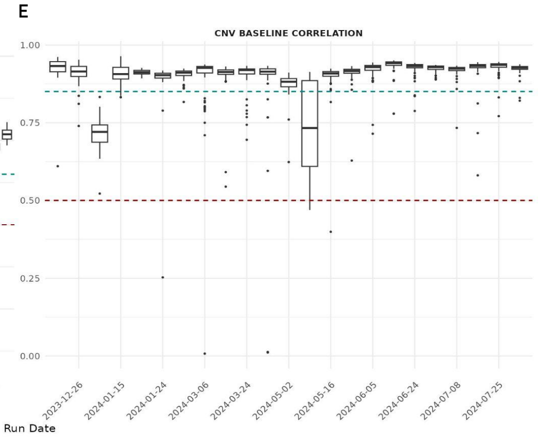


Figure C: Critical metrics across sequencing runs. The distribution of sample values within each run is depicted by boxplots. Thresholds corresponding to outlier (blue) and failed (red) values are added for each metric. **Figure D**: A decrease in target enrichment value following a change in enrichment kit and how technical changes in the lab improved performance. **Figure E**: transitioning into per-batch CNV normalization resulted in better consistency and performance

Conclusions: The bioinformatic quality control process at Clalit's Genomic Center is an integral part of clinical sequencing, enabling multiple parallel checks to detect abnormal metrics and maintain high reliability of process outputs. This process allows for identifying potential failures and addressing them in result analysis, as well as detecting errors in the workflow.

GENOMIC CENTER | By CLALIT HEALTH

CLALIT

CLALIT INNOVATION