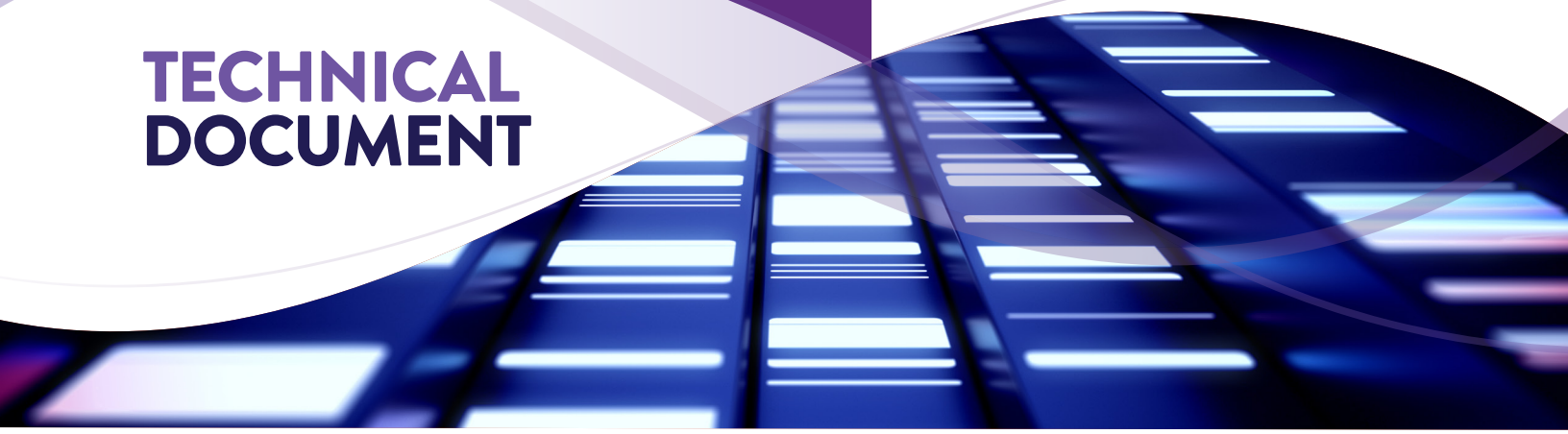
ATCC

TECHNICAL DOCUMENT



THE ATCC GENOME PORTAL: OUR APPROACH TO CELL LINE WHOLE-EXOME AND RNA SEQUENCING

As life science research progresses, the quality of data becomes increasingly more important. As part of our initiative to enhance the authentication of our products, we aim to enrich the characterization of our biological collections by providing authenticated data that corresponds to our physical products.

The purpose of this technical document is to outline the features of the ATCC Genome Portal and provide comprehensive descriptions of the nucleic acid extraction, sequencing, and bioinformatic methods we use to produce high-quality, reference-grade omics data for our cell products.

OUR APPROACH TO WHOLE-EXOME AND RNA SEQUENCING

At ATCC, we are setting the scientific standard in best practices for cell line characterization by offering the whole-exome (WES) and RNA-seq data with our cell lines. WES captures clinically relevant genetic variants in protein-coding regions for disease modeling and drug development, while RNA-seq provides high-quality gene expression profiles for biomarker discovery and pathway analysis. Here, we outline the steps we have taken to produce both WES and RNA-seq data from our authenticated cell lines.

1. Growth of ATCC cell lines according to the ATCC Product Sheet
2. Extraction of nucleic acids from authenticated ATCC cell lines
3. Sequencing of the nucleic acids
4. Illumina Data Quality Control – RNA Sequencing
5. Illumina Data Quality Control – Whole-Exome Sequencing
6. Release data to the ATCC Genome Portal

Each step is accompanied by rigorous quality control methods and criteria to ensure that the data proceeding to the next step is the highest quality possible. Only the data that passes all quality control criteria are published to the ATCC Genome Portal.

GROWTH AND NUCLEIC ACID EXTRACTION

Cell lines are grown in accordance with an item's Product Sheet. For specific growth information, please refer to an item's web page. For each cell line, multiple biological replicates are grown and harvested independently from separate flasks to ensure reproducibility and reduce batch effects.

GENOMIC DNA EXTRACTION AND QUALITY CONTROL (QC)

For whole-exome sequencing, genomic DNA (gDNA) is extracted with the QIAGEN EZ1 Advanced XL automated system and the EZ1&2 DNA Tissue kit. The concentration and purity of the gDNA are assessed using the Invitrogen Qubit dsDNA BR Assay kit and the ThermoFisher Nanodrop. Only samples that satisfy or surpass the criteria below will proceed to the next step.

- Concentration: ≥5 ng/µL
- Purity (A260/A280 ratio): 1.7- 2.1

TOTAL RNA EXTRACTION AND QUALITY CONTROL (QC)

Frozen cell pellets are thawed and prepared for RNA extraction and total RNA is isolated using the QIAGEN QIAcube automated system with the RNeasy Mini QIAcube Kit. RNA is quantified using Invitrogen's Qubit RNA Broad Range Assay kit, RNA purity is determined using the ThermoFisher Nanodrop and RNA integrity and quality are evaluated using the Agilent 4200 TapeStation. Only samples that meet or exceed the following specifications will be moved to the next step:

- ≥3 biological replicates
- Concentration: ≥5 ng/µL
- Purity (A260/A280 ratio): 1.8- 2.2
- RNA integrity (RIN): ≥ 6.5

ILLUMINA SEQUENCING

Illumina Whole-Exome and RNA-seq libraries are prepared using the latest and most reliable library preparation kits available. Whole-Exome Sequencing (WES) libraries are prepared using the Illumina DNA Prep with Exome 2.5 Enrichment kit (Cat # 20077596). RNA-seq libraries are prepared using the Illumina Stranded mRNA Prep kit (Cat # 20040534). Libraries are subsequently sequenced on an Illumina NextSeq2000, producing a paired-end read set per sample. The degree of sample multiplexing is based on the amount of data necessary to generate at least 40X depth of the exome (for WES) or a minimum of 18 million reads (for RNA-seq). Reads are adapter trimmed using the adapter trimming option on the Illumina instrument. Periodic updates to the instruments' software are performed when they are made available by the manufacturer to ensure that the latest version of software is used for base-calling and adapter trimming.

ILLUMINA DATA QUALITY CONTROL – RNA SEQUENCING

All RNA Sequencing (RNA-seq) read sets generated in-house use paired-end technology, which provides forward and reverse reads essential for accurate downstream analysis. To ensure good data quality, FastQC[1] was first used to evaluate the raw paired-end reads, checking metrics such as base quality scores, GC content, and adapter contamination. Occasionally, paired reads become disordered or lose their mates during processing, which can cause issues for tools that rely on properly paired reads. To address this, BBmap's repair.sh[2] was used to re-pair out-of-order reads and remove orphaned mates, restoring correct pairing for subsequent steps. This step is particularly important for software like fastp,[3] which requires properly paired reads to generate accurate statistics. After repairing, fastp was applied to further improve read quality by detecting and removing Illumina adapters and eliminating polyG sequences longer than 10 base pairs. Filtering bases by Phred scores was disabled. To check for potential contamination, Kraken2[16] was used for taxonomic classification using custom databases built internally. These preprocessing steps collectively ensure high-quality, correctly paired reads for reliable alignment and other downstream analyses.

READ MAPPING AND QUALITY CONTROL – RNA SEQUENCING

High-quality paired-end reads were aligned to the appropriate reference genome for each sample (Human: GRCh38.p14; Mouse: GRCm39) using STAR[15] in basic two-pass mode, with output BAM files sorted by coordinate. Samples were considered passing if they contained more than 18 million input reads and achieved greater than 70% uniquely mapped reads.

Aligned and sorted BAM files were processed to generate gene-level counts using the featureCounts function from the Rsubread package,[17] allowing multi-mapping reads. Gene annotations were provided using species-specific GTF files: Human (GRCh38.p14) or Mouse (GRCm39). FPKM and TPM values were calculated using gene lengths obtained from the species-specific GTF file.

Deliverables for RNA-seq of cell line samples available as a downloadable "transcriptome" dataset on the AGP include:

1. Normalized gene-level raw read counts, FPKM, and TPM values. Each replicate's sample name is used as the prefix for each count in the column header. The file has the following naming convention.
   a. <catalog_number>_<lot_number>_<reference_used>...counts.txt
   b. The variant files are available as both a tab-separated values file and excel.
2. JSON-formatted README containing all metadata.

ILLUMINA DATA QUALITY CONTROL – WHOLE-EXOME SEQUENCING

All whole-exome sequencing (WES) read sets generated in-house use paired-end technology, which provides forward and reverse reads essential for accurate downstream analysis. To ensure good data quality, FastQC1 was first used to evaluate the raw paired-end reads, checking metrics such as base quality scores, GC content, and adapter contamination. Occasionally, paired reads become disordered or lose their mates during processing, which can cause issues for tools that rely on properly paired reads. To address this, BBmap's repair.sh[2] was used to re-pair out-of-order reads and remove orphaned mates, restoring correct pairing for subsequent steps. This step is particularly important for software like fastp,[3] which requires properly paired reads to generate accurate statistics. After repairing, fastp was applied to further improve read quality by trimming bases with Phred scores below Q30, detecting and removing Illumina adapters, eliminating polyG and polyX/A sequences longer than 10 base pairs, and performing deduplication of read pairs. These preprocessing steps collectively ensure high-quality, correctly paired reads for reliable alignment and variant calling in downstream analyses.

READ MAPPING AND QUALITY CONTROL – WHOLE-EXOME SEQUENCING

The high-quality paired-end reads were mapped to the respective genome of the sample being analyzed (Human: hg38, Mouse: mm10) using bwa-mem2.[4] The reads in the aligned bam files were sorted by name, assigned read groups, and marked as potential duplicates with samtools sort,[5] Picard AddOrReplaceReadGroups,[6] and MarkDuplicates,[6] respectively. Before variant calling, aligned data for the Human samples were pre-processed using GATK,[7] which includes sorting the BAM file by read coordinates, left-aligning indels, and Base (Quality Score) Recalibration (BQSR) using the dbSNP 138 database.[8] These BAM files were then subject to rigorous QC using qualimap bamqc[9] and variant calling was performed using the GATK Best Practices Pipeline for somatic variant calling in tumor-only mode using Mutect2.[10] Briefly, variants were filtered using The Genome Aggregation Database (gnomAD)[11] and Panel of Normals (PON) from the 1000 genomes project.[12] Variants were annotated using SnpEff (hg38 Human Genome)[13] and filtered with ClinVar. We provide any variant that is annotated by ClinVar and passes our QC metrics (≥ 30X coverage and ≥ 95% allele frequency).[14] Lastly, for human samples, microsatellite instability (MSI) assessment was performed using MSIsensor2, which is designed to analyze tumor-only data. The raw WES BAM file was provided as input to MSIsensor2 following alignment and preprocessing steps. MSI status was determined based on the MSI score reported by the tool, where scores ≥ 0.20 indicate MSI-High. For this analysis, samples with scores ≤ 0.19 were classified as MSI-Stable.

Mouse samples were analyzed with the same pipeline as the Human samples with a few modifications: after alignment with bwa-mem2, sorting the BAM file by read coordinates, left-aligning indels and QC, prior to variant calling, the Base (Quality Score) Recalibration (BQSR) step was modified to accommodate the unavailability of a robust mouse SNP database to mask true variants easily. In the absence of a comprehensive mouse SNP database, BQSR was performed using a bootstrapping approach. Variants were first called from the aligned BAM using bcftools mpileup and bcftools, filtered for high-confidence sites (QUAL ≥ 30), and indexed. This provisional VCF was then provided to GATK's BaseRecalibrator as the known-sites resource to enable recalibration. This high-quality BAM file was used for variant calling using Mutect2. Lastly, variants were annotated with SnpEff against the mm10 Mouse genome.

For both Human and Mouse samples after variant calling with Mutect2, somatic variants were refined using GATK's FilterMutectCalls. This step applies probabilistic models and quality filters to distinguish true somatic mutations from sequencing artifacts. The command uses the reference genome, the raw Variant Call Format (VCF), and optional prior information (-ob-priors) to produce a filtered VCF suitable for downstream analysis.

Deliverables for WES of cell line samples available as a downloadable "exome" dataset on the AGP include:

Human:

- VCF with only ClinVar annotated variants that passed our QC metrics.
  - Variant files for each replicate will start with the following file naming:
    - <sample_name>_<lot_number>_SeqDt_<sequencing_date>
  - The variant files are available as both a tab-separated values file and excel.
- JSON-formatted README containing all metadata.

Mouse:

- VCF with variants annotated with mm10 reference. All variants passing filter are reported, as variants are not referenceable in ClinVar.
  - Variant files for each replicate will start with the following file naming:
    - <sample_name>_<lot_number>_SeqDt_<sequencing_date>
  - The variant files are available as both a tab-separated values file and excel.
- JSON-formatted README containing all metadata.

REFERENCES

1. Babraham Bioinformatics. FastQC: A quality control tool for high throughput sequence data. Babraham Institute. Published 2010. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. Bushnell B. BBTools: repair.sh script. GitHub. Published 2018. https://github.com/bbushnell/BBTools/blob/master/repair.sh
3. Chen S, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34(17): i884–i890, 2018.
4. Vasimuddin Md, et al. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. IEEE Parallel and Distributed Processing Symposium (IPDPS), 2019.
5. Danecek P, et al. Twelve years of SAMtools and BCFtools. GigaScience 10(2): giab008, 2021.
6. Broad Institute. Picard toolkit. Broad Institute. Published 2019. https://broadinstitute.github.io/picard/
7. Van der Auwera GA, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinformatics 43(1110):11, 2013.
8. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1): 308-311, 2001.
9. konechnikov K, et al. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics 32(2): 292-294, 2016.
10. Broad Institute. Mutect2. Genome Analysis Toolkit (GATK). Published 2019. https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2
11. Chen S, et al. A genomic mutational constraint map using variation in 76,156 human genomes. Nature 625(7993): 92-100, 2024.
12. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 526(7571): 68–74, 2015.
13. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) 6(2): 80-92, 2012.
14. Landrum MJ, et al. ClinVar: updates to support classifications of both germline and somatic variants. Nucleic Acids Res 53(D1): D1313-D1321, 2025.
15. Dobin A, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29(1):15-21, 2013.
16. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol 20(1): 257, 2019.
17. Yang Liao, et al. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res 47(8): e47, 2019.

10801 University Boulevard
Manassas, Virginia 20110-2209

703.365.2700 703.365.2701 sales@atcc.org www.atcc.org

CPT-122025-v01