MANUSCRIPT TITLE

ATAC-seq as a versatile tool to portray genomes and epigenomes

AUTHORS

Islem Toumi[1, 2], Laurent Lecam[1, 2] and Pierre-François Roux[1, 2, *]

[1] Institut de Recherche en Cancérologie de Montpellier (IRCM), INSERM U1194, Univ. Montpellier, Institut régional du Cancer de Montpellier (ICM), Montpellier, France.

[2] Equipe labélisée Ligue Contre le Cancer, Paris, France.

* corresponding author.

ABSTRACT

Assay for transposase-accessible chromatin using sequencing (ATAC-seq) is a cornerstone for epigenomic profiling, yet its potential for genomic characterization remains poorly explored. Here, we systematically benchmarked bulk ATAC-seq against whole-genome sequencing (WGS) to assess its capacity for detecting small variants, copy number variations (CNVs), telomeric content, and mitochondrial mutations. Using paired datasets from six melanoma cell lines, we demonstrated that ATAC-seq achieves reasonable precision in SNP detection, robustly profiles CNVs in the nuclear genome, and effectively quantifies mitochondrial heteroplasmy, with strong concordance to WGS at classical sequencing depths. Notably, we present the first systematic evaluation of telomeric length by ATAC-seq, revealing its untapped potential for studying cellular aging and genome stability. By bridging genomic and epigenomic insights into a single genome-wide approach, bulk ATAC-seq emerges as a cost-effective and versatile tool poised to transform cancer and aging research and advance clinical diagnostics through its integrative capabilities.

INTRODUCTION

The dynamic regulation of chromatin accessibility plays a pivotal role in orchestrating gene expression programs essential for cellular identity, development, and responses to environmental signals. As a result, dysregulation of these chromatin landscapes has emerged as a hallmark of cancer, where epigenetic and transcriptional changes often lead to disease onset and progression[1,2]. Advances in next-generation sequencing technologies coupled to genome tagmentation strategies led to the development of ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing), a genome-wide approach that revolutionized our ability to interrogate chromatin states at high resolution. Initially introduced as a streamlined alternative to traditional methods like DNase I hypersensitive sites using sequencing (DNase-seq) or formaldehyde-assisted isolation of regulatory elements using sequencing (FAIRE-seq), ATAC-seq has been validated by numerous laboratories as a versatile tool offering unique insights into chromatin accessibility, transcription factor activity, and enhancer landscapes[3].

Cancer is a multifaceted disease that arises from a complex interplay of genomic and epigenomic alterations, which collectively drive tumor initiation, progression, and resistance to therapy. Genomic aberrations such as mutations, copy number variations (CNVs), structural rearrangements, and chromosomal instabilities are recognized as hallmarks of cancer cells, most of the time acting synergistically. Epigenomic dysregulation—manifested through changes in chromatin accessibility, histone modifications, and DNA methylation—plays an equally critical role in shaping tumor behavior. Aberrant chromatin accessibility, for example, can lead to the activation of oncogenes and repression of tumor suppressors via mechanisms such as enhancer hijacking, super-enhancer reprogramming, and transcription factor rewiring[4–6]. In clinical oncology, genetic profiling of tumors is indispensable for guiding therapeutic decisions, particularly in the context of precision medicine. The identification of actionable mutations, such as those in the BRAF or EGFR oncogenes, enables the use of targeted therapies that improve patient outcomes[7,8]. To achieve this, a combination of sophisticated and often costly technologies is employed, including comparative genomic hybridization (CGH) for CNV detection, karyotyping for chromosomal abnormalities, whole-genome sequencing (WGS), whole-exome sequencing (WES), and fluorescence in situ hybridization (FISH) for specific genetic rearrangements[9]. While these approaches provide critical insights into the tumor genome, they fall short in capturing the epigenomic landscape, which is increasingly recognized as a key determinant of therapeutic response and relapse[10]. The epigenome governs gene expression programs and cellular identity, influencing how tumors evolve and respond to treatment. For instance, chromatin accessibility and enhancer activity are tightly linked to drug sensitivity and resistance, and epigenomic changes can serve as biomarkers for predicting patient outcomes[11,12]. Furthermore, epigenomic profiling offers insights into mechanisms of tumor relapse, as cells surviving therapy often undergo epigenetic reprogramming to adopt a drug-tolerant state. Despite its critical importance, routine clinical assessment of the epigenome is rarely performed, largely due to the lack of accessible and cost-effective technologies that can integrate both genomic and epigenomic data.

The lack of a unified approach to simultaneously profile the genome and epigenome creates a significant gap in cancer diagnostics. Developing a technology that integrates these two layers of information could revolutionize tumor profiling by reducing costs, streamlining workflows, and providing a holistic view of the molecular landscape of cancer cells. ATAC-seq, with its ability to

interrogate chromatin accessibility while capturing genomic information, represents a promising candidate for such an integrated solution. However, systematic benchmarking and validation are essential to unlock its full potential as a dual-purpose tool for genomic and epigenomic analysis. Methods filling these gaps could pave the way for novel diagnosis strategies that will improve cancer treatment and patient outcomes. ATAC-seq is far more than a tool for identifying accessible chromatin regions; its versatility enables a comprehensive exploration of the epigenome. One of its transformative features is transcription factor footprinting, which allows the identification of specific transcription factors bound to DNA by analyzing local chromatin accessibility patterns[13–16]. This provides direct insights into regulatory mechanisms controlling gene expression. Additionally, ATAC-seq can resolve nucleosome positioning, revealing the organization of chromatin at a finer structural level[17,18]. These capabilities make ATAC-seq a cornerstone in understanding the interplay between chromatin structure and gene regulation. The ongoing revolution in single-cell technologies has further expanded the potential of ATAC-seq, enabling chromatin accessibility profiling at the single-cell level. Single-cell ATAC-seq (scATAC-seq) allows researchers to dissect the heterogeneity of cell populations within complex tissues or tumor microenvironments[19]. This advancement is particularly relevant in cancer and aging research, where cell populations often display profound diversity in their epigenomic states. By capturing chromatin accessibility and transcription factor footprints in individual cells, scATAC-seq provides unparalleled resolution of regulatory landscapes[20]. The integration of these high-resolution approaches has opened new frontiers in precision oncology[21]. Single-cell ATAC-seq is already proving invaluable in uncovering novel biomarkers, identifying therapeutic targets, and developing personalized interventions[22,23]. Furthermore, by linking chromatin accessibility to transcription factor activity and chromatin architecture, ATAC-seq is redefining our understanding of genome regulation. The emergence of spatial ATAC-seq further enhances these capabilities by providing spatially resolved chromatin accessibility maps, enabling the study of regulatory landscapes in situ and uncovering spatial heterogeneity in healthy and pathogenic tissues[24].

Despite the breadth of information bulk ATAC-seq can provide, its routine applications remain largely confined to epigenetic studies, such as those identifying accessible chromatin regions and transcriptional regulatory elements[25]. However, bulk ATAC-seq data also contain rich genetic information that has been underutilized in most studies. Recent advances have demonstrated that ATAC-seq can be employed to call single nucleotide polymorphisms (SNPs) and short insertions or deletions (indels) with reasonable accuracy[26], enabling allele-specific and chromatin accessibility quantitative trait (caQTL) analyses[27]. Furthermore, the fragmented nature of ATAC-seq data makes it a promising tool for analyzing structural variations, including copy number variations (CNVs) and loss of heterozygosity (LoH), especially in tumor samples [28]. While these applications have been explored using single-cell ATAC-seq (scATAC-seq), they remain largely uninvestigated in bulk ATAC-seq datasets[6,28,29]. Similarly, ATAC-seq data provide opportunities to study mitochondrial genomes, as it captures mitochondrial DNA (mtDNA) alongside nuclear DNA. Single-cell ATAC-seq has been employed to assess mitochondrial heteroplasmy, including the proportion of mutant and wild-type mtDNA, which has significant implications in cancer and aging research[30,31]. However, these applications have not yet been validated using bulk ATAC-seq data. One particularly overlooked application is the measurement of telomere length using ATAC-seq data. Telomeres, the repetitive sequences at the ends of chromosomes, are crucial markers of cellular aging and genomic stability. Although telomere length is routinely assessed in WGS or dedicated telomere sequencing assays, it

has never been systematically explored using ATAC-seq. Theoretically, the read coverage over telomeric regions in ATAC-seq data could be used to infer telomere length, but no pipeline or methodology has been developed or benchmarked for this purpose. This represents a significant gap in the utilization of ATAC-seq for aging and cancer research. Moreover, a major limitation across all these applications is the lack of systematic and robust benchmarking for extracting genetic information from bulk ATAC-seq. Unlike whole-genome sequencing (WGS), which is specifically designed for comprehensive genomic analyses, bulk ATAC-seq data are inherently sparse and biased toward open chromatin regions. While SNP and CNV calling pipelines have been adapted for ATAC-seq data, their accuracy and reliability remain variable and are highly dependent on sequencing depth, computational methods, and the quality of input data. Furthermore, no standardized protocols exist for analyzing mtDNA variants or telomere content from bulk ATAC-seq data, and comparative benchmarks with WGS or dedicated sequencing approaches have not yet been performed.

In this study, by leveraging paired WGS and bulk ATAC-seq data, we propose a systematic benchmarking approach for detecting small variants, copy number variations (CNVs), telomere length, and mitochondrial variants using ATAC-seq. Our results demonstrate that bulk ATAC-seq is not only a powerful tool for epigenomic profiling but also a versatile platform for genomic characterization of cancer cells. This dual capacity to interrogate both the genome and the epigenome positions bulk ATAC-seq as a transformative technology, opening unprecedented opportunities in clinical settings due to its cost-effectiveness and ease of implementation. By addressing these capabilities, we aim to expand the utility of bulk ATAC-seq as a comprehensive diagnostic and research tool with significant implications for oncology, aging research, and precision medicine.

MATERIAL AND METHODS

Data Pre-processing

Matched raw ATAC-seq, H3K27ac ChIP-seq, RNA-seq, and WGS data from the A375, MM001, MM011, MM029, MM031, MM047, MM057, MM074, MM087, and MM099 patient-derived melanoma cell lines were obtained from the European Genome-Phenome Archive (EGAS00001004136) and GEO (GSE134432, GSE142238, and GSE159965). Cell lines with WGS coverage greater than 30X and ATAC-seq datasets containing more than 50 million reads were retained for further analysis (A375, MM001, MM011, MM031, MM047, and MM074). Following an initial quality assessment, WGS data were randomly down-sampled to achieve 30X genome coverage, while ATAC-seq data were trimmed to 2 × 25 bp using Picard Tools v3.0.0 and iteratively down-sampled with seqtk v1.3 to generate datasets with 10M, 20M, 30M, 40M, and 50M fragments.

ATAC-seq Processing

To investigate genetic features, ATAC-seq data were first processed using the standardized nf-core pipeline, sarek v3.4.0. Reads were trimmed, aligned to the GATK.GRCh38 reference genome, and filtered using default parameters. Small variants, including SNPs and short insertions/deletions, were called using multiple variant callers: DeepVariant, FreeBayes, Strelka, HaplotypeCaller, and mpileup.

For epigenomic analyses, paired-end reads were aligned with bowtie2 v2.5.1 (--very-sensitive-local) to the GATK.GRCh38 reference genome. Alignments falling into blacklisted regions were removed using bedtools v2.30.0. Open chromatin regions were identified with MACS v3.0.0 using default parameters. Super-enhancers were detected with ROSE v1.3.2 (-s 12500, -t 2500), and transcription factor footprints were mapped using TOBIAS v0.16.1 with the HOCOMOCO v11 database.

WGS Processing

WGS data were processed using the standardized nf-core pipeline, sarek v3.4.0. Reads were trimmed, aligned to the GATK.GRCh38 reference genome, and filtered using default parameters. Small variants, including SNPs and short insertions/deletions, were identified using DeepVariant, FreeBayes, Strelka, HaplotypeCaller, and mpileup.

ChIP-seq Processing

H3K27ac ChIP-seq data were analyzed using a standard epigenomic pipeline. Paired-end reads were aligned with bowtie2 v2.5.1 (--very-sensitive-local) to the GATK.GRCh38 reference genome. Alignments overlapping blacklisted regions were excluded using bedtools v2.30.0. H3K27ac peaks were identified with MACS v3.0.0 using default parameters. Super-enhancers were annotated with ROSE v1.3.2 (-s 12500, -t 2500).

RNA-seq Processing

RNA-seq data were processed using the standardized nf-core pipeline, rna-seq v3.11.0. Reads were trimmed and aligned to the GATK.GRCh38 reference transcriptome. For downstream analyses, the pseudocount matrix generated by Salmon was utilized.

Cancer-driving mutational signature

Mutation Annotation Format (MAF) files derived from WGS data for six melanoma cell lines (A375, MM001, MM011, MM031, MM047, and MM074) were analyzed using the *maftools* v2.20.0 R package. Mutations were filtered to prioritize biologically relevant variants based on criteria such as variant type (e.g., SNPs or others), presence of PubMed references, exclusion of the "common_variant" label, and annotations in established variation databases. This filtering reduced the dataset to variants with potential functional or clinical significance. The filtered MAF files were then merged into a single dataset for comparative analysis. Further refinement focused on variants annotated in the COSMIC (Catalogue of Somatic Mutations in Cancer) database, highlighting mutations with established oncogenic relevance. The resulting curated dataset was used to generate an oncoplot, providing a comprehensive summary of the mutational landscape across the melanoma cell lines.

ATAC-seq, H3K27ac and RNA-seq TSS-Associated Signal

To assess the consistency of signal intensities around transcription start sites (TSS) across ATAC-seq, H3K27ac, and RNA-seq samples, a ±1 kb window was defined around each TSS, and signal data from BigWig files were mapped to these regions. Pairwise Pearson correlations were computed, and the correlation matrix was visualized as a heatmap using hierarchical clustering with correlation distance and the Ward D2 method.

Integrative analysis based on Partial Least Square

A block partial least square analysis using mixOmics v6.28.0 R package was performed to integrate and explore relationships between multiple datasets, referred to as "blocks," representing different molecular features of the melanoma cell lines. Each block corresponds to a specific type of biological data. Chromatin accessibility data from ATAC-seq was used to construct two blocks, summarizing signals from super-enhancer regions and enhancer regions called with ATAC-seq data, respectively. Similarly, two blocks were created for H3K27ac ChIP-seq signals, capturing histone modification data for super-enhancers and enhancers called with H3K27ac ChIP-seq data. Transcription factor binding predictions were included as a block, with scaled log-transformed predictions to provide a functional perspective on transcriptional regulation. Another block comprised normalized gene expression data from RNA-seq. Finally, a block was created to summarize mutational profile, to integrate genomic alterations with other regulatory features. The block partial least square analysis was applied to identify latent components that explain the covariance structure across these blocks, revealing coordinated patterns between chromatin accessibility, histone modifications, transcriptional activity, and mutational profiles.

Small variants comparison

To evaluate the consistency and overlap of small variants (SNPs and indels) detected by ATAC-seq and WGS across melanoma cell lines, variant calling was performed and analyzed using multiple tools. Variant Call Format (VCF) files generated from both datasets were processed using the MutationalPatterns v3.14.0 R package. Variants from ATAC-seq were compared against those from WGS, focusing on calls within or without peak regions. Overlapping variants were visualized using Venn diagrams, highlighting common and unique SNPs identified between the datasets. Additionally, precision, recall, and F1 scores were computed using benchmarking summaries computed with hap.py v0.3.15 for both SNPs and indels to assess the accuracy and reliability of ATAC-seq in variant

detection relative to WGS. These metrics were plotted against ATAC-seq sequencing throughput for each cell line.

CNV calling and comparison

CNV calling from ATAC-seq and WGS data was performed using QDNAseq v1.40.0 in R v4.4.0, with the GRCh38 reference genome at resolutions ranging from 1 kb to 1000 kb, and considering various throughput for each ATAC-seq library (from 10M to 50M reads). Fragment counts were binned across the genome, and filters were applied to exclude noise and artifacts, including bins with high residual variance and those overlapping blacklisted regions. Technical biases, such as GC content and mappability, were estimated and corrected. The binned read counts were normalized and smoothed to reduce outliers, followed by segmentation to identify genomic regions with consistent copy number changes. These segments were further normalized using the Anscombe transform. CNVs were called based on predefined log2 thresholds for deletions, losses, gains, and amplifications, ensuring robust detection across varying resolutions. To evaluate the consistency of copy number estimates between ATAC-seq and WGS across different resolutions, throughput and cell lines, density plots were generated using log-transformed CNV data. For each cell line and resolution, the estimated copy numbers from ATAC-seq and WGS were extracted and merged and visualized using hexagonal binning, highlight the relationship between log2-transformed copy number values derived from WGS and ATAC-seq data. Pearson's correlation coefficients were computed to quantify the relationship between copy number estimates from ATAC-seq and WGS for each resolution, throughput and cell line. Sensitivity and specificity of CNV detection were evaluated to determine the performance of ATAC-seq in accurately identifying true positive and true negative CNV calls compared to WGS. Calls from ATAC-seq and WGS were processed to ensure consistent classification, and confusion matrices were computed for each resolution, throughput and cell line. Precision, sensitivity and F1 metrics were extracted and plotted against resolution, with results stratified by CNV class (gain, 2N or loss).

Telomere length computation and comparison

Telomeric repeats and associated motifs were quantified using TelomereHunter v1.1.0, producing data on telomere content and spectral composition for ATAC-seq at varying sequencing depths and for WGS across all cell lines. The spectral data from both platforms were normalized by read count to correct for differences in sequencing coverage. Counts of key telomeric motifs (e.g., TTAGGG and CCCTAA) were extracted, scaled, and merged to facilitate direct comparisons. Pairwise Pearson correlations were calculated to evaluate the consistency of telomere content detection between ATAC-seq and WGS at each sequencing depth and for each cell line. These correlations were visualized as a function of sequencing depth, highlighting the relationship between throughput and concordance across the platforms.

Mutational burden at mitochondrial genome and comparison

Mitochondrial data were processed using mgatk v0.7.0, generating allele counts, coverage statistics, and mutational data for ATAC-seq at varying sequencing depths and WGS across all cell lines. Heteroplasmy—the proportion of alternate alleles at specific positions—was calculated for detected variants to provide a detailed mutational profile. Coverage data for the mitochondrial genome were extracted using deepTools v3.5.2 (bamCoverage), producing per-base signal intensity profiles. These

profiles were normalized to account for variations in sequencing depth, and log-transformed coverage scores were visualized using polar plots to reveal distribution patterns across the mitochondrial genome for each cell line and platform. Mutational profiles were further processed in R, where filtering based on strand correlation, variance-to-mean ratio (VMR), and detection thresholds was applied to identify high-confidence variants. These variants were analyzed to compute heteroplasmy levels, which were visualized through heatmaps, highlighting commonalities and differences between ATAC-seq at various sequencing depths and WGS. Principal component analysis (PCA) of variant allele frequency (VAF) data was performed to evaluate clustering patterns and distinguish mitochondrial profiles across cell lines and sequencing methods.

RESULTS

ATAC-seq captures genetic and epigenetic information to classify melanoma cell lines

This study leveraged multi-omics datasets from 10 melanoma cell lines, including ATAC-seq, RNA-seq, H3K27ac ChIP-seq, and WGS, to systematically evaluate the potential of bulk ATAC-seq for both genomic and epigenomic profiling of populations of cancer cells (Figure 1A). Our primary objective was to benchmark ATAC-seq's ability to detect short variants, assess ploidy (CNVs), measure telomere length, and quantify mitochondrial mutational burden. To ensure robust benchmarking, we selected cell lines with a minimum of 30X WGS coverage and at least 50 million ATAC-seq reads, allowing us to measure the effects of down-sampling to 10M, 20M, 30M, 40M, and 50M reads. This filtering reduced the dataset to six cell lines: A375, MM001, MM011, MM031, MM047, and MM074 (Figure 1B). Notably, the quality of ATAC-seq data varied across the selected cell lines, with Fraction of Reads in Peaks (FRiP) scores ranging from ~20% to ~60%, and library layouts differing between paired-end 2×25 bp, 2×48 bp, and 2×25 bp reads. To ensure consistency, all libraries were trimmed to 2×25 bp. Correlation analyses focused on transcription start sites (TSS) revealed a high degree of concordance across RNA-seq, ATAC-seq, and H3K27ac data, with samples from each platform clustering together, demonstrating data quality and homogeneity (Figure 3C).

Genomic analyses revealed substantial heterogeneity among the patient-derived melanoma cell lines. WGS data identified two major groups: those with BRAF or NRAS mutations, and those without mutations in these genes (Figure 1D). Furthermore, unsupervised integrative multi-omics analysis using partial least squares (PLS) confirmed this genetic segregation. This approach highlighted that the BRAF/NRAS mutation status correlates with gene expression profiles, ATAC-seq signals at super-enhancers and classic enhancers, H3K27ac signals at these regulatory elements, and transcription factor activity inferred from digital footprinting (Figure 1E). These findings demonstrate that ATAC-seq captures not only epigenetic landscapes but also a significant amount of genetic information, transcriptional activity, and epigenetic signals traditionally assessed via H3K27ac ChIP-seq.

Comparison of SNP and Indel Detection Between ATAC-seq and WGS

To evaluate the performance of ATAC-seq in detecting small variants within and outside chromatin-accessible regions, we compared single nucleotide polymorphisms (SNPs) and short insertions/deletions (indels) identified by ATAC-seq and WGS across six melanoma cell lines (A375, MM001, MM011, MM031, MM047, and MM074). Variants detected by ATAC-seq at sequencing depths ranging from 10M to 50M reads were benchmarked against WGS data at 30X coverage. Venn diagrams (Figure 2A) illustrate the overlap of SNPs and indels detected with HaplotypeCaller within ATAC-seq peaks between the two approaches. As expected, WGS consistently identified a greater number of variants overall, while ATAC-seq detected a subset of these, particularly within accessible chromatin regions (Table 1). Sensitivity improved with increased sequencing depth, with ATAC-seq capturing the highest proportion of WGS-identified SNPs at 50M reads. On average, 42% of SNPs detected by WGS were also identified within ATAC-seq peaks, while this number dropped to 6% for indels. Quantitative metrics, including precision, recall, and F1 score, were calculated for SNPs and indels across different sequencing depths (Figure 2B). Precision remained consistently high, indicating that the majority of ATAC-seq-detected variants were true positives. Recall and F1 scores improved gradually as sequencing depth increased, reflecting enhanced sensitivity with higher read

counts. While SNP detection was robust across different sequencing depths, indel detection proved more challenging, with lower recall and F1 scores observed in all cell lines. Additional analyses were performed using DeepVariant, mpileup, and FreeBayes; however, HaplotypeCaller demonstrated the best performance in terms of both precision and recall (Figure S2 A and B, Table S1).

These results highlight that while ATAC-seq has inherent limitations compared to WGS in detecting indels, it performs effectively in identifying SNPs within chromatin-accessible regions, particularly at higher sequencing depths.

Comparative Analysis of CNV Detection Between ATAC-seq and WGS

To assess the capacity of ATAC-seq for detecting copy number variations (CNVs), we compared CNV profiles derived from ATAC-seq at sequencing depths of 10M to 50M reads with those obtained from WGS at 30X coverage using QDNAseq. At a resolution of 100 kb, hexagonal binning plots (Figure 3A) illustrate the relationship between log-transformed CNV values from WGS and ATAC-seq across the cell lines and sequencing depths. The results show increasing correlation between ATAC-seq and WGS estimates as sequencing depth rises, particularly at 40M and 50M reads. Notably, distinct clusters corresponding to diploid bins (2N), loss of heterozygosity (LoH, 1N), and amplifications (3N and higher) were evident even at lower sequencing depths. Pairwise Pearson correlations further confirmed the consistency of CNV estimates between ATAC-seq and WGS across all sequencing depths and cell lines (Figure 3B). Interestingly, correlation values plateaued at approximately 20M ATAC-seq reads, indicating that even lower-depth libraries can achieve a baseline level of reliability in CNV detection.

Quantitative metrics, including precision, sensitivity, and F1 score (Figure 3C), were calculated for CNV detection across three categories: loss, diploid (2N), and gain. ATAC-seq consistently exhibited high precision across all categories and sequencing depths. Sensitivity and F1 scores improved significantly with increased sequencing depth, with the best performance observed for diploid and gain CNVs. Losses, while still detectable, posed a slightly greater challenge in terms of accuracy. Finally, genome-wide CNV profiles for representative cell lines (A375 and MM011, Figure 3D) revealed that log2-transformed CNV signals from ATAC-seq closely mirrored major genomic alterations detected by WGS, including amplifications and deletions. These findings underscore the ability of ATAC-seq to capture genome-wide CNV patterns reliably.

Together, these results establish ATAC-seq as a robust and reliable method for CNV detection, particularly at higher sequencing depths, and highlight its potential as a complementary and cost-effective alternative to WGS in genomic studies.

Correlation of Telomere Content Between ATAC-seq and WGS Across Cell Lines

To evaluate the ability of ATAC-seq to accurately quantify telomeric content computed with TelomereHunter, we compared telomeric repeat signals derived from ATAC-seq at sequencing depths ranging from 10M to 50M reads with those obtained from WGS at 30X coverage across these six melanoma cell lines (Figure 4). Scatterplots of log-transformed telomeric content (Figure 4A) demonstrated a positive correlation between ATAC-seq and WGS datasets, with minimal impact of ATAC-seq throughput on correlation strength. Pearson correlation coefficients (Figure 4B) further highlight the consistency of telomeric content estimates between the two methods. Interestingly,

correlation values plateaued even at lower sequencing depths, suggesting that ATAC-seq provides reliable telomeric profiles regardless of throughput. This result underscores the robustness of ATAC-seq in quantifying telomeric content without requiring high sequencing depths.

Mitochondrial Genome Profiling: Comparison of ATAC-seq and WGS

It is well known that the mitochondrial genome is highly tagmented by the Tn5 tagmentase. This prompted us to evaluate whether we could use ATAC-seq data to assess meaningful information about the mitochondrial genome. Mitochondrial genome coverage and mutational heteroplasmy were analyzed using ATAC-seq at sequencing depths ranging from 10M to 50M reads and compared to WGS data at 30X coverage (Figure 5). Polar plots of log-transformed mitochondrial coverage across six melanoma cell lines (Figure 5A) demonstrated that, despite heterogeneity between cell lines, ATAC-seq achieves comparable or better coverage of the mitochondrial genome than WGS, even at lower sequencing depths. For instance, in the A375 cell line, both WGS at 30X and ATAC-seq at 10M reads achieved mitochondrial genome coverage of approximately 10 000X. Using this dataset, we identified 284 high-quality heteroplasmic variants (Figure 5B) after applying stringent filtering criteria based on strand Pearson correlation and variance-to-mean ratio (VMR). Among these variants, selecting those with the most variability (standard deviation > 0.1) in variant allele frequency (VAF) across different cell lines, platforms, and ATAC-seq depths allowed us to effectively discriminate cell lines and group data by cell of origin. This was evident in the heatmap (Figure 5C) and principal component analysis (PCA, Figure 5D), which showed that robust heteroplasmic variants can be detected even with as few as 10M reads.

DISCUSSION

Historically, bulk ATAC-seq has been employed primarily for investigating chromatin accessibility and transcriptional regulation[3]. In this study, we demonstrate that bulk ATAC-seq is a versatile and powerful tool for simultaneously characterizing genomic and epigenomic features, addressing significant gaps in current cancer research. By systematically benchmarking ATAC-seq against WGS, our analyses confirm that ATAC-seq can reliably identify small variants in chromatin accessible regions, particularly SNPs, with high precision at sequencing depths exceeding 40M reads[25,27]. Here, we also validated its ability to robustly detect copy number variations (CNVs), mitochondrial mutations, and telomeric content, expanding its utility far beyond traditional epigenomic profiling. We also demonstrate that ATAC-seq provides reliable CNV profiles that correlate strongly with WGS data even with as few as 10M fragments, emphasizing its value for analyzing structural variations. These results position bulk ATAC-seq as a cost-effective alternative for CNV detection in clinical and research contexts, particularly in resource-limited settings where WGS may be prohibitively expensive or logistically challenging. Altogether, our findings underline its potential for detailed genomic characterization of cell populations.

Our work also underscores the potential of ATAC-seq for mitochondrial genome profiling. By capturing mitochondrial DNA alongside nuclear DNA, ATAC-seq enables the detection of heteroplasmic mutations and coverage patterns comparable to those obtained with WGS with as few as 10M fragments. This is particularly significant for cancer and aging related studies, where mitochondrial dysfunction plays a critical role in disease etiology, progression and therapeutic responses. Furthermore, we provide the first systematic evaluation of ATAC-seq for quantifying telomeric content, demonstrating its feasibility as an indirect measure of telomere length with as few as 10M fragments. Telomere dynamics are key markers of cellular aging and genomic stability, and the ability to assess this parameter through ATAC-seq opens new avenues for aging research and clinical diagnostics. Recent advances in telomere bioinformatics could further enhance the precision of these analyses.

Our data indicate that the dual capacity of ATAC-seq to interrogate both the genome and epigenome simultaneously offers a streamlined approach to profile tumor heterogeneity, epigenetic reprogramming, and genetic instability, all of which are critical for understanding cancer progression and resistance to therapy. Its cost-effectiveness and ease of implementation open wide perspectives in a clinical setting to characterize large cohorts, a strategy that has so far been mainly restricted to gene expression profiling by RNA-seq. By reducing the reliance on multiple sequencing platforms, ATAC-seq has the potential to democratize access to comprehensive molecular profiling and promote personalized medicine. Additionally, the integration of epigenomic and genomic data from bulk ATAC-seq could accelerate the identification of novel biomarkers and therapeutic targets. Studies have shown that combining chromatin accessibility with genetic profiling can reveal unique regulatory interactions that drive disease states[32–35]. This capability to uncover coordinated patterns across multiple molecular layers highlights the transformative potential of ATAC-seq in precision oncology and geroscience.

Despite its versatility, ATAC-seq has limitations that warrant attention or further improvements. Indeed, our ATAC-seq framework proved to be less efficient than WGS for the detection of indels. However, advancements in computational algorithms and sequencing strategies, including those

implicating machine learning-based approaches, could mitigate this limitation in the future. Moreover, the sparsity and bias of ATAC-seq data toward open chromatin regions can hinder the comprehensive detection of certain genomic features, such as deep intronic mutations or intergenic indels. Additionally, while our benchmarks demonstrate the feasibility of ATAC-seq for telomere and mitochondrial analyses, dedicated pipelines optimized for these applications are needed to improve accuracy and reliability. Future studies should also explore the possibility to integrate bulk and single-cell ATAC-seq datasets to bridge population and cell-specific insights, further enhancing our understanding of tumor heterogeneity and aging-related processes. Moreover, emerging technologies such as multi-modal sequencing approaches combining ATAC-seq with transcriptomics or methylomics could provide deeper insights into disease mechanisms. Nevertheless, our study establishes bulk ATAC-seq as a robust and multifaceted tool for comprehensive genomic and epigenomic profiling. By validating its performance in detecting small variants, CNVs, mitochondrial mutations, and telomeric content, we highlight its potential to transform cancer and aging research.

DATA AVAILABILITY

Matched raw ATAC-seq, H3K27ac ChIP-seq, RNA-seq, and WGS data from the A375, MM001, MM011, MM029, MM031, MM047, MM057, MM074, MM087, and MM099 patient-derived melanoma cell lines were obtained from the European Genome-Phenome Archive (EGAS00001004136) and GEO (GSE134432, GSE142238, and GSE159965)

SUPPLEMENTARY DATA

The code enabling to reproduce all the analyses performed in this study is available as .zip archive.

AUTHOR CONTRIBUTIONS

P.F.R and L.L. conceived the study. I.T. and P.F.R performed the analyses. P.F.R and L.L. wrote the manuscript.

ACKNOWLEDGEMENTS

We thank Stein Aerts from VIB-KU Leuven Center for Brain and Disease Research for kindly providing access to the data.

CONFLICT OF INTEREST

The authors declare no competing interests.

REFERENCES

1. Hanahan, D. Hallmarks of Cancer: New Dimensions. *Cancer Discov* **12**, 31–46 (2022).
2. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in Cancer. *Cell* **168**, 629–643 (2017).
3. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods 2013 10:12* **10**, 1213–1218 (2013).
4. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934 (2013).
5. Lovén, J. *et al.* Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell* **153**, 320–334 (2013).
6. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, (2018).

7. Tan, L. *et al.* A Phase Ib/II Trial of Combined BRAF and EGFR Inhibition in BRAF V600E Positive Metastatic Colorectal Cancer and Other Cancers: The EVICT (Erlotinib and Vemurafenib In Combination Trial) Study. *Clin Cancer Res* **29**, 1017–1030 (2023).

8. Mehlman, C. *et al.* COMPOSIT study: evaluating osimertinib combination with targeted therapies in EGFR-mutated non-small cell lung cancer. *Oncologist* (2024) doi:10.1093/ONCOLO/OYAE312.

9. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

10. Lee, D. Y. Cancer Epigenomics and Beyond: Advancing the Precision Oncology Paradigm. *J Immunother Precis Oncol* **3**, 147 (2020).

11. Wajapeyee, N. & Gupta, R. Epigenetic alterations and mechanisms that drive resistance to targeted cancer therapies. *Cancer Res* **81**, 5589 (2021).

12. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).

13. Schultheis, H., Bentsen, M., Heger, V. & Looso, M. Uncovering uncharacterized binding of transcription factors from ATAC-seq footprinting data. *Sci Rep* **14**, 9275 (2024).

14. Li, Z. *et al.* Identification of transcription factor binding sites using ATAC-seq. *Genome Biol* **20**, 1–21 (2019).

15. Bentsen, M. *et al.* ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. *Nature Communications 2020 11:1* **11**, 1–11 (2020).

16. Ouyang, N. & Boyle, A. P. TRACE: Transcription factor footprinting using chromatin accessibility data and DNA sequence. *Genome Res* **30**, 1040–1046 (2020).

17. Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**, 1757 (2015).

18. Xu, B. *et al.* DeNOPA: decoding nucleosome positions sensitively with sparse ATAC-seq data. *Brief Bioinform* **23**, 1–17 (2022).

19. De Rop, F. V. *et al.* Systematic benchmarking of single-cell ATAC-sequencing protocols. *Nature Biotechnology 2023 42:6* **42**, 916–926 (2023).

20. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nature Methods 2023 20:9* **20**, 1355–1367 (2023).

21. Zhu, C., Preissl, S. & Ren, B. Single-cell multimodal omics: the power of many. *Nature Methods 2020 17:1* **17**, 11–14 (2020).

22. Taavitsainen, S. *et al.* Single-cell ATAC and RNA sequencing reveal pre-existing and persistent cells associated with prostate cancer relapse. *Nat Commun* **12**, (2021).

23. Van de Sande, B. *et al.* Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery 2023 22:6* **22**, 496–520 (2023).

24. Deng, Y. *et al.* Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature 2022 609:7926* **609**, 375–383 (2022).

25. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology 2020 21:1* **21**, 1–16 (2020).

26. Massarat, A. R. *et al.* Discovering single nucleotide variants and indels from bulk and single-cell ATAC-seq. *Nucleic Acids Res* **49**, 7986 (2021).

27. Atak, Z. K. *et al.* Interpretation of allele-specific chromatin accessibility using cell state–aware deep learning. *Genome Res* **31**, 1082–1096 (2021).

28. Ramakrishnan, A. *et al.* epiAneufinder identifies copy number alterations from single-cell ATAC-seq data. *Nature Communications 2023 14:1* **14**, 1–10 (2023).

29. Granja, J. M. *et al.* Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nature Biotechnology 2019 37:12* **37**, 1458–1465 (2019).

30. Lareau, C. A. *et al.* Mitochondrial single-cell ATAC-seq for high-throughput multi-omic detection of mitochondrial genotypes and chromatin accessibility. *Nat Protoc* **18**, 1416 (2023).

31. Lareau, C. A. *et al.* Massively parallel single-cell mitochondrial DNA genotyping and chromatin profiling. *Nature Biotechnology 2020 39:4* **39**, 451–461 (2020).

32. Mack, S. C. *et al.* Therapeutic Targeting of Ependymoma as Informed by Oncogenic Enhancer Profiling. *Nature* **553**, 101 (2017).

33. Liu, T. *et al.* Enhancer Coamplification and Hijacking Promote Oncogene Expression in Liposarcoma. *Cancer Res* **83**, 1517 (2023).

34. Shi, X. *et al.* Integrated profiling of human pancreatic cancer organoids reveals chromatin accessibility features associated with drug sensitivity. *Nat Commun* **13**, 2169 (2022).

35. Wang, H. *et al.* Chromatin accessibility landscape of relapsed pediatric B-lineage acute lymphoblastic leukemia. *Nat Commun* **14**, 6792 (2023).

TABLE AND FIGURES LEGENDS

Figure 1. ATAC-seq as a versatile tool for genetic and epigenetic profiling in melanoma cell lines. **(A)** Overview of the multi-omics strategy employed in this study. Data from WGS, RNA-seq, H3K27ac ChIP-seq, and ATAC-seq were analyzed for 10 melanoma cell lines to benchmark the utility of ATAC-seq in portraying both genomic and epigenomic features.

(B) Summary of the WGS coverage, ATAC-seq read counts, and fraction of reads in peaks (FRiP) for each cell line. Only cell lines with ≥30X WGS coverage and ≥50M ATAC-seq fragments (A375, MM001, MM011, MM031, MM047, and MM074) were included in subsequent analyses.

(C) Pairwise Pearson correlation heatmap of RNA-seq, ATAC-seq, and H3K27ac ChIP-seq data at transcription start sites. The bi-clustering is based on Pearson's correlation and Ward's D^2.

(D) Oncoplot showing mutational signatures across melanoma cell lines for key cancer drivers (e.g., BRAF, TP53, and NRAS). Mutations are categorized into missense, amplification, frameshift, and multi-hit alterations.

(E) Block-PLS analysis integrating mutation data, transcription factor footprinting, chromatin accessibility (enhancers and super-enhancers), histone modifications (H3K27ac at enhancers and super-enhancers), and RNA expression profiles.

Figure 2. Comparison of SNP and indel detection between ATAC-seq and WGS across melanoma cell lines.

(A) Venn diagrams illustrating the overlap of SNPs and indels detected by with HaplotypeCaller on WGS (30X coverage) and ATAC-seq at varying sequencing depths (10M, 20M, 30M, 40M, and 50M reads) for six melanoma cell lines. The outermost circle represents variants detected by WGS, while concentric circles represent variants detected by ATAC-seq at increasing sequencing depths.

(B) Quantitative evaluation of SNP and indel detection performance. The number of SNPs (log10 scale), precision, recall, and F1 score were plotted against ATAC-seq sequencing depth for each cell line. Results demonstrate improved recall and F1 scores with higher sequencing depths, particularly for SNPs. Individual cell lines are color-coded, with data shown separately for SNPs (left) and indels (right).

Figure 3. Comparative evaluation of CNV detection between ATAC-seq and WGS.

(A) Hexagonal binning plots showing the correlation of $\log_2$-transformed CNV values between WGS (30X coverage) and ATAC-seq at different sequencing depths (10M to 50M reads) for six melanoma cell lines.

(B) Pearson correlation coefficients for CNV estimates from ATAC-seq and WGS as a function of ATAC-seq sequencing depth, stratified by cell line.

(C) Precision, sensitivity, and F1 scores for CNV detection by ATAC-seq compared to WGS. Metrics are displayed separately for loss, 2N, and gain CNV classes.

(D) Genome-wide CNV profiles from WGS (top) and ATAC-seq (bottom) for representative cell lines (A375 and MM011). Log_2-transformed signals highlight consistent detection of amplifications (orange) and deletions (purple) by both methods. *cKIT* amplification in MM011 is highlighted.

Figure 4. Comparative analysis of telomeric content detection between ATAC-seq and WGS.

(A) Scatterplots of log_{10}-transformed telomeric repeat content detected by WGS (x-axis) and ATAC-seq (y-axis) across six melanoma cell lines (A375, MM001, MM011, MM031, MM047, MM074). Each column corresponds to a specific ATAC-seq sequencing depth (10M, 20M, 30M, 40M, 50M reads).

(B) Pearson correlation coefficients of telomeric content detected by ATAC-seq and WGS as a function of ATAC-seq sequencing depth, stratified by cell line.

Figure 5. Comparative analysis of mitochondrial genome profiling between ATAC-seq and WGS.

(A) Polar plots of log_{10}-transformed mitochondrial coverage across six melanoma cell lines (A375, MM001, MM011, MM031, MM047, MM099). Coverage patterns from WGS (30X) and ATAC-seq (10M–50M reads) are shown for each cell line.

(B) Scatterplot of mitochondrial variants filtered by strand Pearson correlation (x-axis) and log-transformed variance-to-mean ratio (VMR, y-axis). High-quality heteroplasmic variants are detected in the upper-right quadrant.

(C) Heatmap showing variant allele frequency (VAF) for most variable heteroplasmic mutations (standard deviation > 0.1) across cell lines, sequencing platforms, and ATAC-seq depths. The bi-clustering is based on Pearson's correlation and Ward's D^2.

(D) PCA based on VAF for most variable heteroplasmic mutations called with ATAC-seq (circles) and WGS (triangles) across cell lines.

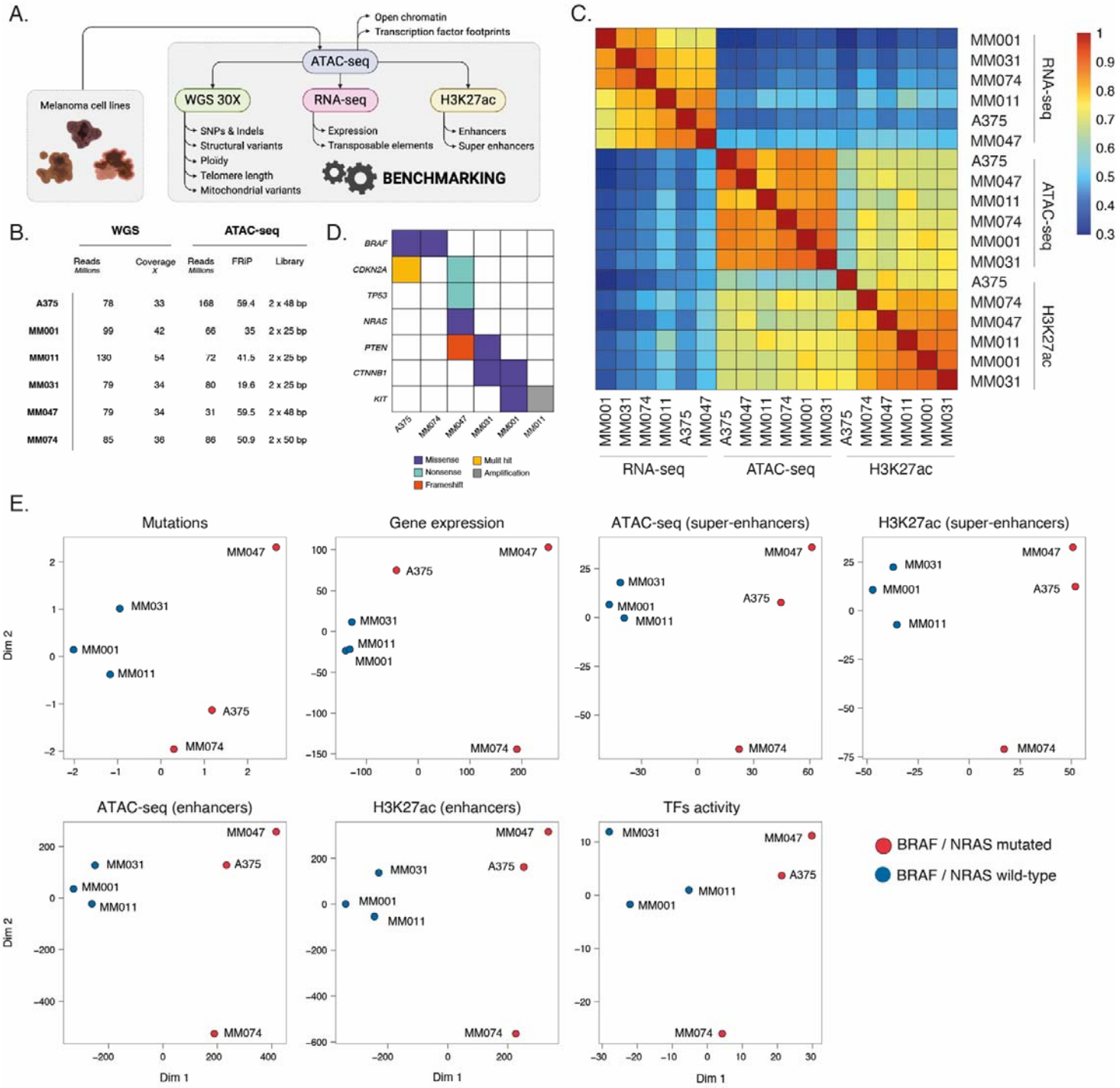
Figure S2. Comparison of SNP and indel detection between ATAC-seq and WGS across melanoma cell lines.

(A) Venn diagrams illustrating the overlap of SNPs and indels detected by with mpileup, FreeBayes and DeepVariant on WGS (30X coverage) and ATAC-seq at varying sequencing depths (10M, 20M, 30M, 40M, and 50M reads) for six melanoma cell lines. The outermost circle represents variants detected by WGS, while concentric circles represent variants detected by ATAC-seq at increasing sequencing depths.

(B) Quantitative evaluation of SNP and indel detection performance. The number of SNPs (log10 scale), precision, recall, and F1 score were plotted against ATAC-seq sequencing depth for each cell line. Results demonstrate improved recall and F1 scores with higher sequencing depths, particularly

for SNPs. Individual cell lines are color-coded, with data shown separately for SNPs (left) and indels (right).

FIGURE 1



	WGS		ATAC-seq		
	Reads Millions	Coverage X	Reads Millions	FRiP	Library
A375	78	33	168	59.4	2 x 48 bp
MM001	99	42	66	35	2 x 25 bp
MM011	130	54	72	41.5	2 x 25 bp
MM031	79	34	80	19.6	2 x 25 bp
MM047	79	34	31	59.5	2 x 48 bp
MM074	85	36	86	50.9	2 x 50 bp

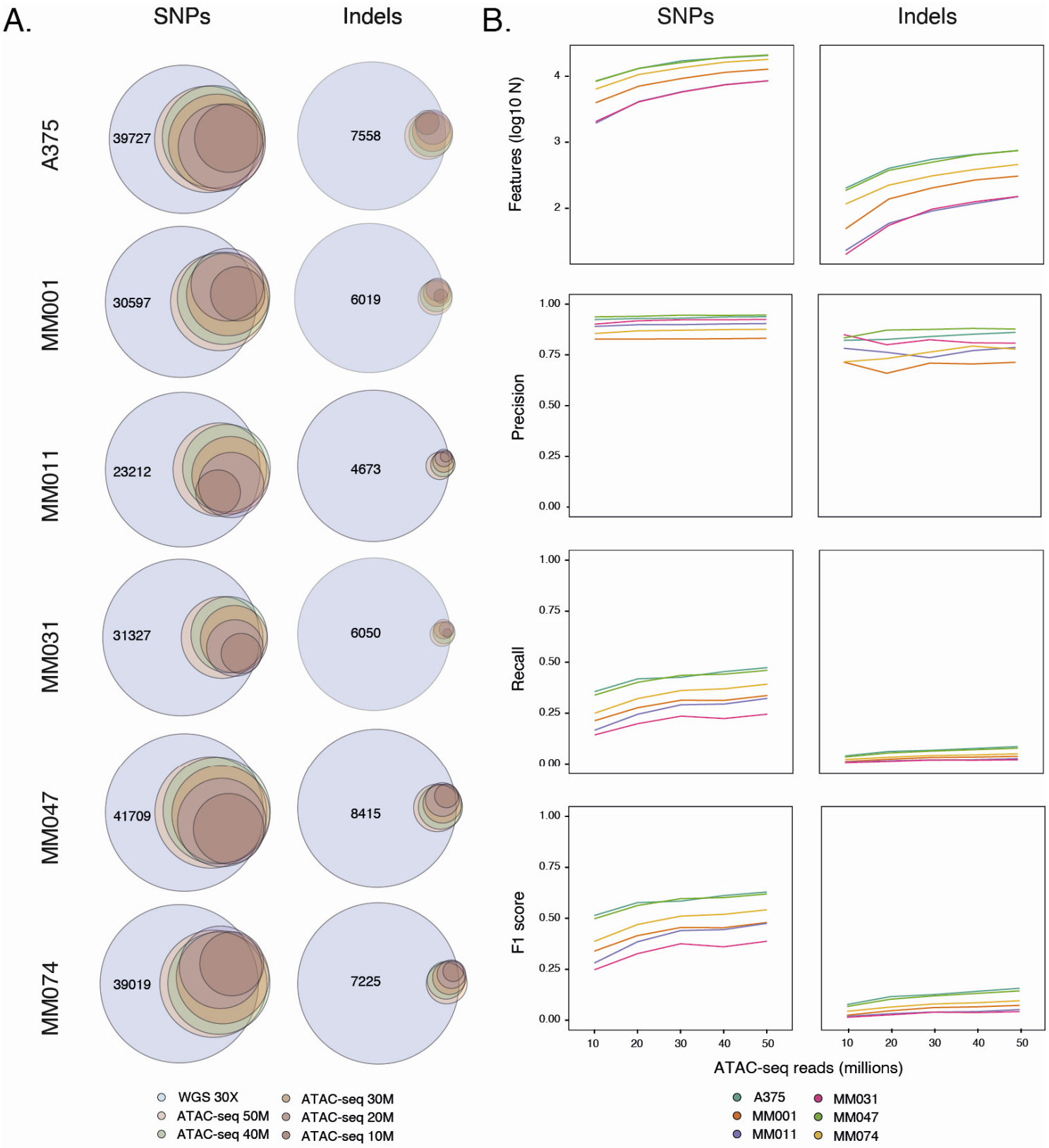

A.
SNPs
Indels
A375
MM001
MM011
MM031
MM047
MM074
39727
7558
30597
6019
23212
4673
31327
6050
41709
8415
39019
7225
WGS 30X
ATAC-seq 50M
ATAC-seq 40M
ATAC-seq 30M
ATAC-seq 20M
ATAC-seq 10M
B.
SNPs
Indels
Features (log10 N)
Precision
Recall
F1 score
ATAC-seq reads (millions)
A375
MM001
MM011
MM031
MM047
MM074

FIGURE 3

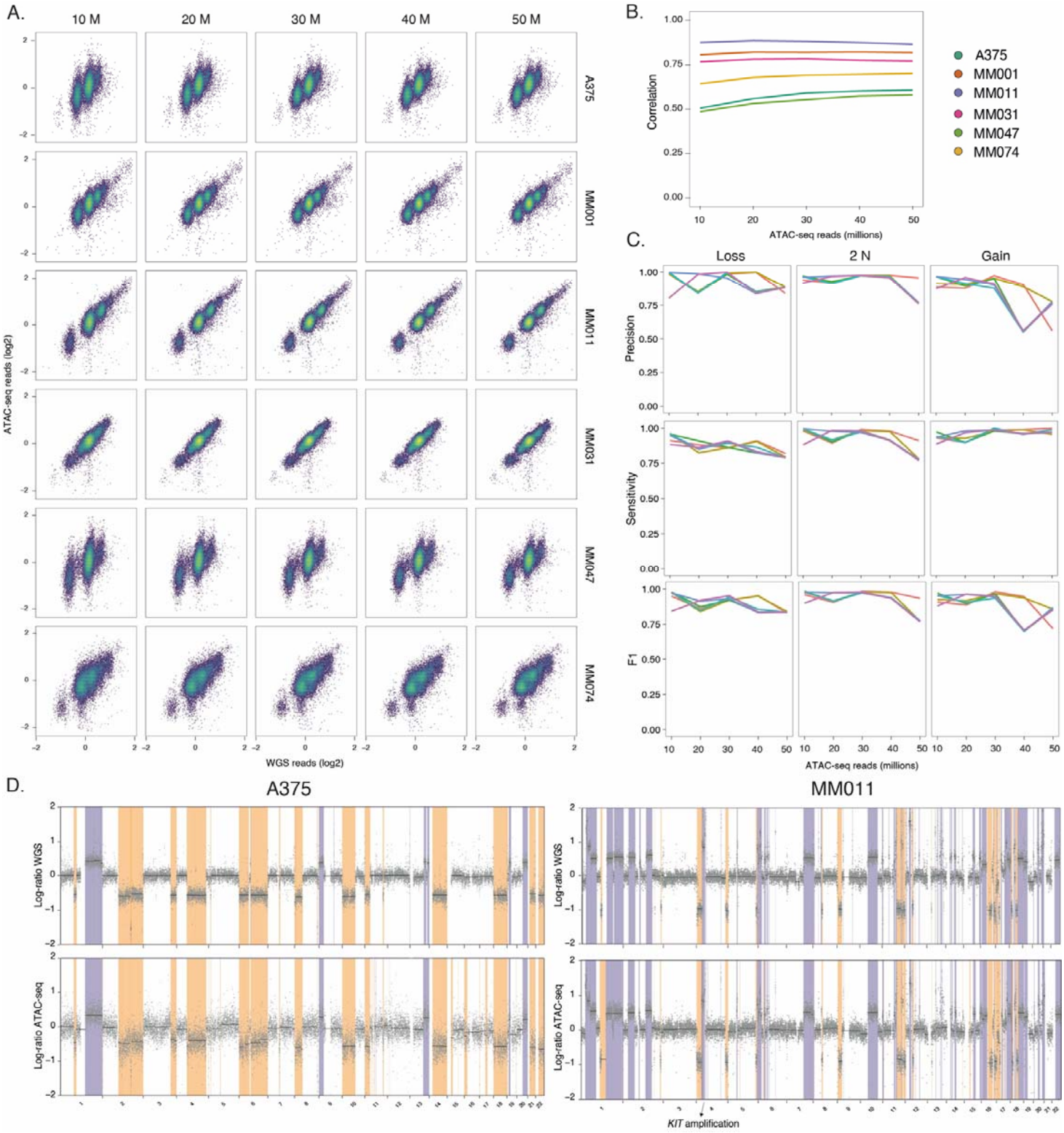

A.
10 M
20 M
30 M
40 M
50 M
A375
MM001
MM011
MM031
MM047
MM074
ATAC-seq reads (log2)
WGS reads (log2)
B.
Correlation
ATAC-seq reads (millions)
A375
MM001
MM011
MM031
MM047
MM074
C.
Loss
2 N
Gain
Precision
Sensitivity
F1
ATAC-seq reads (millions)
D.
A375
MM011
Log-ratio WGS
Log-ratio ATAC-seq
KIT amplification

FIGURE 4

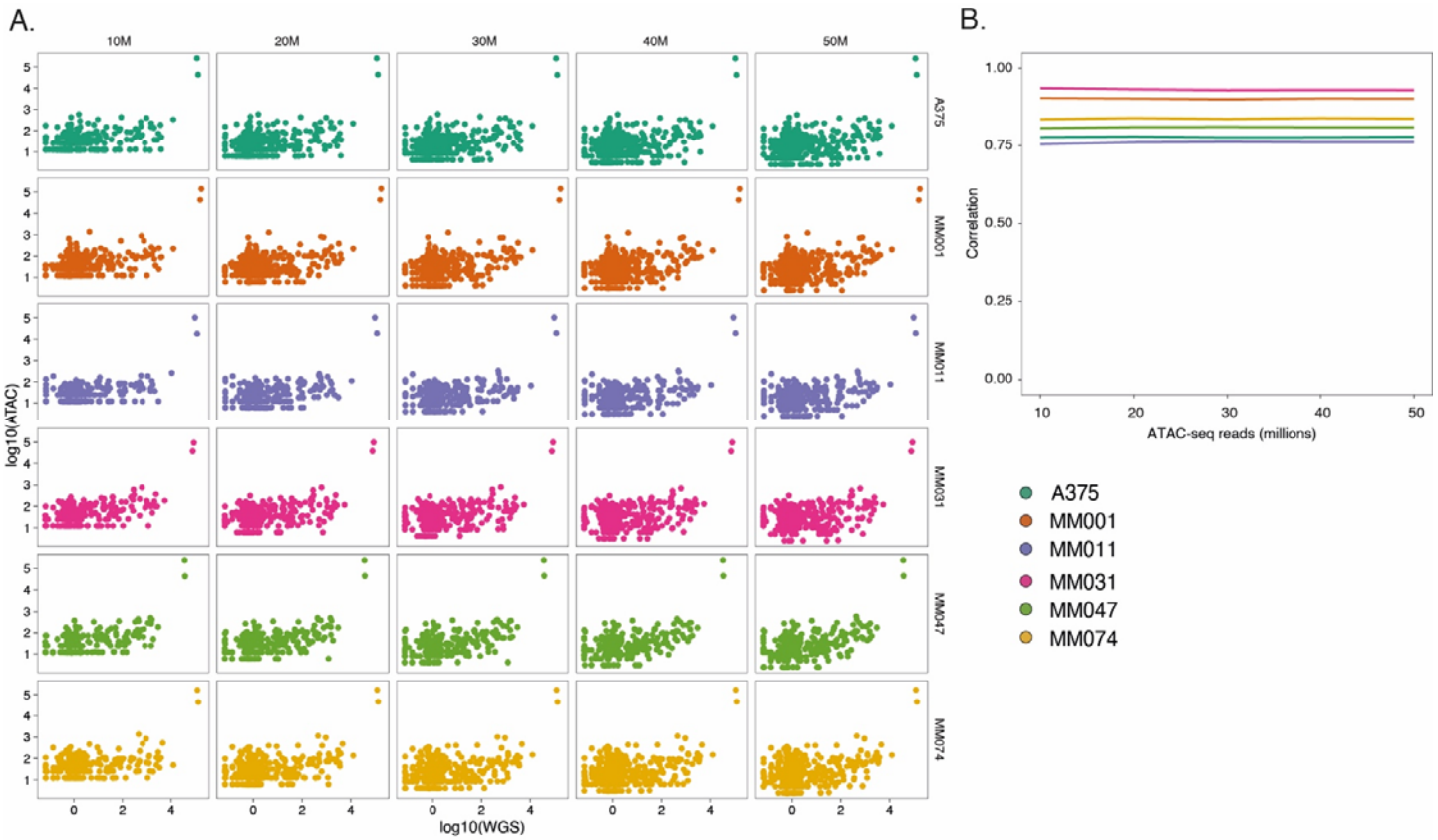
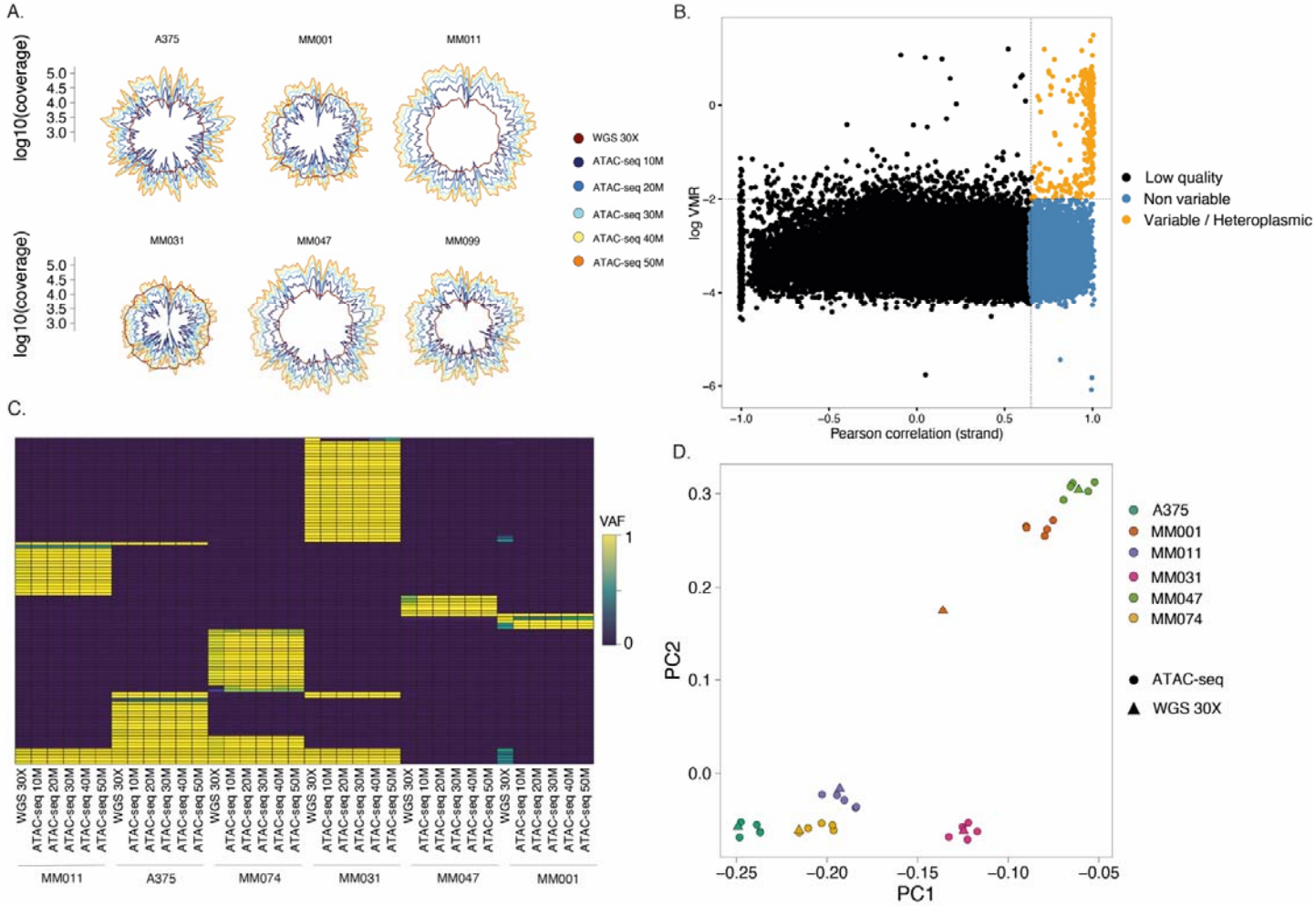

A.
10M
20M
30M
40M
50M
A375
MM001
MM011
MM031
MM047
MM074
log10(ATAC)
log10(WGS)
B.
Correlation
ATAC-seq reads (millions)
A375
MM001
MM011
MM031
MM047
MM074

FIGURE 5


A.
A375
MM001
MM011
MM031
MM047
MM099
log10(coverage)
5.0
4.5
4.0
3.5
3.0
WGS 30X
ATAC-seq 10M
ATAC-seq 20M
ATAC-seq 30M
ATAC-seq 40M
ATAC-seq 50M
B.
log VMR
Pearson correlation (strand)
Low quality
Non variable
Variable / Heteroplasmic
C.
VAF
1
0
MM011
A375
MM074
MM031
MM047
MM001
D.
PC1
PC2
A375
MM001
MM011
MM031
MM047
MM074
ATAC-seq
WGS 30X