Journal of Translational Medicine

RESEARCH

Open Access

Whole-exome evolutionary profiling of osteosarcoma uncovers metastasis-related driver mutations and generates an independently validated predictive classifier

Zhen Wang[1,2], Zhe Wang[1], Ruoyu Wang[1], Zumin Wang[2], Xiangyang Cao[3], Rui Chen[3], Zebing Ma[4], Shanshan Liang[1*] and Shuai Tao[1,2*]

Abstract

Background Osteosarcoma is the most common primary malignant bone tumor, with high invasiveness and metastatic potential and a poor prognosis in patients with metastatic cancer. Despite the rapid advancements in genomics in recent years that provided new perspectives for studying the molecular mechanisms of osteosarcoma, the understanding of its tumor heterogeneity and evolutionary mutation process remains limited.

Methods In this study, whole-exome evolutionary profiling was performed on data from the TARGET database representing 61 osteosarcoma cases. Subclonal architectures were reconstructed to characterize mutational trajectories. Differential mutation analysis was used to identify candidate metastasis-associated mutations. These features were used to build a metastasis-prediction classifier, which was cross-validated and tested on an independent external cohort. Finally, Suppes' probabilistic theory of causality was integrated with cohort data to infer high-frequency evolutionary paths linked to metastasis.

Results A linear evolutionary trajectory was observed in 62% of patients, indicating sequential clonal expansion. Eight key mutations were closely associated with metastatic progression. The classifier achieved 83% accuracy in cross-validation and maintained robust performance on the external validation set. Through causal inference, distinct evolutionary routes underpinning metastasis were uncovered, with *ATRX* mutations frequently occurring as early events that reshaped clonal dynamics and facilitated tumor spread.

Conclusions In this study, the dynamic evolutionary landscape of osteosarcoma metastasis was delineated, an early metastasis classification model was constructed, and the impact of early clonal *ATRX* mutations on metastasis initiation were highlighted. These findings offer potential avenues for the early diagnosis and risk assessment of osteosarcoma.

Keywords Osteosarcoma, Metastasis, Tumor evolution, Machine learning, Cohort analysis

*Correspondence:
Shanshan Liang
liangshanshan@dlu.edu.cn
Shuai Tao
taoshuai@dlu.edu.cn
Full list of author information is available at the end of the article

Background

Osteosarcoma is the most common primary malignant bone tumor, with a peak incidence typically occurring during adolescent growth. It ranks eighth in terms of total incidence among pediatric cancers [1]. The tumor primarily affects the ends of long bones, particularly the femur, tibia, and humerus. Despite the relatively low overall incidence of osteosarcoma, its highly aggressive nature and high metastatic potential make it one of the leading causes of death in adolescents [2, 3]. Approximately 10–20% of osteosarcoma patients have measurable metastatic lesions prior to diagnosis, with the lungs the most common metastatic site (approximately 85%). In the remaining 80–90% of patients, metastasis cannot be accurately detected using current diagnostic methods [4, 5]. The prognosis for osteosarcoma patients is almost entirely dependent on the presence of metastasis and drug resistance; lung metastasis, for example, generally has a 5-year survival rate of less than 20% [6]. Therefore, early detection and treatment of metastasis are crucial for improving the overall survival rate of osteosarcoma patients.

Clinically, osteosarcoma typically presents as localized swelling, persistent pain, and functional impairment. Osteosarcoma symptoms are often misdiagnosed because of their similarity to other benign bone conditions [7]. Although surgical resection remains the primary therapeutic approach for osteosarcoma, its effectiveness significantly decreases with metastasis or recurrence. Chemotherapy, a standard adjuvant treatment, effectively controls the growth of primary tumors. Common chemotherapy regimens include doxorubicin, cisplatin, and high-dose methotrexate; this standard treatment has remained unchanged for over 40 years [3, 8], despite its limited efficacy against metastatic disease. In recent years, targeted therapies and immunotherapies have garnered attention as emerging treatment options [9, 10]. However, their clinical use is still in the exploratory stage, and more fundamental research and clinical data are needed to support their use in osteosarcoma.

As genomics technologies have rapidly advanced, the molecular mechanisms of osteosarcoma, particularly the genomic features associated with metastasis, have gradually become a focus of osteosarcoma research. While osteosarcoma is predominantly driven by structural variants, studies have shown that somatic mutations in genes, including *TP53*, *RB1*, *CDKN2A*, *PTEN*, and *MDM2*, are closely associated with its occurrence, progression, and metastasis [11–14]. In particular, *TP53* mutations have been widely recognized as key driver mutations in osteosarcoma [15–17]. By comparing the genomic features of primary osteosarcoma tumors and metastatic lung tumors, Wang et al. [18] reported that metastatic tumors typically exhibited more complex genomic instability and a greater mutation burden, which may be due to defects in DNA damage response genes. Kinnaman et al. [19] performed longitudinal whole-genome sequencing on eight patients with recurrent or refractory osteosarcoma and revealed limited driver gene heterogeneity in temporally and spatially distinct osteosarcoma sample. Although these studies elucidated the complexity of osteosarcoma genomes, because of their extremely low tumor mutational burden (TMB) and predominance of extensive copy number alterations and structural variants [20], no major oncogenic driver mutations have been identified to date beyond *TP53*. This constitutes one of the key challenges in implementing targeted therapy for osteosarcoma.

Emerging genomic studies have begun integrating evolutionary modeling with machine learning. Wang et al. [21] performed multiregion sequencing on 10 primary osteosarcoma tumors and matched lung metastases by applying an n-dimensional Bayesian Dirichlet process to delineate subclonal populations. They observed pronounced mutation accumulation during metastasis and identified deleterious germline variants in DNA mismatch repair genes in eight patients. Kovac et al. [12] conducted phylogenetic and molecular clock analyses on 12 metastatic osteosarcoma cases to pinpoint when key driver mutations occurred, which underscored G1 cell cycle alterations as particularly promising early therapeutic targets. Liu et al. [22] used multiple machine learning algorithms to interrogate osteosarcoma gene expression datasets, revealing complex tumor microenvironment signatures from which an interpretable predictive network was constructed. Moreover, Hu et al. [23] suggested that early metastasis of breast, colorectal, and lung cancers may occur 2–4 years prior to diagnosis. This implies that in addition to identifying genomic events of the tumor, their evolutionary relationships and critical evolutionary nodes must be explored to reveal the cause of metastasis.

In this study, we comprehensively analyzed whole-exome sequencing data from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) database representing 61 osteosarcoma patients. We proposed a method to calculate tumor purity in cases with limited available data. Additionally, to comprehensively consider both point mutations and copy number variations (CNVs), we calculated the cancer cell fraction (CCF) of each mutation and reconstructed the tumor subclonal structure for each patient. Using the CCFs of the differentially mutated genes, we developed an osteosarcoma metastasis classification model with an accuracy of 83% and performed external validation. By applying Suppes'probabilistic theory of causality at the

cohort level, we revealed significant differences in high-frequency evolutionary mutation paths between metastatic and nonmetastatic patients, particularly for *TP53* and *ATRX*. This study provides new insights into the molecular mechanisms of osteosarcoma metastasis and offers novel approaches for the early diagnosis and personalized treatment of metastatic osteosarcoma.

Methods

Sample and data collection

This study is a reanalysis of existing publicly available datasets. We obtained somatic mutation data of 102 osteosarcoma patients, CNVs data of 81 patients, and clinical information of 97 patients from the TARGET database available on the Genomic Data Commons (GDC) platform (https://portal.gdc.cancer.gov/). We also acquired raw whole-exome sequencing data of primary tumor tissue and matched normal bone marrow tissue of 13 osteosarcoma patients, as published by Xu et al. [13]. All tumor samples included in this study were publicly available with free access, and informed consent was obtained from all patients by the participating centers.

Identification of somatic mutation and CNVs

Using the raw sequencing data of 13 osteosarcoma patients, we aligned the sequencing reads to the human reference genome hg38 via the Burrows–Wheeler Aligner (v.0.7.17) [24]. Following the somatic variant detection best practices guidelines from the Genome Analysis Toolkit (v.4.3.0.0) [25], we marked duplicate reads using MarkDuplicates and performed base quality score recalibration. Somatic mutations were called using Mutect2 and filtered using FilterMutectCalls. After obtaining reliable mutations, we annotated them using Funcotator. CNVKit (v.0.9.10) [26] was used to detect CNVs in each tumor sample, and to identify amplifications and deletions at the gene level. FACETS (v.0.5.14) [27] was used to determine allele-specific copy numbers and tumor purity. To focus on key mutational events, we selected nonsynonymous mutations for subsequent analysis.

We used the feature analysis module in maftools (v.2.14.0) [28] to identify de novo mutational signatures in the TARGET osteosarcoma (TARGET-OS) data for all tumor samples. We then calculated the cosine similarity to map the de novo signatures to known signatures in the Catalogue of Somatic Mutations in Cancer (COSMIC) database. Signatures with a cosine similarity greater than 0.5 were considered interpretable.

Calculation of tumor purity

We followed the method described by Locallo et al. [29] to estimate tumor purity, which used the variant allele frequency (VAF) distribution of SNVs in copy number-neutral tumor regions. Specifically, we first identified copy number-neutral regions using allele-specific copy numbers and then applied the DBSCAN algorithm to cluster the VAFs within these regions. Using DBSCAN, we determined the critical parameter eps (neighborhood radius) using the k-distance method, and the minPts parameter was set empirically to the feature dimension plus 1, which was 2. After obtaining the preliminary clustering results, we performed Wilcoxon tests between clusters and merged the results with no significant differences ($p > 0.05$). Finally, tumor purity was determined by calculating doubled the average VAF of the clone cluster (i.e., the cluster with the highest average VAF). In the TARGET-OS dataset, 29 out of 61 samples did not have mutations in copy number-neutral regions. For these samples, tumor purity was set to the median purity of the other samples.

Calculating the CCF and reconstructing the clonal tree

For each mutation, we calculated its VAF using formula (1):

$$VAF = \frac{alts}{alts + refs} \tag{1}$$

where *alts* is the number of reads supporting the variant allele at the mutation site, and *refs* is the number of reads supporting the reference allele. The VAF reflects the abundance of a mutation in a sample but does not accurately represent the proportion of cells carrying the mutation. Therefore, we calculated the CCF by combining the somatic CNV and tumor purity. Letting ρ represent tumor purity, T_{cn} represent the local copy number in cancer cells (integer), N_{cn} represent the local copy number in normal cells (integer), and m represent mutation multiplicity, the CCF of a mutation was calculated using formula (2):

$$CCF\ c = \frac{VAF}{m \cdot \rho}(T_{cn} \cdot \rho + N_{cn} \cdot (1 - \rho)) \tag{2}$$

The VAF, tumor purity, and local copy numbers can all be quantified from sequencing reads, but the mutation multiplicity m cannot be directly determined from them. We calculated the mutation multiplicity of a single SNV according to formula (3):

$$m = \max\left(1, \text{round}\left(\frac{VAF}{\rho}(T_{cn} \cdot \rho + N_{cn} \cdot (1 - \rho))\right)\right) \tag{3}$$

The function round returns the nearest integer. Using this method, we obtained the CCF for each mutation, which ranged from 0–1. If the CCF of an SNV was close to 1, the mutation was present in nearly all tumor cells and was considered a founder clone mutation. A CCF

less than 1 indicated that the mutation was present in only a subset of tumor cells and was considered a subclone mutation.

To describe the subclonal structure and infer the clonal evolutionary tree, we applied the R package RETCHER [30] to the data from each patient. It used a beta mixture model and Bayesian variational inference to reconstruct the tumor subclonal structure. We set the maximum number of clusters to 10 with at least three mutations required in each cluster. On the basis of the sum condition and the cross rules [31], all possible clonal evolutionary trees were subsequently inferred, and the subclonal proportions were calculated.

Identifying key differentially mutated genes

We divided the data of the 61 patients from the TARGET-OS dataset into two groups on the basis of whether metastasis was present at diagnosis. After selecting nonsynonymous mutations, we performed a differential analysis of the mutations between the two groups using the mafCompare function in the R package maftools, which only considers mutations present in at least two samples. Ultimately, we identified 62 differentially mutated genes. Next, we calculated the CCF of these mutations and used them as feature vectors, and used the metastatic status of each sample as the label. We then applied LASSO regression to select key differential mutations closely associated with metastasis. To ensure the reliability of the results, we applied ten-fold cross-validation and used the mean squared error as the loss function for cross-validation. Ultimately, we identified eight key differentially mutated genes with the lowest mean squared error.

Building the machine learning classification model

We used the CCFs of the eight key differential mutations as feature vectors to construct a machine learning classification model using the sklearn [32] package. First, the features were standardized to ensure balance and improve model stability. Then, we selected four classical machine learning algorithms for modeling: logistic regression (LR), decision tree (DT), support vector machine (SVM), and random forest (RF). To optimize model performance, we used random search and grid search methods to determine the best hyperparameters for each model. During the model training process, we used five-fold cross-validation to ensure the generalizability of the model. To evaluate model performance, we considered multiple metrics, including accuracy, recall, F1 score, and area under the curve (AUC). After comparing the results of the different algorithms, we selected the LR model as the optimal model. In the final training and testing process, we divided the data of 61 patients into a training and a test set at a 7:3 ratio. We used matplotlib [33] and seaborn [34] to visualize the confusion matrix, ROC curve, and feature importance.

Evolutionary pathways of cohort-based inferences of high-frequency mutations

We defined mutations that appeared in at least three samples as high-frequency mutations. We identified 16 high-frequency nonsynonymous mutations, of which 12 were shared by both the metastatic and nonmetastatic groups. Using Suppes' probabilistic theory of causality [35], we inferred the evolutionary relationships of these 12 mutations in both groups. Suppes posited that if event A occurs before event B, and if the occurrence of event A increases the probability of event B, then there exists a causal relationship between A and B. Given this, we let $P(a)$ represent the frequency of mutation a, which is the proportion of samples carrying mutation a among the total samples; similarly, $P(b)$ represents the frequency of mutation b. $P(a, b)$ represents the joint probability of mutations a and b, which is the frequency at which both mutations a and b occur simultaneously. Intuitively, it was reported that earlier genomic events tended to be those that appeared more frequently in a dataset [36]. Therefore, if $P(a) > P(b)$ and $P(b \mid a) > P(b \mid \bar{a})$, we considered there to be a directed edge between mutations a and b, denoted as $a\text{-}>b$. Moreover, following the approach of Szabo et al. [37], we added a weight to each edge using formula (4):

$$w_{a\to b} = \log\left(\frac{P(a)}{P(a)+P(b)} \cdot \frac{P(a,b)}{P(a)P(b)}\right) \tag{4}$$

After all the edges and their corresponding weights were obtained, we first removed redundant direct edges in the graph (those that could be reached through other paths) to simplify the graph structure. Specifically, if mutation x had an edge to mutation y, mutation y had an edge to mutation z, and mutation x also had a direct edge to mutation z, we removed the direct edge from mutation x to mutation z. When a mutation had multiple edges pointing to it, we selected the edge with the largest weight. The resulting structure formed a tree, which we visualized using the R package igraph (V.1.5.0) [38]. The tree represented the evolutionary relationships of mutations within a cohort.

Another key issue in cohort analysis is that a mutation may be present in multiple samples simultaneously, and therefore it is difficult to directly quantify the cellular proportion of that mutation (i.e., the CCF). To solve this problem, we calculated the weighted average CCF of the mutation using the sequencing read depth at the mutation site as the weight. Because the CCF is largely determined by the VAF, which is directly calculated from the ratio of sequencing reads, the reliability of the VAF

is lower in samples with a low sequencing depth. To account for this and evaluate the CCF of mutations at the cohort level, we assigned smaller weights to samples with lower sequencing depths, thereby reducing their influence on the weighted average CCF.

Results

Genomic characterization of TARGET-OS patient data

Initially, we obtained somatic mutation data of 102 osteosarcoma patients, CNVs data of 81 patients, and clinical data of 97 patients from the TARGET database. All the data corresponded to primary tumor sample data. A total of 61 patients had matching data across all three categories, with 19 presenting with metastases at diagnosis and 42 without metastasis (Fig. 1A). The median number of nonsynonymous mutations in the metastatic group was 19 (range, 7–60), whereas it was 18 (range, 2–77) in the nonmetastatic group (Additional file 1: Fig. S1A). The TMB was estimated to be approximately 0.45 non-silent mutations per MB (Fig. 1B). This aligned with previous findings in which pediatric osteosarcomas had a lower TMB, approximately 0.1–0.5 mutations per MB [39], which was lower than the TMB of most cancer types in the TCGA

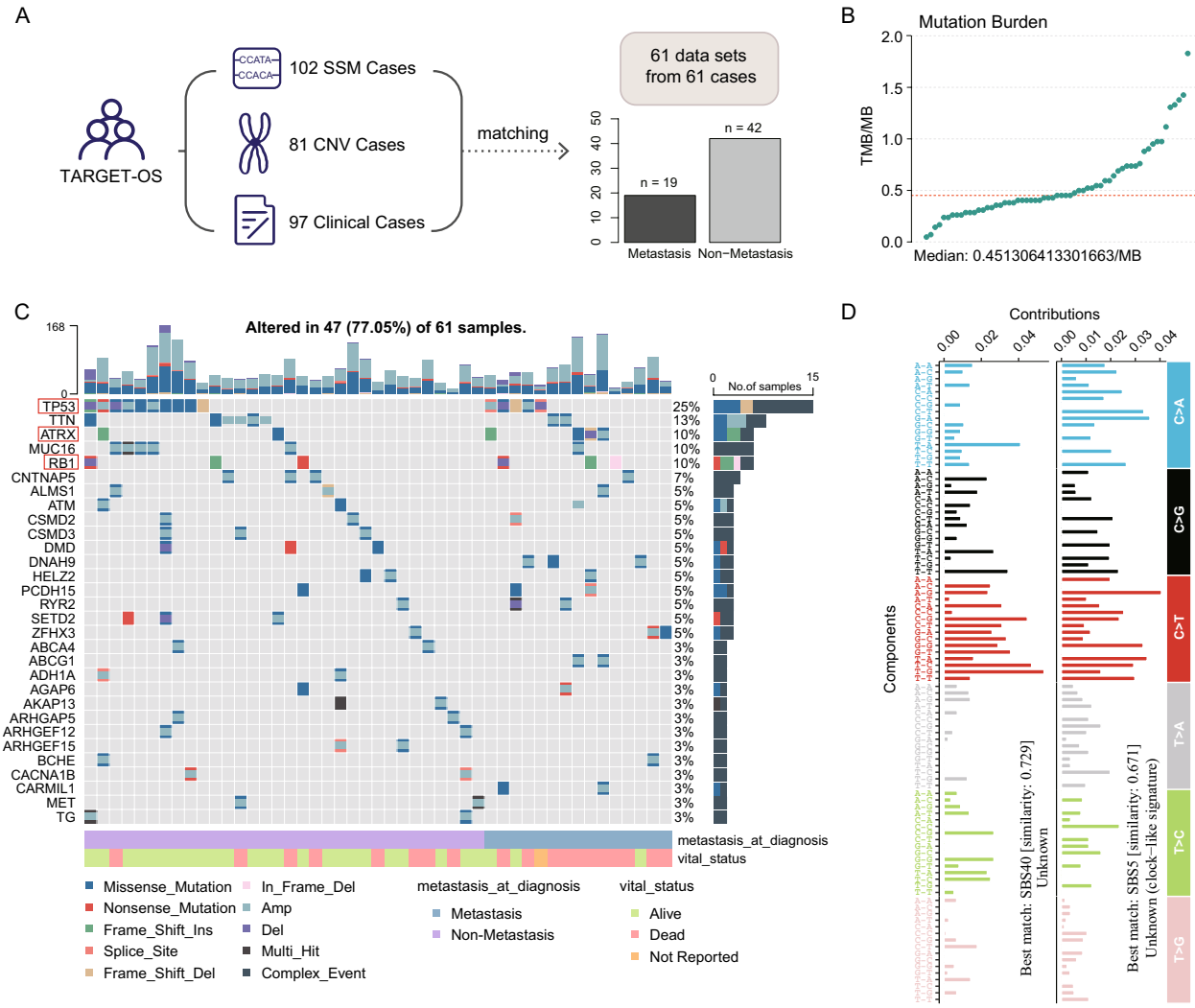

Fig. 1 Genomic landscape of the TARGET-OS data. **A** Data types and metastasis status of the samples. **B** Tumor mutation burden of all samples, represented as the number of mutations per million base pairs. **C** Overview of somatic mutations and copy number variations across all samples. The top panel displayed the total mutation count for each sample. The middle panel showed detailed information on the top 30 most frequent mutations, with large squares representing SNVs and InDels, and small rectangles representing CNVs. The bottom panel labeled the metastasis status and survival outcomes of the patients. **D** Bar chart showed the contribution of the two most significant mutation features and point mutation types

database (Additional file 1: Fig. S1B). The top 30 high-frequency gene mutation profiles of the 61 patients revealed several genes known to be commonly mutated in osteosarcoma, including *TP53*, *RB1*, and *ATRX* [13, 18] (Fig. 1C). Notably, CNV in *TP53* and *ATRX* were more prevalent in the metastatic group. Additionally, based on the reported survival status of the patients, those with metastases at diagnosis generally had a poorer prognosis.

Using a combined nonnegative matrix factorization clustering method, we identified two robust mutation signatures across all the samples (Fig. 1D). According to the COSMIC, these two mutation signatures were annotated as SBS5 and SBS40. Although both are classified as "unknown" in COSMIC (with SBS5 possibly related to a clock-like signature), a study by Venu et al. [40] on 27 pediatric cancer types revealed that SBS5 was present in 96.6% of the tumor samples and significantly correlated with signature activity related to age at diagnosis. SBS40 was active in nine pediatric cancer types, including osteosarcoma, and correlated with age at diagnosis, suggesting an additional clock-like signature.

The structure of tumor tissue is complex and comprises tumor cells, stromal cells, inflammatory cells, the vascular system, and the extracellular matrix, all of which collectively form the tumor microenvironment. The proportion of tumor cells within the tumor tissue represents the tumor purity [41]. Ignoring tumor purity can cause bias in a clonal evolution analysis of tumors. However, the TARGET-OS database does not contain information on tumor purity, and existing methods such as ABSOLUTE [42] and FACETS [27] cannot be directly applied to calculate tumor purity. Therefore, we adopted a method proposed by Locallo et al. [29] and used the VAF of somatic SNVs within copy number-neutral tumor segments to estimate tumor purity (Additional file 1: Fig. S2A, B). The results revealed a median tumor purity of 0.49 (ranging from 0.14 to 0.92) (Additional file 1: Fig. S2C). Because of the difficulty of extracting high-quality DNA from the osteoid-like stroma surrounding live tumor cells, osteosarcoma tumor samples generally tend to have low purity [19]. Of the 61 samples, 29 had no copy number-neutral mutations, and their purity was set to the median purity of the other samples. We compared this method with the FACETS method using whole-exome sequencing samples from 13 osteosarcoma patients published by Xu et al. [13]. We found significant differences in only four samples using the two methods (Additional file 1: Fig. S2D), suggesting that our method provided a viable approach for calculating tumor purity when only limited data are available.

Clonal evolution analysis reveals evolutionary patterns of OS

To integrate the features of somatic mutations and CNV, we performed a clonal evolution analysis on all the tumor samples using RETCHER [30], estimated the CCF of each mutation, and reconstructed the subclonal architectures and evolutionary pathways. Our analysis revealed that 62% (38/61) of patient tumors exhibited a simple linear evolutionary pattern, whereas 38% exhibited a branching evolutionary pattern or included only a single clone cluster (Fig. 2A). Further analysis revealed that linear evolutionary patterns were more prevalent in tumors of nonmetastatic patients, whereas tumors of metastatic patients displayed no clear trend in their evolutionary patterns.

After excluding samples with fewer than 10 mutations, we conducted a more in-depth comparison of the CCF distribution between metastatic and nonmetastatic patients. The results revealed that most mutations had high CCF values, and only a few samples had a median CCF less than 0.8 (Fig. 2B). This finding suggested that most mutations occurred at an early stage of tumor development and were widely present in the tumor cell population. Further analysis of the number of subclonal clusters revealed that 55% (12/40) of nonmetastatic patients had only one or two clusters, whereas this proportion was significantly greater in metastatic patients at 84% (16/19) (Fig. 2C). This difference may be because the tumor in nonmetastatic patients was in a relatively early stage and had not yet undergone significant evolution or selection pressure, producing less clonal differentiation and a relatively simple mutational spectrum; thus, one or two dominant subclones were present.

As an example, we further explored the differences in the clonal evolution of the high-frequency mutation *ATRX* between the nonmetastatic sample PALKDP and the metastatic sample PATMXR. The results revealed that the CCF of *ATRX* mutations in PALKDP was 0.35, indicating its presence in a subclone, whereas in PATMXR, the CCF was 0.93, placing *ATRX* mutations in a clonal cluster, indicating that it was present in almost all tumor cells (Fig. 2D). These findings suggested that *ATRX* mutations may occur early in metastatic tumors and confer a strong clonal selectivity advantage. The *ATRX* gene function is closely related to chromatin remodeling and telomere maintenance [43], and *ATRX* mutations may lead to genomic instability and increased tumor cell survival. These findings suggested that *ATRX* mutations may strongly influence the clonal expansion associated with metastasis, providing a potential genomic basis for the adaptive evolution of metastatic tumor cells.

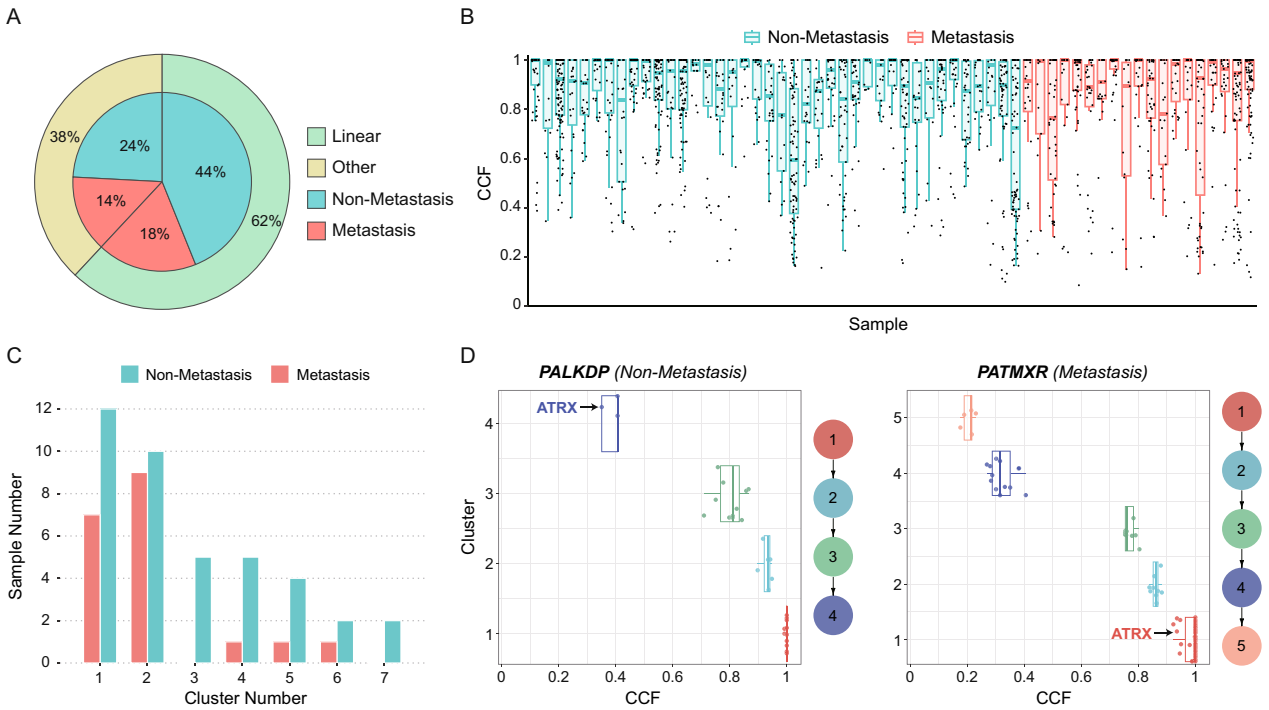

Fig. 2 Tumor clonal evolution analysis reveals the evolutionary patterns and subclonal structure of osteosarcoma. **A** Pie chart showing the proportion of linear evolutionary patterns versus other evolutionary patterns in the samples, as well as the distribution of metastatic and non-metastatic samples within these two categories. **B** Box plot displaying the distribution of the CCF across all samples, with each point representing the CCF value of a mutation, differentiated by color to distinguish metastatic from non-metastatic samples. **C** Grouped bar chart illustrating the distribution of the number of samples with different numbers of subclone clusters. **D** Schematic representation of the subclonal structure of PALKDP and PATMXR patients, with box plots showing the subclone clusters and their corresponding CCF values. The ATRX gene is specifically labeled in both patients. The evolutionary tree displays the clonal relationships between the subclone clusters

Construction of a metastasis classification model using differentially mutated genes

After obtaining the CCF of each mutation, we filtered the genes with nonsynonymous mutations. Specifically, the nonmetastatic group contained 884 genes with nonsynonymous mutations, whereas the metastatic group contained 441 genes with nonsynonymous mutations, with 45 genes overlapping between the two groups. To further compare the mutations between the two groups, we performed a differential mutation analysis on the genes in the samples, excluding the influence of individual mutations. This approach identified 62 differentially mutated genes (Additional file 1: Fig. S3). We then performed LASSO regression on these differentially mutated genes to further filter the mutations associated with metastatic outcomes. We identified eight mutated genes (*ATRX*, *RB1*, *DNAH9*, *RYR2*, *ADH1A*, *ARHGEF15*, *CARMIL1*, *ZMAT1*) with the smallest mean square errors (Fig. 3A, B). Among them, *ATRX*, *RB1*, and *RYR2* were present in both the metastatic and nonmetastatic groups, whereas the other mutations were specific to their respective groups. To investigate the functional characteristics of these mutations, we performed GO and KEGG enrichment analyses (Additional file 1: Fig. S4). Based on the GO analysis, the associated genes were significantly enriched in biological processes related to cytoskeletal organization and chromosomal behavior, suggesting potential roles in cell division and cytoskeletal dynamics. The KEGG pathway analysis further revealed significant enrichment in the cell cycle and various metabolism-related pathways, suggesting involvement in tumor proliferation and metabolic processes. These findings highlighted the potential functions of the differentially mutated genes in maintaining cellular structure, regulating metabolism, and driving tumor progression, providing important insights for future mechanistic studies.

We used the CCFs of these eight mutations as features and applied four classical machine learning algorithms (LR, DT, SVM, and RF) to model the data. The performance of each model was evaluated for accuracy, recall, F1 score, and AUC (Table 1). The results showed that the LR model performed best in terms of accuracy, recall, and F1 score, whereas the support vector machine model achieved the highest AUC. However, there was no significant improvement over the LR model. We divided the data of the 61 patients into a training set (43 cases) and

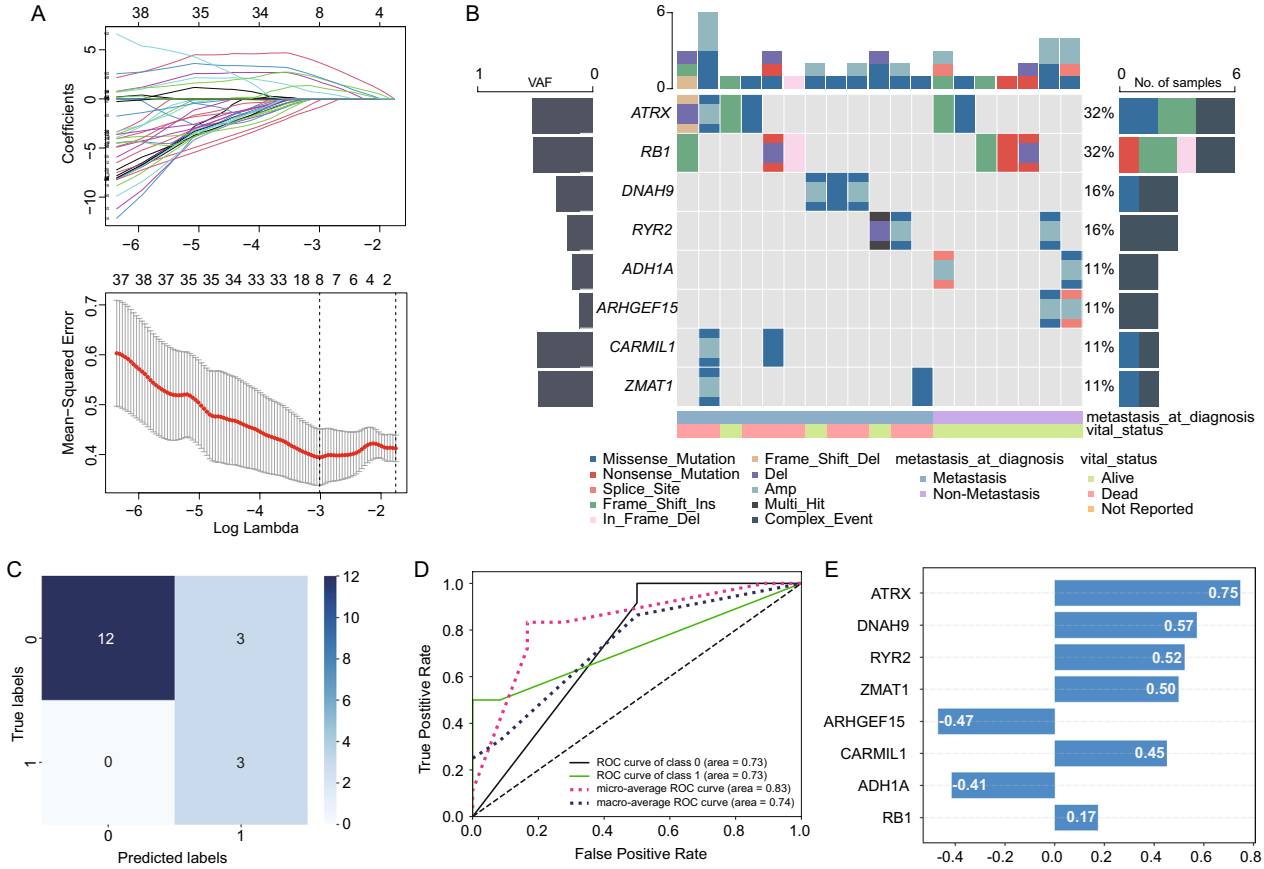

Fig. 3 Construction of the osteosarcoma metastasis classification model. **A** Variable selection process of the Lasso regression model. The top plot showed the changes in variable coefficients at different Lambda values, while the bottom plot illustrated the mean squared error obtained through cross-validation, with the vertical dashed line indicating the optimal Lambda value chosen. **B** Key mutation gene map associated with metastasis. The top part displayed the total number of mutations per sample, the left side showed the VAF of each mutation, and the bottom indicated the metastasis status and survival information of the patients. **C** The confusion matrix showed the performance of the logistic regression model on the test set, with the x-axis representing predicted labels and the y-axis representing true labels. **D** The ROC curve of the logistic regression model displayed the AUC for different categories. **E** The importance coefficients of variables in the logistic regression model. The blue bar chart indicated the magnitude and direction of the regression coefficients for each mutation

Table 1 Performance of the four machine learning models under five-fold cross-validation

Model	Performance			
	Accuracy	Recall	F1-score	ROC-AUC
LR	0.83 (±0.10)	0.58 (±0.49)	0.67 (±0.29)	0.72 (±0.22)
DT	0.72 (±0.22)	0.43 (±0.32)	0.49 (±0.58)	0.58 (±0.48)
SVM	0.80 (±0.17)	0.58 (±0.49)	0.63 (±0.32)	0.76 (±0.25)
RF	0.75 (±0.18)	0.53 (±0.34)	0.61 (±0.21)	0.65 (±0.33)

a test set (18 cases) in a 7:3 ratio and selected LR as the final model. The model achieved an accuracy of 0.83, a recall of 0.5, an F1 score of 0.67, and an AUC of 0.73 on the test set (Fig. 3C, D). We ranked the importance of the eight mutated genes on the basis of the absolute values of the coefficients in the LR equation. The mutated gene with the greatest impact on the prediction was *ATRX*, followed by *DNAH9* (Fig. 3E). *DNAH9* harbored private mutations in the metastatic group, all of which were missense mutations. Two-thirds of these mutations involved copy number amplification.

To validate the broad applicability of our model, we performed external validation using the whole-exome sequencing data from 13 osteosarcoma patients, which is publicly available from Xu et al. [13]. All patients were confirmed to have developed lung metastasis. A genomic analysis revealed that the median number of nonsynonymous mutations was 1729 (ranging from 10–5313), with considerable heterogeneity among patients. Missense mutations were the most frequent type of mutation, and none of the common high-frequency osteosarcoma mutations were observed among

the top 10 most frequent mutations (Fig. 4A). A CNV analysis revealed complex CNV patterns in samples OS2T, OS7T, OS10T, OS12T, and OS13T (Fig. 4B), which may be associated with chromosomal fragmentation and rearrangement [11]. The most common regions of copy number amplification were 8q, 17p, and 20q, whereas the most common regions of copy number deletion were 6q and 13q. A clonal evolution analysis revealed that most mutations clustered at low CCF values, which suggested that selective pressures caused by treatment (e.g., chemotherapy, immunotherapy) likely suppressed the proliferation of certain subclonal mutations that were not treatment resistant. Consistently, mutations that remained in the clone were drug-resistant mutations that survived post-treatment and continued to evolve within the cells. In the OS12T sample, a clonal cluster and four subclonal clusters were observed, with the *ATRX* mutation present in the clonal cluster (Fig. 4C). The subclonal structure of this sample allowed us to infer four distinct clonal evolution pathways (Fig. 4D), indicating the complex intra-tumoral heterogeneity of the tumor. After filtering to remove samples with a tumor purity less than 20% (OS4, OS9, OS12, OS16) and those with fewer than 10 mutations (OS15), we used the remaining eight samples to validate the osteosarcoma metastasis prediction model we constructed. The results showed that, despite differences in data processing pipelines, the model still

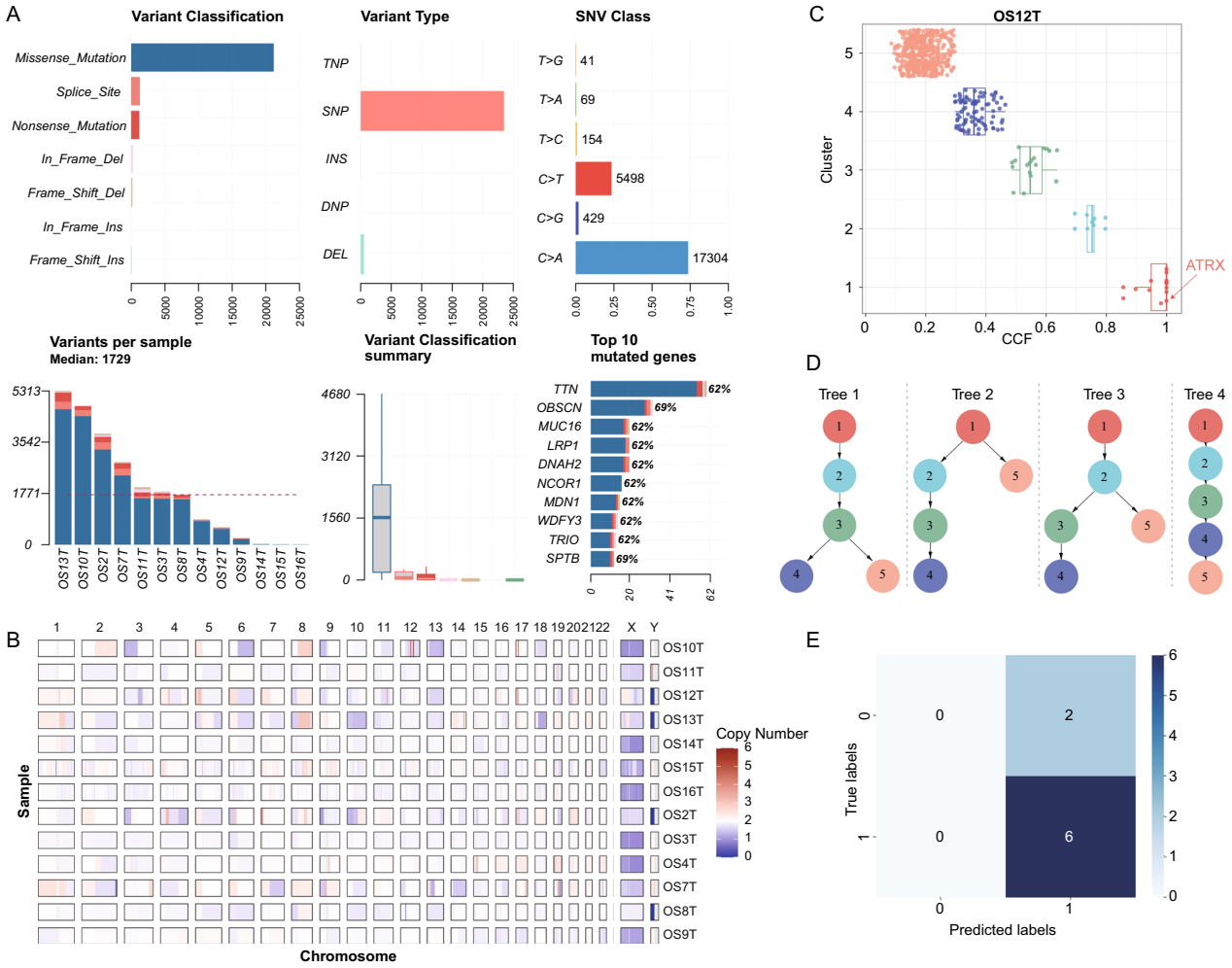

Fig. 4 External validation of the model's performance using osteosarcoma data outside of TARGET-OS. **A** Overview of mutation data. This showed the functional distribution, mutation type distribution, mutation count distribution, and the Top ten most frequent mutations across all samples. **B** CNV heatmap for the samples, which showed the CNV status across chromosome regions for each sample. Red indicated copy number amplification, and blue indicated copy number loss. **C** Subclone structure and CCF distribution for a single sample (OS12T), with the clonal location of the *ATRX* mutation labeled. **D** Clone evolutionary tree inferred from the subclone structure of OS12T, which displayed four possible evolutionary relationships. **E** Confusion matrix showed the prediction results of the osteosarcoma metastasis classification model on this dataset

achieved an accuracy of 0.75 (6/8), which strongly supports the robustness and applicability of the model.

Cohort analysis identifies high-frequency mutations that differ in evolutionary pathways

In addition to the differential mutations, we also focused on the shared mutations between the two groups, attempting to uncover key events in the evolution of metastasis. We identified 45 nonsynonymous mutated genes common to both groups (Fig. 5A). Given the strong heterogeneity observed between osteosarcoma patients, we defined high-frequency mutations as those present in at least three patients. Of the 45 shared mutations, 12 were considered high frequency mutations (Fig. 5B), and therefore may be crucial in the occurrence of metastasis.

We then inferred the evolutionary paths of these 12 mutations present in both the metastatic and nonmetastatic groups at the cohort level on the basis of Suppes'probabilistic theory of causality. In the nonmetastatic group, we identified six initiating mutations, including *TP53* and *RB1*, with *PCDH15* a descendant of *RB1*. The remaining five mutated genes, including *ATRX*, were descendants of *TP53* (Fig. 5C). These findings suggested that the evolutionary relationship driven by *TP53* mutations was most common in nonmetastatic patients. In contrast, in the metastatic group, there were five initiating mutations, including mutations in *TP53* and *ATRX*. Six mutated genes, including *RB1*, were all descendants of *ATRX*, with only one mutation, *RYR2*, a descendant of *TP53* (Fig. 5D). Considering that this was a cohort analysis and that a mutation may be present in multiple samples, we quantified the cellular proportion of mutations by calculating the weighted average CCF of these 12 mutations, using the total number of sequencing reads from each sample as the weight. The results revealed that in the nonmetastatic group, the CCF of the mutations

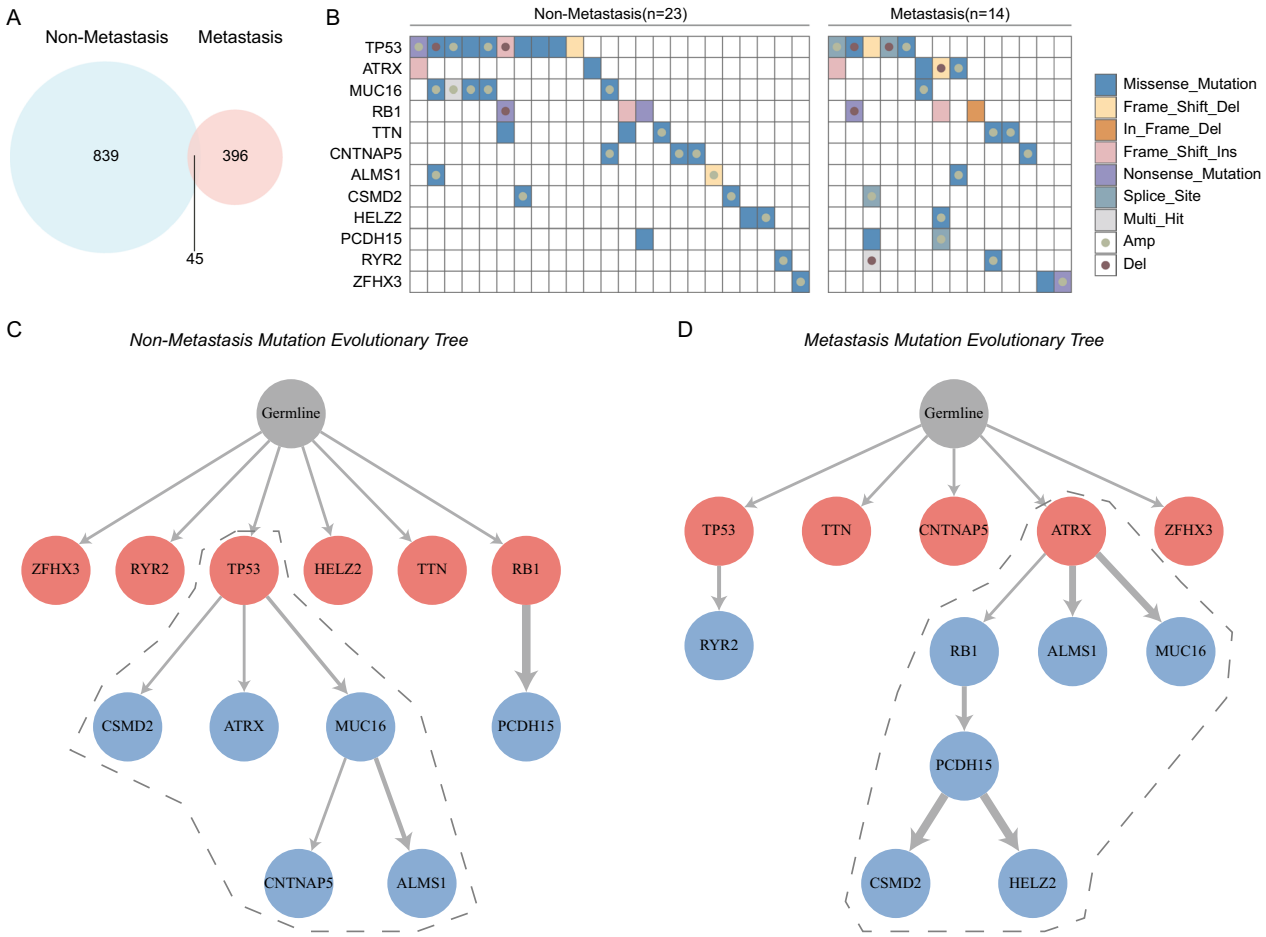

Fig. 5 Cohort analysis revealed differences in the evolutionary pathways of shared high-frequency mutations. **A** Venn diagram showed the intersection and specificity of mutations in the nonmetastatic and metastatic groups. **B** Heatmap displayed the types and distribution of shared high-frequency mutations in the nonmetastatic group (n = 23) and the metastatic group (n = 14). **C**, **D** Evolutionary trees of shared high-frequency mutations were inferred based on the probability theory of causality and cohort analysis, illustrating the hierarchical and evolutionary relationships between mutations. Panel C showed nonmetastatic samples, and Panel D showed metastatic samples

was generally greater than 0.6, with no apparent subclone structure. In contrast, in the metastatic group, two subclone mutations in *PCDH15* and *ZFHX3* were identified, with CCFs of 0.59 and 0.23, respectively (Additional file 1: Fig. S5). We observed that the CCFs of *TP53* and *RB1* were essentially equal to 1 in both groups, whereas the CCF of *ATRX* was 0.72 in the nonmetastatic group and 0.93 in the metastatic group. These findings suggested that *ATRX* mutations in the metastatic group were early initiating mutations, which was consistent with our previous conclusion that early *ATRX* mutations may be critical in metastasis occurrence.

Discussion

This study provides important insights into the mechanisms underlying osteosarcoma metastasis by systematically analyzing the genomic data of osteosarcoma patients. Through a clonal evolution analysis of tumor data from 61 osteosarcoma patients, we found that 62% of the patients exhibited a simple linear evolutionary model, suggesting that their clonal evolution process was relatively straightforward, which may have been associated with a lower tumor mutational burden. These results were consistent with the low TMB that is characteristic of osteosarcoma and aligned with the low mutational burden commonly observed in pediatric solid tumors [44]. Moreover, we observed that metastatic tumors had fewer subclonal clusters than nonmetastatic tumors, and that mutations in metastatic tumors generally had a greater CCF than those in nonmetastatic tumors, indicating that the metastatic mutations had occurred at an earlier evolutionary stage and were widespread in the tumor cells. This early clonal expansion feature was consistent with the high adaptability associated with metastasis. We also assessed the effects of demographic factors (age, sex, and race) on tumor clonal evolutionary patterns and metastatic behavior, but no significant associations were found (Additional file 1: Fig. S6). We believe this lack of association may reflect our limited sample size, the relatively homogeneous racial composition of the patients, and their narrow age range (osteosarcoma predominantly affects adolescents); together, these factors rendered it difficult to identify associations among these variables, tumor evolution, and metastatic behavior. Several studies of various cancers have shown that metastasis occurs earlier than expected [23, 45, 46], suggesting that tumor cells may acquire a significant clonal selection advantage through key mutations that occur early during tumorigenesis. In our study, the high CCF of patients with *ATRX* mutations suggested that these mutations may help drive the metastatic process.

Through differential mutation analysis and LASSO regression, we identified eight key mutated genes closely associated with osteosarcoma metastasis: *ATRX*, *RB1*, *DNAH9*, *RYR2*, *ADH1A*, *ARHGEF15*, *CARMIL1*, and *ZMAT1*. On the basis of these findings, we constructed a predictive classification model for osteosarcoma metastasis and validated its robustness using multiple datasets. The logistic regression model achieved an accuracy of 83% on the test dataset and 75% on the external validation dataset, demonstrating high predictive performance and broad applicability. Features were ranked by importance in the model, with *ATRX* identified as the most influential variable, followed by *DNAH9*. Multiple studies have demonstrated that *ATRX* is a crucial tumor-suppressor in osteosarcoma [3, 11, 18, 19, 47]. The significance of *DNAH9* in osteosarcoma has not yet been shown, although a *DNAH9* mutation was identified as a high-frequency mutation in a study of Ewing sarcoma [48]. These findings provide potential tools for assessing the risk of metastasis and developing personalized treatment strategies for osteosarcoma patients.

Using the probabilistic theory of causality and cohort analysis, we revealed significant differences in the evolutionary pathways of high-frequency mutations between metastatic and nonmetastatic osteosarcoma patients. In nonmetastatic patients, the evolutionary pathway was centered on *TP53* and *RB1*, reflecting the classical tumor suppressor gene-driven model, whereas in metastatic patients, the pathway began with *ATRX* mutations, which further supporting the crucial role of *ATRX* in the initiation of metastasis. Additionally, the subclonal features of *PCDH15* and *ZFHX3* in metastatic patient tumors suggested that these mutations may play a role in the later stages of tumor metastasis. The evolutionary pathway-based analysis revealed key early events related to metastasis and provided important insights for further research into the dynamics and mechanisms of osteosarcoma metastasis.

In recent years, there has been increasing interest in identifying biomarkers that support early disease detection or prognosis [49, 50]. *ATRX* mutations, as the most critical mutated gene in this study, were shown to be high-frequency mutations in metastatic patients and significant drivers in the evolutionary trajectory, according to the evolutionary analysis. *ATRX* is a chromatin remodeling factor, and its functional loss is closely associated with genomic instability. Studies have shown that *ATRX* mutations maintain the stability of chromosome ends through an alternative lengthening of telomere mechanism, increasing the survival and adaptability of tumor cells [51]. The CCF level of *ATRX* mutations and their hierarchical position in the evolutionary tree may serve as important indicators for predicting the metastatic risk of osteosarcoma. By integrating tumor genomic sequencing with clonal evolution analysis, we can more

accurately predict the metastatic risk in patients, and thereby support the development of personalized treatment strategies. For example, patients with high CCFs of *ATRX* mutations may be considered for more proactive follow-up plans or early preventive interventions for metastatic disease.

Despite the significant progress made in this study in revealing the genomic characteristics and evolutionary mechanisms of osteosarcoma metastasis, there are still several limitations that warrant further investigation. First, the sample size of this study was relatively small, particularly the external validation dataset, which may have partially affected the assessment of the performance of the osteosarcoma metastasis classification model. Second, this study lacked longitudinal data, and the cross-sectional nature of the TARGET-OS dataset used in the study precluded an evaluation of temporal changes in the subclonal architecture and metastatic behavior. Third, our analysis was confined to the genomic level and did not incorporate transcriptomic or proteomic data. Furthermore, the considerable heterogeneity among osteosarcoma patient tumors, coupled with the inherent genomic instability of the disease, complicates the development of a universal classification system applicable to all patients.

Future research should aim to expand the sample size and integrate multi-omics data, including transcriptomics, metabolomics, and single-cell analysis, to provide a more comprehensive understanding of the molecular mechanisms and dynamic evolutionary processes underlying osteosarcoma metastasis. Additionally, these findings can be combined with imaging techniques and biomarkers, such as circulating tumor DNA, to develop noninvasive clinical detection tools that may help capture early signals of metastatic risk, ultimately improving the prognosis of osteosarcoma patients and advancing precision medicine.

Conclusions

Our study delineated the dynamic evolutionary landscape of osteosarcoma metastasis by reconstructing subclonal architectures, identifying key mutations associated with metastasis, and establishing an early metastasis classification model. In addition, we propose that early clonal *ATRX* mutations may represent critical events that promote the initiation of metastatic progression. These findings provide valuable insights into the temporal sequence of tumor evolution and offer potential avenues for early diagnosis and risk stratification. However, further validation in larger, prospective cohorts and functional studies is needed to confirm the causal role of key mutations and refine the clinical applicability of the model.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-025-06796-6.

Additional file 1.

Acknowledgements
We gratefully acknowledge the support of our colleagues and funding agencies.

Author contributions
S.T. and Z.W. conceived the study. S.S.L., Z.W., and R.Y.W. further verified the experimental design. Z.W. conducted the study and wrote the first draft. R.C. and Z.B.M. reviewed the first draft. Z.M.W. and X.Y.C. supervised the study. All the authors have read and approved the submission of the final draft.

Funding
This project was supported by grants from the National Natural Science Foundation of China (No.82172822), the Liaoning Provincial International Science and Technology Cooperation Project (2024JH2/101900006), the Revitalizing Liaoning Talents Program-Medical Experts Project (YXMJ-JC-10), and the Liaoning Provincial Science and Technology Plan-Joint Program-Key Technology Research Project (2024JH2/102600067).

Data availability
The datasets generated and analyzed during the current study are available in the GitHub repository, https://github.com/zlsys3/data_code.

Declarations

Ethics approval and consent to participate
This study was conducted with publicly available, deidentified data from a public database. The original data collection received the necessary ethical approval and informed consent from the participants, and no additional approval was required for this secondary analysis.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]The Key Laboratory of Biomarker High-Throughput Screening and Target Translation of Breast and Gastrointestinal Tumors, Affiliated Zhongshan Hospital of Dalian University, Dalian, Liaoning, China. [2]College of Information Engineering, Dalian University, Dalian, Liaoning, China. [3]Luoyang Orthopedic Hospital of Henan Province, Orthopedic Hospital of Henan Province, Zhengzhou, Henan, China. [4]Hunan University of Chinese Medicine, Changsha, Hunan, China.

Received: 29 April 2025 Accepted: 22 June 2025
Published online: 07 July 2025

References

1. Ottaviani G, Jaffe N. The epidemiology of osteosarcoma. Cancer Treat Res. 2009;152:3–13.
2. Sheng G, Gao Y, Yang Y, et al. Osteosarcoma and metastasis. Front Oncol. 2021;11:780264.
3. Brady SW, Ma X, Bahrami A, et al. The clonal evolution of metastatic osteosarcoma as shaped by cisplatin treatment. Mol Cancer Res. 2019;17:895–906.
4. Zhao X, Wu Q, Gong X, et al. Osteosarcoma: a review of current and future therapeutic approaches. Biomed Eng Online. 2021;20:24.
5. Harris MA, Hawkins CJ. Recent and ongoing research into metastatic osteosarcoma treatments. Int J Mol Sci. 2022;23:3817.

6. PosthumaDeBoer J, Witlox MA, Kaspers GJL, et al. Molecular alterations as target for therapy in metastatic osteosarcoma: a review of literature. Clin Exp Metas. 2011;28:493–503.
7. Aran V, Devalle S, Meohas W, et al. Osteosarcoma, chondrosarcoma and Ewing sarcoma: clinical aspects, biomarker discovery and liquid biopsy. Crit Rev Oncol Hematol. 2021;162:103340.
8. Rathore R, Van Tine BA. Pathogenesis and current treatment of osteosarcoma: perspectives for future therapies. J Clin Med. 2021;10:1182.
9. Gill J, Gorlick R. Advancing therapy for osteosarcoma. Nat Rev Clin Oncol. 2021;18:609–24.
10. Aghapour SA, Torabizadeh M, Bahreiny SS, et al. Investigating the dynamic interplay between cellular immunity and tumor cells in the fight against cancer: an updated comprehensive review. Iran J Blood Cancer. 2024;16:84–101.
11. Meijer DM, Ruano D, Briaire-de Bruijn IH, et al. The variable genomic landscape during osteosarcoma progression: insights from a longitudinal WGS analysis. Genes Chromosom Cancer. 2024;63:e23253.
12. Kovac M, Ameline B, Ribi S, et al. The early evolutionary landscape of osteosarcoma provides clues for targeted treatment strategies. J Pathol. 2021;254:556–66.
13. Xu H, Zhu X, Bao H, et al. Genetic and clonal dissection of osteosarcoma progression and lung metastasis. Int J Cancer. 2018;143:1134–42.
14. Chiappetta C, Mancini M, Lessi F, et al. Whole-exome analysis in osteosarcoma to identify a personalized therapy. Oncotarget. 2017;8:80416.
15. Martin JW, Squire JA, Zielenska M. The genetics of osteosarcoma. Sarcoma. 2012;2012:627254.
16. Chen X, Bahrami A, Pappo A, et al. Recurrent somatic structural variations contribute to tumorigenesis in pediatric osteosarcoma. Cell Rep. 2014;7:104–12.
17. Synoradzki KJ, Bartnik E, Czarnecka AM, et al. TP53 in biology and treatment of osteosarcoma. Cancers (Basel). 2021;13:4284.
18. Wang D, Niu X, Wang Z, et al. Multiregion sequencing reveals the genetic heterogeneity and evolutionary history of osteosarcoma and matched pulmonary metastases. Cancer Res. 2019;79:7–20.
19. Kinnaman MD, Zaccaria S, Makohon-Moore A, et al. Subclonal somatic copy-number alterations emerge and dominate in recurrent osteosarcoma. Can Res. 2023;83:3796–812.
20. Espejo Valle-Inclan J, De Noon S, Trevers K, et al. Ongoing chromothripsis underpins osteosarcoma genome complexity and clonal evolution. Cell. 2025;188(2):352-370.e22.
21. Wang D, Wang J, Niu X et al. Clone evolution and genomic alteration analysis of osteosarcoma and matched lung metastasis. American Society of Clinical Oncology, 2017.
22. Liu G, Wang S, Liu J, et al. Using machine learning methods to study the tumour microenvironment and its biomarkers in osteosarcoma metastasis. Heliyon. 2024;10:e29322.
23. Hu Z, Li Z, Ma Z, et al. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. Nat Genet. 2020;52:701–8.
24. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, ArXiv 2013;1303.
25. van der Auwera G, O'Connor BD. Genomics in the cloud: using Docker, GATK, and WDL in Terra. Sebastopol: O'Reilly Media, Incorporated; 2020.
26. Talevich E, Shain AH, Botton T, et al. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLoS Comput Biol. 2016;12:e1004873.
27. Shen R, Seshan VE. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. Nucleic Acids Res. 2016;44:e131–e131.
28. Mayakonda A, Lin DC, Assenov Y, et al. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. 2018;28:1747–56.
29. Locallo A, Prandi D, Fedrizzi T, et al. TPES: tumor purity estimation from SNVs. Bioinformatics. 2019;35:4433–5.
30. Wang Z, Fang Y, Wang R, et al. Reconstructing tumor clonal heterogeneity and evolutionary relationships based on tumor DNA sequencing data. Briefings Bioinform. 2024;25(6): bbae516.
31. Grigoriadis K, Huebner A, Bunkum A, et al. CONIPHER: a computational framework for scalable phylogenetic reconstruction with error correction. Nat Protoc. 2024;19:159–83.
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
33. Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9:90–5.
34. Waskom ML. Seaborn: statistical data visualization. J Open Sour Softw. 2021;6:3021.
35. Suppes P. A probabilistic theory of causality. British J Philos Sci. 1973;24:409.
36. Ramazzotti D, Caravagna G, Olde Loohuis L, et al. CAPRI: efficient inference of cancer progression models from cross-sectional data. Bioinformatics. 2015;31:3016–26.
37. Szabo A, Boucher K. Estimating an oncogenetic tree when false negatives and positives are present. Math Biosci. 2002;176:219–36.
38. Csardi G, Nepusz T. The igraph software. Complex Syst. 2006;1695:1–9.
39. Reed DE, Shokat KM. Targeting osteosarcoma. Proc Natl Acad Sci U S A. 2014;111:18100–1.
40. Thatikonda V, Islam SMA, Autry RJ, et al. Comprehensive analysis of mutational signatures reveals distinct patterns and molecular processes across 27 pediatric cancers. Nature Cancer. 2023;4:276–89.
41. Li L, Xu B, Zha L. Tumor purity: a potential ignored confounder in tumor research. Cancer Res Prev Treat. 2019;46:570–4.
42. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21.
43. Aguilera P, López-Contreras AJ. ATRX, a guardian of chromatin. Trends Genet. 2023;39:505–19.
44. Gröbner SN, Worst BC, Weischenfeldt J, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555:321–7.
45. Deryugina EI, Kiosses WB. Intratumoral cancer cell intravasation can occur independent of invasion into the adjacent stroma. Cell Rep. 2017;19:601–16.
46. Podsypanina K, Du YC, Jechlinger M, et al. Seeding and propagation of untransformed mouse mammary cells in the lung. Science. 2008;321:1841–4.
47. DeWitt SB, Plumlee SH, Brighton HE, et al. Loss of ATRX promotes aggressive features of osteosarcoma with increased NF-κB signaling and integrin binding. JCI insight. 2022;7:e151583.
48. Crompton BD, Stewart C, Taylor-Weiner A, et al. The genomic landscape of pediatric Ewing sarcoma. Cancer Discov. 2014;4:1326–41.
49. Torfi E, Bahreiny SS, Saki N, et al. Evaluation of Pro-BNP biomarker in heart failure patients and its relationship with complete blood count parameters: a case–control study. Health Sci Rep. 2024;7:e70083.
50. Eftekhar Z, Aghaei M, Saki N. DNA damage repair in megakaryopoiesis: molecular and clinical aspects. Expert Rev Hematol. 2024;17:705–12.
51. Lovejoy CA, Li W, Reisenweber S, et al. Loss of ATRX, genome instability, and an altered DNA damage response are hallmarks of the alternative lengthening of telomeres pathway. PLoS Genet. 2012;8:e1002772.

Publisher's Note