Collaborative Understanding of the Range and Evolution of Osteosarcoma Single-cells
MIB AGENTS OSTEOSARCOMA ALLIANCE

Single-cell Data Aggregation and Hosting Plan
WORKING DRAFT

Plan for Centralized Visualization and Analysis of Osteosarcoma Single-Cell Data Using the Broad Single Cell Portal

Overview and Goals

To maximize translational impact from existing and emerging single-cell datasets in osteosarcoma (OS), we propose building a centralized, standardized, and interactive data repository and visualization platform using the Broad Single Cell Portal. This will serve as both a reference atlas and an analysis hub to:

1. Enable cross-study comparisons and identification of shared or divergent transcriptional cell states across samples.
2. Compute clinically actionable correlations between inferred cellular features and phenotypes.
3. Identify underrepresented clinical or molecular subtypes to guide future data generation efforts.

Phase I: Data Aggregation and Standardization

Objective

Create a curated table of existing normal bone and OS single-cell RNA-seq datasets with consistent metadata and processed features.

Dataset	GEO Accession	# Cells	# Patients	Clinical Metadata	Technology	Publication
Example 1	GSEXXXXX	15,000	5	Diagnosis, site, outcome	10x v3	Author et al. 2021
Example 2	GSEYYYYY	25,000	8	Age, treatment, relapse	SMART-seq 2	Another et al. 2022
...	...	...	...	...	...	...

Key Features for Standardization

- Deposition of raw and uniformly processed data (counts, metadata) on GEO or other open-access repositories.
- Harmonized clinical annotation: age, sex, site (primary/metastatic), treatment history, outcome.
- Data dictionary: specific definitions for each clinical attribute

---

Collaborative Understanding of the Range and Evolution of Osteosarcoma Single-cells
MIB AGENTS OSTEOSARCOMA ALLIANCE

Single-cell Data Aggregation and Hosting Plan
WORKING DRAFT

Phase II: Feature Calculation

Each dataset will be preprocessed and annotated with the following computed features, which will be made explorable per cell, per cluster, and per sample.

Per-Cell Features to Store and Visualize

- UMAP coordinates
- Cell cycle phase
- Pathway enrichment scores (e.g., hallmark gene sets, metabolic pathways)
- Transcription factor activity scores
- Top high-variance genes
- Differentiation scores (e.g., archetypes)
- Cluster membership
- Infer copy number variants

Cluster-Level Summaries

- Top 10% of differentially expressed genes per cluster with p-values, any FDR/qval
- Canonical marker gene expression for various cell types
- Pathway enrichment
- ....

Patient-Level Summaries

- Relative abundances of cell states
- Clinical Metadata, including timepoint of collection
- Experimental/Sequencing Metadata
- ....

Collaborative Understanding of the Range and Evolution of Osteosarcoma Single-cells
MIB AGENTS OSTEOSARCOMA ALLIANCE

Single-cell Data Aggregation and Hosting Plan
WORKING DRAFT

Phase III: Interactive Visualization and Analytical Tools on the Portal

We will build interactive visualizations in the Broad Single Cell Portal, supporting:

1. Per-sample UMAPs:
    - Colored by clinical metadata
    - Overlayed with computed scores (pathways, TF activity, etc.)
2. Cluster Explorer:
    - Select individual clusters to view cell composition, top genes, and pathway scores.
    - Compare selected clusters across patients or cohorts.
3. Sample-Level Summaries:
    - Archetype composition per sample
    - Relative abundance of predefined cell states
    - Clinical and genomic annotations
4. Correlation Matrix Table:
    - Compute correlations between:
        - Cell state abundances
        - Archetype compositions
        - Sample-level clinical features (e.g., response, survival, demographics, subtype, etc.)
        - Computed transcriptional features (e.g., proliferation, inflammation)

This matrix will help prioritize biological features for downstream validation and potential therapeutic targeting.

Phase IV: Gap Analysis and Future Data Generation

Using the centralized data:

- Summarize the clinical and molecular diversity covered by existing datasets.
- Highlight missing categories (e.g., underrepresented metastases, relapse samples, treatment-exposed tumors).
- Guide future sample acquisition and sequencing to maximize coverage and translational relevance.

Collaborative Understanding of the Range and Evolution of Osteosarcoma Single-cells
MIB AGENTS OSTEOSARCOMA ALLIANCE

Single-cell Data Aggregation and Hosting Plan
WORKING DRAFT

Limitations and Considerations

- Sample diversity: Current datasets may be biased toward untreated primary tumors or specific patient demographics.
- Batch effects and platform variability: Integration across different sequencing technologies (e.g., 10x vs SMART-seq2) may introduce artifacts.
- Incomplete clinical annotation: Some public datasets lack survival, treatment, or progression information necessary for clinical correlation.
- Resolution of cell state: Cell states inferred from transcriptomic data may differ across datasets and depend on integration fidelity. Report sequencing modalities as correlates.

Conclusion

This plan provides a roadmap for creating a high-value, interactive resource that democratizes access to curated osteosarcoma single-cell data, enables meaningful translational analyses, and informs future data collection. The Broad Single Cell Portal provides the necessary infrastructure to make this vision feasible and impactful.
