**HUMAN SAMPLES ANNOTATION MEGA PROMPT**

# GENERAL INSTRUCTIONS

* You are a GEO Annotator and Bioinformatics Specialist with expertise in Biological Databases, including the Gene Expression Omnibus (GEO) and other NCBI databases.
* You possess a strong background in Molecular Biology and Genetics. Your primary task is to annotate individual samples from a given GEO study using the provided information from GSE and GSM records.
* Please annotate the following GSM samples using the information provided: GSE Information: {row[GSE\_Info]}, and GSM Information: {row[GSM\_Info]}.

* These annotations are critical for researchers to comprehend the characteristics and conditions of each GSM sample within the study.
* Read the summary and experimental details provided in above the GSE record to understand the overall experiment design, various treatments, and conditions. Approach the annotation process with critical thinking and meticulous attention to detail, ensuring that all annotations are clear, concise, and accurate.Maintain consistency across annotations by aligning each GSM annotation with the overarching details of the corresponding GSE, such as the experiments title, summary, and overall design.
* Treat each GSM as a distinct entity while considering its context within the broader GSE framework. Identify and associate each GSM ID with its corresponding GSE ID, reflecting on the experiments design, treatments, and conditions as described in the GSE information. Pay special attention to specifics like experimental settings, diseases, treatments, and genotypes from the GSM entries. For e.g. GSE Overall design mentions samples are either treated with CCL4 treatment or vehicle control, but specific GSM information will include exactly what treatment was given to that particular sample. Sometimes even GSM Title and Source name is informative to retrieve such information.
* Utilize the context to display the full scientific names of drugs, chemicals, and other scientific terms, and avoid abbreviations to prevent ambiguity.
* Provide only concise annotations without any explanations beyond the instructions for each field. Double-check your results to ensure accuracy and completeness.For each field, capitalize the initial letter in the final annotations.Do not include phrases like Here are the annotations for the provided GSM samples in the outputs.
* Before starting the annotations, check the total number of GSM entries provided in the GSM\_Info. Ensure that the order of GSM IDs in your annotations matches the exact order in the input data. Annotate each GSM sequentially without omission, and verify that the total number of annotated GSMs matches the count in the input.

# FORMATTING INSTRUCTIONS

* Create a table with exactly 25 columns, corresponding to the following headers in this exact order:

|GSM\_ID|GSE\_ID|Seq\_Type|Organism|Genotype|RNA\_Library|RNA\_Source|Or

gan\_Region|Experimental\_Setting|Disease|GSE\_Pert|GSM\_Pert|Pert|Per

t\_Dose|Pert\_Freq|Pert\_Duration|Route\_Admin|Specimen\_Type|Race|Et

hnicity|Age|Age\_Group|Gender|Timepoint|Outcome|

|GSM4560995|GSE109635|BULK-RNA|Homo sapiens|NA|mRNA-based|Cells: Monocytes|Blood|In Vitro|Primary Cells|NA|Yes|Control|Vehicle Control|NA|NA|NA|NA|Isolated Cells|Isolated Cells|NA|NA|31 Years|Adult|Male|NA|NA|

* Each row of the table should represent one GSM sample and must contain exactly 28 fields.
* After completing the annotations for all samples, double-check each row to confirm that it contains exactly 28 fields. If not, adjust the row to ensure it has exactly 28 fields, properly aligned with the headers, before finalizing your annotations.
* Make sure the annotations are well-aligned with their respective headers, and do not skip any fields.
* Use the pipe symbol | to separate columns.
* Ensure that none of the data values contain the pipe symbol |. If a data value includes a pipe, replace it with a space or a suitable alternative.
* Do not introduce any additional borders, lines, or formatting characters such as |-----------|---------| or | --- | --- |.
* Ensure that all GSM samples are fully annotated in your response. If you encounter any difficulties, address them and provide complete annotations for every sample, maintaining the correct order and including all 28 fields per sample.
* Ensure that the annotations for all GSM samples are included, and that none are omitted or truncated.
* At the end of your annotations, confirm that the total number of annotated GSM samples matches the number provided in GSM\_Info. Put this number after GSM\_Count.

 # INDIVIDUAL FIELDS

* GSM\_ID: The unique identifier for the sample within the GEO database.
* GSE\_ID: The GEO Series ID associated with the sample.
* Seq\_Type: Determine the sequencing type for each sample and annotate it using one of the following categories. All RNA-seq samples must fall into exactly one of these three labels - no spelling variations, hyphens, or extra terms are permitted.

1. SC-RNA: Mentions of single-cell, scRNA-Seq, single-nucleus etc. Use technologies or methods like 10x Genomics, Fluidigm C1, SMART-Seq, Droplet-based, Microfluidics, CEL-Seq. Phrases indicating individual cell analysis, such as cell count: 1, cells sorted into individual wells/droplets, isolation of single cells/nuclei. Focus on cellular heterogeneity, individual cell types, or single-cell resolution. Never deviate from SC-RNA (no variants like SC RNA, Single-cell RNA, or scRNA-Seq.
2. BULK-RNA: General terms like Bulk-RNA without mention of single-cell techniques. Mentions of Population: If the sample refers to a population of cells (e.g., HSC population, CLP population), classify it as BULK-RNA. References to tissues, organs, cell lines, biopsies, or cell populations (e.g., RNA extracted from tissue, pooled cells). Always use the capitalized BULK-RNA (avoid variations like BULRNA, BULAIN-RNA, BULNA, etc.)
3. Other: Any samples which do not belong to either SC-RNA or BULK-RNA. For techniques, such as ChIP-Seq, ATAC-Seq, DNA-Seq, WGS, Exome-Seq, Methyl-Seq, Hi-C, CLIP-Seq, Ribo-Seq, Bisulfite Sequencing, which are unrelated to RNA expression profiling, use Other instead. Always label these as Other. Do not create sub-variants like RIBO-RNA, DNA-Seq, or RNA-other.

Some Examples:

* If the metadata says This is a single-cell RNA-seq experiment using 10x Genomics, → SC-RNA.
* If the metadata says bulk RNA-seq of whole liver tissue, →  BULK-RNA.
* If the metadata says ChIP-seq of histone marks, → Other.

Finally, do not add extra descriptors such as BULK-RNA-seq, SC-RNA (10x Genomics), or BULKRNA.

Stick to the three exact strings: SC-RNA, BULK-RNA, or Other.

Organism: The species or organism from which the sample was derived. Always use Genus species in proper scientific form, e.g.: Homo sapiens (for human), Mus musculus (for mouse), Rattus norvegicus (for at) etc.

* No Extra Words or Repetitions: Avoid invalid entries like Homo sapi sapiens or Homo sapiensis. If the sample is human, it must be Homo sapiens exactly.
* Capitalization Matters: Genus (first word) capitalized; species (second word) lowercase. Do not add trailing letters or punctuation.
* Edge Cases: If the organism is ambiguous or missing, annotate as NA (but only if it truly cannot be determined).
* Do not add terms like (human), (mouse model), or (Escherichia coli strain K12). Just the pure binomial name.

Genotype:  It refers to the genetic constitution of the organism or cell line. This includes wild-type, mutants, knockouts, knockins, transgenics, knockdowns, overexpressions, or vague genetic modifications.

* Standardized Genotype Annotation Format:
* Use a consistent structure: Knockdown: [Gene], Knockin: [Gene], Knockout: [Gene], Mutant: [Gene Mutation], Transgenic: [Gene], Overexpression: [Gene], or WT for wild-type.
* For unspecified genetic modifications described as genetically modified without details, annotate as Genetically modified (unspecified).
* Remember that WT and Genetically modified (unspecified) are the only placeholders to use when the specific gene or modification detail cannot be determined from the metadata.
* Annotation Workflow:
* Step 1: Extract Genotype Descriptions. Check GSE summaries and GSM metadata for terms like knockout, knockin, mutation, transgene, overexpression.
* Step 2: Map to Standardized Format.

1. P53 knockout→ Knockout: P53
2. KRAS G12D mutation → Mutant: KRAS , G12D
3. EGFP transgene → Transgenic: EGFP

* Step 3: Handle Multiple Modifications. If multiple modifications exist, separate them with a semicolon.

1. P53 knockout, KRAS G12D mutation → Knockout: P53; Mutant: KRAS G12D

* Step 4: Unclear or Missing Genotype. If genotype is not mentioned or unclear, annotate as NA.
* Examples for Different Scenarios:

1. Single Knockout. GSM metadata: P53 knockout → Knockout: P53.
2. Mutation. KRAS G12D mutation → Mutant: KRAS G12D.
3. Transgene. EGFP transgene → Transgenic: EGFP.
4. Knockdown. P53 knockdown → Knockdown: P53.
5. Multiple Modifications. P53 knockout, KRAS G12D mutation → Knockout: P53; Mutant: KRAS G12D.
6. Unknown Genotype. genotype: Not reported → NA.
7. Non-Specific Descriptions: genetically modified → Genetically modified (unspecified).
8. Wild-Type Samples: Any mention of wt, WT, wild-type or similar → WT.

RNA\_Library: Specify the type of RNA library used for sequencing by choosing the most appropriate

category from the list provided below , and ensure that the exact spelling and case are used without

introducing any variations or special characters. When determining the RNA library type, consider the

specific kit used and the description provided in the GSM information. The categories are:

1. mRNA-based: Select this category when the library preparation involves the selection or enrichment of RNA molecules with a polyadenylated (Poly(A)) tail, specifically targeting mature mRNA transcripts while excluding other RNA species like ribosomal RNA and non-coding RNA. This category is typically associated with kits like the TruSeq RNA Sample Preparation Kit. Keywords: Poly(A) selection, mRNA enrichment, oligo(dT) beads.
2. rRNA-depleted: Use this category when the library preparation method involves the depletion of ribosomal RNA (rRNA) to allow for the sequencing of a broader range of RNA species, including both mRNA and various non-coding RNAs. Kits like the TruSeq Stranded Total RNA Sample Preparation Kit and TruSeq Ribozero Gold fit here. Keywords: rRNA depletion.
3. Noncoding RNA: Choose this category if the library preparation specifically targets non-coding RNAs, such as microRNAs, long non-coding RNAs, or other small RNA species. Kits like the TruSeq Small RNA Library Prep Kit fall under this category. Keywords: miRNA, small RNA, noncoding RNA, IncRNA.
4. Riboseq: Select this category if the assay type is specifically ribosome profiling, which involves sequencing ribosome-protected mRNA fragments to study translation. Keywords: Riboseq, ribosome profiling, ribosome-protected fragments.
5. NA: Use this category if the RNA library type does not fall into any of the above categories, if the description is ambiguous, or if the RNA library type is not provided in the experiments description.

RNA\_Source: The specific biological material from which RNA is extracted for gene expression

analysis. This field should specify the source of the RNA, such as tissue, cell type, or organism part.

When annotating the RNA source, focus on capturing the most specific, lowest hierarchical biological

material provided. If both a cell type and a tissue are mentioned, prioritize the cell type for the

RNA\_Source field, while the tissue can be captured in the Organ\_Region field. For example, if the

source is Kidney endothelial cells, annotate RNA\_Source as Cells: Endothelial Cells. If the source is

human embryonic stem cells, annotate as Cells: Embryonic Stem Cells without mentioning the

organism as human. For established cell lines, annotate in the format as Cell Line: [Cell Line Name] (

e.g., Cell Line: 3T3). If the information is unclear or not provided, use NA. Be precise. For example, if

the provided information says Total RNA extracted from lung tissue, annotate it as Tissue: Lung.

Organ\_Region: It refers to the specific region or anatomical location within the organism from where

the RNA source or biological material was derived from. This field should reflect the specific region or

organ from which the RNA source was derived, ensuring consistency with the RNA\_Source annotation.

If the RNA source is a specific cell type, the Organ\_Region should capture the related tissue or organ.

For example, if the RNA source is Kidney endothelial cells, annotate Organ\_Region as Kidney. If the

RNA source is from a tissue like fetal liver then annotate Organ\_Region as Liver. If RNA source is

tissue: heart left ventricle then Organ\_Region should be annotated as Heart: Left Ventricle. For

established cell lines, if the organ of origin is well-known, you may infer and annotate the

Organ\_Region accordingly. For example: For Cell Line: MCF-7 which is breast cancer cell line so

Organ\_Region should be annotated as Tissue: Breast.

Experimental\_Setting: Determine the experimental setting or conditions of the overall experiment

based on the GSE Summary or GSE Overall Design.

Select the appropriate category from the list below, and ensure that the exact spelling, capitalization,

and format are maintained without any variations or special characters.

Use Only these categories. In Vivo, Ex Vivoand In Vitro.

Always refer to the GSE Summary or GSE Overall Design to accurately determine the experimental

setting. You can refer to the following definitions to guide your selection. Please determine one type

among the above three, and do not provide NA.

1. In Vivo: Experimentation conducted within a living organism. This category applies to studies where biological processes are examined in the context of an intact, living system. It includes research involving the interactions between different organs, tissues, and systems within the body of the organism. This setting is often used to study the effects of drugs, treatments, or genetic modifications in live animals or humans.
2. Ex Vivo: Experimentation conducted outside of a living organism, but using tissues, cells, or organs that have been derived from an organism. This category is used when the biological material is extracted from the organism and studied in a controlled environment, while still retaining some aspects of the complexity found in a living system. Examples include tissue culture studies or organ slice experiments.
3. In Vitro: Experimentation conducted in a controlled environment, such as a cell culture, biochemical assay, or similar laboratory setting. This category is used when the study investigates biological processes in isolated systems, allowing precise control over experimental conditions and variables. Common examples include experiments performed in test tubes, petri dishes, or culture plates.

 Disease: When annotating the Disease field, carefully review both GSE and GSM details to

determine the specific disease or medical condition associated with the given GSM sample. Use the exact disease names as provided or commonly accepted. Capitalize the first letter of each word in the disease name. Avoid abbreviations.

* Explicitly Mentioned Disease: If the disease or condition is explicitly mentioned in the provided information, the disease should be annotated as [Disease Name] (Extracted). For example, study title: Gene Level Expression Profiling in human tongue squamous carcinoma cell line (SAS), the Disease annotation: Squamous Carcinoma (Extracted).
* Disease Inferred from Context: If the disease is not directly mentioned but can be reasonably inferred from the context, the disease should be annotated as [Disease Name] (Inferred). For example, Transcriptional profiling of GIF-5 mouse gastric epithelial cells comparing CD133-positive and CD133-negative cells. The former formed CD133-positive and CD133-negative cells while the latter only CD133-negative cells, suggesting that CD133-positive cells are mother cells. The former produced differentiated type tumors while the latter undifferentiated types in vivo, indicating a relationship between CD133-expression and glandular structure formation. Disease field should include Gastric Tumor (Inferred) as disease information is not directly provided. Only infer diseases when there is strong evidence. Do not guess or assume diseases based on limited information.
* Multiple Diseases in a Study: annotate each GSM sample with the specific disease it represents. For example, in a study that includes both lung cancer and chronic obstructive pulmonary disease (COPD) samples, the disease should be annotated as Lung Cancer (Extracted) for lung cancer samples and Chronic Obstructive Pulmonary Disease (Extracted) for COPD samples. Separate distinct diseases with a semicolon if the GSM truly represents more than one disease in this format: for example, Disease A (Extracted); Disease B (Extracted). Similarly, when dealing with disease progression or staging, annotate samples based on their specific disease stage. For instance, in a study on liver disease, samples should be annotated as Fatty Liver (Extracted) or Cirrhosis (Extracted) depending on the stage of liver disease each GSM sample represents, while healthy liver samples should be annotated as NA.
* Healthy or Disease-Free Samples: For GSM samples from healthy, disease-free tissue, annotate as NA. For example, in a study where some alveolospheres are treated with bleomycin to induce pulmonary fibrosis and others are left untreated, the disease for bleomycin-treated samples should be Pulmonary Fibrosis (Extracted) and for untreated samples, the disease should be NA.

* Genetic Disease Models: annotate as [Disease Name] (Extracted) for the disease model and NA for the wild-type controls. For example, in a study comparing knockin zQ175 Huntingtons disease model mice and Wild Type (WT) mice, the disease for zQ175 samples should be Huntingtons Disease (Extracted) and for WT samples, the disease should be NA.

* For treatment controls derived from diseased sources: the disease should be retained as [Disease Name] (Extracted). For example, in a study involving AML cell lines where some samples are drug-treated and others are treated with DMSO as a control, the Disease for both treated and control samples should be Acute Myeloid Leukemia (Extracted).
* Unclear or Multiple Possible Diseases: annotate as NA. For instance, if a study description is vague or could relate to various conditions without clear evidence, the Disease should be annotated as NA.

# PERTURBATION RELATED FIELDS

Next identify the presence of Perturbations, which may include details such as dose, frequency, and duration.

* Prioritizing Information Sources: GSM Information should be your main reference for sample-specific details. Use GSE Information for general experiment context or when GSM details are missing.
* Maintain Consistent Formatting: Use the exact units and terms throughout all annotations. Ensure that entries for multiple Perturbations are in the correct order across all fields.
* For multiple Perturbations, list doses, frequencies and durations in the same order as the Perturbations, separated by + .
* Avoid Assumptions and not inferring information. Only use information explicitly provided in the GSE or GSM records. Do not assume doses, frequencies, or durations.
* Annotate as NA for each field when information for those specific fields is not available or not applicable. Do not leave fields blank.

GSE\_Pert: It indicates whether the overall experiment described in the GSE information involves any Perturbations. It reflects whether the study includes any deliberate interventions applied to the samples. Carefully review GSE Summary and GSE Overall Design to determine if the study includes Perturbations.

* Perturbations can be:

                 - Genetic: gene knockouts, knockdowns, overexpression.

                 - Chemical: drug treatments, toxin exposure.

                 - Environmental: changes in temperature, lighting, housing conditions.

                 - Physiological: surgical interventions, induced injuries.

* If any part of the GSE involves Perturbations, annotate as Yes even if not all samples are perturbed. If the study is purely observational with no interventions, annotate as No.
* Examples:

                 - Yes: A study where patients are treated with a drug or placebo

                 - No: A study analyzing gene expression in healthy human tissues without any treatments.

GSM\_Pert: It specifies whether an individual GSM sample is a Perturbed sample or a Control

sample within the GSE.

* Examine the GSM Information to determine the samples status. Look for

indications of treatments, interventions, or genetic manipulations in the GSM information.

* If the sample is untreated or received a placebo/control substance or served as baseline or control then annotate as Control.

Control: A cell line treated with vehicle only (e.g., DMSO) or left untreated.

* If the sample received any form of Perturbation, annotate as Perturbed.

                 Perturbed: A cell line treated with a chemotherapeutic agent.

 Pert: Extract information from the GSM record about the treatments or interventions.

* For Perturbed Samples:  List the specific perturbations applied. Use precise terms, including drug names, genetic modifications, or environmental changes.
* For Control Samples: Indicate the control substance or condition (e.g., DMSO, Vehicle Control, Untreated). Sometimes GSM\_Info only mentions Control or Vehicle but GSE\_Info includes specific details about control samples so carefully review and extract the Perturbations.
* List multiple Perturbations separated by + . Use exact names of drugs or agents, and avoid abbreviations unless they are standard and unambiguous. For genetic Perturbations, specify the gene and type of modification (e.g., Knockout: PTEN).

Pert\_Dose: It specifies the dose of the perturbation applied to the sample.

* Provide the dose amount along with appropriate units (e.g., 20 mg/kg, 10 μM) to maintain clarity.
* Exclude perturbation names from this field and use NA for unspecified dosages.
* Examples:

1. Sample Treated with CCL4 at 20 mg/kg: Pert\_Dose: 20 mg/kg.
2. Sample Treated with Drug A 5 μM + Drug B 10 μM: Pert\_Dose: 5 μM + 10 μM.
3. Control Sample with DMSO at 0.1%: Pert\_Dose: 0.1%.
4. Dose Not Specified: Pert\_Dose: NA

     Pert\_Freq: It indicates the frequency at which the perturbation was applied.

* Describe the frequency using standard expressions (e.g., Once daily, Twice weekly, Single dose).
* Use consistent terminology for clarity. Examples of Standard Expressions:Once daily, Every 12 hours, Twice weekly, Single dose.
* Examples:

1. Daily Treatment for a Week: Pert\_Freq: Once daily.
2. Administered Every Other Day: Pert\_Freq: Every other day.
3. Single Injection: Pert\_Freq: Single dose.
4. Frequency Not Specified: Pert\_Freq: NA.

Pert\_Duration: It specifies the total duration over which the Perturbation was applied. Provide the duration along with units (e.g., 1 week, 48 hours, 6 months).

                Pert\_Duration: 7 days

                Treatment Lasting 3 Weeks: Pert\_Duration: 3 weeks.

                Single Dose Sampled After 24 Hours: Pert\_Duration: 24 hours.

                Duration Not Specified: Pert\_Duration: NA

Some Examples Incorporating All Pert Fields:

1. For Perturbed Sample with Complete Information:

                GSE\_Pert: Yes

                GSM\_Pert: Perturbed

                Pert: CCL4

                Pert\_Dose: 20 mg/kg

                Pert\_Freq: Once daily

                Pert\_Duration: 1 week

1. For Control Sample Treated with DMSO:

                GSE\_Pert: Yes

                GSM\_Pert: Control

                Pert: DMSO

                Pert\_Dose: 0.1%

                Pert\_Freq: Once daily

                Pert\_Duration: 1 week

1. For Control Sample Without Treatment:

                GSE\_Pert: Yes

                GSM\_Pert: Control

                Pert: Untreated

                Pert\_Dose: NA

                Pert\_Freq: NA

                Pert\_Duration: NA

1. For Perturbed Sample with Multiple Perturbations:

                GSE\_Pert: Yes

                GSM\_Pert: Perturbed

                Pert: Drug A + Drug B

                Pert\_Dose: 5 mg/kg + 10 mg/kg

                Pert\_Freq: Once daily + Once daily

                Pert\_Duration: 2 weeks + 2 weeks

Route\_Admin: When annotating the Route of Administration through which a unique perturbation or drug was administered to the sample, select from the following standardized options.

1. Intraperitoneal for injection into the peritoneal cavity
2. Intravenous for injection directly into a vein
3. Oral for administration by mouth
4. Subcutaneous for injection into the tissue layer between the skin and muscle
5. Intramuscular for injection directly into a muscle
6. Topical for application directly to the skin or mucous membranes
7. Inhalation for administration through the respiratory tract
8. Intranasal for administration through the nasal passages
9. Intracerebral for injection directly into brain tissue
10. Intraventricular for injection into the brains ventricles
11. Intrathecal for injection into the spinal canal
12. Ocular for administration directly to the eye
13. Transdermal for absorption through the skin, typically via a patch.
14. If the drug was added to the culture media, annotate as In Media without specifying the drug name.

* If multiple drugs were used, list the corresponding routes of administration in the order they appear, separated by +
* If only injection is mentioned without specifying the route, annotate as Injection (unspecified) and seek further clarification if possible.
* Use Other [specify] if the route does not fit into any of the predefined categories, and replace [specify] with the specific route used.
* If no unique route is applicable or if no drug or pertturbation was used, annotate as NA.
* It is crucial to maintain the exact case format and avoid variations in upper and lower case, as well as special characters, to ensure consistency across all annotations.

Specimen\_Type: Use the Specimen\_Type field to categorize the biological source or model system used in a study. It provides clear and standardized descriptions of the specimen, ensuring consistency in annotation across diverse datasets (e.g., patient tissues, animal models, cell lines, and cultured systems).

The field captures Directly obtained tissues, Cultured or engineered biological models, Cellular populations, Fetal samples.

* Choose one from following Specimen\_Type Categories:

1. Primary Tissue: Tissue obtained directly from an organism (human, animal, or other), either fresh or previously frozen, while maintaining its original structure and microenvironment. For example : Human lung tissue from a cadaver

Key Distinction: Not cultured; excludes PDX, organoids, or precision-cut tissue slices.

1. PDX (Patient-Derived Xenograft): Tissue or tumors obtained from a human patient that are implanted into an animal model (e.g., mouse, rat) for disease modeling and treatment studies.

Examples:

1. Human breast tumor implanted into a mouse
2. Patient-derived colon carcinoma in a rat model

Key Distinction: Implanted into animals; excludes directly isolated tissues (Primary Tissue) and cultured systems (Organoids, Tissue Culture).

1. Cell Line: Immortalized or primary cells maintained in vitro as stable, reproducible cultures. Includes human, animal, or microbial cell lines, embryonic stem cell lines, and induced pluripotent stem cells.

Examples:

1. HeLa (human cervical cancer cell line)
2. MCF7 (human breast cancer cell line)

Key Distinction: 2D culture; excludes organoids (3D cultures) and tissue cultures.

1. Organoid: In vitro 3D culture systems that mimic tissue or organ architecture and function. Derived from stem cells or tissues.

Examples:

1. Intestinal organoids derived from human stem cells
2. Brain organoids developed from pluripotent stem cells
3. Lung spheroids used for drug testing

Key Distinction: 3D structure; excludes monolayer cell lines and precision-cut tissue slices.

1. Isolated Cells: Specific cell populations isolated or enriched from tissues, blood, or fluids using techniques like density gradient centrifugation or cell sorting.

Examples:

1. PBMCs isolated from whole blood
2. Macrophages isolated from bronchoalveolar lavage fluid

Key Distinction: Individual cell populations; excludes whole tissues (Primary Tissue) and cultured systems (Cell Line, Organoid).

1. Fetus: Biological material, tissues, or cells collected directly from a developing fetus (human or animal). Examples:
2. Human placenta or amniotic fluid
3. Human fetus used for analysis

Key Distinction: Fetal origin; excludes adult tissues, isolated cells, or cultured systems.

1. Tissue Culture: Tissue sections cultured ex vivo under controlled conditions, often used to study tissue responses while retaining original architecture.

  Examples:

1. Precision-cut liver tissue slices cultured for 48 hours
2. Cultured skin tissue for drug toxicity assays

Key Distinction: Cultured ex vivo tissue; excludes organoids (3D stem cell-derived) and cell lines

(immortalized cultures).

1. NA

Definition: No specimen information is available.

Examples:

No description of specimen provided

* Additional Rules for Resolving Overlaps or Ambiguities:

1. Primary Tissue vs. Tissue Culture

Rule: If the tissue is directly obtained (fresh or previously frozen) without ex vivo culture, annotate as Primary Tissue. If the tissue is cultured ex vivo, annotate as Tissue Culture. Example:

a. human liver tissue directly used for RNA-seq → Primary Tissue

b. Precision-cut liver tissue slices cultured for 24 hours → Tissue Culture

B. Cell Line vs. Organoid

Rule: If the sample is an immortalized 2D culture, annotate as Cell Line. If it is a 3D culture mimicking organ structure, annotate as Organoid.

Example:

1. MCF7 breast cancer cells grown in vitro → Cell Line
2. Intestinal organoids for disease modeling → Organoid

C. Primary Tissue vs. PDX

Rule: If the tissue is directly isolated from a patient or organism, annotate as Primary Tissue. If it

has been implanted into an animal host, annotate as PDX.

Example:

1. Human colon biopsy → Primary Tissue
2. Human tumor implanted into a mouse model → PDX

D. Isolated Cells vs. Primary Tissue

Rule: If specific cells have been isolated or enriched, annotate as Isolated Cells. If the sample retains tissue integrity, annotate as Primary Tissue.

Example:

1. PBMCs isolated from blood → Isolated Cells
2. Freshly obtained human liver tissue → Primary Tissue
3. Tissue Culture vs. Organoid

Rule: If the culture involves thin tissue slices or ex vivo tissues, annotate as Tissue Culture. If it is a 3D structure grown to mimic organ architecture, annotate as Organoid.

Example:

1. Precision-cut lung slices in culture → Tissue Culture
2. Lung organoids for drug screening → Organoid

* Lastly, here are the final notes for annotations:
* Always Prioritize Specificity: Always annotate using the most specific category.
* Apply Resolution Rules: Use the provided overlap rules to resolve ambiguous cases.
* Document Unusual Cases: If you use Other or NA, note why it was chosen if possible.
* Consistency: Ensure consistent usage across all samples to avoid misclassification.

# RACE AND ETHNICITY FIELD

Next, annotate Race and Ethnicity. First of all, here is the Race vs. Ethnicity Distinction.

* At a high level, Race refers to broad categories based on shared physical or social traits (e.g., ‘White,’ ‘Black or African American,’ ‘Asian’), while Ethnicity focuses on cultural, national, religious, or linguistic identity (e.g., ‘Hispanic,’ ‘Han’).
* In GEO metadata, you may occasionally see ‘ethnicity: African American’ or ‘ethnicity: Han,’ but these are more accurately annotated as Race: 'Black or African American' (with Ethnicity: 'Not Specified') or Race: 'Asian' (with Ethnicity: 'Han'), respectively. If you find a Race descriptor incorrectly placed under ‘Ethnicity,’ re-map it to the Race field and mark Ethnicity as appropriate (e.g., ‘Not Specified’).
* Key Note on 'Not Specified' vs. 'NA'.
* 'Not Specified' is used only for Patient Specimen samples when relevant information (Race or Ethnicity) is missing or not provided.
* 'NA' is used for samples that do not originate from a patient (e.g., cell lines, organoids, animal tissues), where Race/Ethnicity simply does not apply.

Race: Use the 'Race' field to classify Patient Specimen samples into one of the U.S. Census Bureau race categories or related terms provided. Race refers to broad population groups often defined by physical characteristics and geographic ancestry. If no information is available or the sample is not a Patient Specimen, follow the rules below.

* Applicability: Annotate 'Race' only if SampleType is 'Patient Specimen'. If SampleType is not 'Patient Specimen' (e.g., 'Cell Line', 'Organoid'), annotate 'NA'.
* Determining Race:
* Refer to GSE-level metadata first. If a single race is stated for all Patient Specimens, apply that category uniformly.
* If multiple races are mentioned at the GSE level, check GSM-level metadata for sample-specific race information.
* Use the category that most closely matches the provided race information. If a population does not neatly fit into one of the categories below, annotate as 'Not Specified' or provide the closest equivalent with additional specificity if possible. When you see synonyms or older terms like Caucasian, map them to ‘White’; African American or African descent → ‘Black or African American’; Han → ‘Asian’.
* Categories: Use these exact case-sensitive categories whenever possible:

1. 'American Indian or Alaska Native'
2. 'Asian'
3. 'Black or African American'
4. 'Hispanic or Latino'
5. 'Native Hawaiian or Pacific Islander'
6. 'White'
7. 'Two or More Races'
8. 'Not Specified'
9. 'NA' (for non-Patient Specimen samples)

* Missing or Ambiguous Race:
* If no race information is available for a Patient Specimen, annotate as 'Not Specified'.
* For non-Patient Specimen samples, annotate as 'NA'.
* If the metadata suggests a specific race not listed, provide the closest category and if unclear, 'Not Specified'.
* Examples:

1. GSE Info: 'All patients are of European ancestry' → Closest category: 'White'.
2. GSE mentions multiple donors: African, European, and Mixed.

Map 'African' to 'Black or African American', 'European' to 'White', and if a donor is

described as mixed ancestry from multiple categories, use 'Two or More Races'.

* GSM001 (Patient Specimen, African donor): Race: Black or African American
* GSM002 (Patient Specimen, European donor): Race: White
* GSM003 (Patient Specimen, described as mixed from multiple groups): Race: Two or More Races

1. No race data at GSE or GSM level.

* GSM001 (Patient Specimen): Race: Not Specified
* GSM002 (Organoid): Race: NA

Ethnicity: Use the 'Ethnicity' field to capture cultural, national, religious, or linguistic affiliations that distinguish groups such as 'Hispanic' or 'Non-Hispanic'.

* Applicability: Only annotate 'Ethnicity' if SampleType is 'Patient Specimen'. If SampleType is not 'Patient Specimen', annotate 'NA'.
* Determining Ethnicity:
* Check GSE-level metadata: If all Patient Specimens are of a single ethnicity, apply that to all.
* If multiple ethnicities are mentioned, refer to GSM-level metadata for sample-specific details.
* If the original GEO labeling incorrectly uses Ethnicity for Race (e.g., ethnicity: African American), reassign appropriately to the Race field, and set Ethnicity to ‘Not Specified’ if no true ethnic descriptor is given.
* Categories: Basic categories: 'Hispanic', 'Non-Hispanic', 'Not Specified', 'NA'.
* If more specific ethnic details are provided (e.g., 'Asian-Han Chinese', 'Asian: Indian'), annotate them as given to maintain specificity.
* Missing or Ambiguous Ethnicity: If no ethnicity information is available for a Patient Specimen, use 'Not Specified'.
* For non-Patient Specimen samples, use 'NA'.
* Examples:

1. GSE Info: 'All European samples are Non-Hispanic'.
2. GSM Info shows a donor listed as Hispanic.

GSM001 (Patient Specimen): Ethnicity: Hispanic

* Summary of 'Not Specified' vs. 'NA'

1. Not Specified: Used for Patient Specimen samples when the metadata lacks race or ethnicity details.
2. NA: Used for samples not derived from a patient (e.g., cell lines, organoids, animal models).

Age: Age records the exact age of the sample source with numeric values and standardized units, avoiding any inference or assumption.

* Capturing Age Record numeric age only if the metadata explicitly states it as age. For example: 'age: 25', '25 years old', or 'age: 6 Months.'
* Do not use numeric fields that simply appear in the GSM record but are unrelated to age (e.g., sample numbers, replicate IDs, 'bmi: 22.6').
* Include units in the final annotation: '25 Years', '6 Months', '7 Days', '5 Hours'.
* Leave a space between the number and the unit (e.g., 8 Days, 20 Years).
* Do not use 'old' or other variations; maintain the strict Number + Unit format.
* If no numeric age is provided or it cannot be confirmed from the metadata, annotate as Unknown.
* If age does not apply (e.g., a cell line or organoid sample), annotate as NA.
* Important Notes:
* Confirm that the numeric value truly refers to age in the GSM/GSE description. Any numeric entry lacking explicit mention of 'age,' 'years,' 'months,' etc. must not be assumed to be age.
* Allowed units: 'Years', 'Months', 'Days', 'Hours.'
* Avoid abbreviations like 'yrs,' 'mos,' 'hr.'
* Do not infer age from file names or sample IDs.
* If the numeric value references 'bmi' or any other measurement (e.g., 'sample #' or numerical values from the titles), you must not conflate it with age.

Age\_Group: 'Age\_Group' categorizes the sample based on the provided or known age, fitting it into one of the predefined categories listed below.

* Categories:

1. 'Embryo': Developmental stage before fetal development.
2. 'Fetus': After embryonic stage and before birth.
3. 'Neonate': Newborn, up to 28 days old.
4. 'Infant': Post-neonate to 1 year of age.
5. 'Pediatric': 1 to <12 years.
6. 'Adolescent': 12 to 18 years.
7. 'Adult': 18+ to 64 years.
8. 'Elderly': Older than 65 years.
9. 'Unknown': Age not provided or cannot be determined.
10. 'NA': Age not applicable (e.g., Cell line, Organoid).

* Step-by-Step Instructions:
* Refer to GSE and GSM Information:
* Check GSE-level metadata for general age-related details (e.g., 'adults aged 20-40 years').
* Check GSM-level metadata for specific numeric ages or direct age statements.
* Annotate Age and Age\_Group: If numeric age is provided, record it in the 'Age' field with the correct format. Determine 'Age\_Group' based on the numeric age or descriptive information.
* f numeric age is not given and no inference can be made, 'Age' → 'Unknown' and 'Age\_Group' → 'Unknown'.
* If the sample type does not require an age (e.g., Cell line, Organoid), 'Age' → 'NA' and 'Age\_Group' → 'NA'.
* Handle Missing/Ambiguous Data:

- If neither numeric age nor descriptive category is provided: 'Age' → 'Unknown' and 'Age\_Group' → 'Unknown'.

- If the sample type inherently has no age (e.g., Cell line): 'Age' → 'NA' and 'Age\_Group' → 'NA'.

* Example Scenarios:

1. Exact Age Provided

* GSM001 (Patient Specimen): 25 Years → Age\_Group: Adult
* GSM002 (Patient Specimen): 6 Months → Age\_Group: Infant
* GSM003 (Patient Specimen): 68 Years → Age\_Group: Elderly

1. Age Range in GSE

* GSE states 'pediatric donors (2-10 years)'

GSM004: Age: Unknown, Age\_Group: Pediatric

1. Age Not Applicable

* GSM005 (Cell line): Age: NA, Age\_Group: NA
* GSM006 (Organoid): Age: NA, Age\_Group: NA

Gender : Gender must be accurately extracted, annotated, and standardized based on the GSE summary, overall design, and GSM-specific metadata.

* The annotations should reflect the study's experimental design and the terminology used in the metadata.
* Where to Extract Gender Information:
* GSE Overall Design and Summary': Look for explicit mentions of gender in the study-level description. Examples include '72 males and 72 females' or 'Male patients treated with drug A', or references to gender-specific biological processes like 'Effects of testosterone in males'.
* GSM-Specific Metadata': Search for explicit gender mentions in sample-level fields. For instance, 'Sex: Male' or 'Gender: Female'. Gender may also appear indirectly, such as 'pregnant' or 'ovarian cancer patient' (implying 'Female') or 'patient with prostate enlargement' (implying 'Male').
* Implicit Gender Clues: If not explicitly stated, infer gender from biological or experimental context. For example, 'ovarian tissue' implies 'Female', 'prostate tissue' implies 'Male', and 'pregnant' implies 'Female'. If no inference is possible, use 'Unknown'.
* Strategy for Extracting and Annotating Gender:
* Parse the GSE Summary and Overall Design. If the GSE states something like '72 male', annotate gender as 'Male'. If it states 'samples collected from females', annotate as 'Female'. Standardize all explicitly mentioned genders.
* Parse GSM Metadata. Look for explicit fields like 'Sex: M' or 'Gender: F' and map them to 'Male' or 'Female'. If only abbreviations or indirect clues are given, interpret them accordingly.
* Infer Gender When Necessary. Use tissue type or experimental context if direct mentions are absent. If no inference is possible, use 'Unknown'.
* Standardize Gender Information. The final annotation should be 'Gender: Male', 'Gender: Female', 'Gender: Unknown', or if not applicable (e.g., for cell lines without donor info), 'Gender: NA'.
* How to Handle Gender for Cell Lines:
* In general, cell lines are not gender-specific, so annotate 'Gender: NA'.
* If donor information is provided, use it to assign 'Male' or 'Female'. If the cell line originates from ovarian carcinoma, 'Female'; from prostate carcinoma, 'Male'. If multiple donors of different genders were used to create a hybrid cell line, annotate 'Gender: Mixed'.
* Examples :

1. A cell line 'A549' (lung carcinoma, no gender info): 'Gender: NA'.
2. A cell line 'HeLa' derived from a female cervical carcinoma: 'Gender: Female'.
3. A cell line 'LNCaP' (prostate carcinoma): 'Gender: Male'.
4. Example 4: A hybrid cell line from male and female donors: 'Gender: Mixed'.

Timepoint

* Annotate and standardize timepoint information from GEO metadata.
* Identify when a sample was collected relative to events, stages, or factors.
* Timepoints can be 'NA' if no time dimension applies (e.g., purely cross-sectional data with no mention of collection timing). Always verify that any 'Day X' or 'Week X' matches the GSE-level conventions (e.g., '7 days post-inoculation' → 'Post-inoculation Day: 7').
* Potential Timepoint Types:

1. Pre/Post-Event: Defined relative to a key event, such as inoculation, treatment, surgery, or drug administration. (for example, Pre-treatment Day: -1, Post-treatment Hour: 24; Post-surgery Week: 2)
2. Developmental: Based on the biological or developmental stage of the organism. (for example, Gestational Day: 14; Postnatal Day: 7; Embryonic Day: 10)
3. Chronological: Absolute time markers, often tied to the experimental design (Day: 7; Week: 2; Month: 1)
4. Follow-Up: Used in longitudinal or clinical studies to indicate repeated measurements. (Baseline; Follow-Up Month: 3; Follow-Up Year: 5)
5. Event-Driven: Associated with external stimuli or interventions. (Post-inoculation Day: 14; Post-challenge Hour: 4)
6. Behavioral/Physiological: Defined by symptoms or biological responses. (First Symptom Day; Onset of Fever)
7. Cyclic/Rhythmic: Relevant for circadian rhythm studies or periodic sampling. (Circadian Phase: ZT12; Diurnal Hour: 16)
8. Age-Based: Biological age of the subject used explicitly or inferred as a timepoint. (Age: 12 Weeks)

* Lastly, please follow the following Strategies for Annotating Timepoints:

1. Review GSE Overall Design and Summary

* Look for explicit descriptions of timepoints in the study design (e.g., 48 samples collected at 7, 14, and 28 days post-inoculation).
* Define standardized terms for timepoints based on this overview:

Examples:

1. 7 days post-inoculation → Post-inoculation Day: 7
2. 14 days post-inoculation → Post-inoculation Day: 14
3. Examine individual sample metadata fields (e.g., timepoint, day, dpi) to extract timepoint information.

* GSE overall summary says  48 samples collected at 7, 14, and 28 days 'post-inoculation'). And specific GSM Metadata includes  timepoint: D7 → Annotation: Post-inoculation Day: 7
* GSM Metadata: timepoint: baseline → Annotation: Baseline
* For missing or ambiguous timepoints, infer from replicates or GSE-level descriptions, or label as unknown.

* Standardize Timepoint Information
* Map all extracted timepoints to a consistent format: [Context] [Unit]: [Value]
* Examples:
* 7 days post-inoculation → Post-inoculation Day: 7
* 14 days post-treatment → Post-treatment Day: 14
* Address Age-Related Timepoints : When timepoints can be derived from age:
* Example: Baseline age: 12 weeks
* Timepoint: 14 days post-inoculation
* Effective age: 14 weeks
* If no timepoint or it is truly not relevant, annotate ‘NA.’ If ambiguous or missing: 'NA'.

Outcome : The Outcome field is a unified annotation field designed to capture all relevant treatment-related and prognostic outcomes for both patient-level and cell line data. It provides a structured and standardized way to represent treatment response, survival status, and prognosis information in a single, parsable format.'

* The Outcome field integrates three components (if available): 'Response', 'Survival', and 'Prognosis'.

1. Response: Describes the subject's or sample's response to treatment.

For patients:

* Responder: Complete remission or significant response.
* Partial Responder: Partial remission or partial improvement.
* Stable Disease: No progression or improvement.
* Non-Responder: Disease progression or no response.
* Unknown: Ambiguous or unclear response.

For cell lines:

* Sensitive: Effective response (e.g., growth inhibition).
* Partially Sensitive: Moderate or partial response.
* Resistant: No response or lack of inhibition.
* Unknown: Ambiguous or unclear response.

1. Survival: Represents survival data for patients, including survival time and status.'

* Format: '[Survival\_Time]: [Status]'
* 'Survival\_Time': Duration reported in months (preferred) or years.
* 'Status': 'Alive', 'Deceased', or 'Unknown'.
* Example: '36 Months: Alive', '12 Months: Deceased'.
* If sample is from a living patient but no survival data is provided, annotate 'Unknown'. If it’s a cell line or purely animal model with no survival aspect, use 'NA'.

1. Prognosis: Provides the predicted clinical outcome for the patient based on available metadata.'

* 'Good Prognosis': Favorable or low-risk outcome expected.
* 'Poor Prognosis': Unfavorable or high-risk outcome expected.
* 'Unknown': Prognosis unclear or not reported.
* Combining Multiple Aspects: You may combine Response, Survival, and Prognosis in the same cell by separating them with semicolons. For example Responder; 36 Months: Alive; Good Prognosis.
* Examples: Below are example scenarios demonstrating how to annotate Response, Survival, and Prognosis within the Outcome field for both patient-level and cell-line samples. Each scenario provides a short description of the sample’s situation, followed by the recommended Outcome Annotation.

Patient-Level Examples

1. Scenario: Complete remission, survived 36 months  
   Outcome Annotation: Responder; 36 Months: Alive
2. Scenario: Partial remission, survived 24 months  
   Outcome Annotation: Partial Responder; 24 Months: Alive
3. Scenario: Disease progression, deceased at 12 months  
   Outcome Annotation: Non-Responder; 12 Months: Deceased
4. Scenario: Stable disease, no survival or prognosis  
   Outcome Annotation: Stable Disease
5. Scenario: Poor prognosis only  
   Outcome Annotation: Poor Prognosis
6. Scenario: Ambiguous response, unclear status  
   Outcome Annotation: Unknown
7. Scenario with No Outcome Data Provided Outcome Annotation: Unknown

Cell Line Examples

1. Scenario: Sensitive to treatment, no survival/prognosis data  
   Outcome Annotation: Sensitive
2. Scenario: Partial sensitivity to treatment  
   Outcome Annotation: Partially Sensitive
3. Scenario: Resistant to treatment  
   Outcome Annotation: Resistant
4. Scenario: Ambiguous response  
   Outcome Annotation: Unknown
5. Scenario: No treatment outcome  
   Outcome Annotation: NA