AML_vanGalen_2019_NanoWell_Anndata.h5ad obs metadata meaning from model perspective

Dataset Analysis Reference

December 12, 2025

Column Name	Description
BUCKET 1: CONDITIONING INPUTS (The 'Levers')	
Variables used to CONTROL the generation (Batch keys, Labels, Biological States).	
DiseaseCode	AML for Acute Myeloid Leucemia or NBM for Normal Bone Marrow.
DiseaseType	Cancer or Healthy.
DiseaseTypeDetailed	AcuteMyeloidLeucemia or NormalBoneMarrow. (Basically Diseasecode but spelled out)
Treatment	None, Chemotherapy, Azacitidine_Venetoclax or HiDAC.
TreatedWithDrug	Treatment status (Yes/No).
CellType	Cell type annotation.
CellType_Level_0	Malignant/Immune/Progenitor/Stem-cell/Unknown.
CellType_Level_1	More granular Celltyping than CellType_Level_0.
CellType_Level_2	More granular Celltyping than CellType_Level_1.
CellType_Level_3	More granular Celltyping than CellType_Level_2.
CellType_Level_4	More granular Celltyping than CellType_Level_3. (Same values as CellType_Level_3 but more accurate labels)
curated_celltypes	Additional cell type level, manually curated from existing annotations and scores..
Age	Patient age.
Sex	Patient sex.
SampleID	Identifier for the tissue sample. There are 43 samples in this datset.
PatientID	Identifier for the patient. There are 23 patients in this dataset.
SampleSite	Cell origin from either BoneMarrow (extracted from patient) or CellLine (lab grown).
SampleType	Tumor or Normal.
SampleTypeDetailed	PrimaryTumor or RecurrentTumor or Normal.

Continued on next page

Continued from previous page

Column Name	Description
CyclingScore	Cell cycle score. (A higher score means the cell is likely proliferating and a lower score means its more likely resting)
CyclingBinary	The final decision if the cell is cycling or not. (yes/no/unlabeled)
Days.from.diagnosis	Time in days from diagnosis to collection. Format: D followed by number of days.
InputTissue	Frozen or Fresh.
CellSelection	Method used for cell selection/sorting. This is important because it affects the cell population analyzed. For most patients all cells in the Sample were sequenced however there are the following exceptions in format (Method:PatientID): 1. AML: MUTZ3, OCI-AML3 2. CD34+CD8-: BM5 3. CD34+: BM5

BUCKET 2: VALIDATION & AUDIT (Hidden from Model)

Variables HIDDEN during training and used ONLY to verify the output after generation.

MutTranscripts	Amount of transcripts mapped to a known genetic mutation.
WtTranscripts	Amount of transcripts mapped to the wild-type allele.
Blast.count	Blast cell fraction in the bone marrow. (Basically a very high myeloid blast percentage is indicative of AML because it's a cancer that develops from this cell type.)
PredictionRF2	Likely an output of a random forest model used to classify malignant vs normal cells. (normal/malignant/Unlabeled)
PredictionRefined	Same as PredictionRF2 but added label unclear.
Score_HSC	Score for Hematopoietic Stem Cell based on similarity to reference profile.
Score_Prog	Score for Progenitor based on similarity to reference profile.
Score_GMP	Score for Granulocyte-Monocyte Progenitor based on similarity to reference profile.
Score_ProMono	Score for Promonocyte based on similarity to reference profile.
Score_Mono	Score for Monocyte based on similarity to reference profile.
Score_cDC	Score for conventional Dendritic Cell based on similarity to reference profile.
Score_pDC	Score for Plasmacytoid Dendritic Cell based on similarity to reference profile.
Score_earlyEry	Score for Early Erythroid based on similarity to reference profile.

Continued on next page

Continued from previous page

Column Name	Description
Score_lateEry	Score for Late Erythroid based on similarity to reference profile.
Score_ProB	Score for Pro-B cell based on similarity to reference profile.
Score_B	Score for B cell based on similarity to reference profile.
Score_Plasma	Score for Plasma cell based on similarity to reference profile.
Score_T	Score for T cell based on similarity to reference profile.
Score_CTL	Score for Cytotoxic T Lymphocyte based on similarity to reference profile.
Score_NK	Score for Natural Killer cell based on similarity to reference profile.
percent.mt	Mitochondrial percentage. (This is accurate. Use this!)
total_counts_mt	Absolute mitochondrial counts. (Accurate)
nFeature_RNA	Number of Genes per cell. (Only accurate column for this information. Use this!)
nCount_RNA	Number of counts per cell. (Only accurate column for this information. Use this!)
DoubletScores	Doublet prediction score derived from prior analysis. (Basically an extra QC filter that's already been applied)
leiden	Output of previous leiden clustering. (For cell type annotation)
seurat_clusters	Seurat cluster assignments. (also for cell type annotation. seems they used different methods for different label columns)

BUCKET 3: DISCARD (Remove to prevent leakage)

Redundant, inaccurate, or pure noise. These will be deleted.

Cell	Unique Cell identifier.
orig.ident	Another cell identifier.
SharePointID	Source study. In this case just AML_vanGalen_2019_NanoWell.
Cell.number	Likely another sample identifier since there are 43 of these just as there are 43 samples.
NanoporeTranscripts	Genotype of possible driving mutations detected by Nanopore sequencing. (Basically a result of another sequencing method we can ignore for now)
Author_CellType	Completely identical to CellType.
RNA_snn_res.1	Seurat clustering remnant. (Just a technical leftover)
_scvi_batch	scVI leftover. (Delete before rerunning scVI)

Continued on next page

Continued from previous page

Column Name	Description
_scvi_labels	scVI leftover. (Delete before rerunning scVI)
NumberOfReads	Raw reads per cell pre UMI conversion. (When our methods ask for raw reads do not use these. These still contain the full PCR-bias. Use the UMIs in nCount_RNA)
AlignedToGenome	Raw reads successfully aligned to the human reference genome.(Another technical leftover just like NumberOfReads)
AlignedToTranscriptome	Raw reads successfully mapped to Exons. (Technical leftover just like NumberOfReads and AlignedToGenome)
TranscriptomeUMIs	Outdated UMI data. Use nCount_RNA instead.
NumberOfGenes	Outdated gene count data. Use nFeature_RNA instead.
n_genes	Inaccurate measurement of the numbers of genes per cell. Use nFeature_RNA instead.
percent.mito	Inaccurate measurement of mitochondrial percentage. (Do not use)
UMIs.min	Minimum UMI count in the sample of origin of this cell.
UMIs.mean	Mean UMI count in the sample of origin of this cell.
UMIs.max	Maximum UMI count in the sample of origin of this cell.
Genes.min	Minimum Gene count in the sample of origin of this cell.
Genes.mean	Mean Gene count in the sample of origin of this cell.
Genes.max	Maximum Gene count in the sample of origin of this cell.
SequencingTechnology	Every sample was analyzed using Seq-Well.
SequencingTechnologyDetailed	Exact same as SequencingTechnology. Just says Seq-Well.
InputMaterial	Only Cells were sequenced so every entry says Cell.