

大規模言語モデルを用いた BioSample データベースメタデータの品質向上



○池田秀也¹, 末竹裕貴², 大田達郎^{1,3,4}

1) 情報・システム研究機構 国立遺伝学研究所 バイオデータ研究拠点 ライフサイエンス統合データベース部門 (DBCLS), 2) 株式会社 Sator, 3) 千葉大学大学院医学研究院人工知能 (AI) 医学,
4) 千葉大学国際高等研究基幹

BioSample は、実験に用いられた生物学的サンプルに関するメタデータを蓄積する公共データベースである。メタデータの記述方法の多くは投稿者の裁量に委ねられているため、同一の実験条件であっても投稿者によって異なる表現が用いられることが多く、データの再利用性を低下させる要因となっている。我々はこの問題に対処するため、ローカル環境で動作する大規模言語モデル (LLM) を用いてメタデータを解釈し、細胞株、疾患、ノックアウト遺伝子などの概念を表す文字列を抽出し、対応するオントロジータームへマッピングするシステムを開発した。マッピングの正確性の評価は、細胞株サンプルを対象としたマッ

ピングを定義した正解データセットを作成して行った。さらに、マッピングの正誤判定や誤答の分類についても LLM を使用する手法の検討を進めている。これにより、マッピング対象の概念を拡大した場合の正解セット作成の手間を軽減しつつ、誤りを減らすことを目指している。実行条件を最適化した結果、現在では NVIDIA H200 GPU を 8 基用いることで、1週間当たり 200 万件を超えるサンプルを処理可能となった。今後は、マッピング結果を検索可能なインターフェースを通じて公開する予定である。また、マッピング結果を活用したサンプル統計情報の可視化システムについても開発を進めている。

BioSample の課題

属性名とその値のペアの形でサンプルメタデータを記述
- 同じものを表すのにシノニムが使われており検索しにくい
- 属性名が統一されておらず管理しにくい

sample name	iPSC_1390G3	source name	Induced pluripotent stem cell
cell line	1390G3-526	biomaterial provider	parental cell line from Coriell
cell type	iPSC	tissue	Induced pluripotent stem cell
sex	female	derived from cell line	NA19193

Attribute name	# of Sample	
cell line	4734	同じ文字列でも違うことを意味しているかもしれない ← 値に“H1”という文字列を含む属性の属性名 (一部)。“H1”は細胞株名、ヒストン、サンプル ID などを表す可能性がある
source_name	4360	
cell type	1043	
cell line	635	
well	263	
isolate	192	
strain	160	
tissue	77	
genotype	59	
treatment	54	

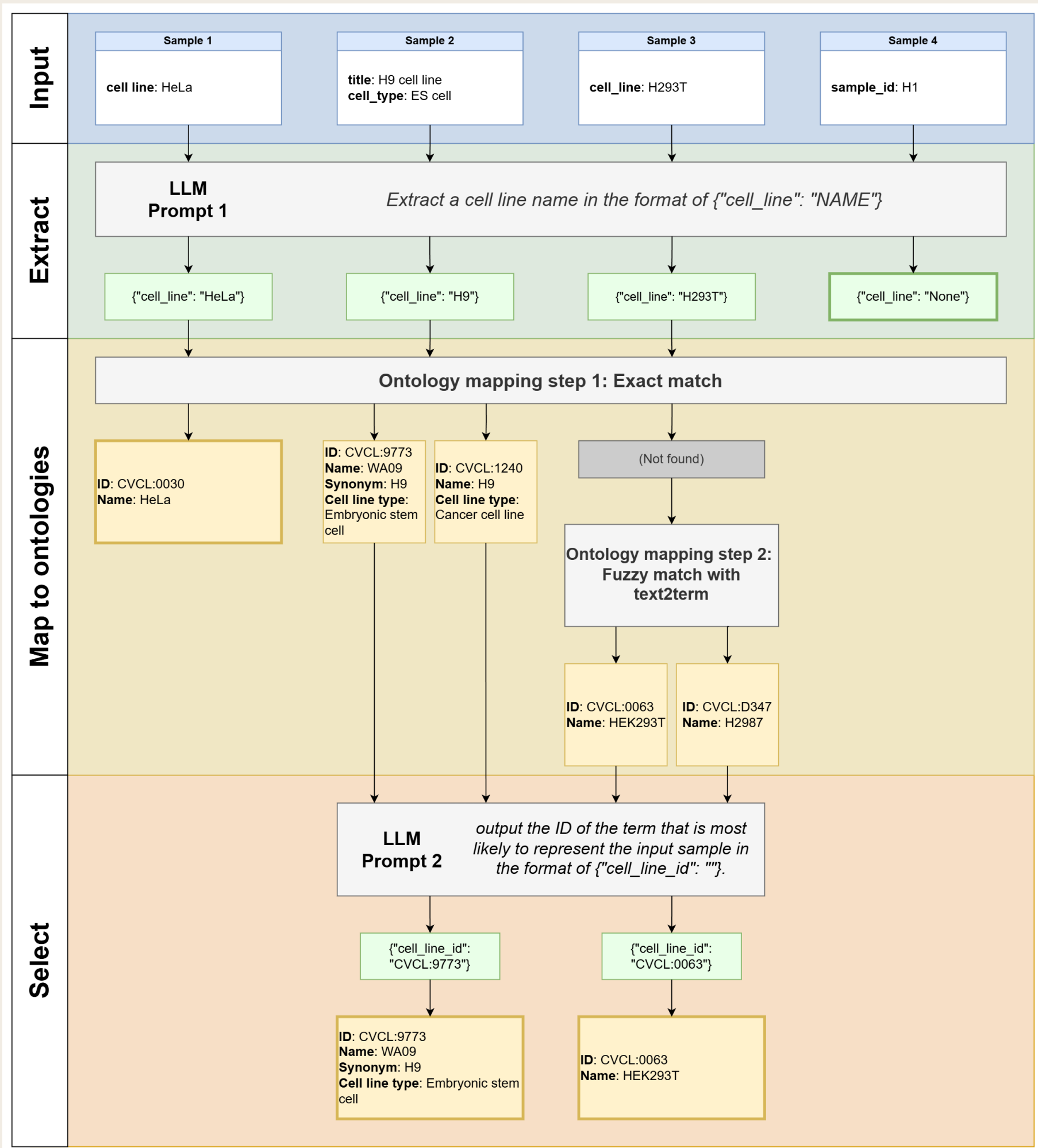
6000 万レコード近くあり、手動での網羅的なキュレーションは困難
既存手法 MetaSRA [1]: 文字列一致ベースでオントロジーにマッピング
→ 表記揺れを吸収し検索性を向上

Attribute	Value	Ontology term
cell_line	H1	→ Cellosaurus ID: CVCL_9771
cell_type	WA01	→ name: WA01
source_name	H1	→ synonym: H1
sample_id	H1	→

細胞株らしくない属性名の場合は細胞株のタームにマップしない

課題: ミスマップを軽減するため、事前に許容した属性の値しか使っていない
→ **LLM による改善の可能性**

方針



正解セット作成

ChIP-Atlas [3] のマニュアルキュレーションの成果を利用し、評価用の正解セットを作成。

細胞株名を対象、オントロジーは Cellosaurus を使用。
322 細胞株、278 非細胞株の正解を定義

BioSample ID	BioSample Attributes	String to be extracted	Term to be mapped
SAMN1347801	chip antibody: rK44m3 Abcam ab8580 genotype: PRIC-1M1 source_name: Patient derived tissue: Primary tumor	ST28T	CVCL-8917
SAMN0911769	chip antibody: rK477e Active Motif 39133, lot 31814008 source_name: Hepatocellular carcinoma tissue: BMAC-721 H9K27ac, ChIPSeq, DNase-Seq, H3K27ac, ChIPSeq, DNase-Seq, H3K27ac, ChIPSeq, DNase-Seq	BMAC-721	CVCL-0534
SAMN0248158	source_name: diploid fibroblast factor: p53 tissue: Adipose tissue	p53	

正解セットを用いた評価

Pipeline	Model	Accuracy	Time
MetaSRA	-	0.819	
LLM-assisted	mistral-small3.1:24b	0.922	27m
	qwen3:32b	0.920	34m
	llama3.3:70b	0.917	51m

- LLM を用いた手法は従来手法よりも高い性能を示した
 - 正解セットはモデル比較にも活用
 - 出力 JSON スキーマを指定することで処理時間を大幅に削減
 - Think-step-by-step 法はあまり効果的ではなかった
- 8 基の NVIDIA H200 GPU で週 200 万件以上のサンプルを処理可能

Method	Model	Input	Time (sec)	Accuracy
TSBS	qwen3:32b	600	1288	0.890
Format	qwen3:32b	600	93	0.913

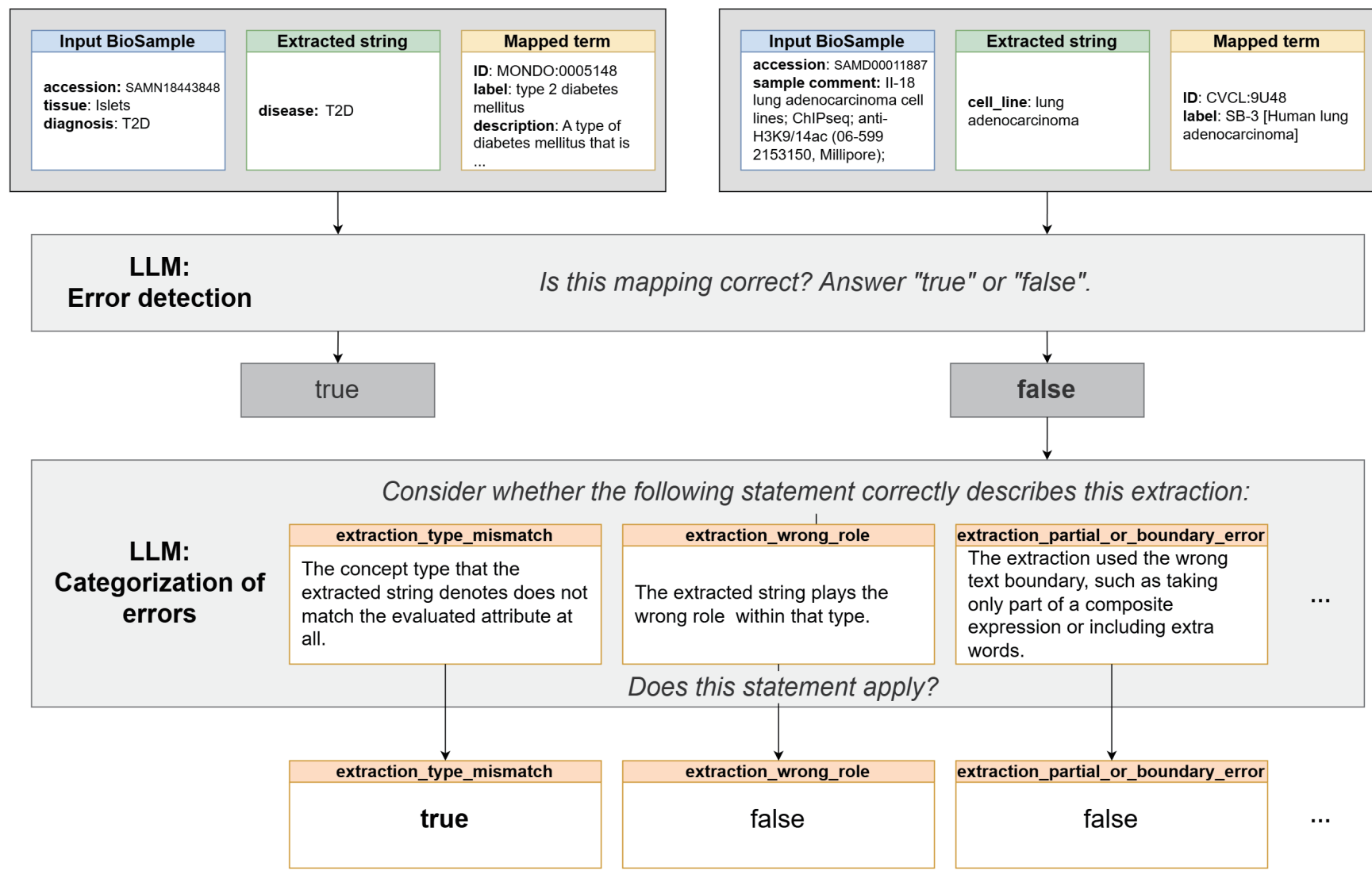
対象の拡大

Category	Target ontology	Epigenomic		RNA-seq	
		human	mouse	human	mouse
cell line	Cellosaurus	95859	19699	612262	96424
cell type	Cell Ontology	70510	97598	693398	680962
tissue	UBERON	55568	103129	1039687	1278533
disease	MONDO	36848	3852	462374	60519
drug	CHEBI	22136	13948	351113	148868
knockout gene		11131	32719	61033	306818
knockdown gene	NCBI Gene	6356	2429	49098	15117
overexpressed gene		9930	10296	53529	83105
total samples		179015	188122	2108453	1745491

- ヒトおよびマウスのエピゲノム実験サンプル (ChIP-Atlas に含まれるもの) と RNA-seq サンプル 400 万件以上について、各種概念へのオントロジーへのマッピングを実施
- 植物および節足動物サンプルへのマッピングも検討中

LLM による評価

- ・細胞株については正解セットでマッピングの品質を確認したが、他の概念への手法適用の妥当性は未評価
- ・すべての概念について正解セットを作成するのは労力がかかりすぎる → LLM による評価は可能か？



細胞株正解セットによる判定と比較。
LLM によって「正しい」と判定されたマッピングの正解率は 219/221 (0.991) と高い

80% 以上のマッピングが LLM の検証によって「正しい」と判定 (ChIP-Atlas のヒトサンプルを用い、抽出語とマッピング語の組が同一のものを除いて算出)

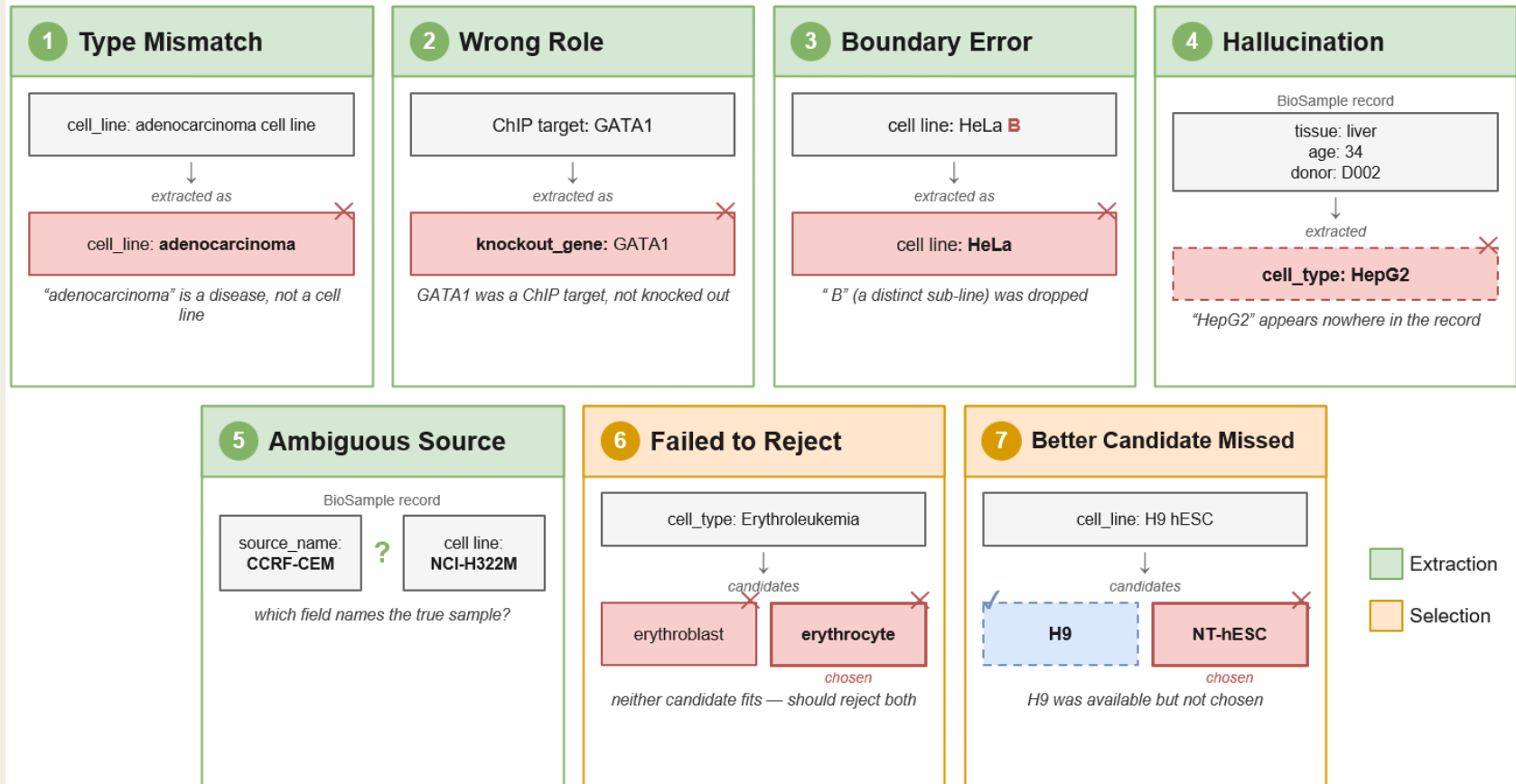
LLM evaluation	Gold-standard evaluation	
	Correct	Wrong
Correct	219	2
Wrong	59	38

model: Qwen3.8-27B-UD-Q4_K_XL

target	% of mappings judged as correct by the error detector	% of mappings judged as correct after detailed categorization
cell_line	67.6	86.0
cell_type	75.7	80.0
disease	91.0	94.5
drug	77.2	81.8
knockdown_gene	88.0	91.5
knockout_gene	80.0	87.9
overexpressed_gene	86.5	89.1
tissue	82.8	85.6

LLM で誤答のカテゴリ分け

現在用いているカテゴリ



target	type mismatch	wrong role	boundary error	hallucination	ambiguous source	failed to reject	better candidate missed	total false	total
cell_line	2.36	16.22	3.94	0.00	0.86	7.48	6.12	1046	3251
cell_type	14.09	5.93	9.50	1.78	1.34	19.89	36.94	674	2777
disease	2.78	1.39	1.39	9.72	1.39	23.61	23.61	72	738
drug	2.34	6.25	7.42	0.78	0.39	32.81	33.88	256	1125
knockdown_gene	4.85	30.10	10.88	2.91	4.85		3.88	103	850
knockout_gene	0.95	29.21	13.02	5.08	0.00	0.95	1.27	315	1576
overexpressed_gene	0.93	63.26	12.09	4.19	0.93	3.72	2.79	215	1588
tissue	22.70	6.13	5.52	3.69	1.84	11.04	25.77	163	950

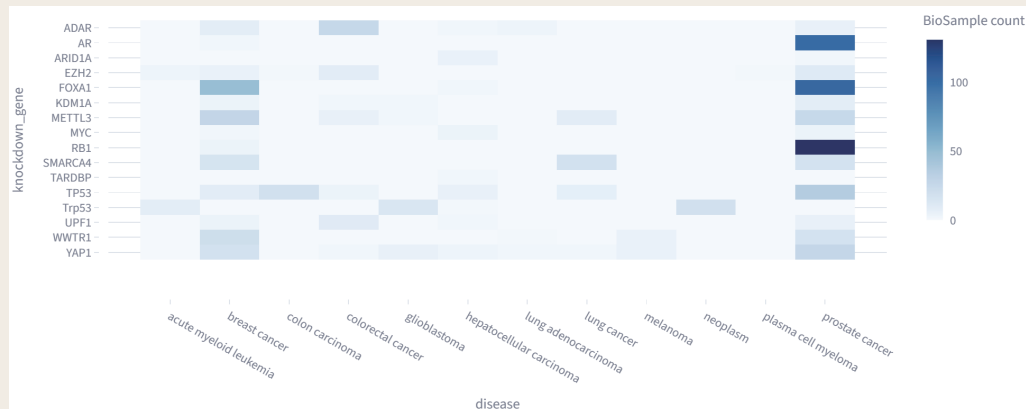
オントロジーへのフィードバック

抽出されたがマッピングされなかった単語は、オントロジーにおける語彙の不足を示している可能性がある
→ 各オントロジーへの新規ターム・シノニムの追加を提案

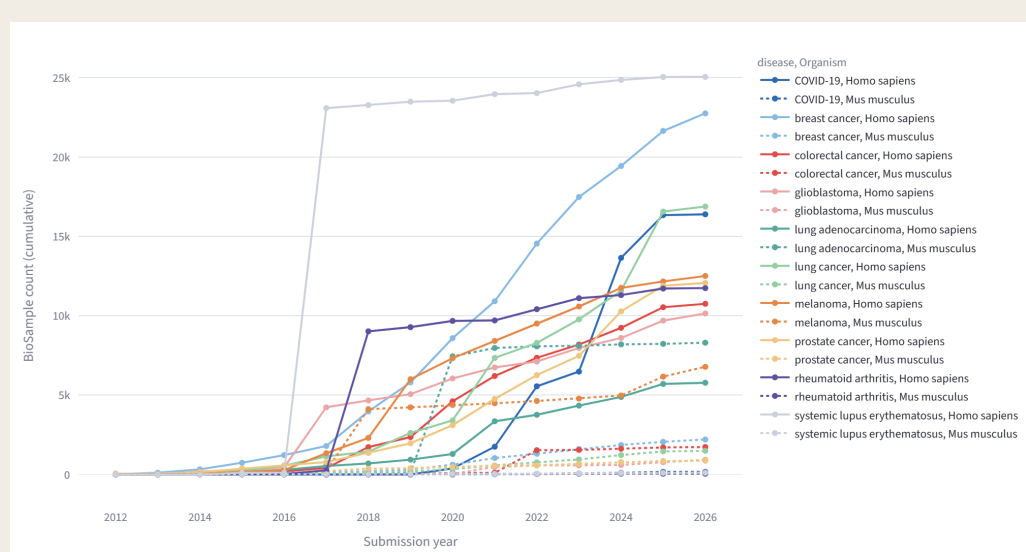
Word	Ontology	Suggestion	Status
needle	Plant Ontology	New narrow synonym of "vascular leaf"	Accepted
RPE	Cell Ontology	New exact synonym of "retinal pigment epithelial cell"	Accepted
EpILC	Experimental Factor Ontology	New term	Pending

可視化

2 つの属性を選択し、各条件におけるサンプル数をヒートマップとして可視化



特定の特徴を持つサンプル数の年ごとの推移



論文・レポジトリ

JOURNAL ARTICLE
Extraction of biological terms using large language models enhances the usability of metadata in the BioSample database
Shuya Ikeda, Zhaonan Zou, Hidemasa Bono, Yuki Moriya, Shuichi Kawashima, Toshiaki Katayama, Shinya Oki, Tazuo Ohta
GigaScience, Volume 14, 2025, gfa070, <https://doi.org/10.1093/gigascience/gfa070>
Published: 23 June 2025 Article history



<https://github.com/dbcls/bsllmner-mk2>
<https://github.com/sh-ikeda/bsllmner-evaluator>

展望

課題: 意味を完全に捉えることの難しさ
- 「分化した幹細胞株由来のサンプル」と「幹細胞株そのもののサンプル」の区別
- 「実験の焦点となる化合物名」と「実験プロトコル内で単に使用されているだけの化合物名」の区別
- BioProject 情報の活用が有効である可能性がある
- 数百万件のサンプルを処理する必要があるため、正確性と速度のバランスが重要

[1] Bernstein MN et al. (2017). MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive.

[2] Gonçalves RS et al. (2024) The text2term tool to map free-text descriptions of biomedical terms to ontologies.

[3] Zou Z et al. (2024). ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data.

