Using Literature-Based Knowledge Extraction to Develop a Disease-Specific Ontology for Skin Dysbiosis

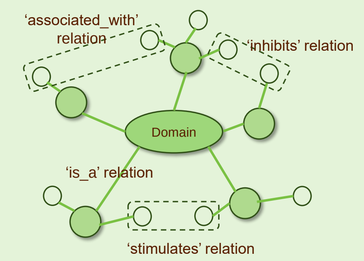
Cynthia J Grondin, Oleg Stroganov, Daniel Cooper, Yaw Nti-Addae
Rancho Biosciences, LLC, Rancho Santa Fe, CA USA

Abstract Highlights

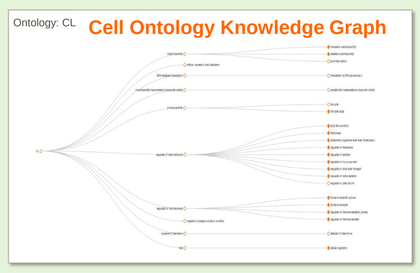
- **Custom ontology needed for Skin Dysbiosis**
  - Identify relevant sources from peer-reviewed literature using Rancho's DataCrawler
  - Extract tuples from the scientific literature with 'subject-predicate-object' triplets
  - Map, clean and harmonize terms to existing ontologies using Rancho's TMS tool
  - Incorporate MESH, NCIT, GO, CL and BTO, definitions, synonyms and identifiers from PubChem, ChEBI, UniProt, and NCBI Taxonomy
  - Add new terms with uniquely generated identifiers
  - Generate Biolink predicates using OpenAI's GPT-4 model to assist relationship classification
  - Build SD-Ontology using R from cleaned terms using linked ontology hierarchies and Biolink predicates
  - n-triples (*.nt), turtle (*.ttl) and OWL (*.OWL) formats using rdf conversion tool rapper and Protégé
- **Custom SD-Ontology**
  - **Provides insights on relationships among terms**
  - **Increases understanding of the pathophysiology of this disease**
  - **Advances strategies for improved treatment**
  - Provides proof-of-concept of successful utilization of Rancho's DataCrawler and TMS tools as innovative and effective methods to identify and prioritize peer-reviewed literature from public databases and clean disparate terms to facilitate knowledge extraction and organization to advance understanding of relationships across various disciplines

Importance of Ontologies and Knowledge Graphs

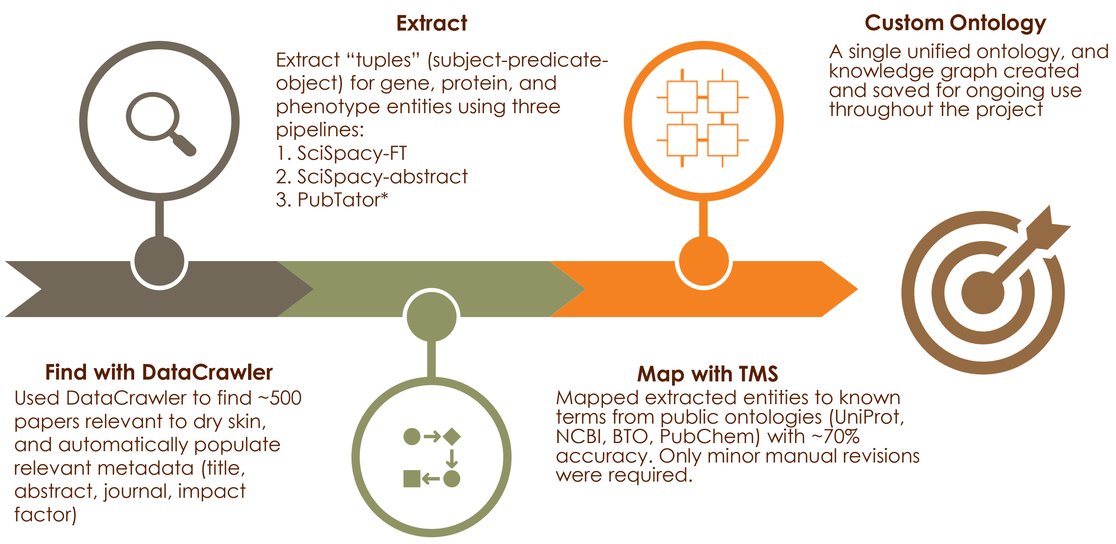
- **Ontologies**
  - Knowledge classification of a domain, where the relationships between concepts are formally defined and logically related
  - Text (human readable) and Logical (machine readable) definitions
  - Hierarchical arrangement of defined terms and relationships
  - Centralizes and harmonizes data
  - Facilitates computational reasoning
  - Promotes logical inferences and sophisticated data queries
  - Facilitates knowledge extraction for therapeutics
  - Over 700 biomedical ontologies in BioPortal
  - 262 ontologies in EMBL-EBI Ontology Lookup Service (8,584,670 classes, 44,633 properties, 687,082 individuals (updated Sep 2024)
  - Language formats in RDFS, OBO, or OWL
  - **Organize, Filter and Connect Data to Suggest New Relationships**



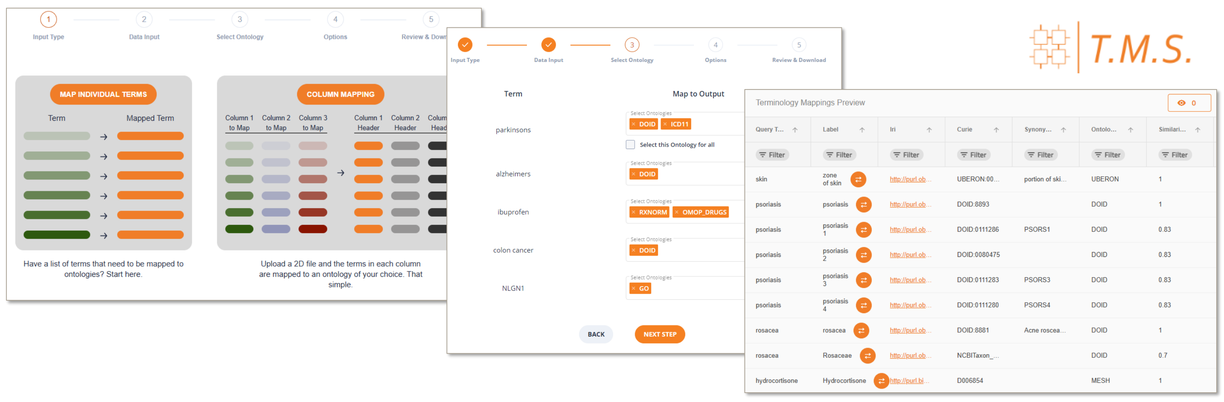
- **Knowledge Graphs (KGs)**
  - Coupling of hierarchical knowledge with direct relationships
  - Results in entity "Nodes" connected to relationships "Edges"
  - Stored and presented as triplet subject-->predicate-->object
  - While entities are primary information for other modalities, edges are key for KGs
  - Enables complex data and logical flow queries
  - **Enables data-centric hypothesis testing**



Ontology Development Workflow



3. Map to Existing Ontologies - Clean- Harmonize



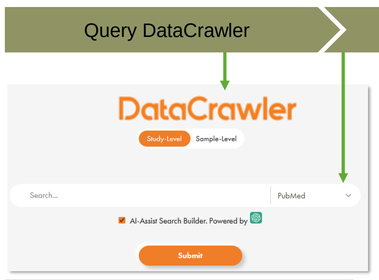
TMS Supported Ontologies and Dictionaries

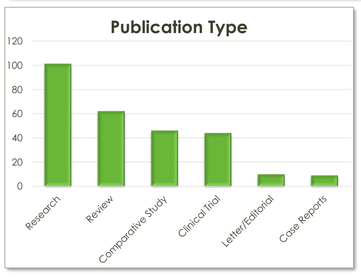
Disease/ Phenotype	Multi-Modal	Assay	Strain	Other
DOID HPO ICDO3 ICD10CM ICD11 MONDO ORDO	***BTO*** EFO FMA GARD ***MeSH*** ***NCIT*** OBA OMIM OMOP SNOMEDCT UMLS MedDRA	BAO CHMO OBI	***NCBITAXON*** RS	AFO CDISC ***GO*** HGNC LOINC PATO PR UO
Drug				
CHEBI RXNORM VO				
Tissue/Cell				
BTO ***CL*** CLO ***UBERON***				

Bold = Ontologies used in Skin Dysbiosis Ontology

1. Identify Relevant Literature

Query DataCrawler → Crawl PubMed → Analyze Output → QC Check





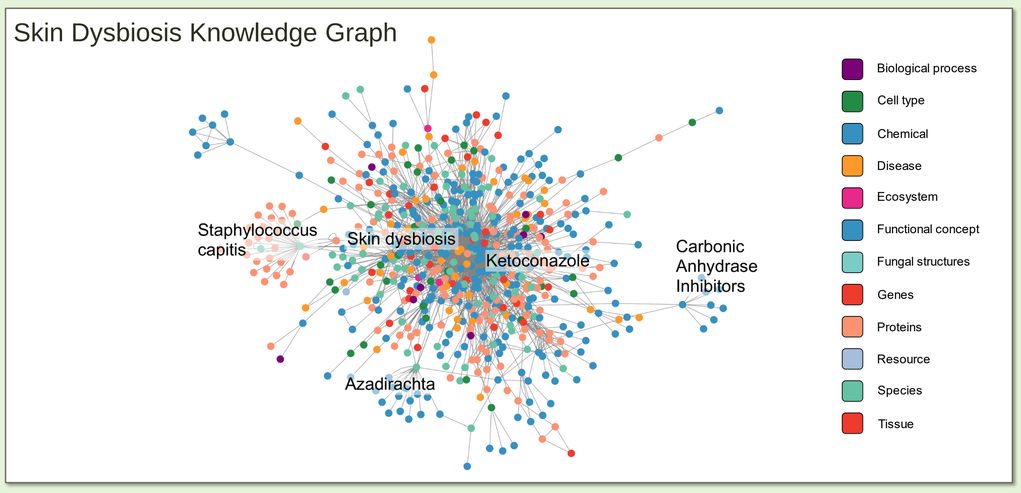
pubmed_id	Title	Abstract	Keywords	AIScore
30002814	Recent advance	The ichthyoses are a heteroger	Corneocyte lipid envelope | Ichthy	81.69
16788637	Inherited mouse	Nearly 100 mouse mutations have been described as causing some		81.48
32147743	Ichthyosis: A Ro	The understanding of monoge	ARCI | ceramides | congenital | e	81.35
6358130	[Genodermatos	Fourteen monogenic cutaneous disorders of man are compared to		81.35
21944047	Mutations in CS	Autosomal-recessive exfoliative ichthyosis presents shortly after		81.18
15607112	Mutations in X-l	X-linked ichthyosis is an inherited genetic disorder of the skin tha		81
15953139	Neonatal skin ba	The development of the human skin from intrauterine to extraute		80.88
31121896	3D-Organotypic	Atopic dermatitis (AD) is chara	atopic dermatitis | epidermal equ	80.76
1002598	Ichthyosis in tw	Ichthyosis in 2 dogs was characterized by extreme hyperkeratosis		80.62
28976107	S1 guidelines fo	Ichthyoses are a group of rare genetic skin disorders that pose nur		80.6
10754460	Comparative stu	The skin of the back of the ASR hairless miniature swine shows a r		80.58
12427540	Genetic reversio	Human epidermis is a squamous stratified epithelium whose inte		80.54
25803941	[Spontaneous m	Ichthyoses encompass a heterogeneous group of genodermatoses		80.52
11428941	A novel in vivo r	BACKGROUND/AIMS: The lack of a suitable, validated animal mod		80.37
34996433	DSP missense va	Ichthyosis describes a localized	Cattle | Corneal ulcers | Dermato	80.29
39136317	Heterozygous D	Ichthyoses comprise a large he	Canis lupus familiaris | dermatolo	80.28
17164798	Filaggrin mutati	Ichthyosis vulgaris (IV) is the most common hereditary disorder of		80.24
15344228	Defective repair	Most forms of the human hereditary disease xeroderma pigmenta		80.22
2141868	Papulosquamou	We review the spectrum of papulosquamous disorders in the sett		80.2
36967672	SDR9C7 missens	Ichthyoses represent a hetero	Canis lupus familiaris | animal mo	80.12
2384688	Effects of aging	Amino acid compositions of skin samples from young and old subj		80.1
21709432	Dermocosmetic	Dry skin is associated with a disturbed skin barrier and reduced fo		80.08
28249031	A de novo variar	Ichthyoses are a heterogeneous group of inherited cornification d		80.07

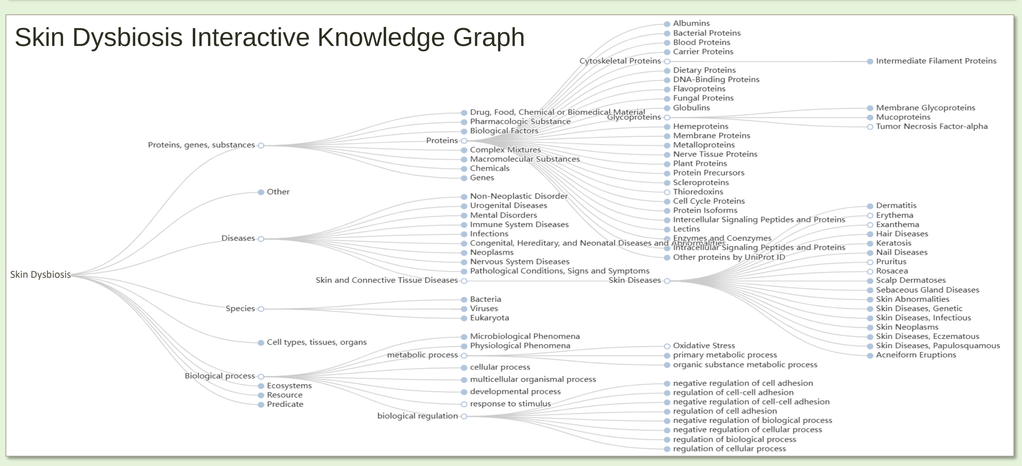
All relevant papers were found using DataCrawler + AI assist.

- Use functions and macros to clean/remove non-ascii and non-printable characters extracted from literature

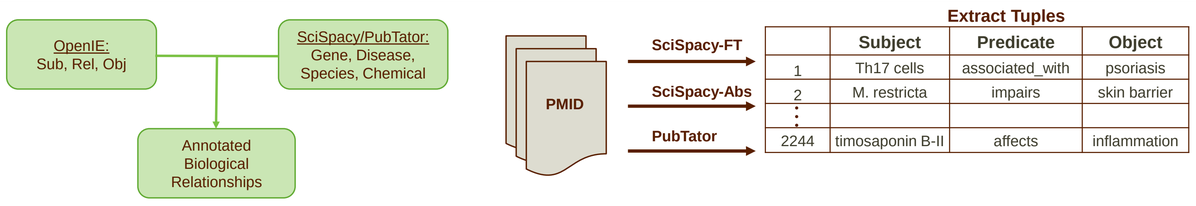
4. Build Ontology and Knowledge Graphs

- Skin Dysbiosis Ontology was built in R, with ontology represented in n-triples format (*.nt)
- (*.nt) format was converted to turtle (*.ttl) and owl (*.owl) formats using rdf conversion tool *rapper* and *Protégé* UI
- Skin Dysbiosis Ontology combined standard ontologies and an additional ontology with terms which were missing in standard ontology





2. Extract Tuples



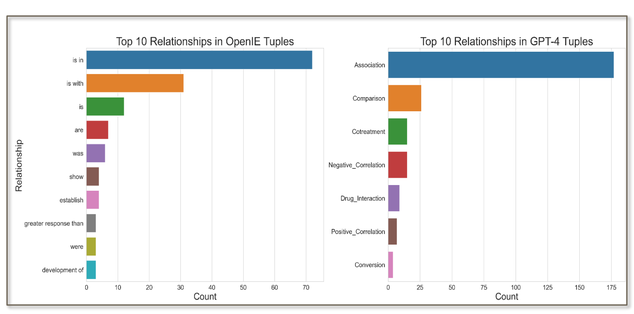
Extract Tuples

	Subject	Predicate	Object
1	Th17 cells	associated_with	psoriasis
2	M. restricta	impairs	skin barrier
⋮			
2244	timosaponin B-II	affects	inflammation

Compare OpenIE and other non-GPT based NLP tools to GPT-4 for biological/biomedical relationship extraction

- GPT-4 pipeline with OpenAI function calling and Few-Shot prompting is cautious in labeling relationships and entities as seen in the high precision
- In this task, biggest error is in recall; improve by allowing the models to extract multiple relationships per sentence, fine-tuning, prompt engineering

Metric	RE	Sub_NER	Obj_NER	NER
Precision	0.86	0.94	0.97	0.96
Recall	0.20	0.46	0.44	0.45
F1	0.32	0.62	0.61	0.62



Top Chemicals and Genes in Skin Dysbiosis Extracted Tuples