TransMed consortium in Translator Translator and Fanconi Anemia

CG Chute, Hopkins
Maureen Hoatlin, OHSU
Julie McMurry, OHSU
Chris Mungall, Lawrence Berkeley
Melissa Haendel, OHSU

Chris Chute
Clinical terminologies,
frameworks & text mining

Melissa Haendel
Developmental bio, model
organisms & ontologies

Chris Mungall
Semantic engineering &
similarity algorithms

Hongfang Liu
Text mining, pathway
modeling, data
normalization

Ben Good
Crowdsourcing,
community curation,
semantic web

Andrew Su
Data integration, data
services

Peter Robinson
Clinical ontologies &
diagnostic algorithms,
Medical genetics

Shannon McWeeney
cancer imaging, omics,
drugs, flow cytometry,
machine learning

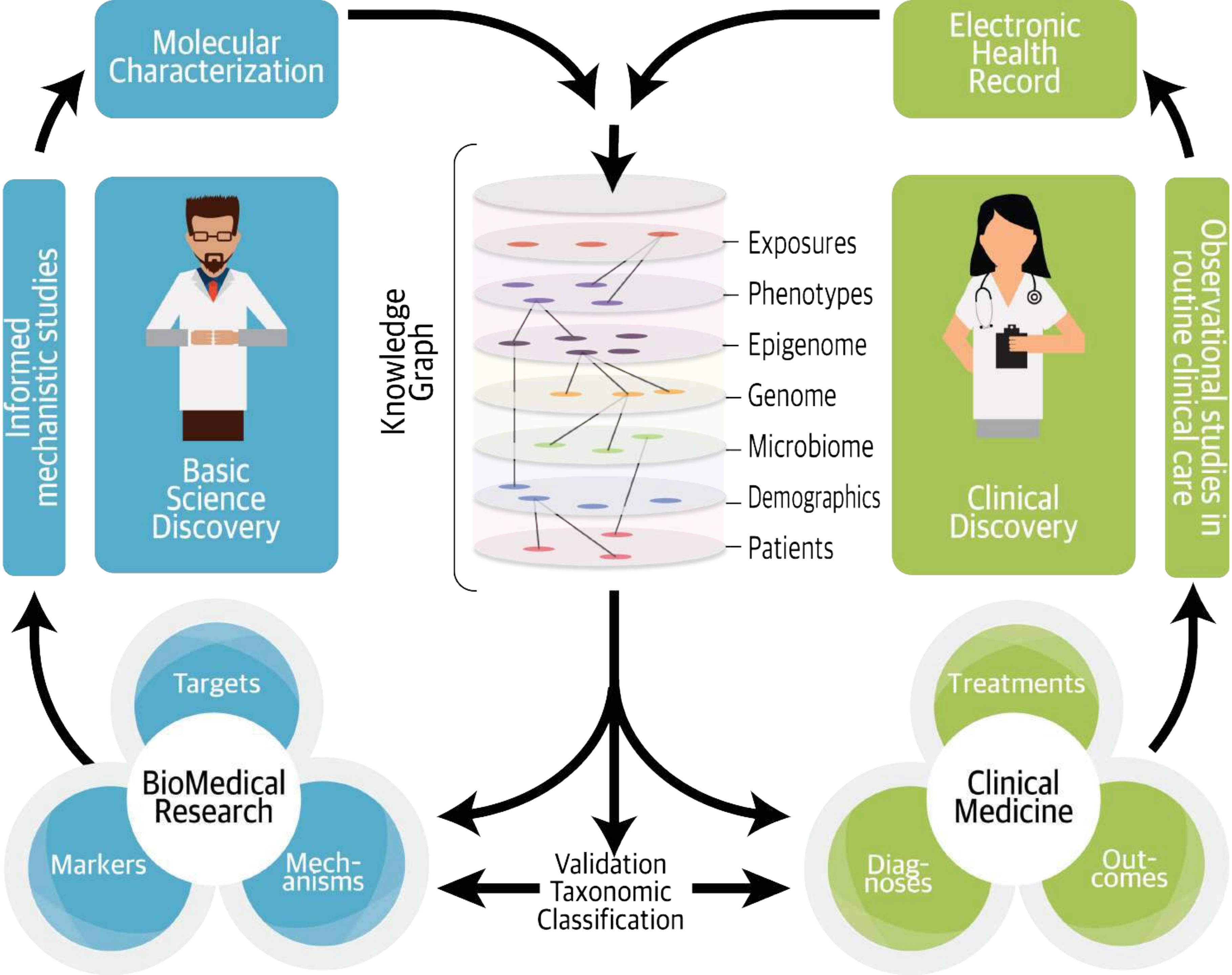
Maureen Hoatlin
Fanconi Anemia, rare
disease, biochemistry,
basic research

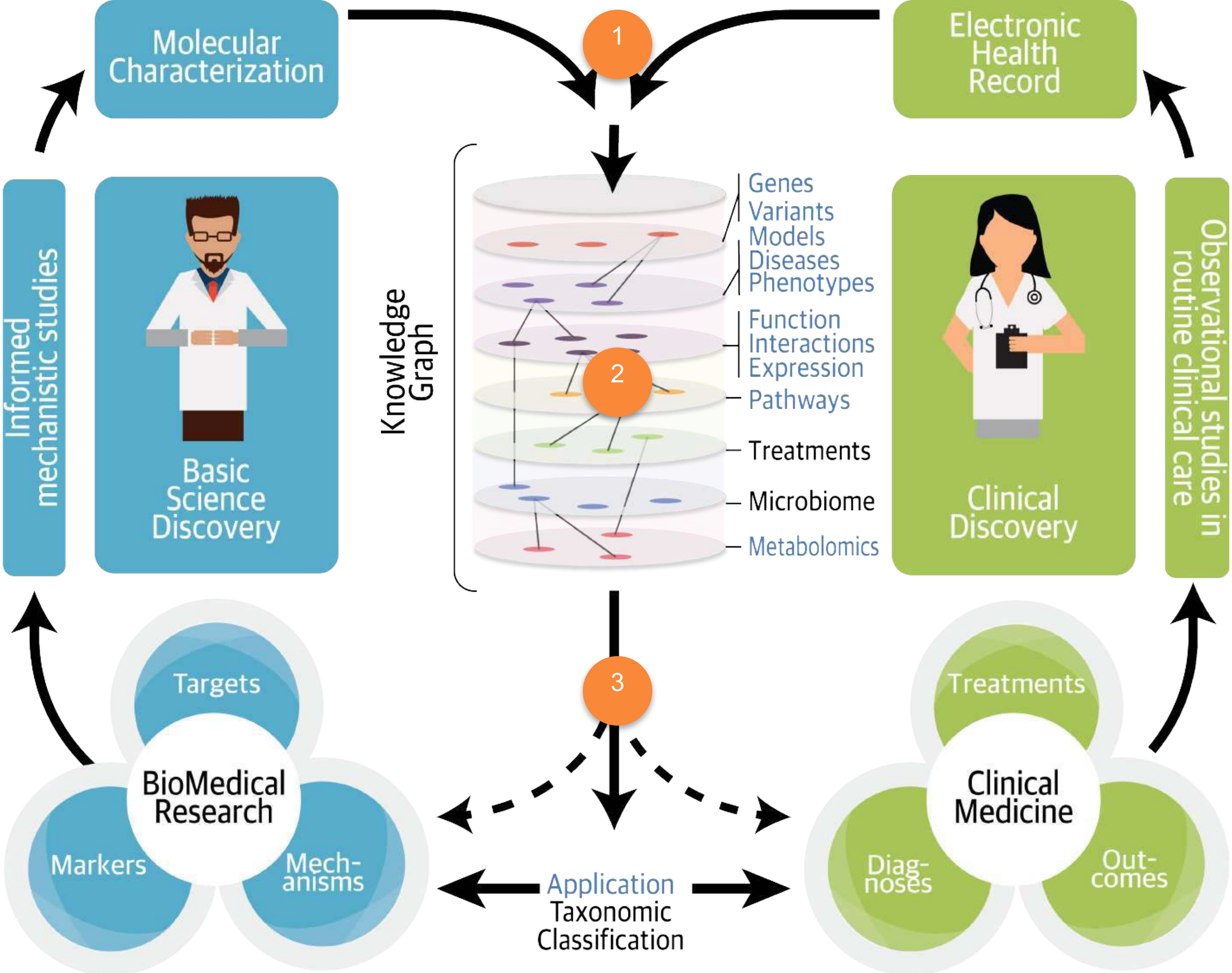
Guoqian Jiang
Clinical data standards &
Interoperability, semantic
modeling & validation

David Koeller
Clinician of inborn
errors of metabolism,
Undiagnosed diseases

Casey Overby
Knowledge-based methods
& evaluation of precision
medicine applications

Julie McMurry
Project Management,
User Experience,
Public Health


Molecular Characterization
Electronic Health Record
Informed mechanistic studies
Basic Science Discovery
Knowledge Graph
Exposures
Phenotypes
Epigenome
Genome
Microbiome
Demographics
Patients
Clinical Discovery
Observational studies in routine clinical care
Targets
BioMedical Research
Markers
Mech-anisms
Validation Taxonomic Classification
Treatments
Clinical Medicine
Diag-noses
Out-comes

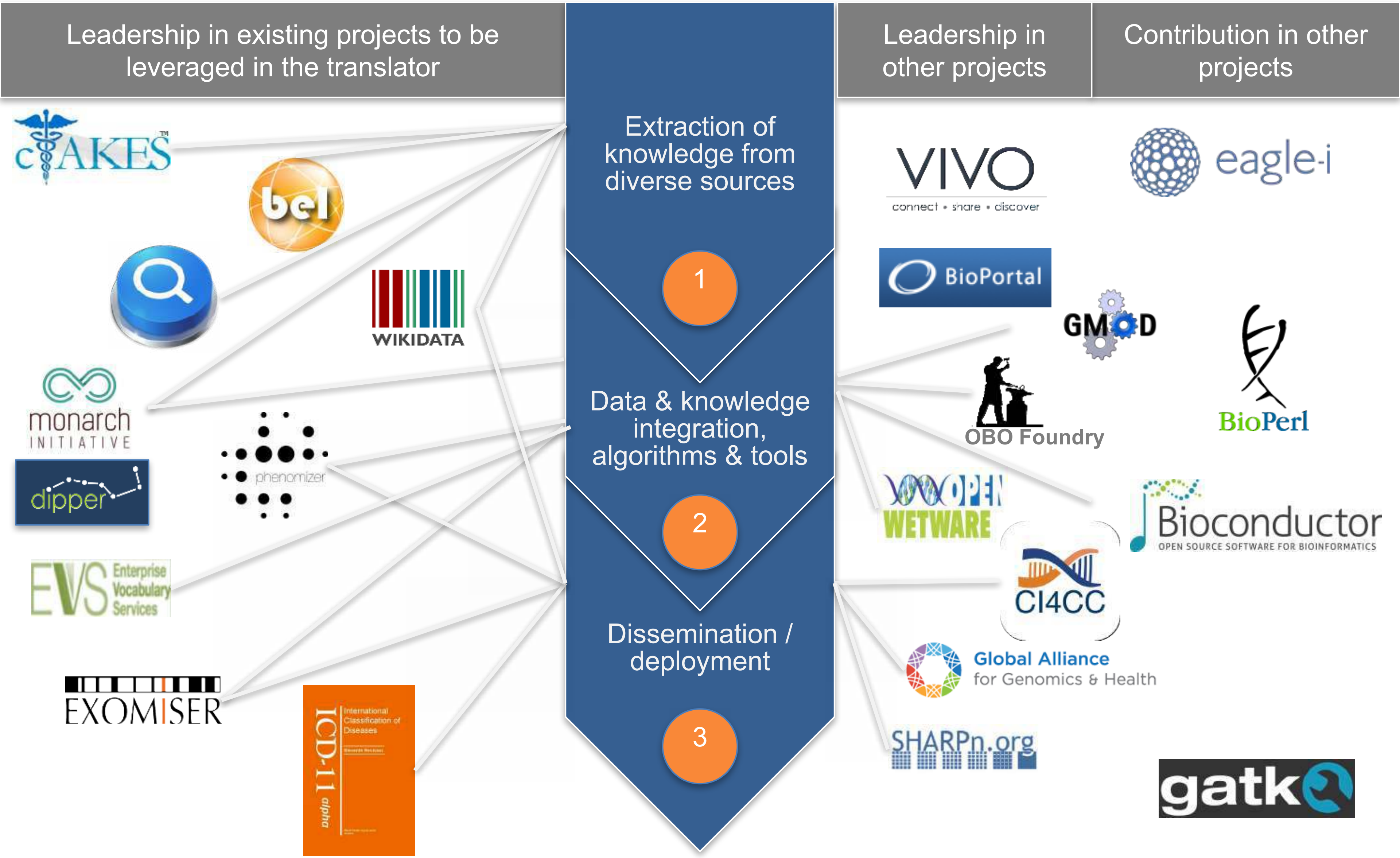

Molecular Characterization
1
Electronic Health Record
Informed mechanistic studies
Basic Science Discovery
Knowledge Graph
Genes
Variants
Models
Diseases
Phenotypes
Function
Interactions
Expression
2
Pathways
Treatments
Microbiome
Metabolomics
Clinical Discovery
Observational studies in routine clinical care
Targets
BioMedical Research
Markers
Mech-anisms
3
Application
Taxonomic Classification
Treatments
Clinical Medicine
Diag-noses
Out-comes

Deep open source experience, each step of the way

Leadership in existing projects to be leveraged in the translator

Leadership in other projects

Contribution in other projects



Extraction of knowledge from diverse sources

1

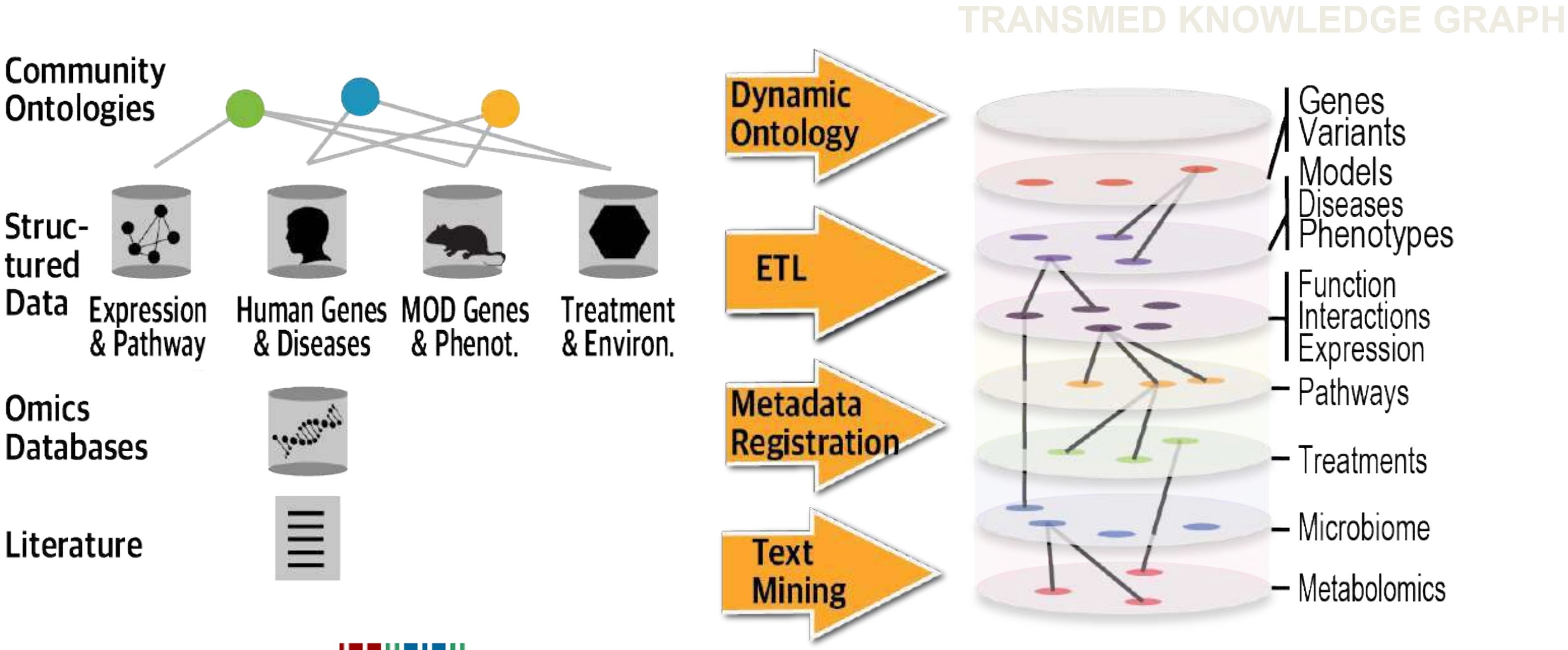
Data & knowledge integration, algorithms & tools

2

Dissemination / deployment

3

(1) Extraction of knowledge from diverse sources



kBOOM

EVS Enterprise Vocabulary Services

WIKIDATA

My*

dipper

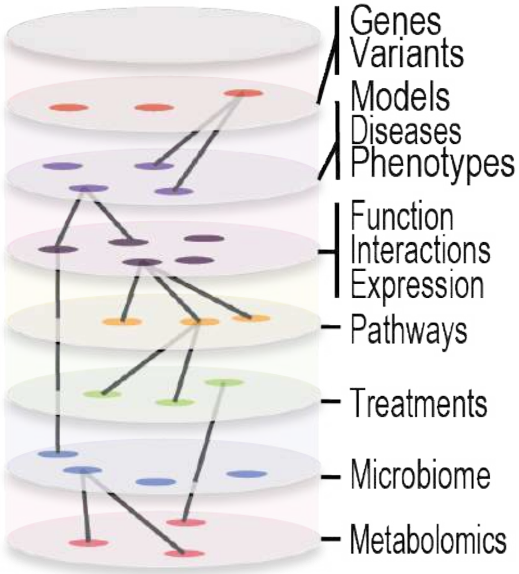
bel

cTAKES™

- Tools to integrate vocabularies
- Tools to identify equivalencies:
  Identifier and synonym alignment
  Conceptual alignment based upon logic determinations and prior probabilities
- Text mining for clinical concepts and pathway fragments

(2) Data & knowledge integration, algorithms & tools

TRANSMED KNOWLEDGE GRAPH



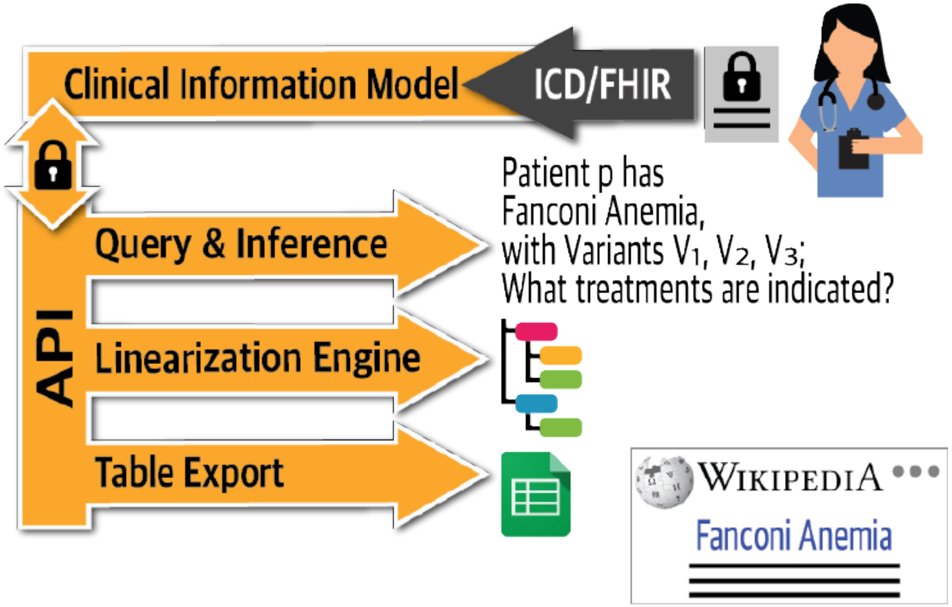
GRAPH INFERENCE

- OWL-based reasoning over large graphs
- BELpathway ‘chains of causation’
- Probabilistic inference across disease-phenotype-gene
    - Bayesian Ontology Query Algorithm (Boqa)

GRAPH QUERY

- Query related entities within/across species, sources
- Query similar sets of entities (OWLsim)
    - Sets of phenotypes
    - Expression patterns
    - Pathway modules

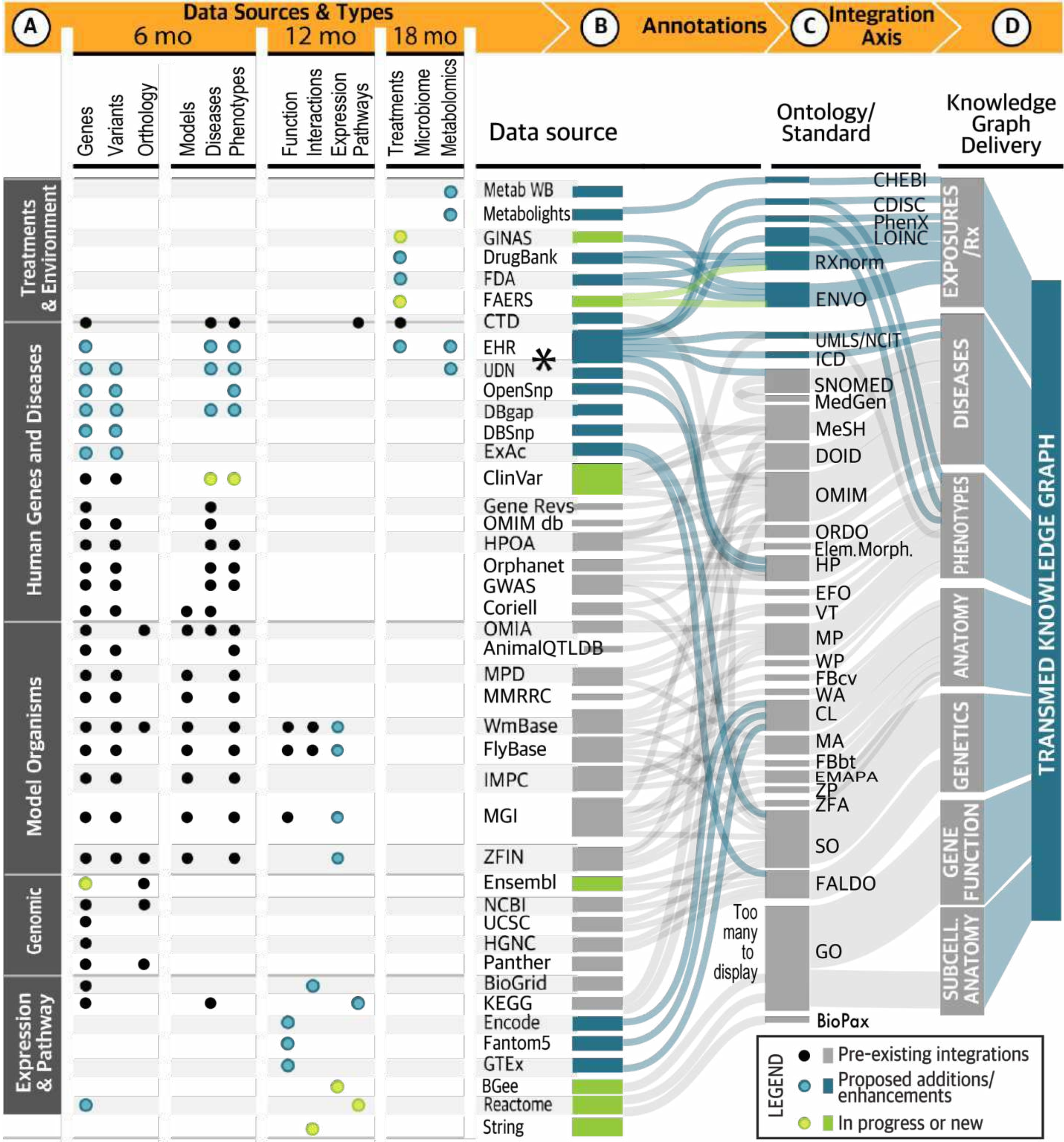
(3) Dissemination, deployment & validation



EVALUATION

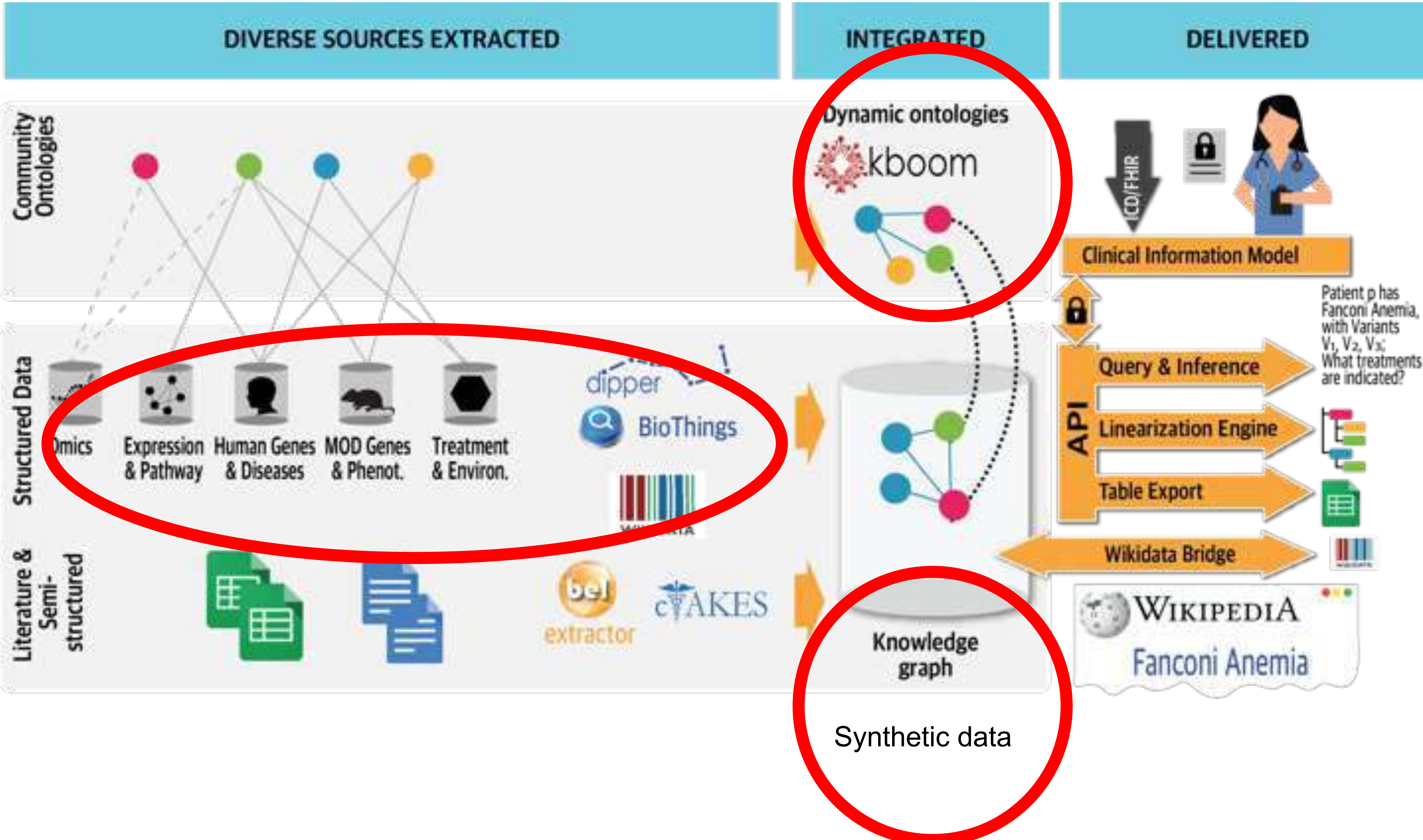
- Purpose built for clinicians, researchers, & bioinformaticians
- TransMed outputs based on condition set and competency questions are iteratively compared against current phenomonological nosology
- Comparisons of integrated delivery of mechanistic modules against single source inquiries
- T-Score measuring connectivity of clinical and basic sources

Monarch (via Dipper) ingest so far in TransMed

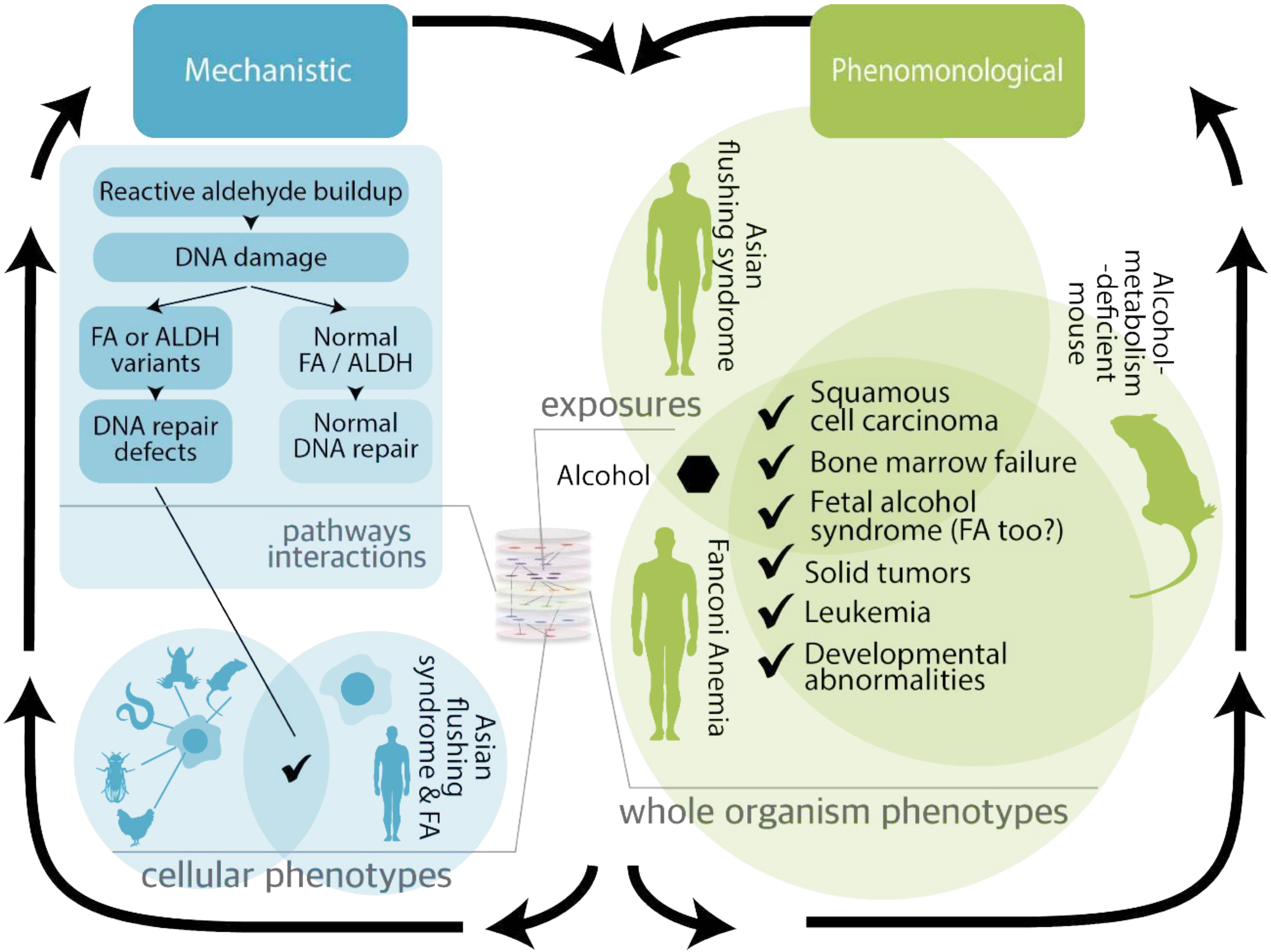


http://bit.ly/monarch-data-dashboard

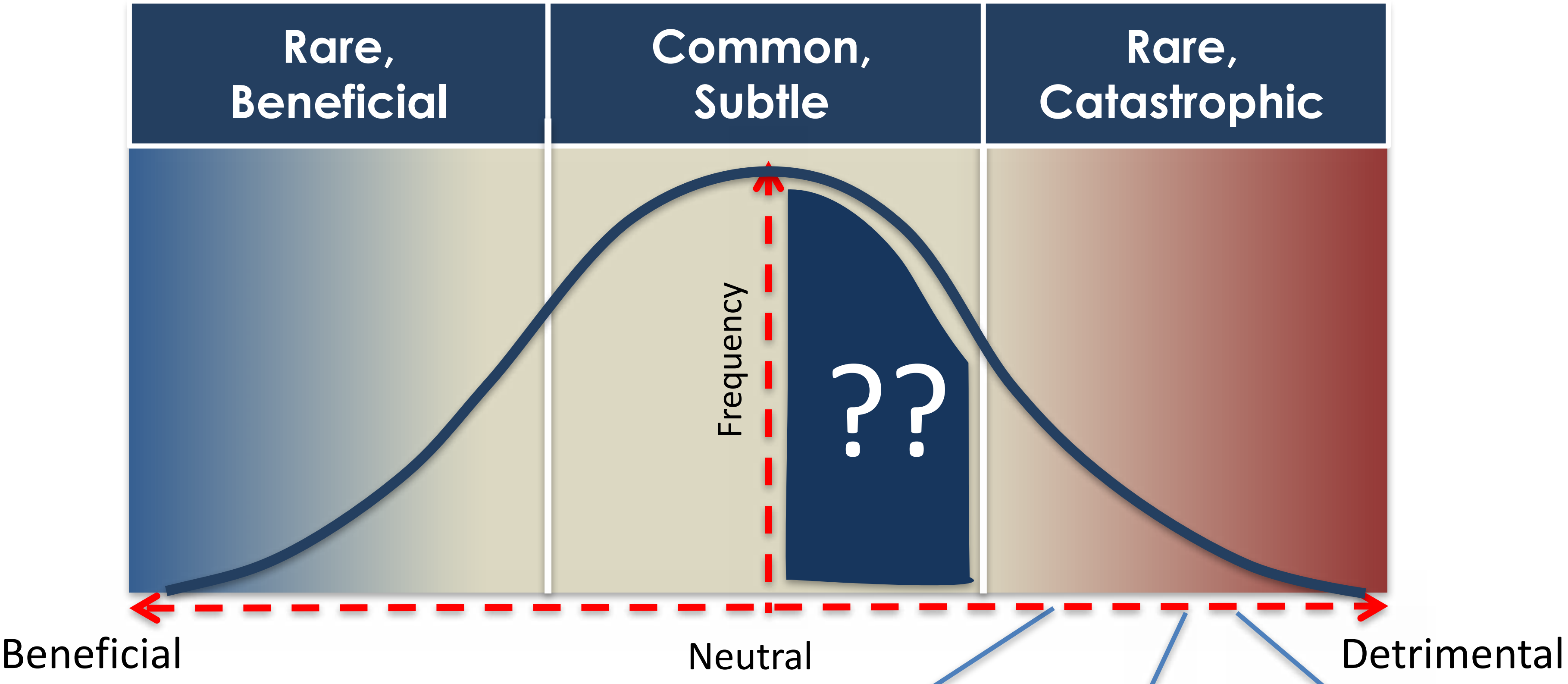
Overview of TransMed components



Fanconi mechanistic inference



Common mechanistic underpinnings of rare & common/complex disease

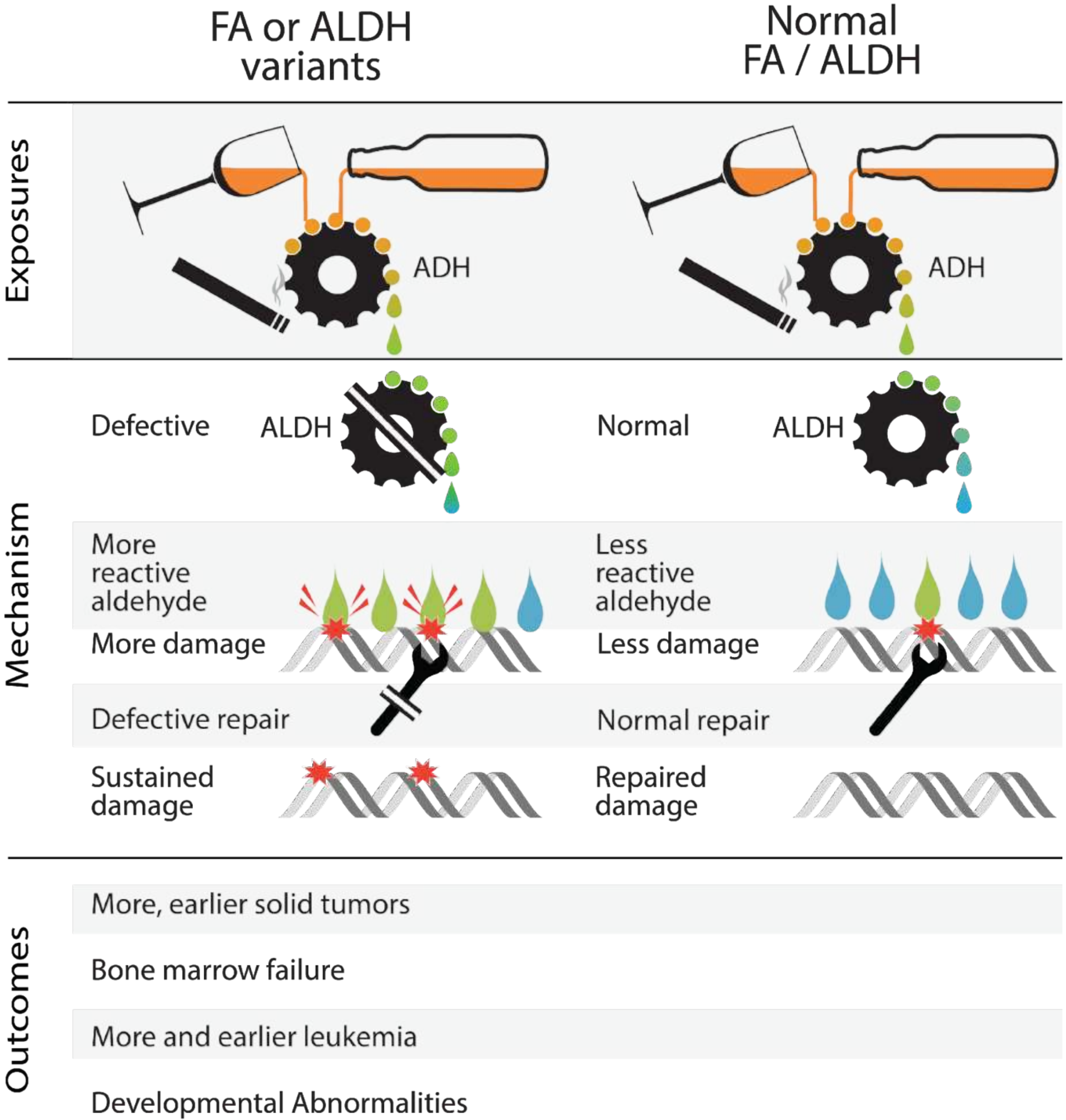


Acetaldehyde metabolism mutations:
500 Million People Affected

Fanconi Anemia:
1 in 160,000 individuals worldwide

ALDH KO: ?

The Fanconi Anemia mechanism



FA or ALDH variants

Normal FA / ALDH

Exposures

Mechanism

Defective ALDH

Normal ALDH

More reactive aldehyde

Less reactive aldehyde

More damage

Less damage

Defective repair

Normal repair

Sustained damage

Repaired damage

Outcomes

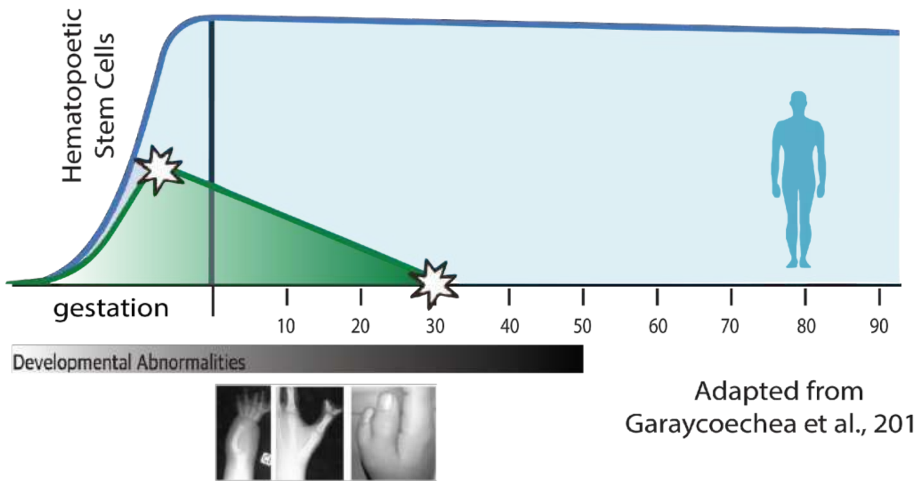
More, earlier solid tumors

Bone marrow failure

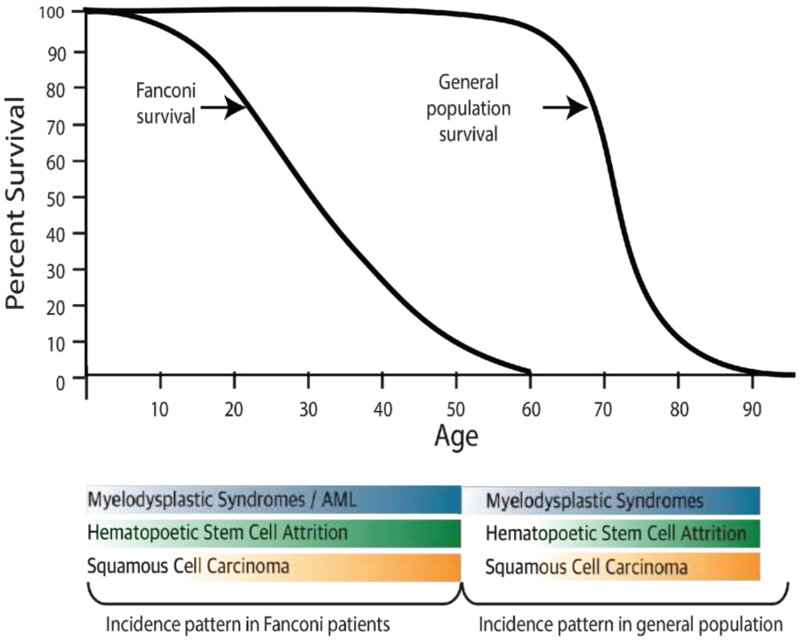
More and earlier leukemia

Developmental Abnormalities

Fanconi Phenotypes



Adapted from
Garaycoechea et al., 2012



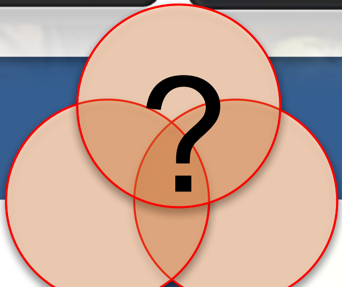
1) Environmental risk factors

Alcohol

Smoking

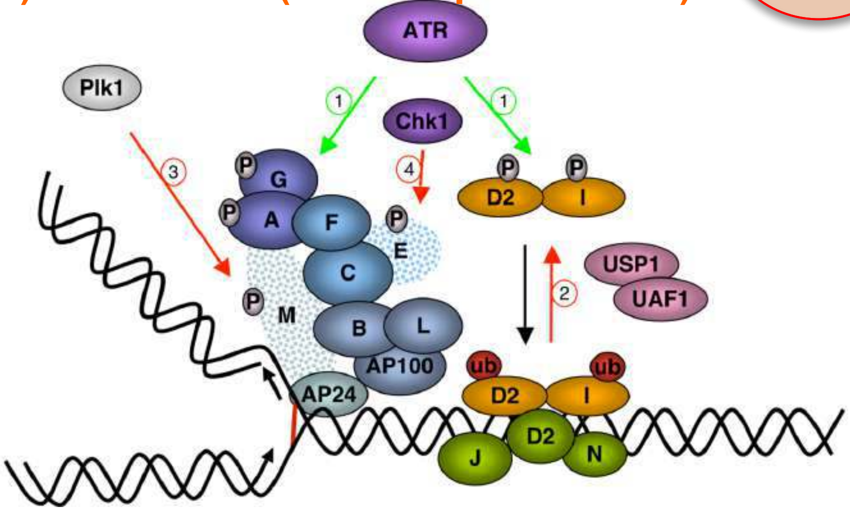
Fermented foods

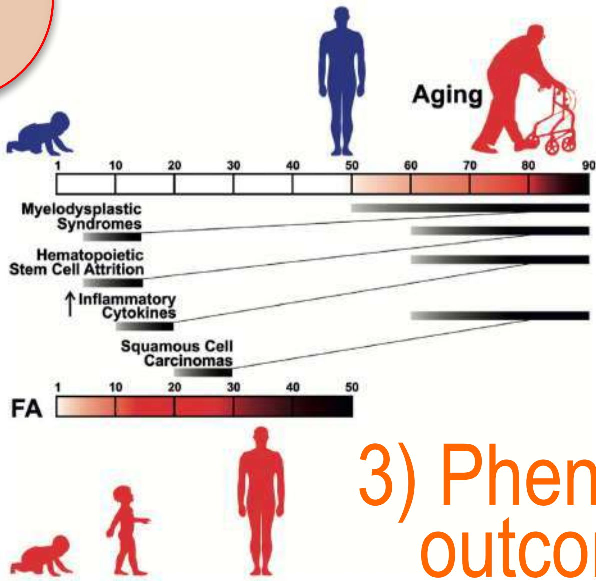
Proximity to aldehyde dumping site



bit.ly/fanconi-cq

2) Genes (22 implicated)



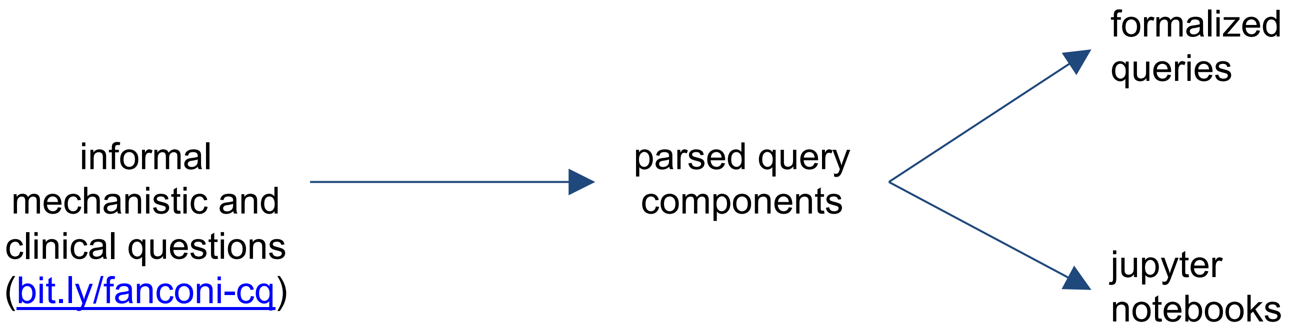


3) Phenotypic outcomes

		FA genes	
		KO	WT
aldh2	KO	mostly embryonic lethal	full FA disease phenotype profile*
	WT	partial phenotype profile (hallmark chromosomal defects)*	normal*

*in presence of moderate ethanol to pregnant mouse

FA competency question corpus

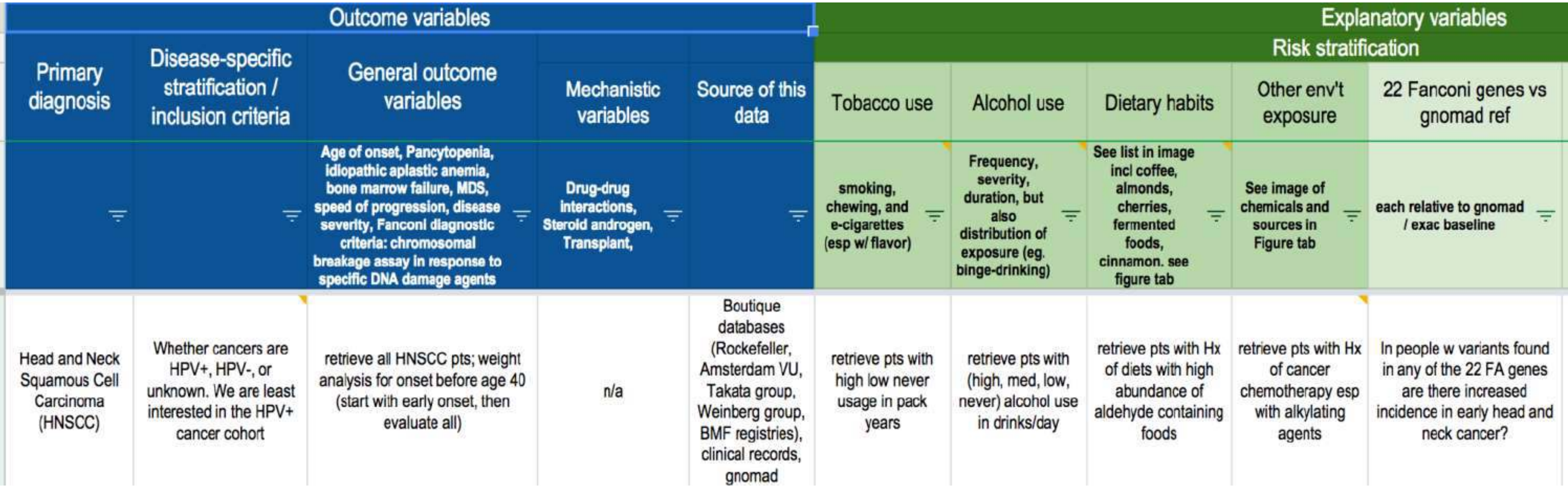


informal mechanistic and clinical questions (bit.ly/fanconi-cq)

parsed query components

formalized queries

jupyter notebooks



Outcome variables					Explanatory variables				
					Risk stratification				
Primary diagnosis	Disease-specific stratification / inclusion criteria	General outcome variables	Mechanistic variables	Source of this data	Tobacco use	Alcohol use	Dietary habits	Other env't exposure	22 Fanconi genes vs gnomad ref
		Age of onset, Pancytopenia, idiopathic aplastic anemia, bone marrow failure, MDS, speed of progression, disease severity, Fanconi diagnostic criteria: chromosomal breakage assay in response to specific DNA damage agents	**Drug-drug interactions, Steroid androgen, Transplant,**		**smoking, chewing, and e-cigarettes (esp w/ flavor)**	**Frequency, severity, duration, but also distribution of exposure (eg. binge-drinking)**	**See list in image incl coffee, almonds, cherries, fermented foods, cinnamon. see figure tab**	**See image of chemicals and sources in Figure tab**	**each relative to gnomad / exac baseline**
Head and Neck Squamous Cell Carcinoma (HNSCC)	Whether cancers are HPV+, HPV-, or unknown. We are least interested in the HPV+ cancer cohort	retrieve all HNSCC pts; weight analysis for onset before age 40 (start with early onset, then evaluate all)	n/a	Boutique databases (Rockefeller, Amsterdam VU, Takata group, Weinberg group, BMF registries), clinical records, gnomad	retrieve pts with high low never usage in pack years	retrieve pts with (high, med, low, never) alcohol use in drinks/day	retrieve pts with Hx of diets with high abundance of aldehyde containing foods	retrieve pts with Hx of cancer chemotherapy esp with alkylating agents	In people w variants found in any of the 22 FA genes are there increased incidence in early head and neck cancer?

Example query

What pathways are uniquely targeted by stem cell therapy pre-conditioning drugs that are well- vs poorly- tolerated by FA patients?

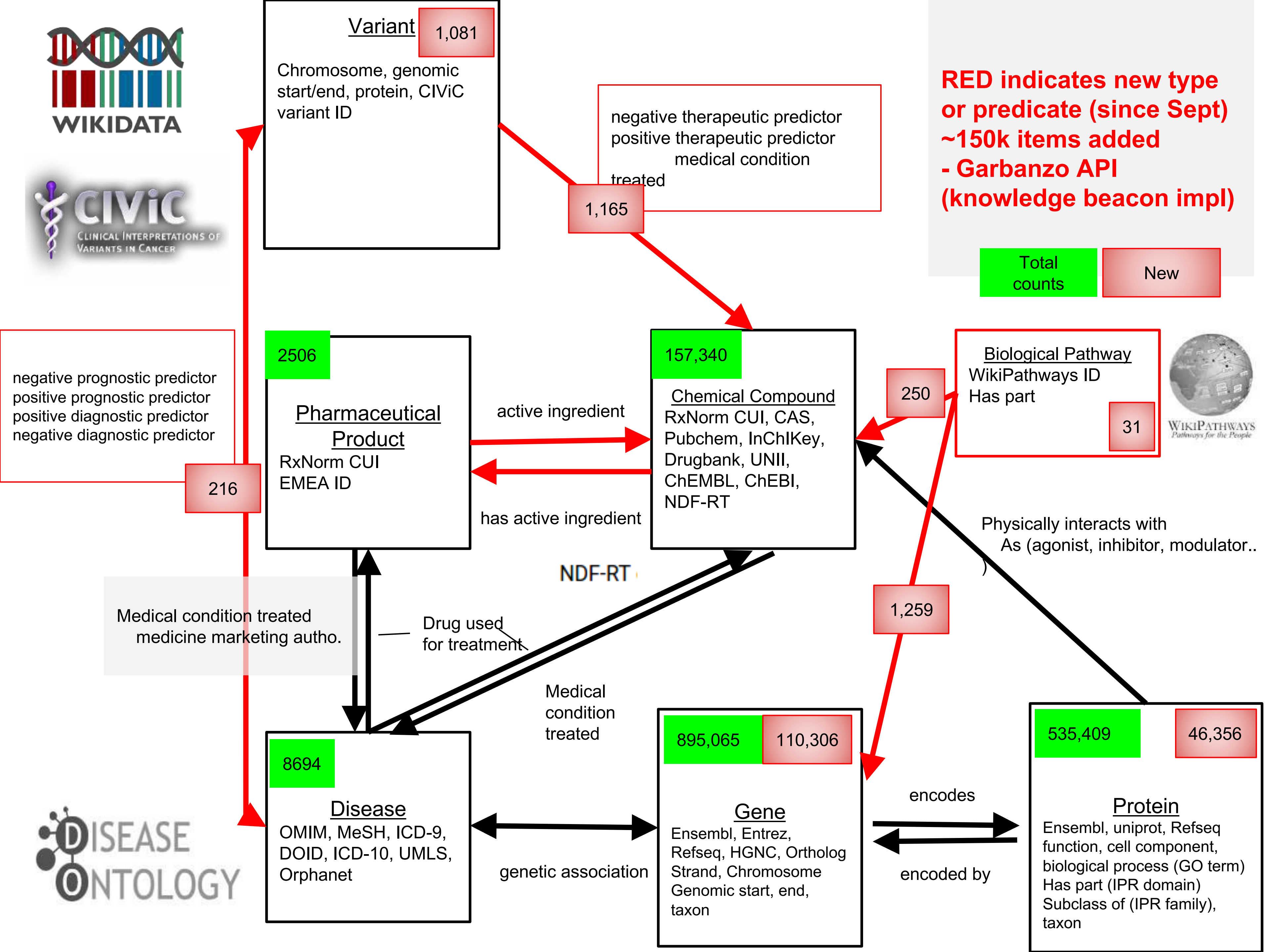
Query Path with Semantic Types

Drug	- [molecularly_controls] ->	**Protein**	- [encoded_by] ->	**Gene**	- [member_of] ->	**Pathway**
|		|		|		|
drugbank:DB01073 -	[molecularly_controls] ->	**Uniprot:P09884** -	[encoded_by] ->	**HGNC:9173**	- [member_of] ->	**WikiPath:WP2446**
(Fludarabine)		***(POLA1 protein)***		***(POLA1 gene)***		***(Rb Pathway)***

Data Types and Sources:

1. **Drug-Protein Interactions**
   a. from DrugBank via BioThings API
   b. from DGIdb, via DGIdb API
2. **Protein-Gene Associations**
   a. from Ensembl via Ensembl API or BioLink API
   b. from Uniprot via Wikidata API
3. **Gene-Pathway Membership**
   a. from Wikipathway via Wikidata API
   b. from Reactome via BioLink API

(Query implemented in OrangeQ2.4_Drug_Gene_Pathway Jupyter notebook)



RED indicates new type or predicate (since Sept)
~150k items added
- Garbanzo API (knowledge beacon impl)

🟩 Total counts
🟥 New

Variant 1,081
Chromosome, genomic start/end, protein, CIViC variant ID

negative therapeutic predictor
positive therapeutic predictor
medical condition treated
1,165

negative prognostic predictor
positive prognostic predictor
positive diagnostic predictor
negative diagnostic predictor
216

2506
Pharmaceutical Product
RxNorm CUI
EMEA ID

active ingredient
has active ingredient

157,340
Chemical Compound
RxNorm CUI, CAS, Pubchem, InChIKey, Drugbank, UNII, ChEMBL, ChEBI, NDF-RT

250
Biological Pathway
WikiPathways ID
Has part
31

Physically interacts with
As (agonist, inhibitor, modulator..)

1,259

NDF-RT

Medical condition treated
medicine marketing autho.

Drug used for treatment

Medical condition treated

8694
Disease
OMIM, MeSH, ICD-9, DOID, ICD-10, UMLS, Orphanet

genetic association

895,065 110,306
Gene
Ensembl, Entrez, Refseq, HGNC, Ortholog Strand, Chromosome Genomic start, end, taxon

encodes
encoded by

535,409 46,356
Protein
Ensembl, uniprot, Refseq function, cell component, biological process (GO term)
Has part (IPR domain)
Subclass of (IPR family), taxon

Observations

- Biomedical databases are hugely rich and growing
- Information and knowledge remains latent
- Integration of these resources is a necessary first step
- Query and exploration frameworks follow
- TransMed is one of the *pilot efforts* in NCATS Translator to develop and demonstrate infrastructure