CLINICAL DATA FOR GENOMIC ANALYSIS

Séminaire BioinfoDiag 2024
2024/05/15

Kévin Yauy

MD in Medical Genetics

PhD in Bioinformatics & ML

LAPEYRONIE
ARNAUD DE VILLENEUVE
LA COLOMBIÈRE
SAINT ELOI
GUI DE CHAULIAC
ANTONIN BALMÈS
BELLEVUE
CENTRE DE SOINS DENTAIRES
SITE EUROMÉDECINE
CENTRE ADMINISTRATIF ANDRÉ BÉNECH

MONTPELLIER CHU CENTRE HOSPITALIER UNIVERSITAIRE

ERIOS MONTPELLIER

Université de Montpellier Faculté de Médecine Montpellier-Nîmes

Diagnostic challenge in rare diseases

A (very) large knowledge

- There are more than **10000** rare diseases
- A geneticist only has **1** brain

Linkeropathies

Ciliopathies

Kabuki Syndrome

Familial mediterranean fever

22q11.2 Syndrome

Williams-Beuren Syndrome

Increasing numbers of discoveries

2023	Jan	Feb	Mar	Apr	May	Jun
New	54	38	41	53	48	39
Updated	422	257	531	415	310	380

Jul	Aug	Sep	Oct	Nov	Dec
20	42	32	37	46	37
415	449	316	411	306	334

OMIM.org

MONTPELLIER CHU CENTRE HOSPITALIER UNIVERSITAIRE

Physicians needs computer helps since long time…

Clinical geneticists were early adopters of software as clinical decision support

"For precision medicine human and artificial intelligence need to join efforts." (Peter Krawitz)

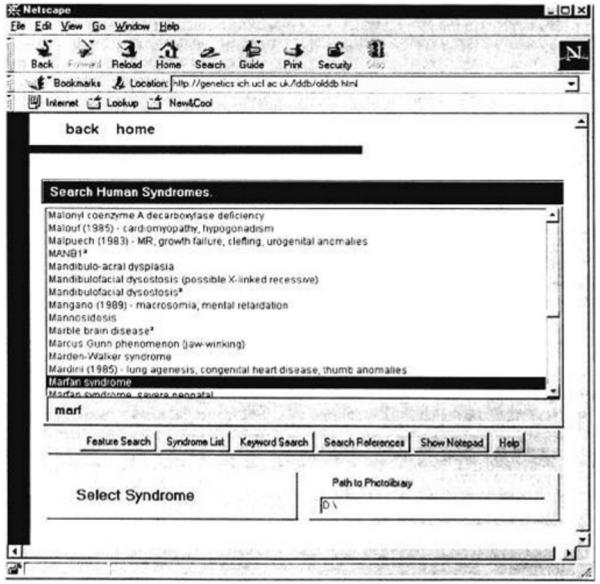
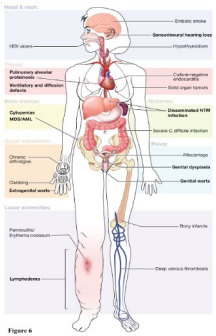
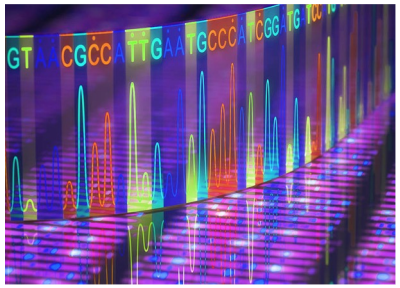


Fig. 1 Syndrome list search screen.

Physicians : how we used to do



Physicians select a targeted sequencing analysis according to their knowledge



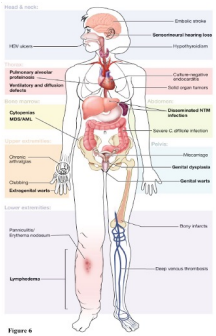
"Phenotype - first"

1. Recognition of "gestalt" and 2. association of symptoms : "phenotype"

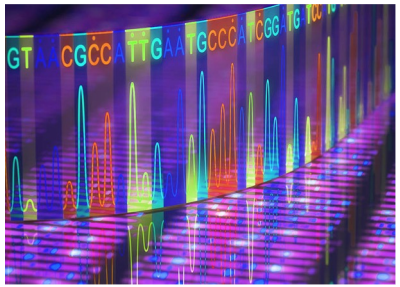
3. DNA analysis/sequencing : "genotype"

Which rare diseases do I think ? (precise query)

Physicians : today in the genomic medicine era



Look into the complete genome to identify variants that causes patient's disease: precision medicine



"Genotype - first"

1. Recognition of "gestalt" and 2. association of symptoms : "phenotype"

3. DNA analysis/sequencing : "genotype"

Which indication do I choose when I prescribe ? (large query)

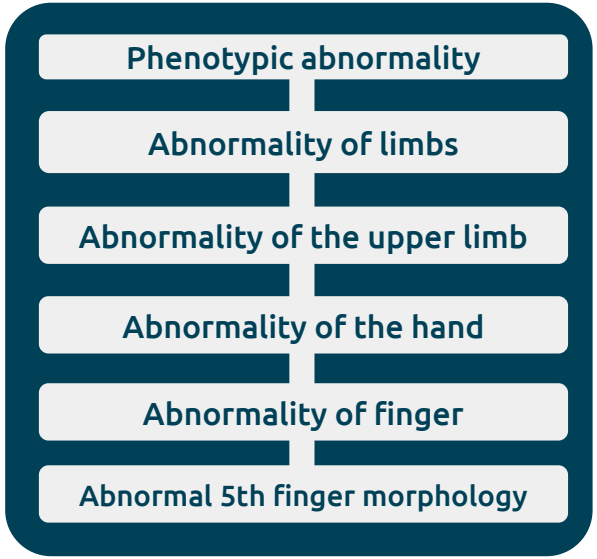
The need for computational phenotype analysis

Phenotyping :

Physicians identification of characteristics deviating from normal morphology, physiology, and behavior

Ontology :

Standardized symptom terms linked according to the human development architecture



human phenotype ontology

Computational phenotype analysis :

Identification of diagnostic hypothesis, clinically relevant groups of patients, …

A common language between human and machine is necessary for computer support (for now)

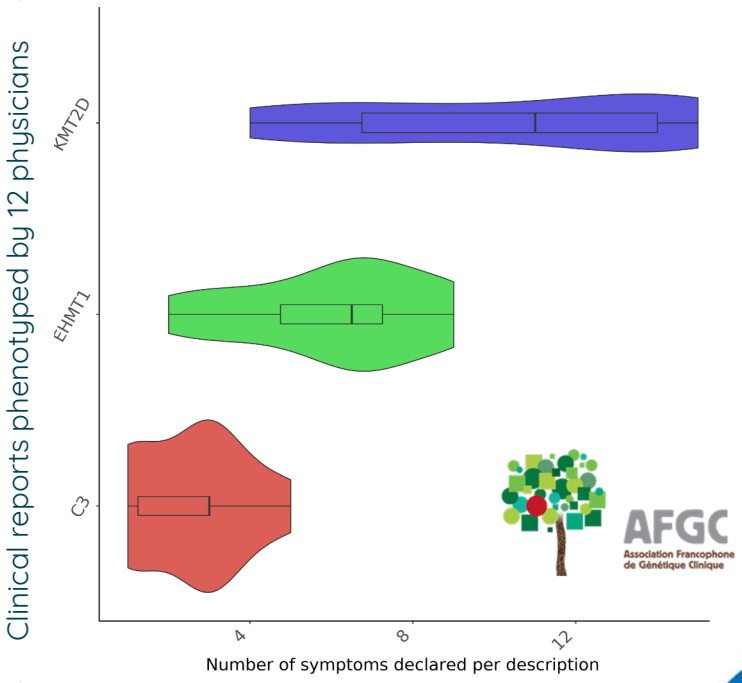
MONTPELLIER CHU CENTRE HOSPITALIER UNIVERSITAIRE

Very heterogeneous and sparse phenotyping practices…

- Phenotypes described in a cohort of **1686 patients**
- **47%** of HPO terms declared only once

Three clinical reports described by 12 physicians

Physicians phenotyping diversity explains the observed heterogeneity



MONTPELLIER CHU CENTRE HOSPITALIER UNIVERSITAIRE

The clinical data pipeline for genomic analysis

PDF
PNG
TXT

Clinical data warehouse

The clinical data pipeline for genomic analysis

PDF
PNG
TXT

Clinical data warehouse

Extract data

Data access
Optical character recognition (OCR)

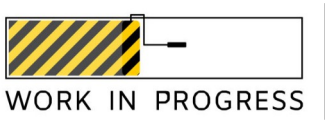
Transform data

Ontology terms extraction (HPO, OMOP…)

Exploit data

Ontology to variant prioritization and more…

The clinical data pipeline for genomic analysis



Extract data

Data access
Optical character recognition (OCR)

Transform data

Ontology terms extraction (HPO, OMOP…)

Exploit data

Ontology to variant prioritization and more…

Challenges in data extraction and sharing

1. Tous les outils d'analyse phénotypique computationnelle et d'extractions de termes d'ontologie n'utilisent que l'anglais comme langue.

2. Il est très difficile d'obtenir des médecins des descriptions cliniques en termes de HPO

3. Difficile de partager des informations cliniques anonymes avec d'autres médecins dans le monde entier (GeneMatcher)

What we dream to improve clinical data extraction and sharing

"Enzo DUPONT"
"né le 10/12/2021"
"mesure 75cm (-2DS)"
"retard psychomoteur"

Traduire, Anonymiser et Résumer les rapports médicaux

I. Traduire

"Enzo DUPONT"
"born in 10/12/2021"
"measures 75cm (-2DS)"
"global developmental delay"

II. Anonymiser

"CAS INDEX"
"born in *DATE*"
"measures 75cm (-2SD)
"global developmental delay"

III. Résumer en format HPO

"HP:0001263: Short stature"
"HP:0004322: Global developmental delay"

Let’s try !

ClinFly

Traduire, Anonymiser et Résumer les rapports médicaux

I. Traduire

Développer les abréviations

Traduction avec OpenNMT

Correction des erreurs courantes

OpenNMT:

Open-Source Toolkit for Neural Machine Translation

II. Anonymiser

Une liste de termes PHI français à rendre anonymes

Une liste de termes cliniques à ne pas anonymiser

Anonymiser

microsoft/**presidio**

Context aware, pluggable and customizable data protection and de-identification SDK for text and images

III. Résumer en format HPO

Extraire les biométries

Flags des phrases qui ne concernent pas le cas d'index

Résumer

ClinPhen

Résultats sur 50 courriers

1. ClinFly identifie les PHI et **gère la protection des données de santé**, atteignant :

- **Recall rate of 99%**
- **Precision rate of 77%**

449 PHI

443 PHI anonymisé
- 51 PHI de localisation
- 231 PHI temporel
- 166 PHI de personnes

6 FN PHI
- 3 noms de jeune fille
- 3 territoires français

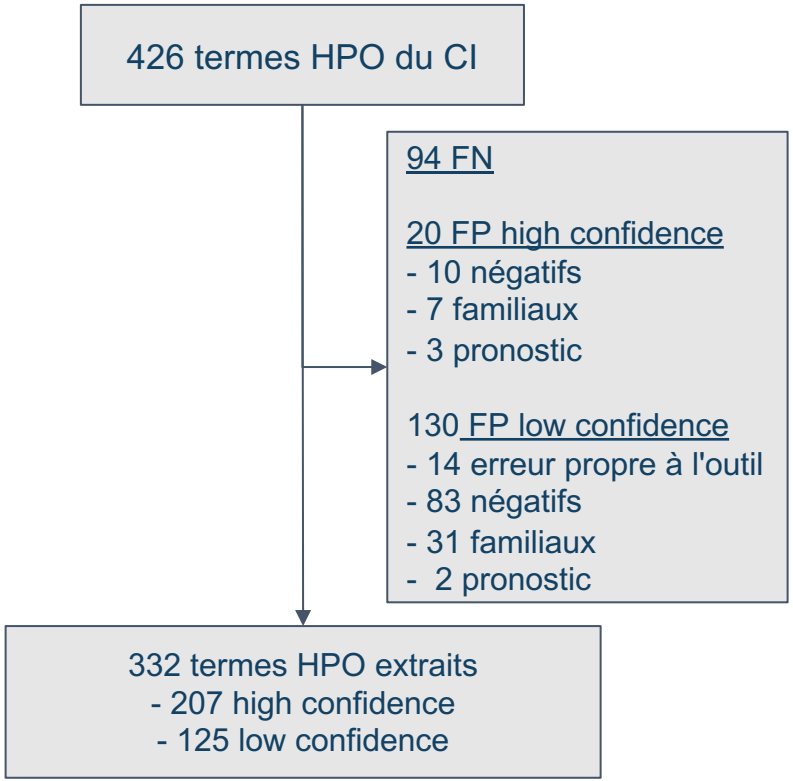
133 FP PHI
- 9 non-PHI diagnostic/motif
- 55 non-PHI symptômes
- 69 non-PHI cliniquement peu pertinent

Lucas Gauthier
CHU Lyon

Résultats sur 50 courriers

2. Les performances de ClinFly pour le **résumé d'un rapport médical par le format HPO** :

- **Precision of 92%**
- Maintained an average of **4.5 HPO** terms per report = **49% recall**



Graphical interface and command line

Graphical User Interface

For single report usage with interactive analysis, ClinFly provides a web application accessible at https://huggingface.co/spaces/kyauy/ClinFly.

To run the Streamlit application on your local computer, activate the poetry shell and run the `clinfly_app_st.py` file:

```
poetry shell
streamlit run clinfly_app_st.py
```

Command Line Interface

For processing multiple reports with offline options, use the command line interface provided by `clinfly_app_cli.py`.

The input should be a TSV .txt file structured as follows (see `data/test.tsv` for an example):

```
Report_id_1   Doe  John  Report text
...
Report_id_X   Doe  John  Report text
```

Outputs will be placed in the `results` folder according to the file extension, using first three columns in filename.

- The deidentify report will be generated and placed in the `results/Reports` folder.
- Three HPO extraction outputs will be generated in `TSV`, `TXT` and `JSON` folders.

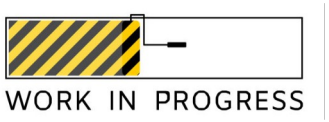
To run the CLI application on your local computer :

```
poetry shell
<python running version> clinfly_app_cli.py --file <input txt file with the reports> --languag
```

Enody
Gernet
CHU
Montpellier


ClinFly

The clinical data pipeline for genomic analysis



Extract data

Data access
Optical character recognition (OCR)

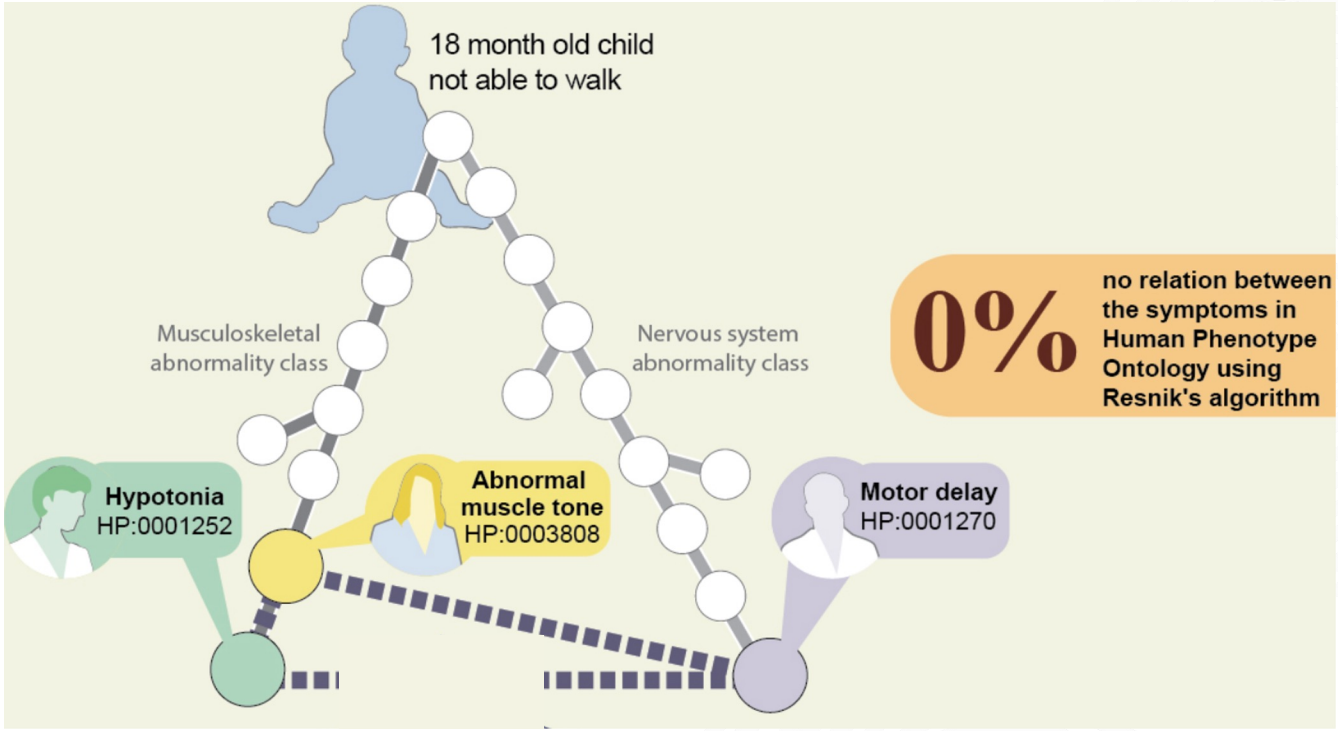
Transform data

Ontology terms extraction (HPO, OMOP…)

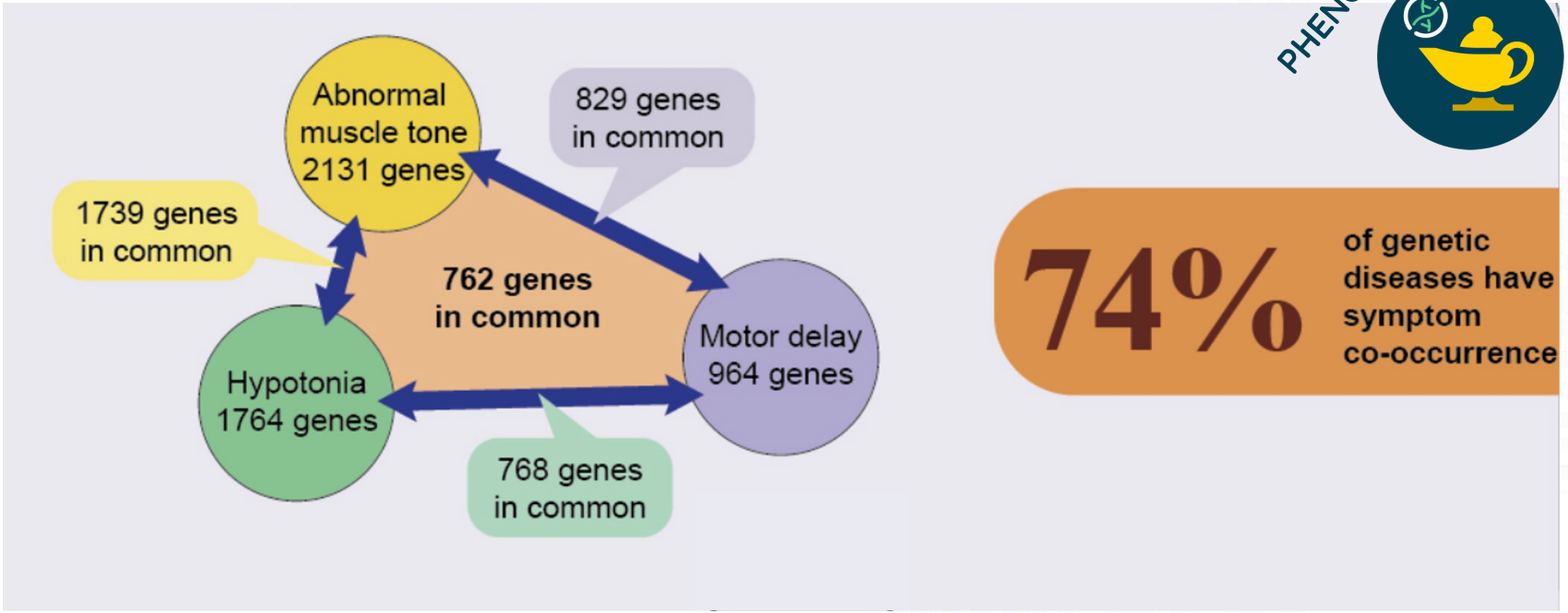
Exploit data

Ontology to variant prioritization and more…

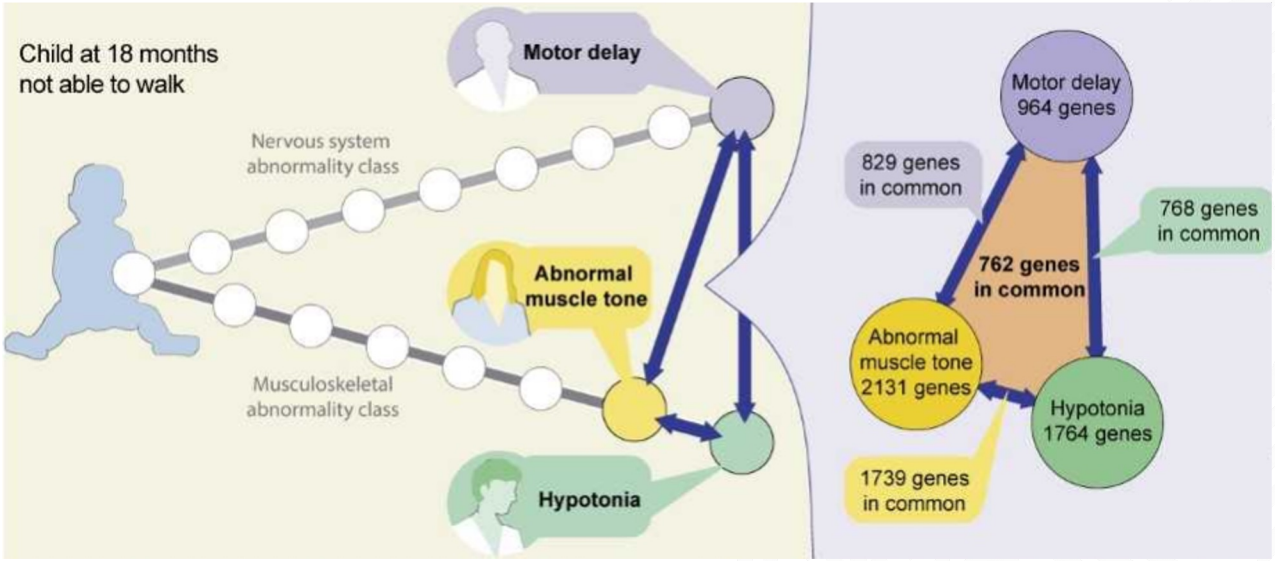
PROBLEM Heterogeneity of clinical descriptions between physicians, which cannot be captured by the HPO ontology



SOLUTION Learning phenotypic similarity using a knowledge graph based on gene-HPO associations



Multiple ways to describe patients



Physicians acquire intricate cognitive frameworks to solve diagnostic problems

→ Link between symptoms are very different from Human Phenotype Ontology structure

Inductive reasoning

Microcephaly (HP:0000252)

Delayed ability to walk (HP:0031936)

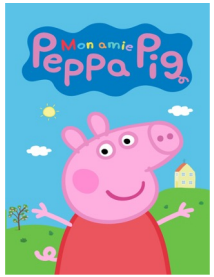
Disease /Phenotypic patterns I know

Intellectual deficiency (HP:0031936)

Shin. J Med Educ 2019

Inductive reasoning... through modeling

- Superhero
- Action
- US
- Robert Downey Jr



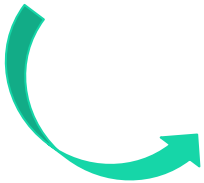
- Child
- Anime
- UK
- Animals

NETFLIX

Recommendation algorithms are really efficient

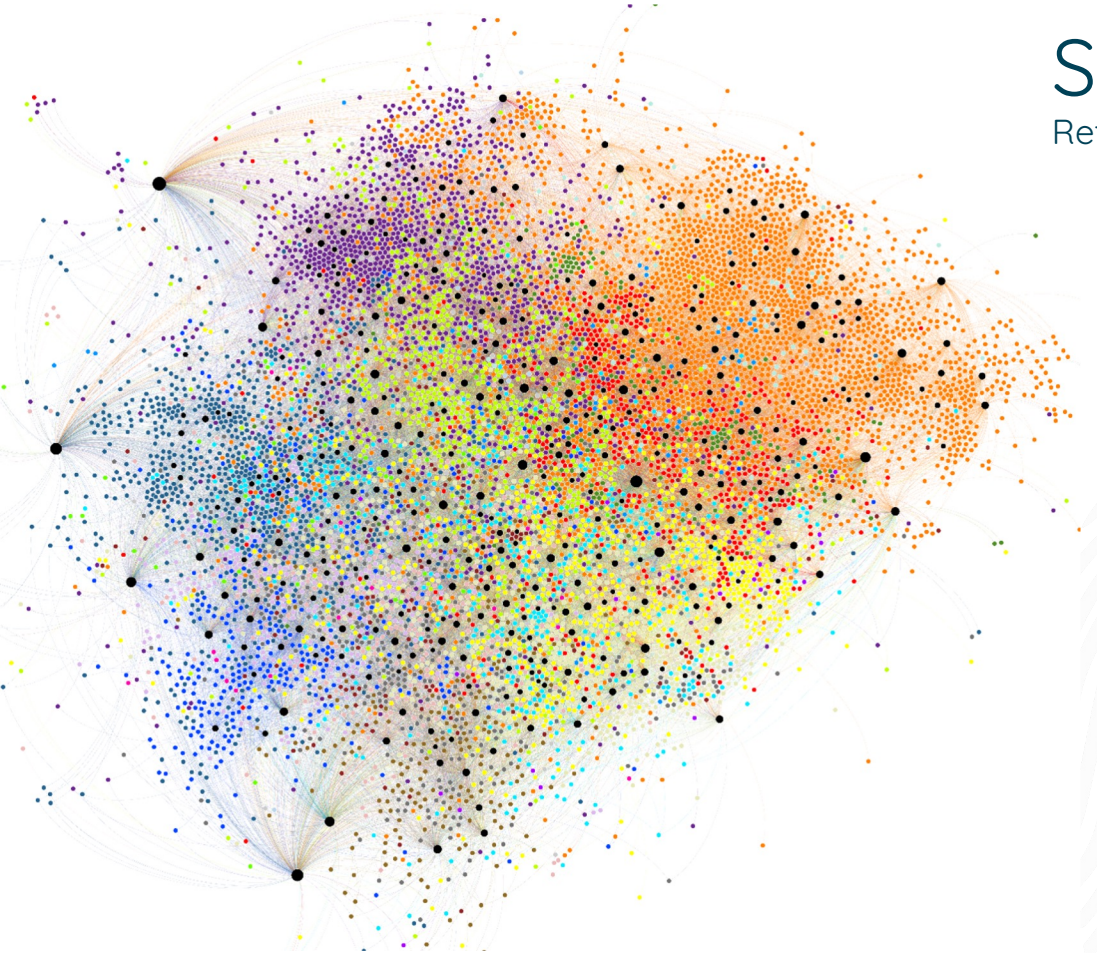
Apply inductive reasoning in symptom-gene associations

	Unstructured data			Structured databases			
	Text-matching via elasticsearch *			Merge all HPO-gene association available			HP:0000006 Probability
Gene	OMIM	PubMed	NCBI National Center for Biotechnology Information	The Jackson Laboratory	orphanet	G2P	VOTE
BRCA1	1	1	0	0			0.5



	HP01	HP02	HP03	...	...	...	HP15785
Gene 1	0.23	0.41	0	...	...	...	0
Gene 2	0	0.21	0.32	...	...	...	0.42
...	...	...	...	...	...	...	...
Gene 4531	0.11	0.27	0.42	...	...	...	0.42

Building a symptom-gene association matrix with ~16000 symptoms

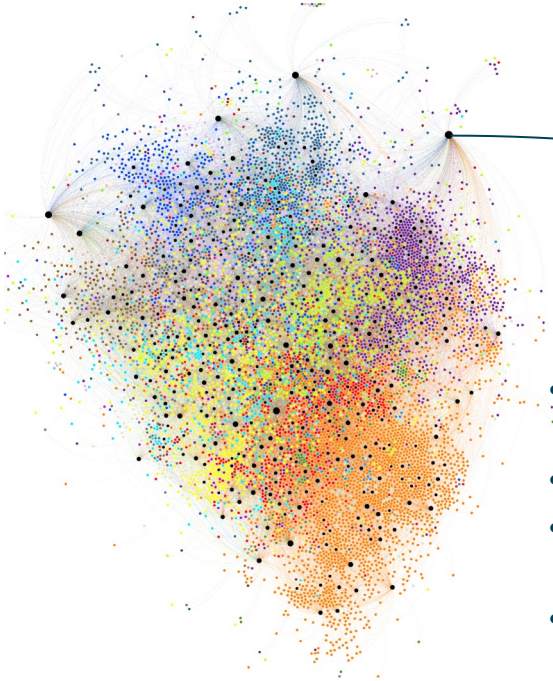


Symptom-gene graph

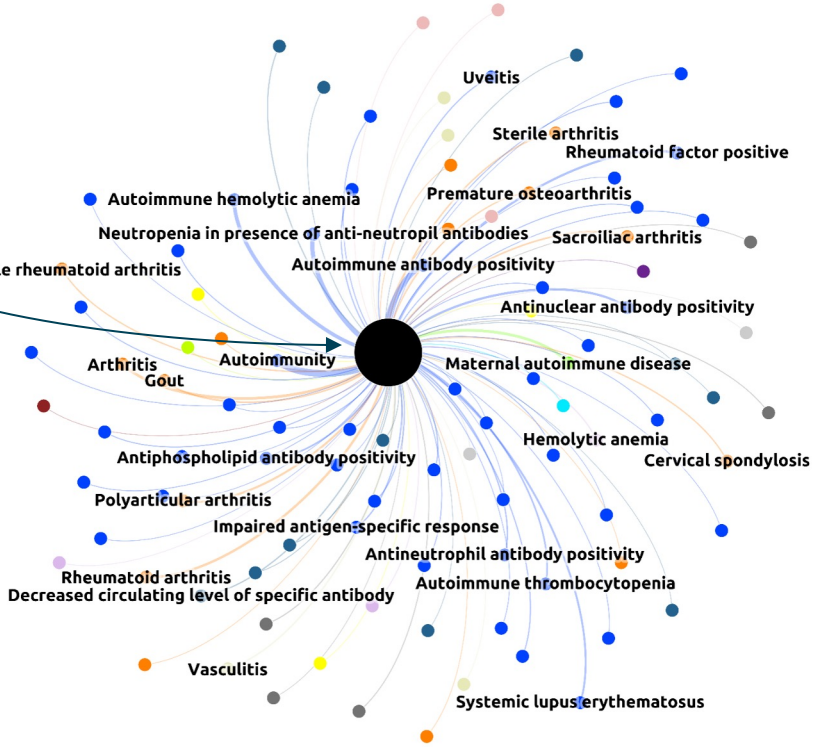
Retrieving Symptom-gene associations with graphs

- **390 groups** of symptoms (n=43,308, 10% of symptom-group associations)
- 5,971,755 pairs of symptoms
- **3,222,053 additional** NMF-based symptom-gene associations
- **“only” 2% cohort symptom-gene association missing !**

An example: Autoimmunity group

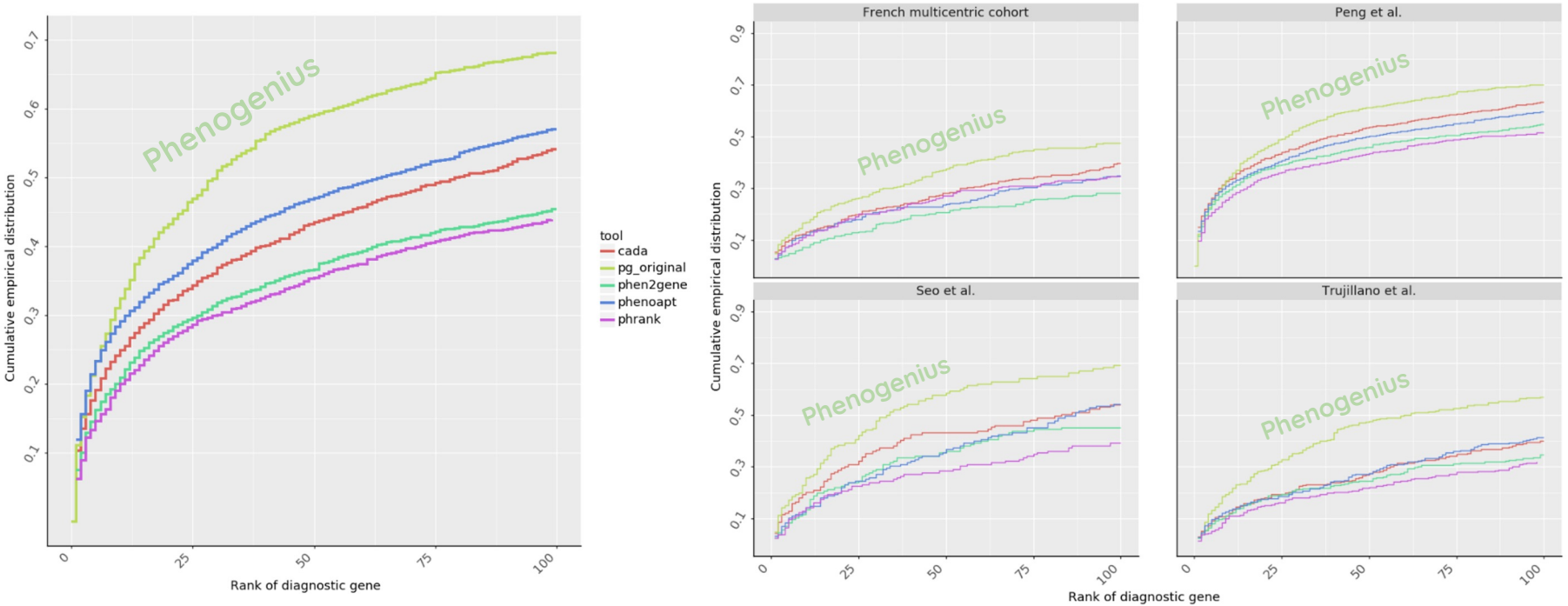


- Main symptom : Autoimmunity
- 99 HPOs included
- 14 different HPO classes
- 545 genes associated to the group



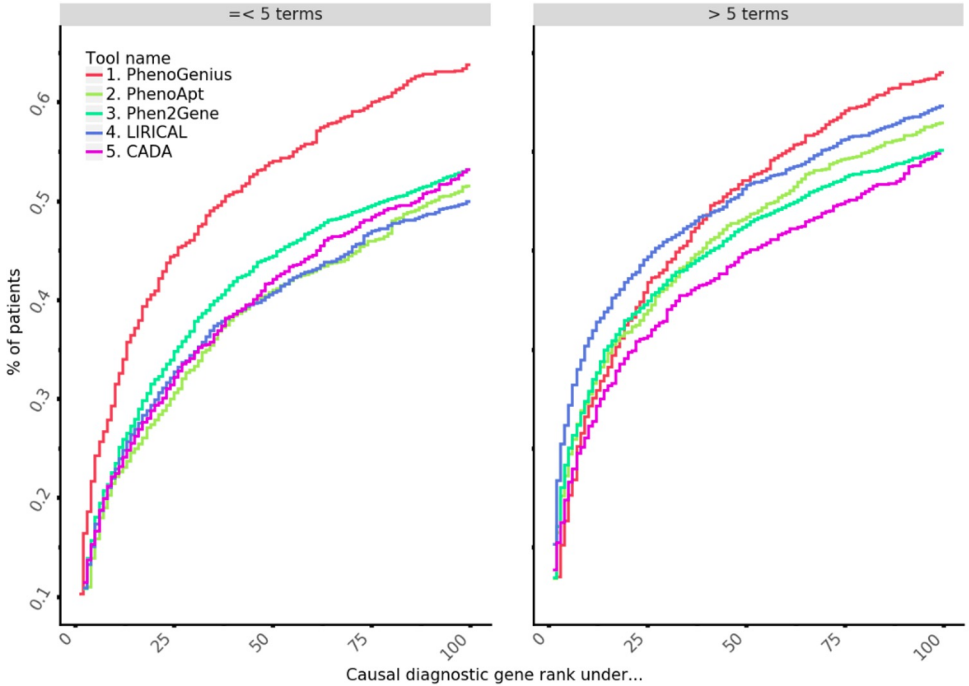
Performance on 4 cohorts (1686 patients)

Outperforming 2022 state-of-the-art ranking methods



Performance on 4 cohorts (1686 patients)

A way better semantic similarity using disease and not ontology architecture

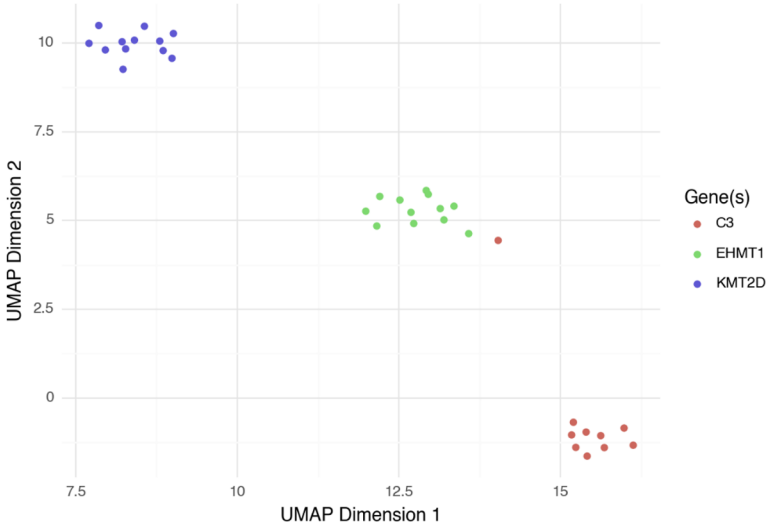


Median diagnostic rank =< 5 terms :
PhenoGenius : 37
Phen2Gene (best competitor): 78

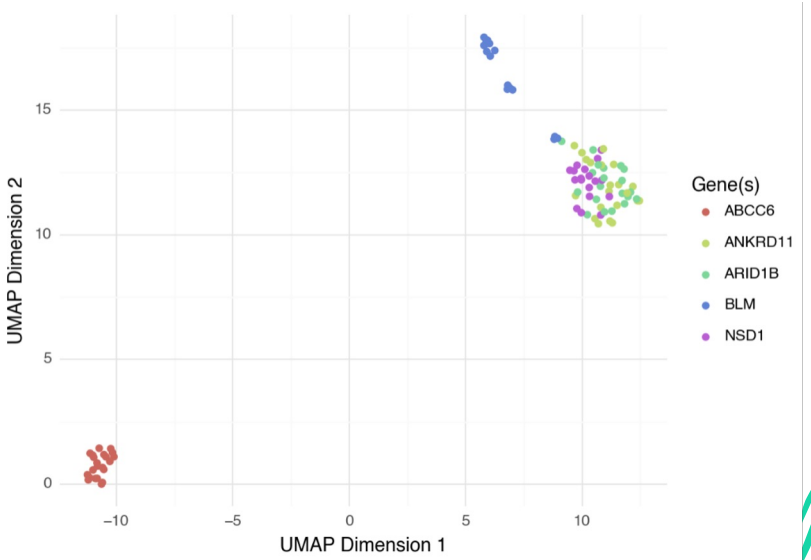
% of cohort with <= 5 terms :
50% (853/1686)

Use phenotypes as phenomics

Three clinical reports described by 12 physicians

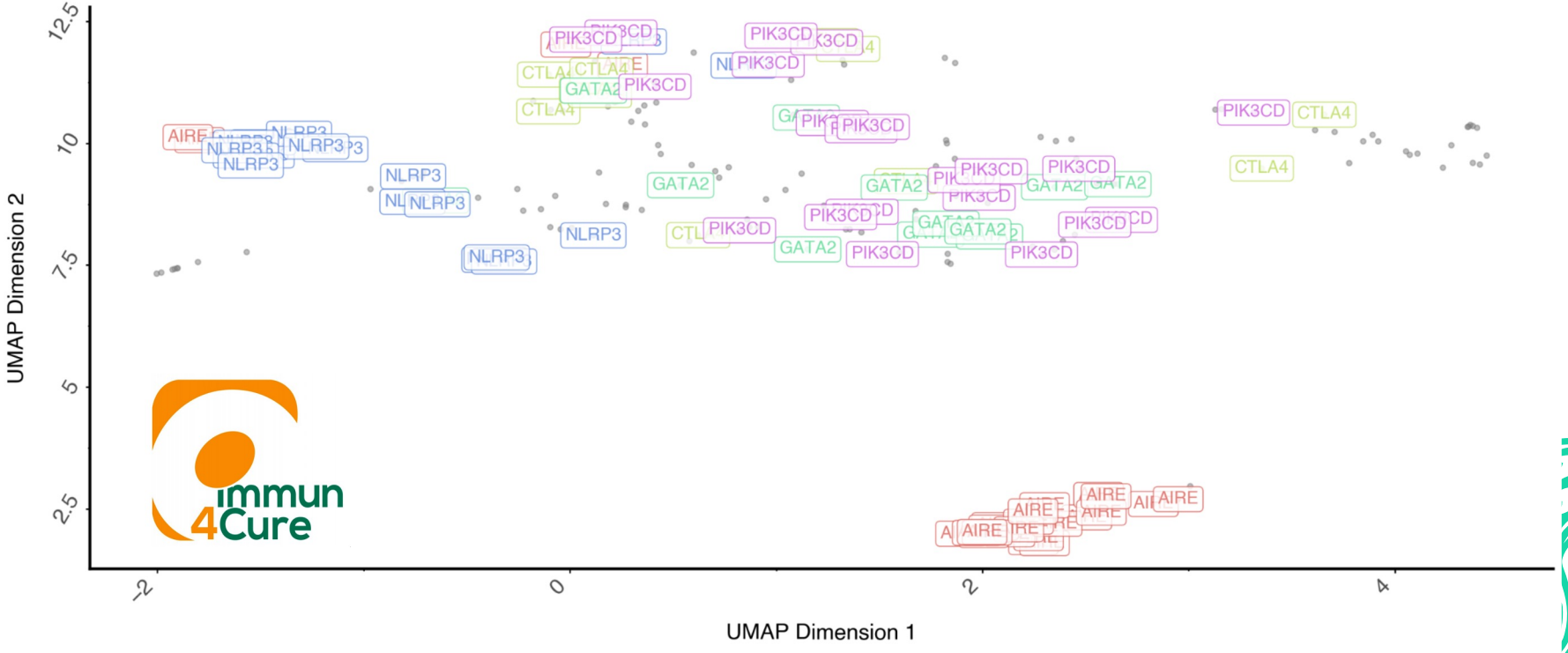


Cohort : 5 most frequent diagnostic genes



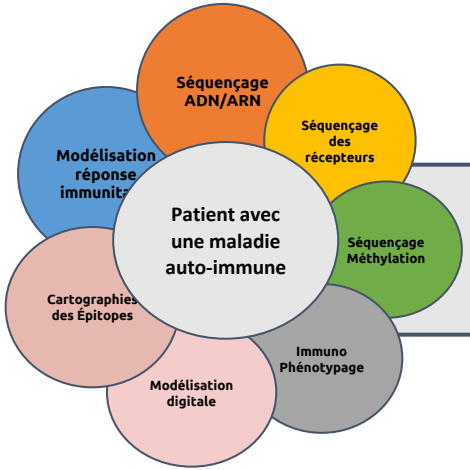
-> Vectorial projection available for all patients

ICohort phenomics projection



Clinical data as Phen-Omics, an omic type among others
A 5 to 10 years vision ?

Integrative immunology

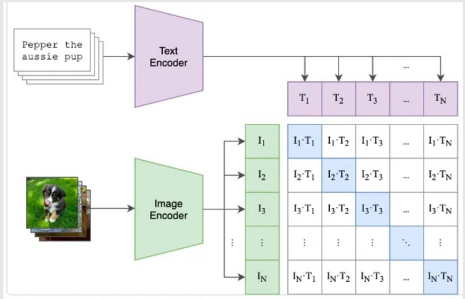


Multi-omics analysis using structured data

Generative AI to structure clinical data

Multimodal embeddings into vector database

Omics data wrangling to structured text or imaging



Inference & clinical applications

Chat Agents using LLM

Patient stratification

Thanks for your attention!

kyauy/ClinFly
kyauy/PhenoGenius

✉️ : kevin.yauy@chu-montpellier.fr
𝕏 : @KevinYauy
in : kyauy

Warm thanks to all contributors and collaborators !

⭐**CHU Montpellier: Bioinformatics & Data science team**
✨**CHU Grenoble Alpes**