Initiatives to Speed up Data Mining

Ségolène Aymé

ICORD 2014, Ede, Netherlands

9 October 2014

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Objectives

- Make the most of remarkable advances in the molecular basis of human diseases
  - dissect the physiological pathways
  - to improve diagnosis
  - to develop treatments

- Make rare diseases visible in health information systems
  - to gain insight into them
  - to access real life data already collected

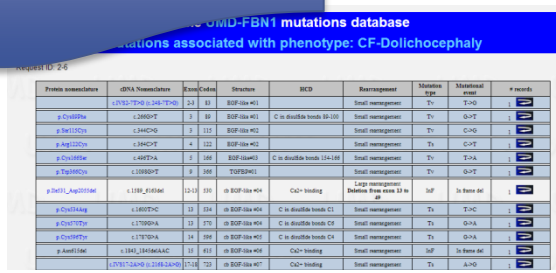
IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

What do you mean?

scaphocephaly

Long narrow head

dolichocephaly



Protein nomenclature	cDNA Nomenclature	Exon	Codon	Structure	HCD	Rearrangement	Mutation type	Mutational event	# records
	c.IVS2-7T>G (c.248-7T>G)	2-3	83	EGF-like #01		Small rearrangement	Tv	T->G	1
p.Cys109Ser	c.326G>T	5	89	EGF-like #01	C in disulfide bonds 88-100	Small rearrangement	Tv	G->T	1
p.Ser115Cys	c.344C>G	5	115	EGF-like #02		Small rearrangement	Tv	C->G	1
p.Arg122Cys	c.364C>T	4	122	EGF-like #02		Small rearrangement	Ts	C->T	1
p.Cys168Ser	c.499T>A	5	168	EGF-like#03	C in disulfide bonds 154-168	Small rearrangement	Tv	T->A	1
p.Trp366Cys	c.1098G>T	9	366	TGFB2#01		Small rearrangement	Tv	G->T	1
p.Ile231_Asp235del	c.1189_1216del	12-13	231	cb EGF-like #04	Ca2+ binding	Large rearrangement Deletion from exon 12 to 49	InF	In frame del	1
p.Cys234Arg	c.1603T>C	13	234	cb EGF-like #04	C in disulfide bonds C1	Small rearrangement	Ts	T->C	1
p.Cys270Tyr	c.1709G>A	13	270	cb EGF-like #04	C in disulfide bonds C8	Small rearrangement	Ts	G->A	1
p.Cys396Tyr	c.1787G>A	14	396	cb EGF-like #05	C in disulfide bonds C4	Small rearrangement	Ts	G->A	1
p.Asn915del	c.1843_1845delAAC	15	915	cb EGF-like #09	Ca2+ binding	Small rearrangement	InF	In frame del	1
	c.IVS17-2A>G (c.2198-2A>G)	17-18	723	cb EGF-like #07	Ca2+ binding	Small rearrangement	Ts	A->G	1

elementsofmorphology.nih.gov

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

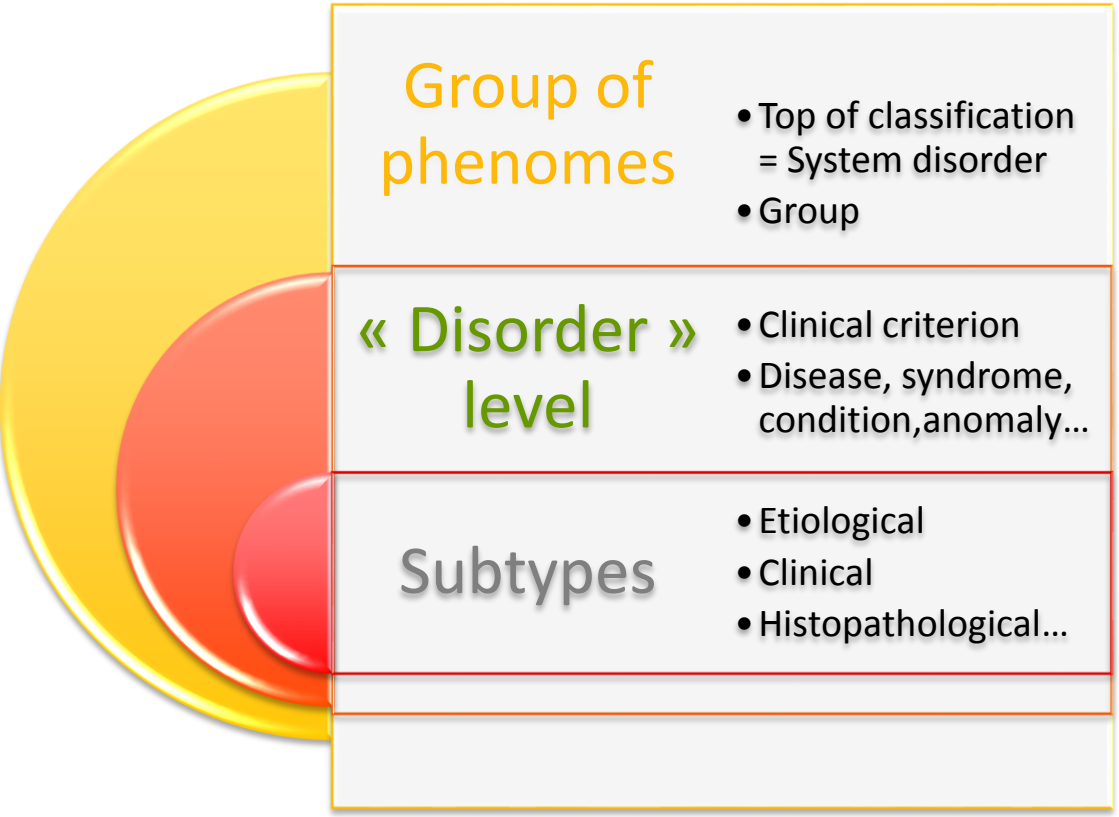
What is the problem ? Computers are not smart enough….

- Phenotypic descriptions that are very evocative for humans: “myopathic electromyography » or « still walking 25 years after onset”
  - Unfortunately computers, being signicantly dumber than humans, don't quite get it...

- The following descriptions mean the same thing to you: \generalized amyotrophy", \generalized muscle atrophy", \muscular atrophy, generalized" (etc)
  - But your computer thinks they're completely unrelated

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

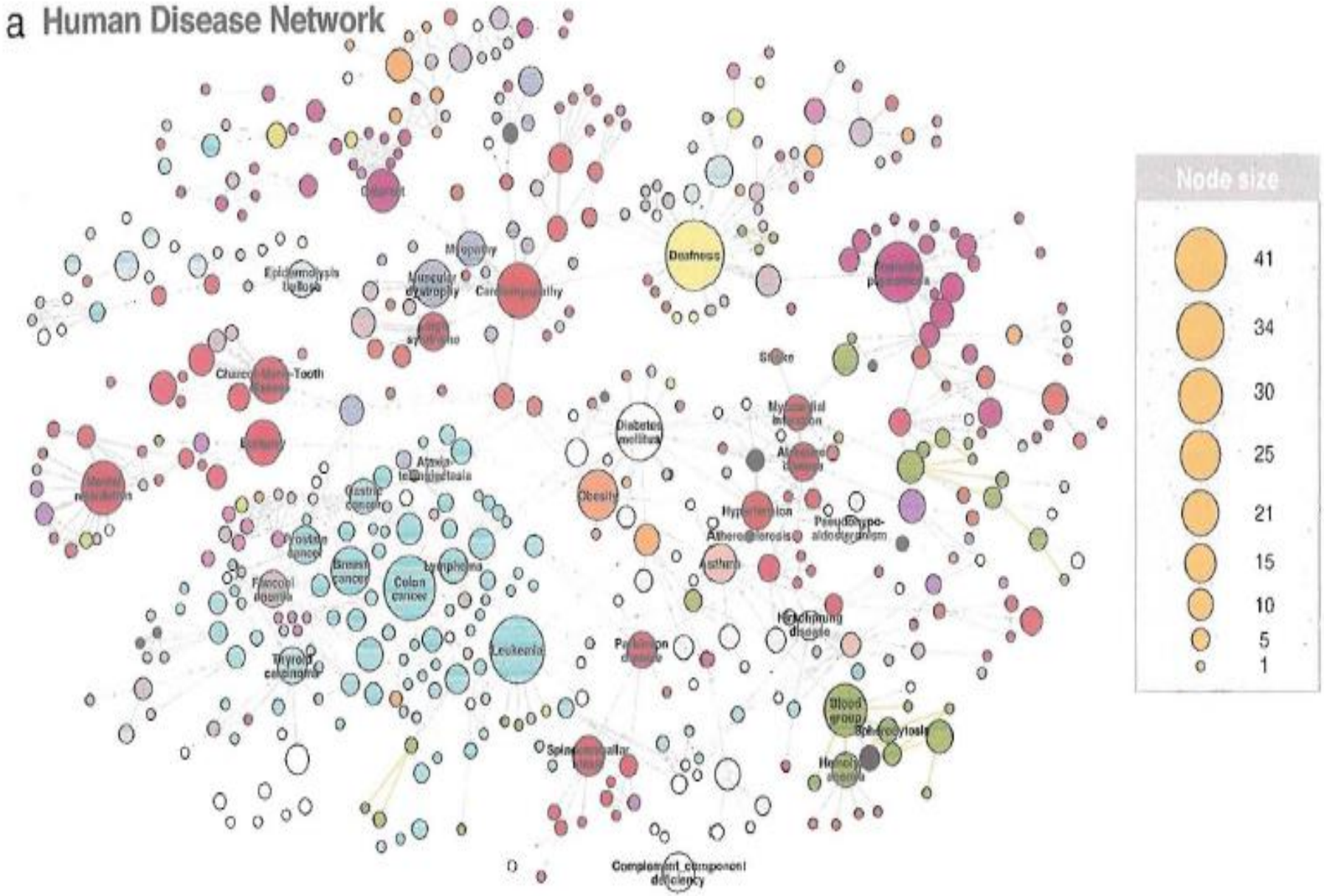
Phenomes: a continuum



- Disease
- Malformation syndrome
- Morphological anomaly
- Biological anomaly
- Clinical syndrome
- Particular clinical situation

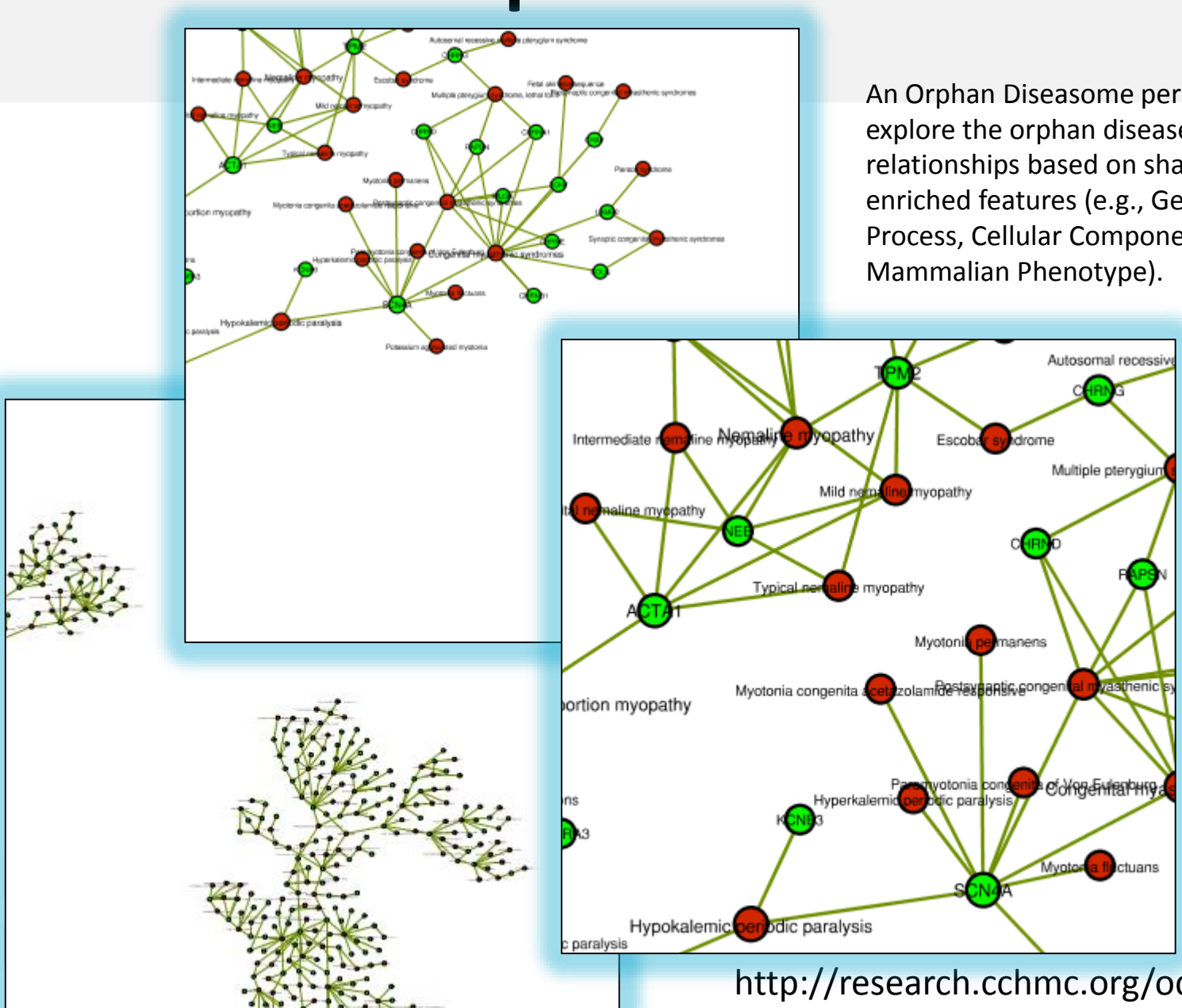
- No type: waiting to have a type attributed

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM



INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Orphan Diseasome



An Orphan Diseasome permits investigators to explore the orphan disease (OD) or rare disease relationships based on shared genes and shared enriched features (e.g., Gene Ontology Biological Process, Cellular Component, Pathways, Mammalian Phenotype).

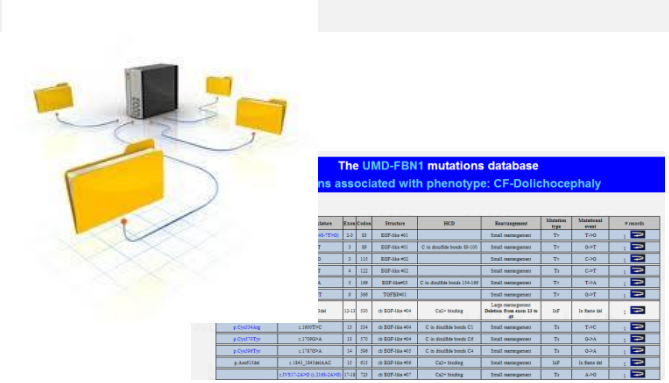
The red nodes represent the orphan diseases and the green ones the related genes. A disease is connected to a gene if and only if a mutation which is responsible of the disease has been identified on this gene.

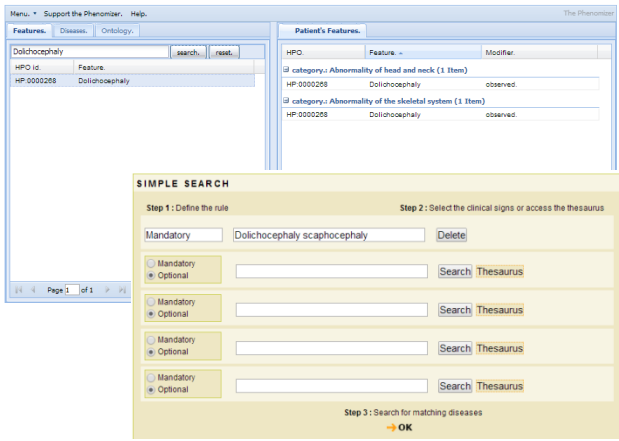
http://research.cchmc.org/od/01/index.html

UMLS = Unified Medical Language System

- **ICD = International Classification of Diseases**
  - Since 1863 by WHO
  - Used by most countries to code medical activity, mortality data
- **MeSH = Medical Subject Headings**
  - controlled vocabulary thesaurus used for indexing articles for PubMed by National Library of Medicine (USA)
- **SnoMed CT = Systematized Nomenclature of Medicine--Clinical Terms**
  - clinical terminology by the International Health Terminology Standards Development Organisation (IHTSDO) in Denmark
  - Used in the USA and a few other countries
- **MedDRA = Medical Dictionary for Regulatory Activities**
  - medical terminology to classify adverse event information associated with the use of medical products
  - by the International Federation of Pharmaceutical Manufacturers and Associations (IFPMA)

Different resources, different terminologies





(e)HR:

SNOMED CT

Others?

Free text

Mutation/patient registries, databases:

HPO

LDDB

PhenoDB

Elements of morphology

Others? Free text?

Tools for diagnosis:

HPO

LDDB

Orphanet

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Each terminology has a purpose–driven approach

- Indexing health status of individual patients for health management (SnoMED)
  - Detailed, focus on manifestations and complaints
  - Adapted to clinical habits
  - Analytical approach
- Indexing health status of individual patients for statistical purpose in public health (ICD)
  - More agregated, interpreted phenotypic features
  - Agregated concepts
  - Unambiguous to avoid blanks

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Purpose–driven approach (2)

- Indexing health status of individual patients for clinical **research purpose** (HPO / PhenoDB / Elements of morphology)
  - Highly detailed to fit with the research questions
  - Specific terminologies developed for disease-specific patient registries
- Indexing health status of individual patients for retrieving **possible diagnoses** (LDDB,POSSUM,Orphanet)
  - Agregated concepts
  - Requires a judgement of clinicians about phenomic expressions that are relevant
  - Unambiguous to avoid blanks

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

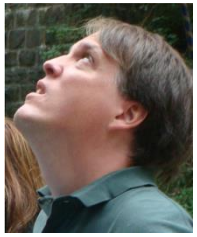
HOW TO MAKE ALL THESE TERMINOLOGIES INTER-OPERABLE ?

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Overview of project progress

- Sept 2012: start of mappings (Orphanet)
- EUGT2 – EUCERD workshop (Paris, September 2012)

PhenoDB



HPO

Orphanet

LDDB
Elements of Morphology
POSSUM
SNOMED CT (IHTSDO)

DECIPHER
IRDiRC

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

- ICHPT workshop (ASHG, Boston, October 2013)

Ground Work

- Orphanet group aligned terms from HPO, PhenoDB, Orphanet terminology, MESH, MedDRA, UMLS, LDDB, SNOMed-CT, and Elements of Morphology

- PhenoDB group compared data on use of terms in 3 large scale projects: DDD/DECIPHER, FORGE/Canada using HPO and BHCMG using PhenoDB. Mapped all terms.

Phenotype terminology project

- Aims:
  - **Map** commonly used clinical terminologies (Orphanet, LDDB, HPO, Elements of morphology, PhenoDB, UMLS, SNOMED-CT, MESH, MedDRA):
    - automatic map, expert validation, detection and correction of inconsistencies
  - **Find common terms** in the terminologies
  - Produce a **core terminology**
    - Common denominator allowing to share/exchange phenotypic data between databases

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

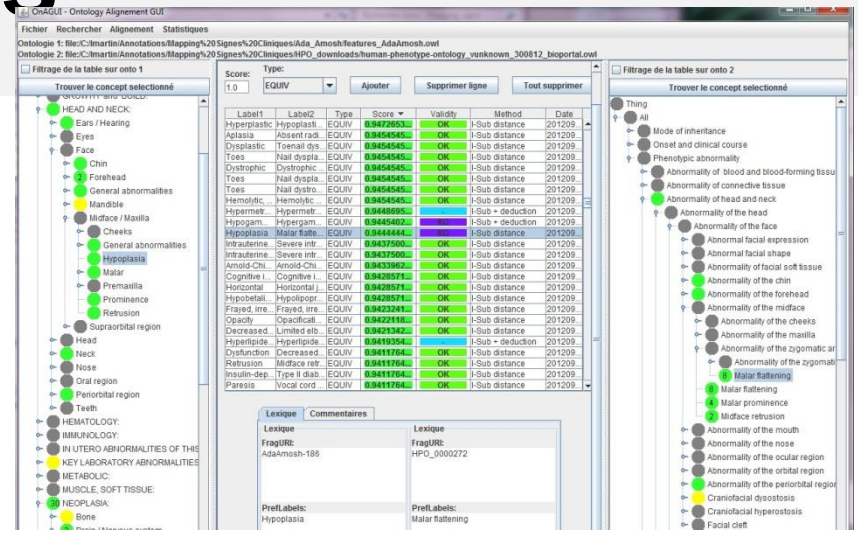
Mapping Terminologies

- **Orphanet**: 1357 terms (Orphanet database, version 2008)
- **LDDB**: 1348 dysmorphological terms (Installation CD)
- **Elements of Morphology**: 423 terms (retrieved manually from publication AJMG, January 2009)
- **HPO**: 9895 terms (download bioportal, obo format, 30/08/12)
- **PhenoDB**: 2846 terms (given in obo format, 02/05/2012)
- UMLS: (version 2012AA) (integrating **MeSH**, **MedDra**, **SNOMED** CT)

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Tools

- OnaGUI (INSERM U729): ontology alignment tool
  - Work with file in owl format
  - I-Sub algorithm: detect syntaxic similarity
  - Graphical interface to check automatic mappings and manually add ones



- Metamap (National Library of Medicine): a tool to map biomedical text to the UMLS Metathesaurus

```
Phrase: "total anosmia"
>>>>> Phrase
total anosmia
<<<<< Phrase
>>>>> Candidates
Meta Candidates (1):
   833 C0003126:Anosmia {MDR, MSH, SNOMEDCT} [Sign or Symptom]
<<<<< Candidates
```

- Perl scripts: format conversion, launching Metamap, co
results…

A	C	D	E	F	G	H	I	J	K	L	N
Sign_Id	ORPHA PT	SCORE	CORRI	ORPHA SYN	UMLS MATCH	UMLS PT	UMLS CODE	TERMINOLO	CODE	PT/corr CONCEPT	SEMANTIC TYPI
10.01.05	Facial cleft	1000		facial cleft	Facial cleft	Cleft face	C0685787	UMLS	C0685787	Cleft face	Congenital Abr
10.01.05	Facial cleft	1000		facial cleft	Facial cleft	Cleft face	C0685787	SNOMEDCT	92821006	Cleft face (disorder)	Congenital Abr
11.02.00	Lower lip, ge	770		general abnc	Lip	Benign neopla	C0153932	UMLS	C0153932	Benign neoplasm of tl	Neoplastic Pro
11.02.00	Lower lip, ge	770		general abnc	Lip	Benign neopla	C0153932	MDR	10024556	Lip neoplasm benign	Neoplastic Pro
11.02.00	Lower lip, ge	770		general abnc	Lip	Benign neopla	C0153932	SNOMEDCT	92185002	Benign neoplasm of li	Neoplastic Pro
11.02.00	Lower lip, ge	770		general abnc	LIP	Lymphoid inte	C0264511	UMLS	C0264511	Lymphoid interstitial	Disease or Syn
11.02.00	Lower lip, ge	770		general abnc	LIP	Lymphoid inte	C0264511	MDR	10062997	Lymphoid interstitial	Disease or Syn

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Comparison of mappings and deduction

- Perl script to compare all the mappings and infer mappings of non-Orphanet terminologies

  Eg: Orphanet ID XX mapped to YY in HPO and ZZ in LDDB -> deduction: YY and ZZ should probably map

- Retrieve HPO mappings versus UMLS, MeSH

	LDDB	El. Morpho	PhenoDB	HPO	UMLS…
Orphanet	E: 1062	E: 416	E: 978	E: 2228	E: 6948
LDDB		D: 275	D: 533	D: 1123	D:2678
El. Morpho			D: 177	D: 716	D: 409
PhenoDB				D: 1045	D:3268
HPO					D: 6307+4800
UMLS…					

Mapping of non-Orphanet terminologies

- Automatic and infered mappings were checked by experts
  - Using OnaGUI for all, except UMLS
    - ✓ Automatic I-Sub: 7.0 + deduction

	El. Morpho	PhenoDB	HPO	UMLS…
LDDB	D: 257 +23 added	D:528, 92%E A:674, 38%E	D: 1105, 87%E A: 2084, 23%E	D: 2654, 83%E A: 11731
El. Morpho		D:174, 50%E A:189, 74%E	D:393, 93%E A: 436, 16%E	D:405, 84%E A:1248
PhenoDB			D:1018, 91%E A: 4168, 6%E	D: 3222, 82%E A: 18776
HPO				D: 7389 A: 65535
UMLS…				

INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

First list of common terms

- Present in at least 3 terminologies
- Definition of rules for nomenclature
- Addition of terms present in each terminology as synonyms

1	Identifier	Preferred term	Synonyms		
2	T0001	Anomaly of the skull	Skull anomalies	Cranial bones, general abnormalities	Cranium, gener
3	T0003	Cranial hyperostosis	Thick skull	Thickened skull	Dense skull
4	T0004	Basilar hyperostosis	Sclerosis of skull base		
5	T0005	Calvarial hyperostosis	Thick calvaria	Thickened calvaria	Dense calvaria
6	T0006	Decreased skull ossification	Poorly ossified skull	Ossification defects of skull	Undermineraliz
7	T0008	Decreased calvarial ossificatio	Thin calvaria	Absent ossification of calvaria	Thin calvarium
8	T0010	Anomaly of the cranial sutures	Cranial sutures, general abnormalit	Head Sutures anomalies	Abnormality of
9	T0011	Wide cranial sutures	Cranial sutures, wide	Wide cranial sutures (finding)	
10	T0012	Ridged cranial sutures	Cranial sutures, ridged		
11	T0013	Anomaly of the sella turcica	Sella turcica anomaly		
12	T0014	Large sella turcica	Sella turcica, large		
13	T0015	J-shaped sella turcica	Sella turcica, J-shaped	Shoe-shaped sella turcica	
14	T0016	Small sella turcica	Sella turcica, small		
15	T0018	Anomaly of the temporal bone	Abnormality of the temporal bone		
16	T0019	Anomaly of the mastoid proces	Mastoids, general abnormalities	Abnormality of the mastoid	
17	T0020	Small foramen magnum	Foramen magnum, small	Foramen magnum stenosis	
18	T0021	Large foramen magnum	Foramen magnum, large		
19	T0022	Delayed pneumatization of the	Delayed pneumatization of mastoids		
20	T0023	Advanced pneumatization of th	Advanced pneumatization of mastoids		

Inter-operability based on mappings

- Syntaxic:
  - The terms are identical
    - Can be done by machines
- Semantic:
  - The concepts are identical
    - Should be done by humans
- Structural:
  - The comprehension of the concepts is identical
    - Impossible to maintain

Problems encountered

- Structure
  - Criteria for top hierarchy
  - Granularity
- Terminology
  - Synonymy
  - Ambiguity
  - Uniqueness
- Versioning

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Granularity

- Differences between depth degrees between terminologies
- Even if very granular, none of the terminologies is exhaustive
- Differences in granularity makes it necessary to qualify the relationships between mapping terms in order to allow interoperability
  - E = Exact mapping
  - NTBT = narrower-to-broader
  - BTNT= broader-to-narrower

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

International Consortium for Human Phenotype Terminologies (ICHPT)

- Workshop on 21-22 October 2013 in Boston
- Participants: HPO, DDD/DECIPHER, FORGE/Canada, PhenoDB, Orphanet, Elements of Morphology, POSSUM, ClinGen, OMIM, Mouse Genome Informatics, SimulConsult
- Committed to agree on ~2000 high level terms (with definitions and synonyms)
- Make sure that these are used, mapped, and have behind them an ontology
- WHO and SNOMed-CT are committed to adopting these core terms

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Success!

- Reviewed 2736 terms appearing 2 or more times in the 6 terminologies in 17 hours
- 2302 terms chosen, including preferred term
- Synonyms are clear from the list
- Definitions are from Elements of Morphology if available, and HPO/Stedman's Medical Dictionary, if not
- List of terms, mapping to HPO, PhenoDB, Elements of Morphology will be available at http://ichpt.org by January 2015.
- All tools will map to this terminology to allow interoperability among resources

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Consensus

- New tools being developed to allow data sharing of unsolved exomes/genomes will use and/or incorporate this terminology
  - GeneMatcher
  - LOVD
  - PhenomeCentral
  - DECIPHER

Adoption of a core set of >2,300 terms common to all terminologies

Workshop on Terminologies for RD – Paris, 12 September 2012

- Many terminologies in use to describe phenomes - No interoperability
- Joint EuroGenTest and EUCERD workshop
- Organized by Ségolène Aymé
- Agreement to define a core set of terms common to all terminologies and a methodology
- Core set identified by cross referencing
  - HPO
  - PhenoDB
  - Orphanet
  - UMLS: MeSH, MedDRA, SnoMed CT
  - LDDB
  - Elements of morphology

Workshop of validation, Boston 21-22 October 2013

- Workshop supported by HVP and EuroGenTest
- Organized by Ada Hamosh
- Expert review of the initial proposal
- Selection of 2,370 terms
- Decision to propose them for adoption by all terminologies
- Establishment of the International Consortium for Human Phenotype Terminologies – ICHPT
- **Publication on the IRDiRC website** with definitions from
  - HPO
  - Elements of morphology

IRDiRC
INTERNATIONAL RARE DISEASES RESEARCH CONSORTIUM

COMPUTERS ARE NOT SMART

FROM A TERMINOLOGY TO AN ONTOLOGY

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Why ontologies are needed ?

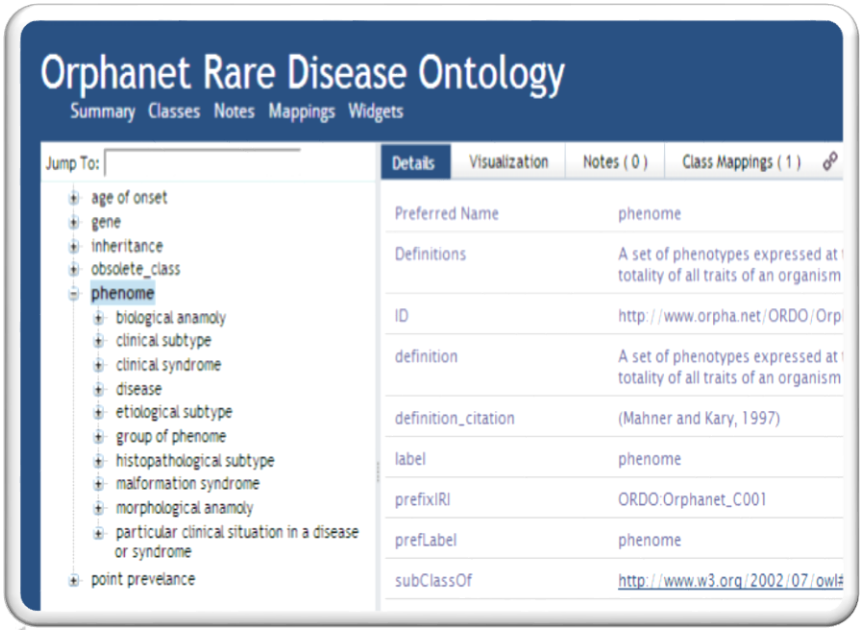
- Ontologies are representations of the knowledge in a way which is directly understandable by computers
- Ontologies allow reasoning
- Ontologies define the objects AND the relationship between the objects
  - Is a ….is part of…. Is a cause of….
- Anemias
  - Shistosomias

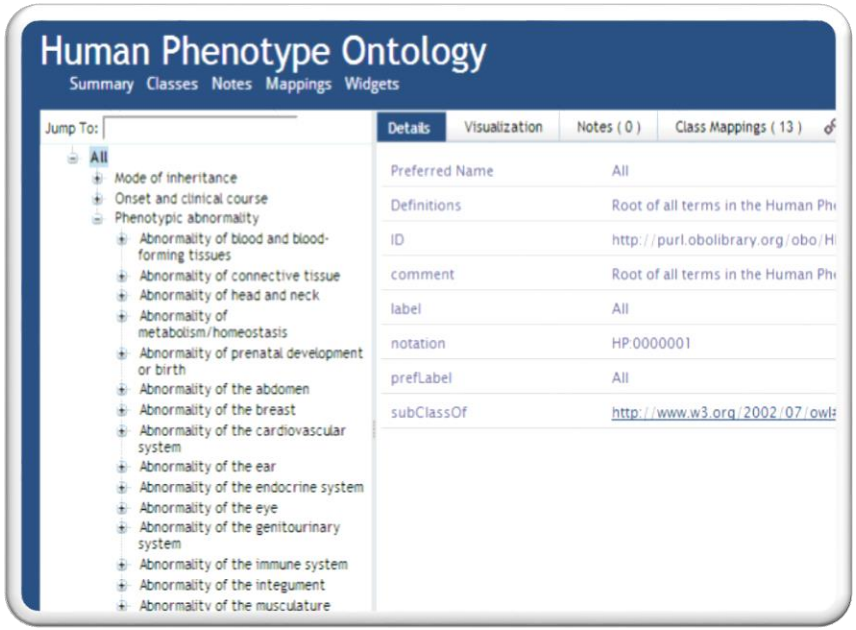
  Is a possible cause of anemia but is not an anemia

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Standardization of Phenotype Ontologies

Workshop Sympathy, 19 Apr 2013, Dublin
Organized by IRDiRC, supported by the University of Dublin, Forge and EuroGenTest
Conclusion: Adopt HPO & ORDO & cross-reference with OMIM

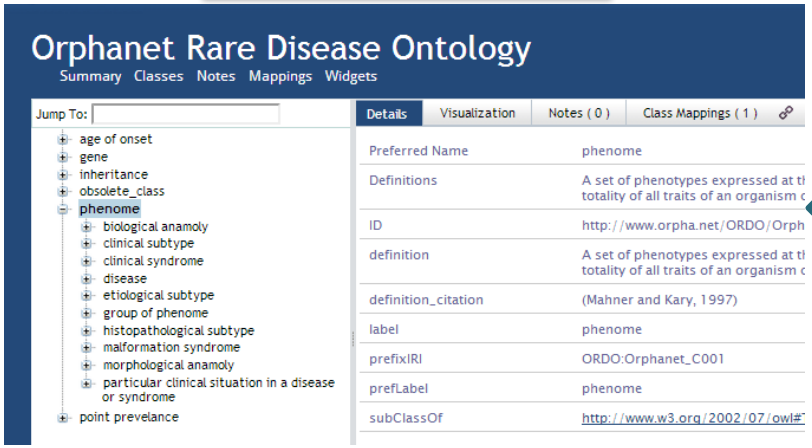




IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

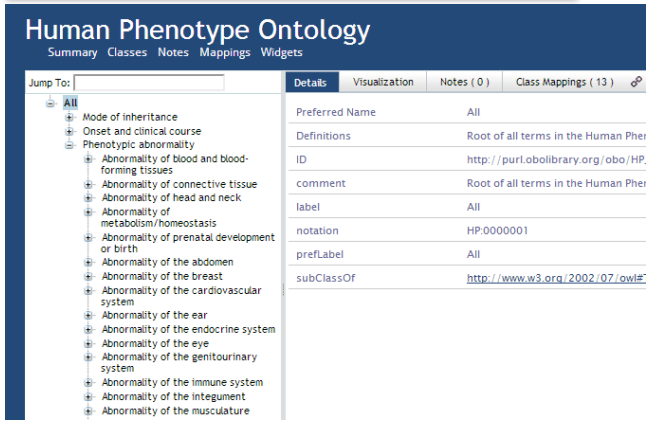
Standardisation of Phenotype Ontologies

Rare Diseases



bioportal.bioontology.org/ontologies/ORDO

Phenotypic Features



bioportal.bioontology.org/ontologies/HP

Based on Orphanet multi-hierarchical classification of RD

Genes– diseases relationships

Cross-references:

-For RD nomenclature : OMIM, SNOMED CT, ICD10, MeSH, MedDRA, UMLS

-For genes : OMIM, HGNC, UniProtKB, IUPHAR, ensembl, Reactome

ICHPT

(International Consortium for Human Phenotype Terminologies)

2,307 terms- core terminology

Mapped to:

HPO Elements of Morphology
Orphanet LDDB
SNOMED CT Pheno-DB (OMIM)
MeSH UMLS

Available soon for download at ichpt.org

INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Orphanet Ontology Browser

- obsolete_class
- gene
- age of onset
- phenome
  - biological anamoly
  - clinical subtype
  - clinical syndrome
  - disease
  - etiological subtype
  - group of phenome
    - Rare abdominal surgical disease
    - Rare allergic disease
    - Rare bone disease
    - Rare cardiac disease
      - Cardiomyopathy
        - Arrhythmogenic right ventricular dysplasia
        - Dilated cardiomyopathy
        - Hypertrophic cardiomyopathy
        - Idiopathic giant cell myocarditis
        - Restrictive cardiomyopathy
        - Unclassified cardiomyopathy
      - Coronary artery disease - hyperlipidemia - hypertension - dial
      - LMNA-related cardiocutaneous progeria syndrome
      - Pericarditis
      - Rare cardiac rythm disease
      - Rare cardiac tumor
    - Rare circulatory system disease

Legend:

- is a
- part of
- develops from
- other

Help (hide)

Double-click a term to see its children. The ontology browser is populated dynamically. If there are many children for a given term, there may be a small delay while the browser fetches **Click** to highlight a term to see any information associated with it. **Hover** over a term to see its relation with its immediate parent. Root terms will not display any relational information.

Relations

Term Information

ID: Orphanet:329874 Zoom

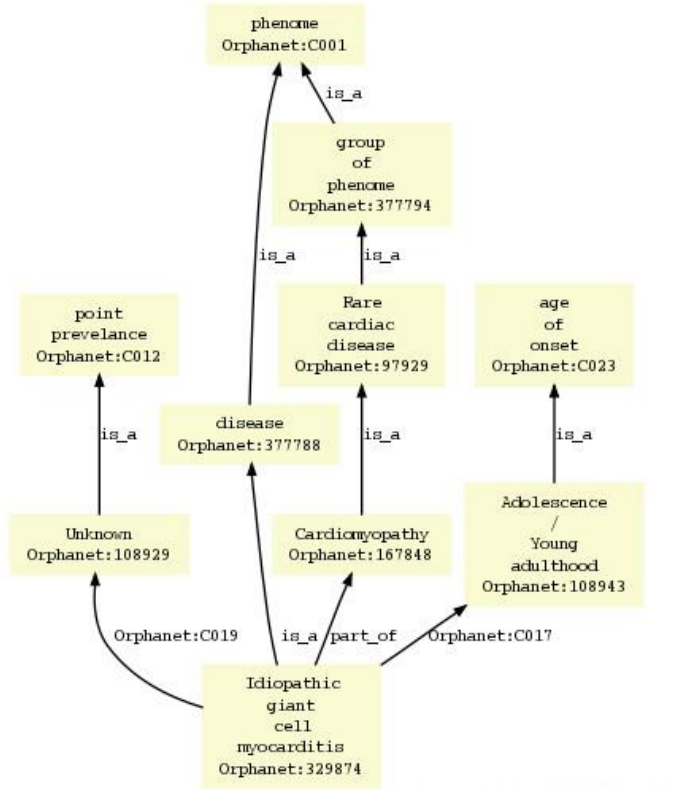
Name: Idiopathic giant cell myocarditis

Associated information

alternative:term	IGCM

Term Hierarchy

Paths to Root: Child relationships:



You can zoom the ontology browser by clicking on a term in the graph.

Consensus at IRDiRC level

- New tools being developed to allow data sharing of unsolved exomes/genomes should use and/or incorporate this terminology and these ontologies
  - GeneMatcher
  - LOVD
  - PhenomeCentral
  - DECIPHER
- ORDO and HPO formally adopted by IRDiRC

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM

Please disseminate these tools
to speed up R&D
to the benefit of the patients

Thank you for your attention

IRDiRC
INTERNATIONAL
RARE DISEASES RESEARCH
CONSORTIUM