

# FAIRifying genotype and phenotype data for rare diseases

Friederike Ehrhart  
Helis Academy – FAIR data stewardship  
05.11.19

1. FAIRification of clinical phenotype data
2. Interoperability of genetic variant data across databases
3. (Linking knowledge by linking identifiers for rare genetic diseases)

# 1. Example project: MolData2

**Task 1**  
Status quo of rare  
disease molecular  
data

**Task 3**  
Harmonizing data  
across different  
sources

**Task 2**  
Interoperability  
status and  
workflows for  
FAIRification

**Task 4**  
Creation and  
improvement of  
resources

**Task 5**  
Application use  
case: what can be  
done with the  
data?



Chris Evelo

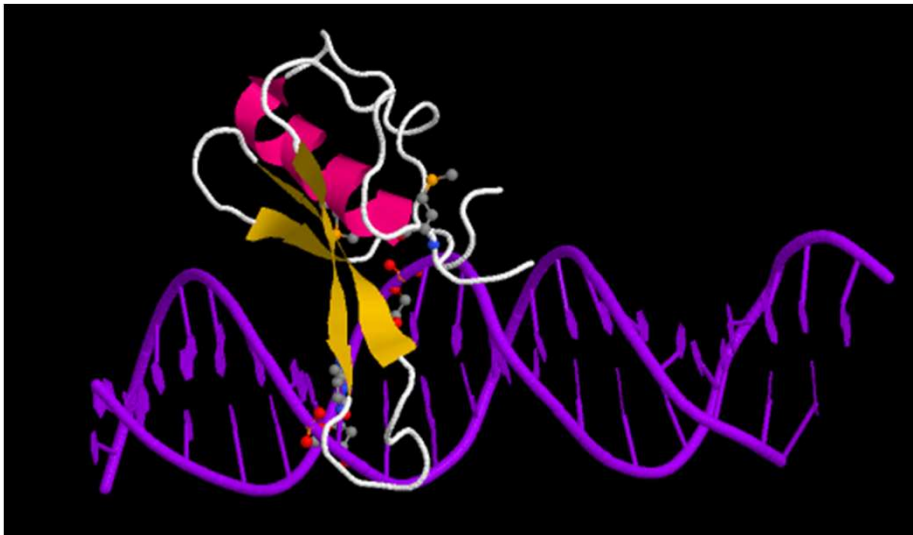


Marco Roos  
Annika Jacobsen

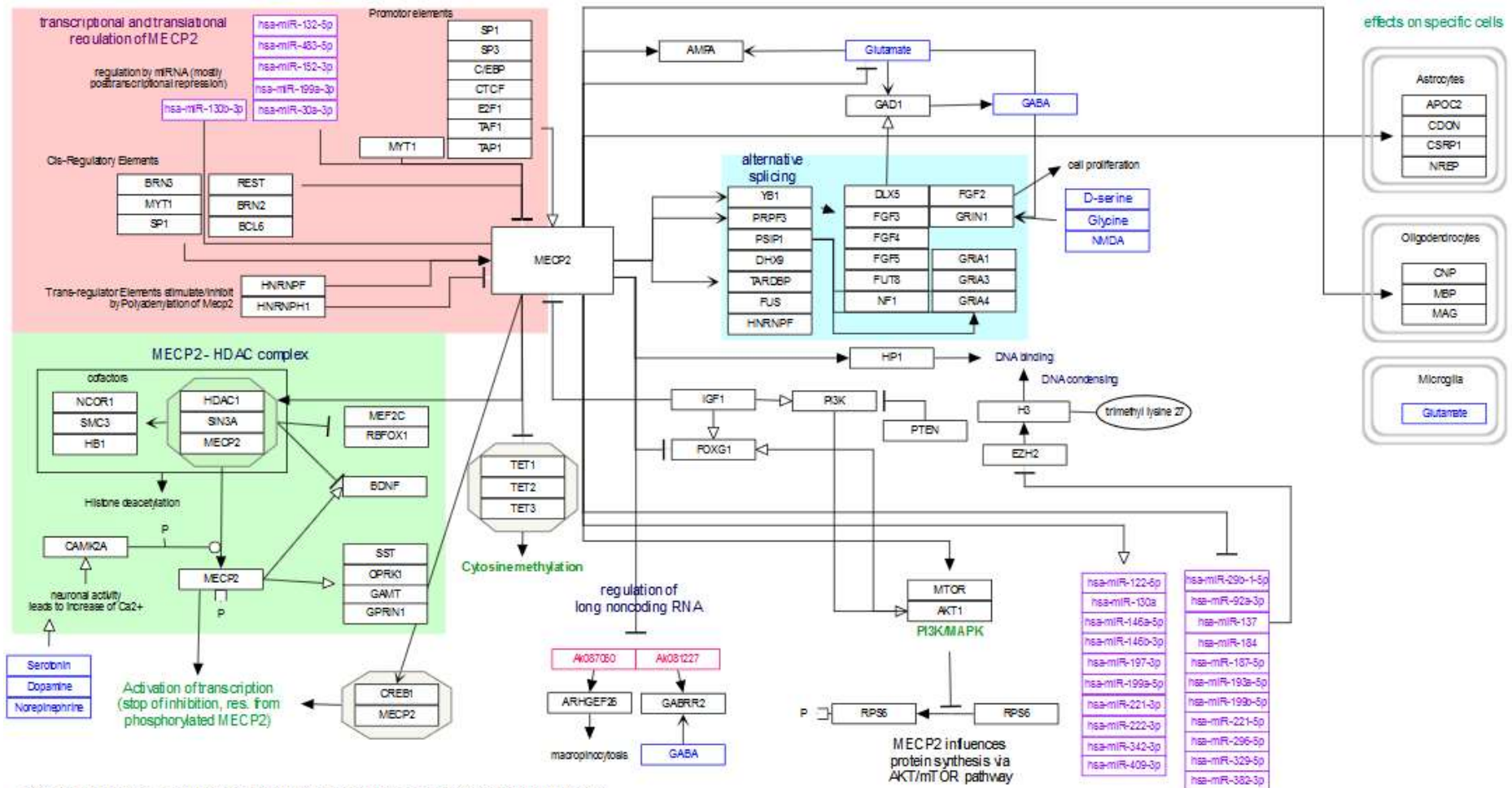


# Rare Disease data from Rett Syndrome patients

- Rare ( 1 : 10-15.000), monogenic (MECP2) disorder
- Diagnosis criteria: stagnation and loss of acquired motoric and communication skills at the age of 6 – 18 months
- Phenotype: severe mental and physical disability, stereotypic movements, epilepsy, muscle tonus, scoliosis,...

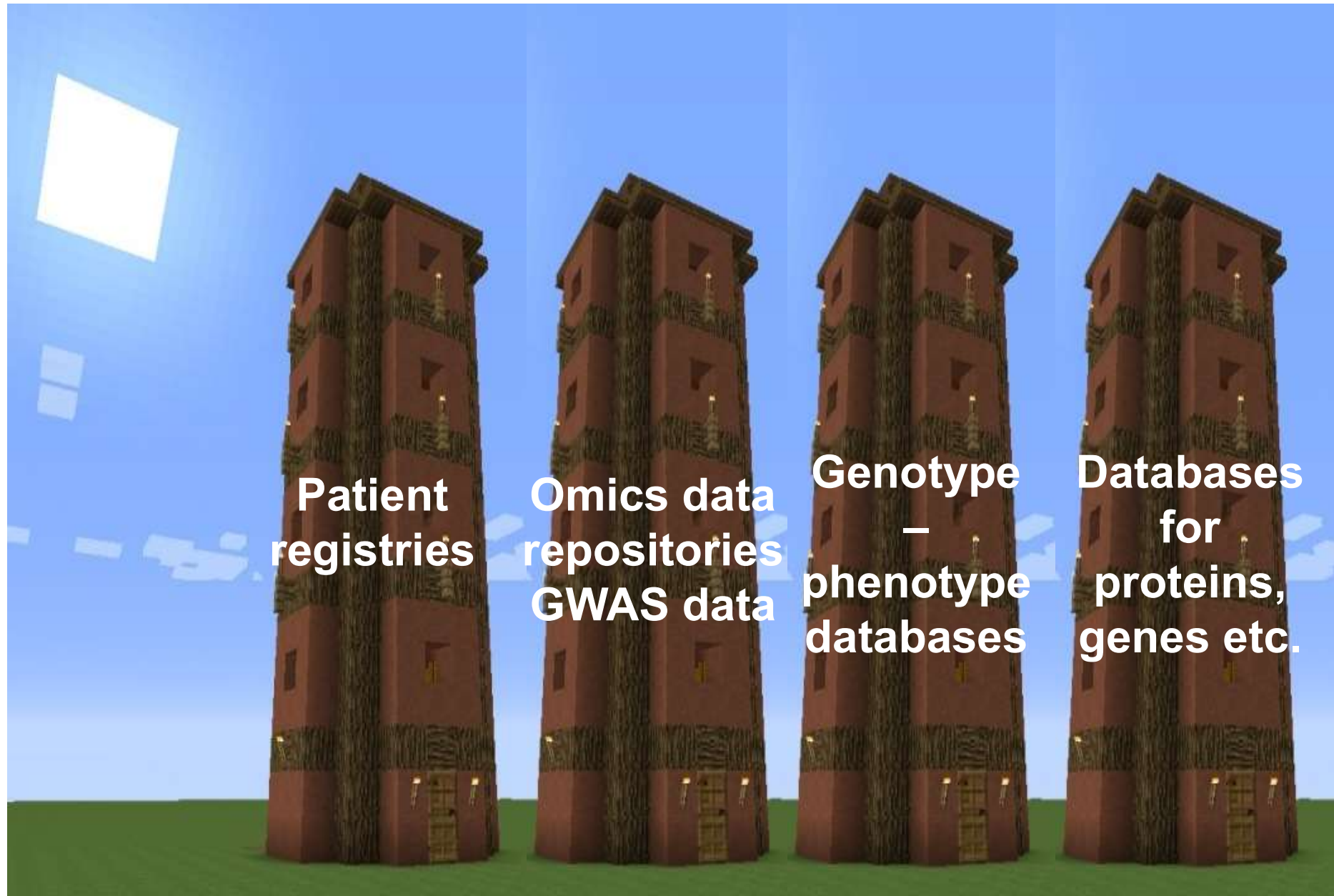


# Rare diseases teach us about gene function!



List of genes and metabolites which are up or downregulated in Rett syndrome

# Molecular Data in Rare Diseases





# Starting point

Spreadsheet from a retiring pediatric neurologist with patient data, genetic information, diagnosis, Kerr score, personal notes from specialist



## Task

- FAIRify this!
- Share in RD-connect/Orphanet

## • Genotype-phenotype data



## Task

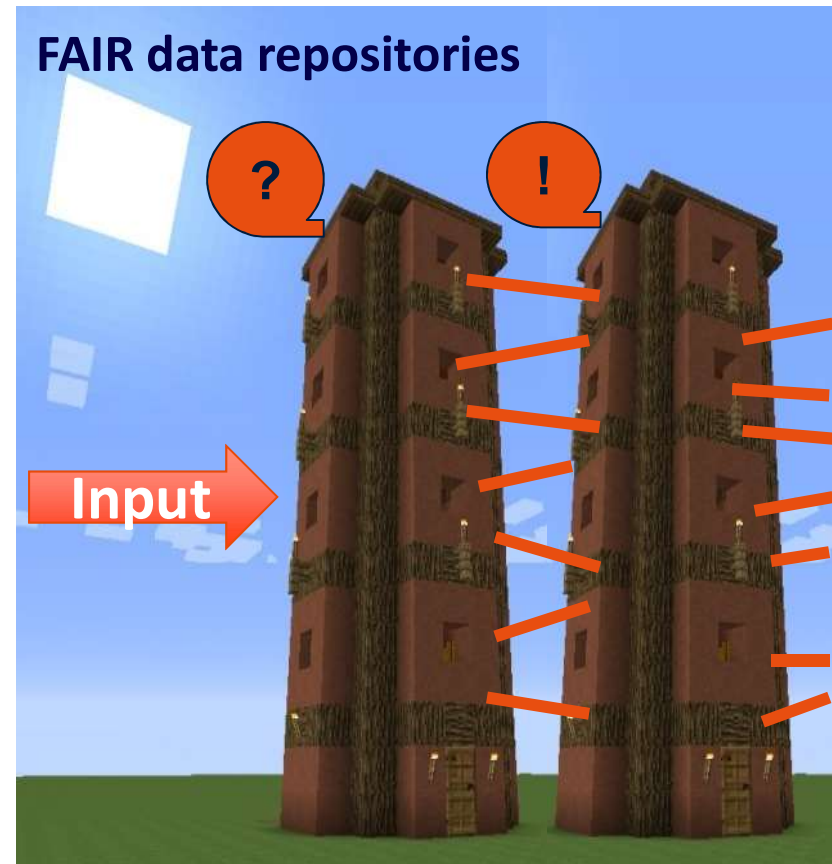
- Upload to genotype/phenotype database
- Download existing information
- Modelling

## • Resource update/improvement

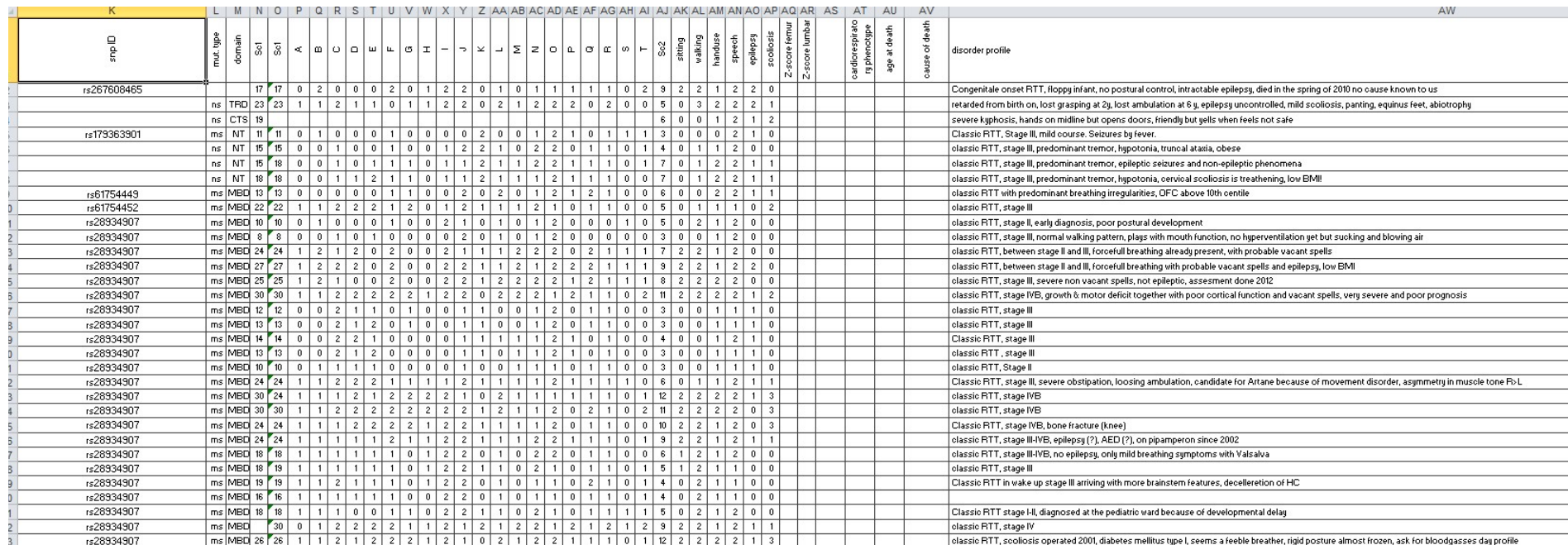


## Task

- WikiPathways for rare diseases
- gene-variant mapping database
- OMIM gene-disease linkset
- ChEMBL drug-target linkset



- 324 patients (+ 2-5 per year)
- Genetic data + diagnosis and phenotype data
- Several patients visited hospital multiple times (500+ data points)



# What would you use?

Female patient

Gender	Year of birth
?	1970



# Why using ontologies instead of plain text?

Female

female

F

01

Weiblich

W

femme

Preferred Name

Female

Synonyms

Females

ID

<http://purl.bioontology.org/ontology/MESH/D005260>

altLabel

Females

check tag only for female organs, diseases, physiologic processes, genetics, etc.; do not confuse with WOMEN as a social, cultural, political, economic force;  
CATALOGER: Do not use

AN

cui

C0086287

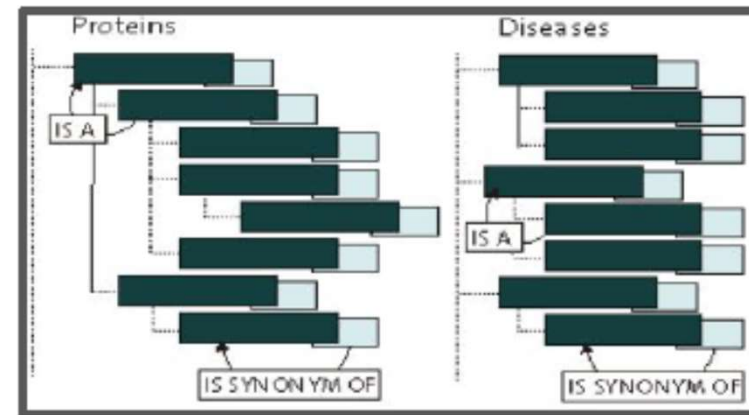
# Structure matters: but we do not define it a priori

**Controlled Vocabulary (CV):**  
An authoritative list of terms



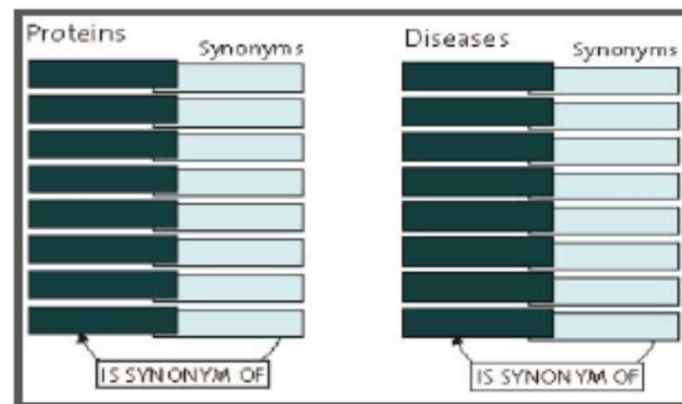
**Taxonomy:**

A CV with a tree-hierarchical (parent/child term) structure



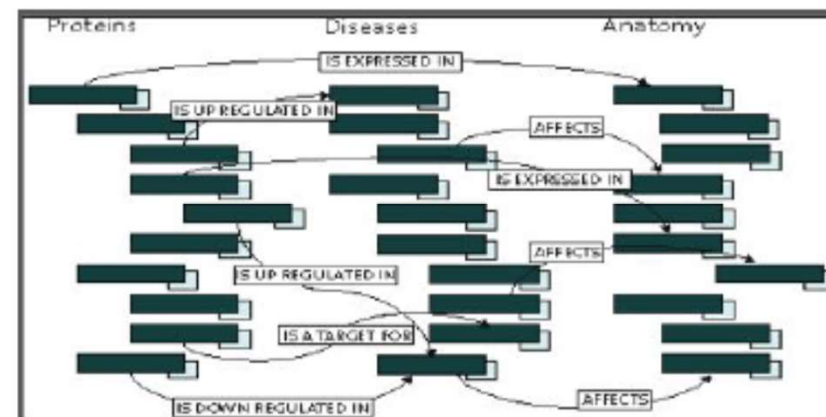
**Thesaurus:**

A kind of taxonomy with structure and specific types of relationships between terms



**Ontology:**

is a kind of taxonomy, but the types of relationships are greater in number and more specific in their function



No matter which structure we will create, it will be machine-readable and FAIR!!

# Kerr score

Score	Kerr score Description	HPO	HPO
<b>A</b>	<b>Head circumference during the first year</b>		
<b>2</b>	Already below the third percentile at birth	<a href="http://purl.obolibrary.org/obo/HP_0011451">http://purl.obolibrary.org/obo/HP_0011451</a>	Congenital microcephaly
<b>1</b>	Normal at birth but decelerating	<a href="http://purl.obolibrary.org/obo/HP_0000253">http://purl.obolibrary.org/obo/HP_0000253</a>	progressive microcephaly
<b>0</b>	Normal at birth with no deceleration	-	
<b>B</b>	<b>Early developmental progress(birth to 12 months)</b>		
<b>2</b>	No or virtually no progress	<a href="http://purl.obolibrary.org/obo/HP_0007281">http://purl.obolibrary.org/obo/HP_0007281</a>	developmental stagnation
<b>1</b>	Suboptimal progress	<a href="http://purl.obolibrary.org/obo/HP_0012758">http://purl.obolibrary.org/obo/HP_0012758</a>	neurodevelopmental delay
<b>0</b>	Normal progress	-	
<b>C</b>	<b>Present head circumference (Pc/SD)</b>		
<b>2</b>	Below 3 <sup>rd</sup> percentile	<a href="http://purl.obolibrary.org/obo/HP_0005484">http://purl.obolibrary.org/obo/HP_0005484</a>	postnatal microcephaly
<b>1</b>	3 to 10 <sup>th</sup> percentile	<a href="http://purl.obolibrary.org/obo/HP_0000253">http://purl.obolibrary.org/obo/HP_0000253</a>	progressive microcephaly
<b>0</b>	Above the 10th percentile	-	
<b>D</b>	<b>Weight (kg)</b>		
<b>2</b>	Below the 3 <sup>rd</sup> percentile	<a href="http://purl.obolibrary.org/obo/HP_0045082">http://purl.obolibrary.org/obo/HP_0045082</a>	<b>decreased BMI</b>
<b>1</b>	3 to 10 <sup>th</sup> percentile	<a href="http://purl.obolibrary.org/obo/HP_0045082">http://purl.obolibrary.org/obo/HP_0045082</a>	<b>decreased BMI</b>
<b>0</b>	Above the 10th percentile	-	

# Manual annotations

name as found in Rett database	ID	ID name (if different)
Phelan Mc Dermid syndrome	<a href="http://identifiers.org/OMIM/606232">http://identifiers.org/OMIM/606232</a>	Phelan Mc Dermid syndrome
Lesch-Nyhan syndrome	<a href="http://identifiers.org/OMIM/300322">http://identifiers.org/OMIM/300322</a>	Lesch-Nyhan syndrome
stereotypical hand wringing	<a href="http://identifiers.org/hpo/HP:0012171">http://identifiers.org/hpo/HP:0012171</a>	stereotypical hand wringing
abiotrophy	<a href="http://purl.obolibrary.org/obo/HP_0007495">http://purl.obolibrary.org/obo/HP_0007495</a>	Prematurely aged appearance
epilepsy	<a href="http://identifiers.org/hpo/HP:0001250">http://identifiers.org/hpo/HP:0001250</a>	seizures
seizures by fever	<a href="http://identifiers.org/hpo/HP:0002373">http://identifiers.org/hpo/HP:0002373</a>	febrile seizures
tremor	<a href="http://identifiers.org/hpo/HP:0001337">http://identifiers.org/hpo/HP:0001337</a>	tremor
lethargic	<a href="http://identifiers.org/hpo/HP:0001254">http://identifiers.org/hpo/HP:0001254</a>	lethargy
autistiform behaviour	<a href="http://identifiers.org/hpo/HP:0000729">http://identifiers.org/hpo/HP:0000729</a>	autistic behaviour
sleep disorder	<a href="http://identifiers.org/hpo/HP:0002360">http://identifiers.org/hpo/HP:0002360</a>	sleep disturbance
hyperactivity	<a href="http://identifiers.org/hpo/HP:0000752">http://identifiers.org/hpo/HP:0000752</a>	hyperactivity
<b>preserved speech</b>		0
<b>friendly and quiet behaviour</b>		0
night crying	<a href="http://identifiers.org/hpo/HP:0030215">http://identifiers.org/hpo/HP:0030215</a>	inappropriate crying
mood changes	<a href="http://identifiers.org/hpo/HP:0001575">http://identifiers.org/hpo/HP:0001575</a>	mood changes

## Ontologies used

- OMIM
- HPO

# Drugs

Drug	drugbank ID	drug name (if different)
VITD3	<a href="http://identifiers.org/drugbank/DB00169">http://identifiers.org/drugbank/DB00169</a>	Cholecalciferol
Baclofen	<a href="http://identifiers.org/drugbank/DB00181">http://identifiers.org/drugbank/DB00181</a>	
sinemet/Carbidopa	<a href="http://identifiers.org/drugbank/DB00190">http://identifiers.org/drugbank/DB00190</a>	
R-Tramadol	<a href="http://identifiers.org/drugbank/DB00193">http://identifiers.org/drugbank/DB00193</a>	Tramadol
pantomed	<a href="http://identifiers.org/drugbank/DB00213">http://identifiers.org/drugbank/DB00213</a>	Pantoprazol
diphantoine	<a href="http://identifiers.org/drugbank/DB00252">http://identifiers.org/drugbank/DB00252</a>	Phenytoin
Depakine	<a href="http://identifiers.org/drugbank/DB00313">http://identifiers.org/drugbank/DB00313</a>	Valproic Acid
omeprazole	<a href="http://identifiers.org/drugbank/DB00338">http://identifiers.org/drugbank/DB00338</a>	
frisium	<a href="http://identifiers.org/drugbank/DB00349">http://identifiers.org/drugbank/DB00349</a>	Clobazam
artane	<a href="http://identifiers.org/drugbank/DB00376">http://identifiers.org/drugbank/DB00376</a>	Trihexyphenidyl
loretidine	<a href="http://identifiers.org/drugbank/DB00455">http://identifiers.org/drugbank/DB00455</a>	Loratadine
Fluoxetine/Prozac	<a href="http://identifiers.org/drugbank/DB00472">http://identifiers.org/drugbank/DB00472</a>	
Buspiron/Buspar	<a href="http://identifiers.org/drugbank/DB00490">http://identifiers.org/drugbank/DB00490</a>	
dextrometorphan	<a href="http://identifiers.org/drugbank/DB00514">http://identifiers.org/drugbank/DB00514</a>	
Lamictal	<a href="http://identifiers.org/drugbank/DB00555">http://identifiers.org/drugbank/DB00555</a>	Lamotrigine

# The (almost) FAIR Maastricht Rett dataset

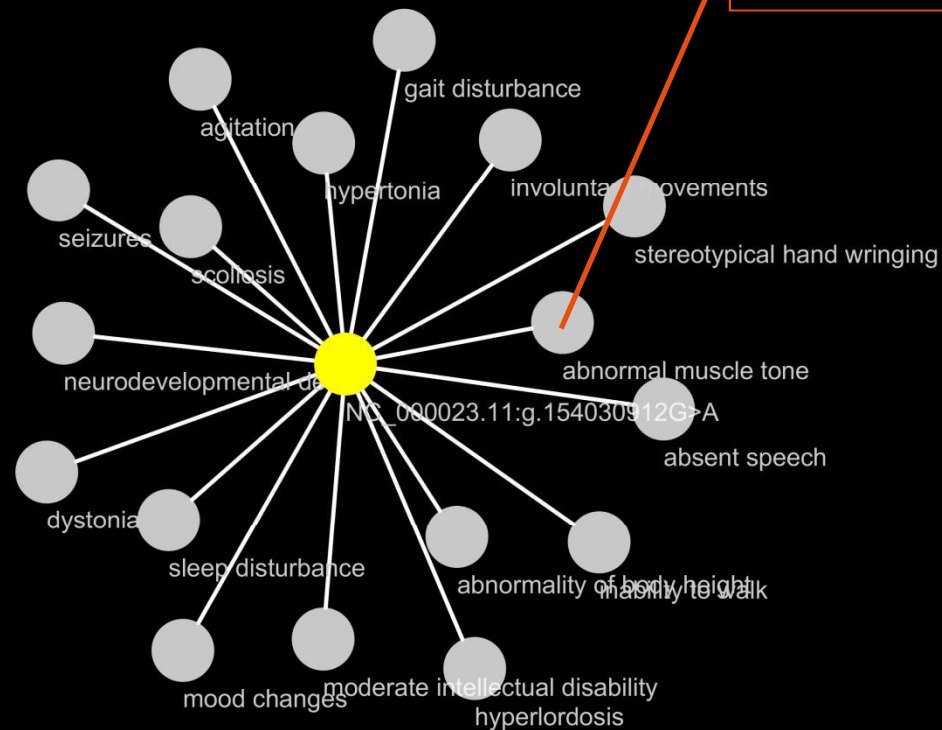
- Not allowed to share patient data without ethical approval and informed consent of care givers
  - But publication of anonymous genotype-phenotype data was allowed after ethical approval
- FAIRification
  - Genetic information – standardized HGNC format
  - Diagnosis information – annotated with OMIM
  - Phenotype information – Kerr score mapped to HPO
  - Individual annotations and observations – annotated with HPO, medication with Drugbank ID
- Problems experienced:
  - Data capture spreadsheets from clinicians come with different input information
  - Broad variety of formats for genetic information – “lab slang”, historic formats, different RefSeq
  - Medical technical language, “slang”



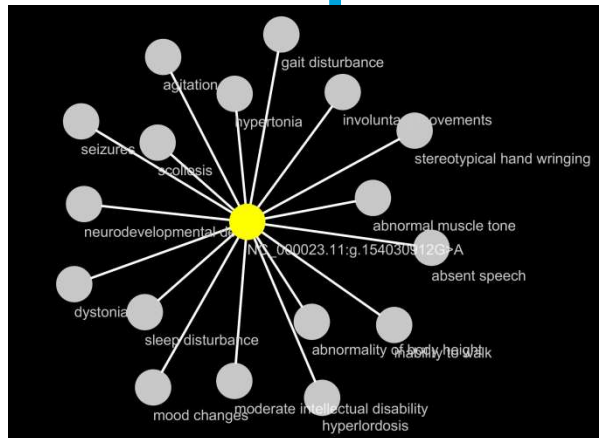
# The gain of interoperable data

- Genotype – phenotype of one patient c.925C>T mutation

[http://purl.obolibrary.org/obo/HP\\_0003808](http://purl.obolibrary.org/obo/HP_0003808)

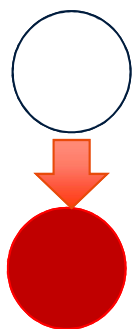


# More patients, more information



1 mutation, 1 patient

1 mutation, multiple patients

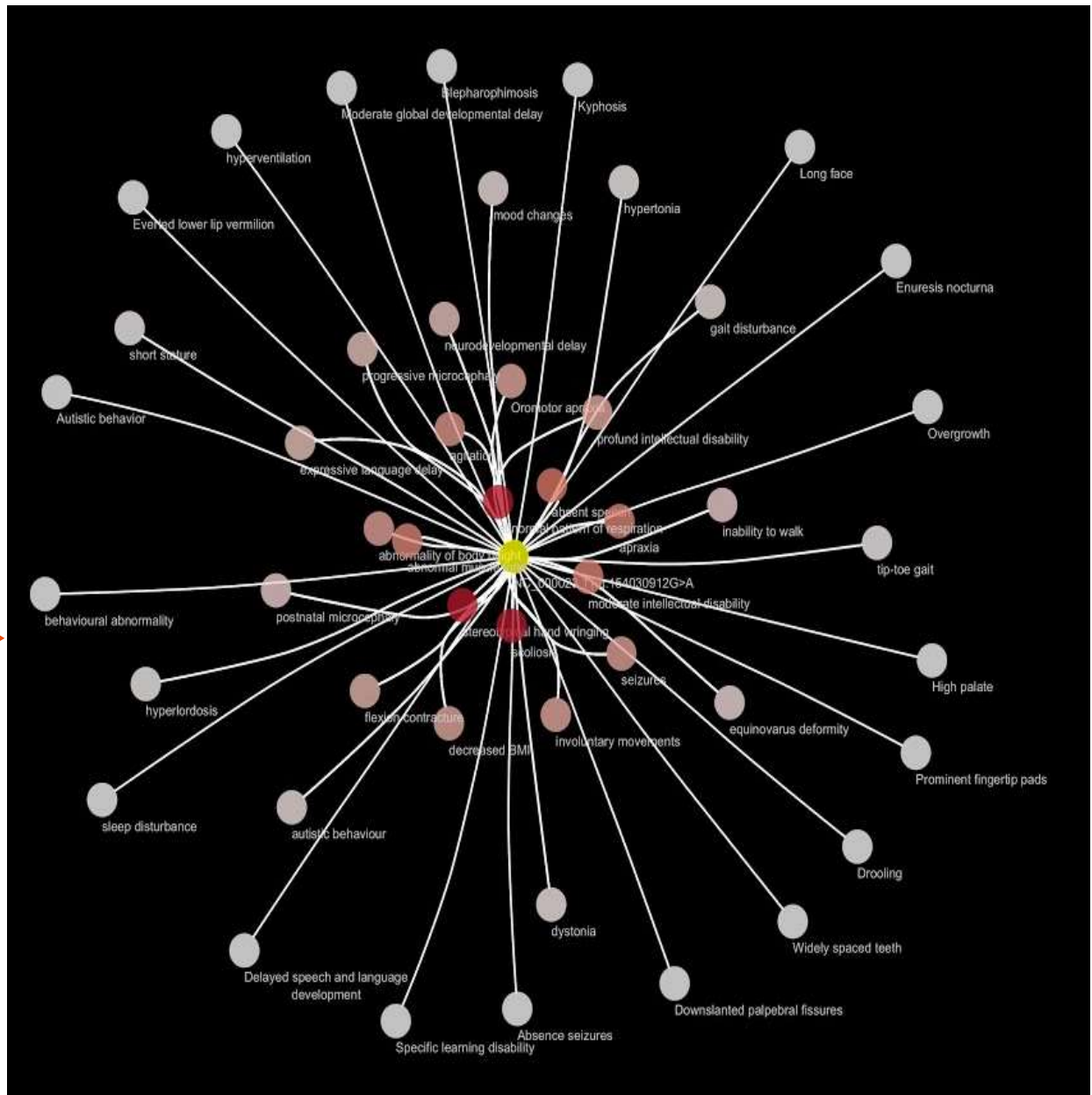


Observed once

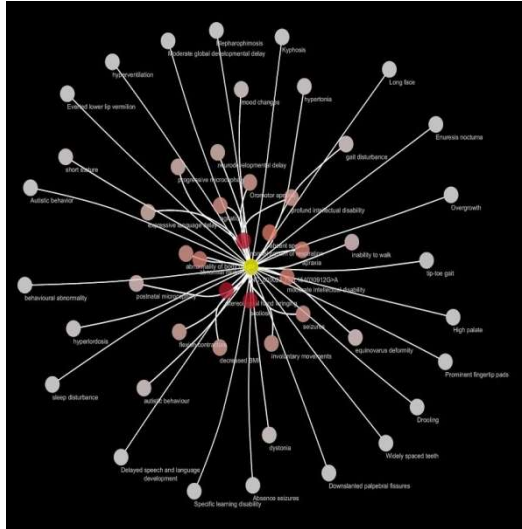
Observed in many patients



Maastricht University



# Statistics possible?



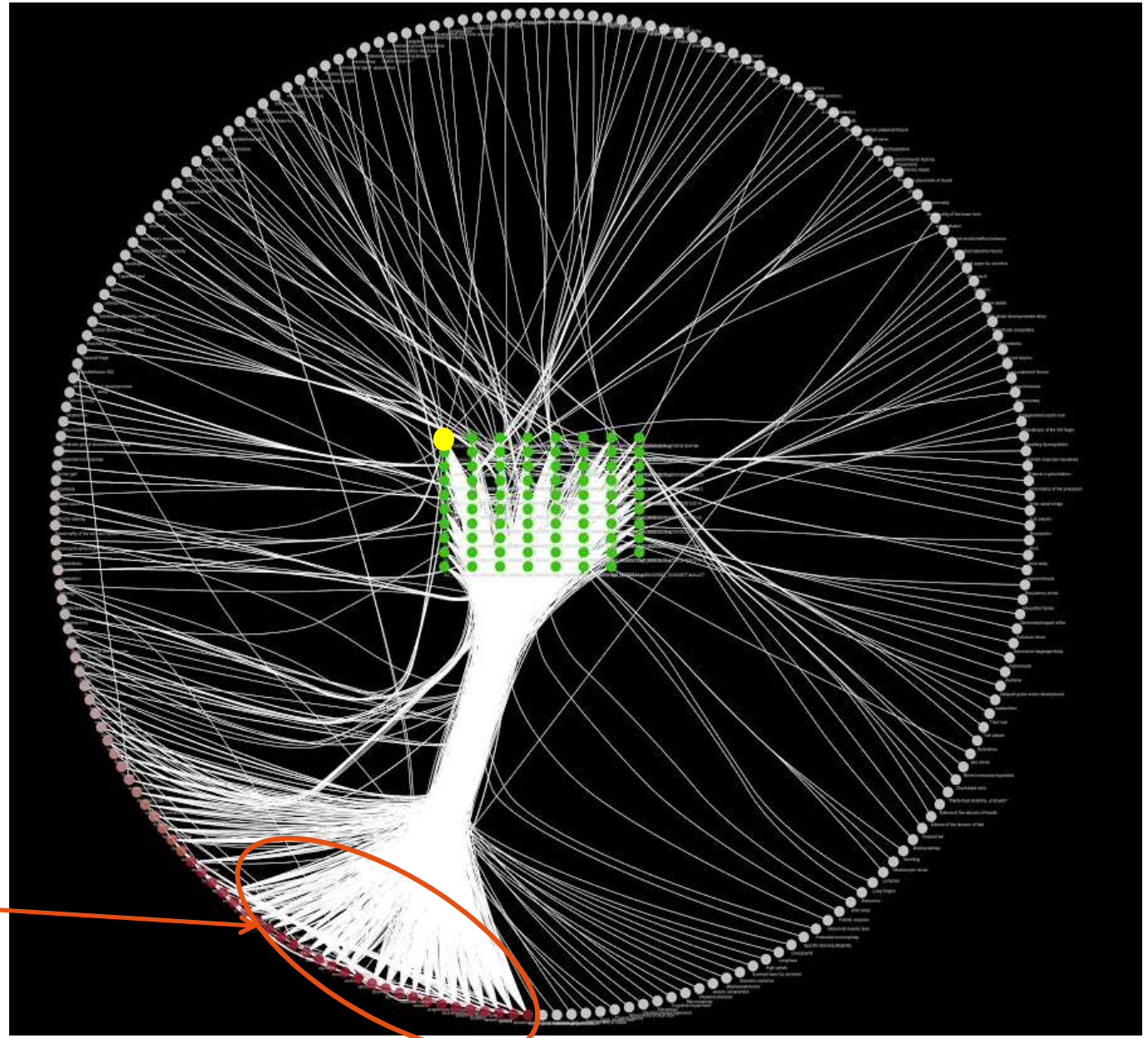
1 mutation, multiple patients

79 mutations, 356 patients

**Diagnosis criteria**



Maastricht University

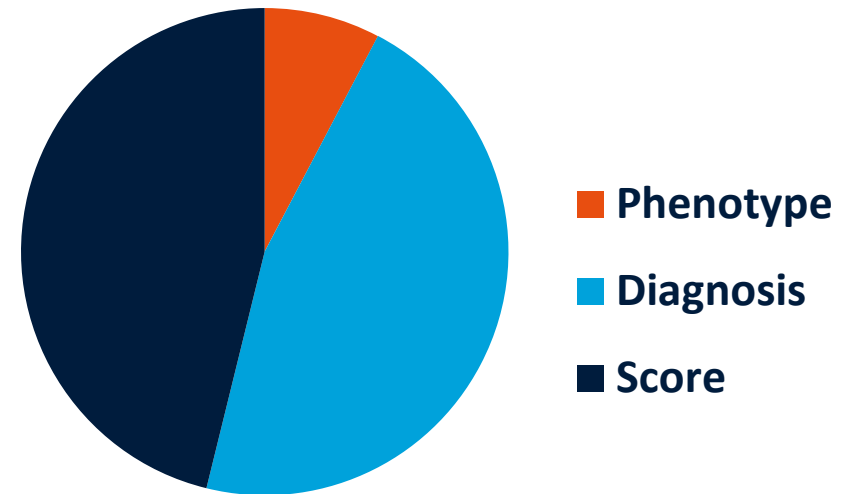


## 2. Genotype-phenotype databases

Database	# MECP2 variant entries
RettBase	4705 (892 unique)
KMD (Korean Mutation Database)	35
ClinVar	1103
LOVD (Leiden Open Variation Database)	4472 (806 unique)
DECIPHER (Database of genomic variation and Phenotype in Humans using Ensembl Resources)	196 (32 annotated)
EVS (Exome Variant Server)	95
EVA (European Variation Archive)	423
ExAC Browser (Exome Aggregation Consortium)	594
Cafe Variome (Meta database)	PhenCode: 809; UniProtKB: 71; Human Gene Mutation Database: 249; Locus-specific Databases: 10
dbSNP	12105 (790 unique exonic)
dbVAR	492
*Townend et al. 2018 (in revision)	Status: July 2017

# Phenotype?!

- RettBase
- KMD (Korean Mutation Database)
- ClinVar
- Cafe Variome
  - PhenCode, Uniprot...
- **DECIPHER**
- **LOVD** (Leiden Open Variation Database)
- **EVS** (Exome Variant Server)
- **EVA** (European Variation Archive)
- **ExAC Browser** (Exome Aggregation Consortium)
- **dbSNP**
- **dbVAR**



Standard recommended:

**DOID** – human disease ontology

**ORDO** – Orphanet rare disease ontology

**HPO** – human phenotype ontology



# Problems in data integration (1)

## Genotype data

- Broad variety of formats
  - Actual e.g. HGVS, RS, different RefSeq
  - Historic formats

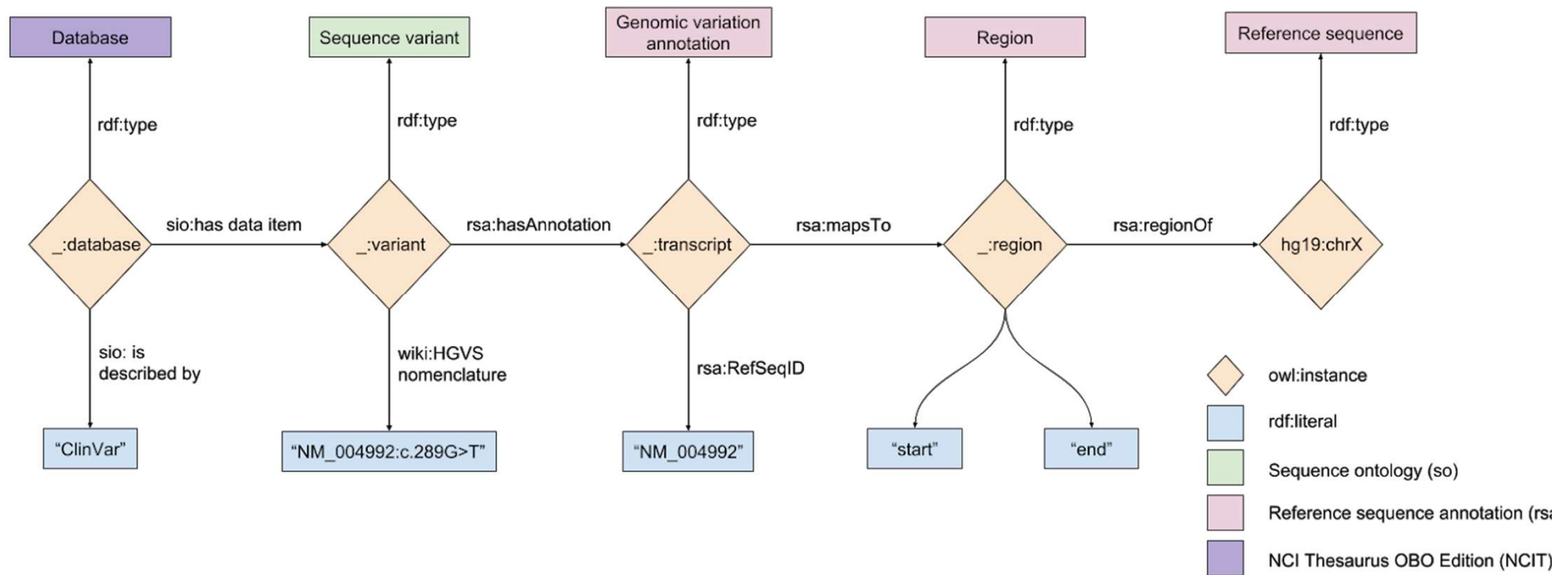


**Mutalyzer**

[www.mutalyzer.nl](http://www.mutalyzer.nl)



# Semantic model variant description



NM\_004992:c.289G>T

# Problems in data integration (1)

## Genotype data

- Broad variety of formats
  - Actual e.g. HGVS, RS, different RefSeq
  - Historic formats

## Consequent application of these formats

# Mutalyzer and manual curation

<https://mutalyzer.nl/>

Database	Number of variations	Criteria	Number of variations meeting criteria	Mutalyzer		Manual curation	
				n	%	n	%
Maastricht Rett dataset	429	Rett syndrome, Rett syndrome preserved speech variant, congenital variant	428	388	<b>90.7</b>	393	<b>91.8</b>
ClinVar	1134	Rett syndrome or X-linked mental retardation 13 ("male" Rett syndrome) or encephalopathy, clinical significance NOT benign	562	500	<b>89.0</b>	545	<b>97.0</b>
RettBase	4705	Rett syndrome (including variants and atypical forms) or X-linked mental retardation, with MECP2 variation likely causing disorder	3544	3372	<b>95.1</b>	3542	<b>99.9</b>
KMD	35	Rett syndrome or Rett variant (OMIM)	35	35	<b>100</b>	35	<b>100</b>
EVS	190	MECP2 - pathogenic	22	22	<b>100</b>	22	<b>100</b>

# Problems in data integration (2)

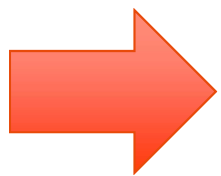
## Data

- Broad variety of formats
  - Actual e.g. HGVS, RS, different RefSeq
  - Historic formats

## Consequent application of these formats

## Low degree of interoperability of phenotypic data (use of ontologies), disease information or pathogenicity scores

## Re-use or data accession permission often hidden



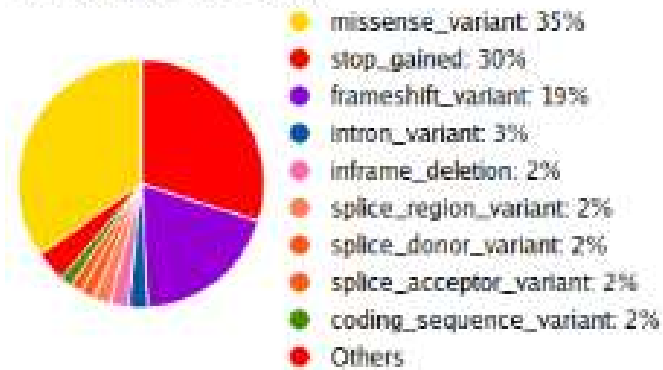
Dataset with:

- about 4000 RTT causing variants (800 unique)
- about 400 benign MECP2 variations (X unique)  
(available as spreadsheet or RDF (soon))

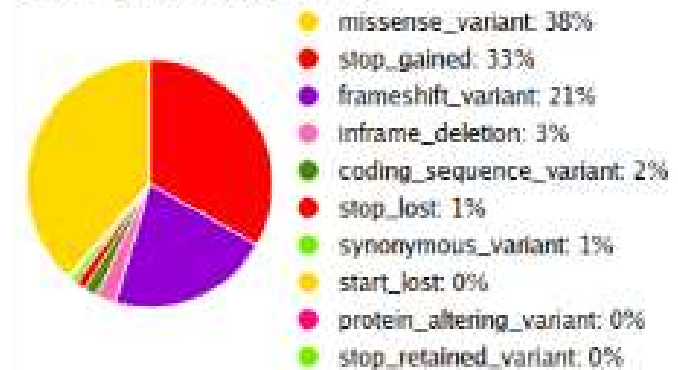
# Variant Effect Predictor (VEP) analysis

## 4000 RTT causing variants

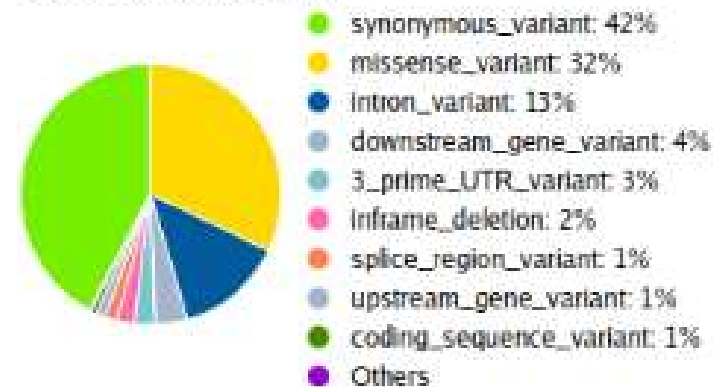
Consequences (all)



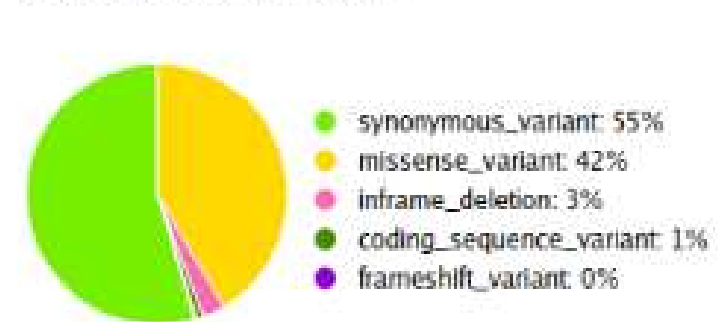
Coding consequences



Consequences (all)

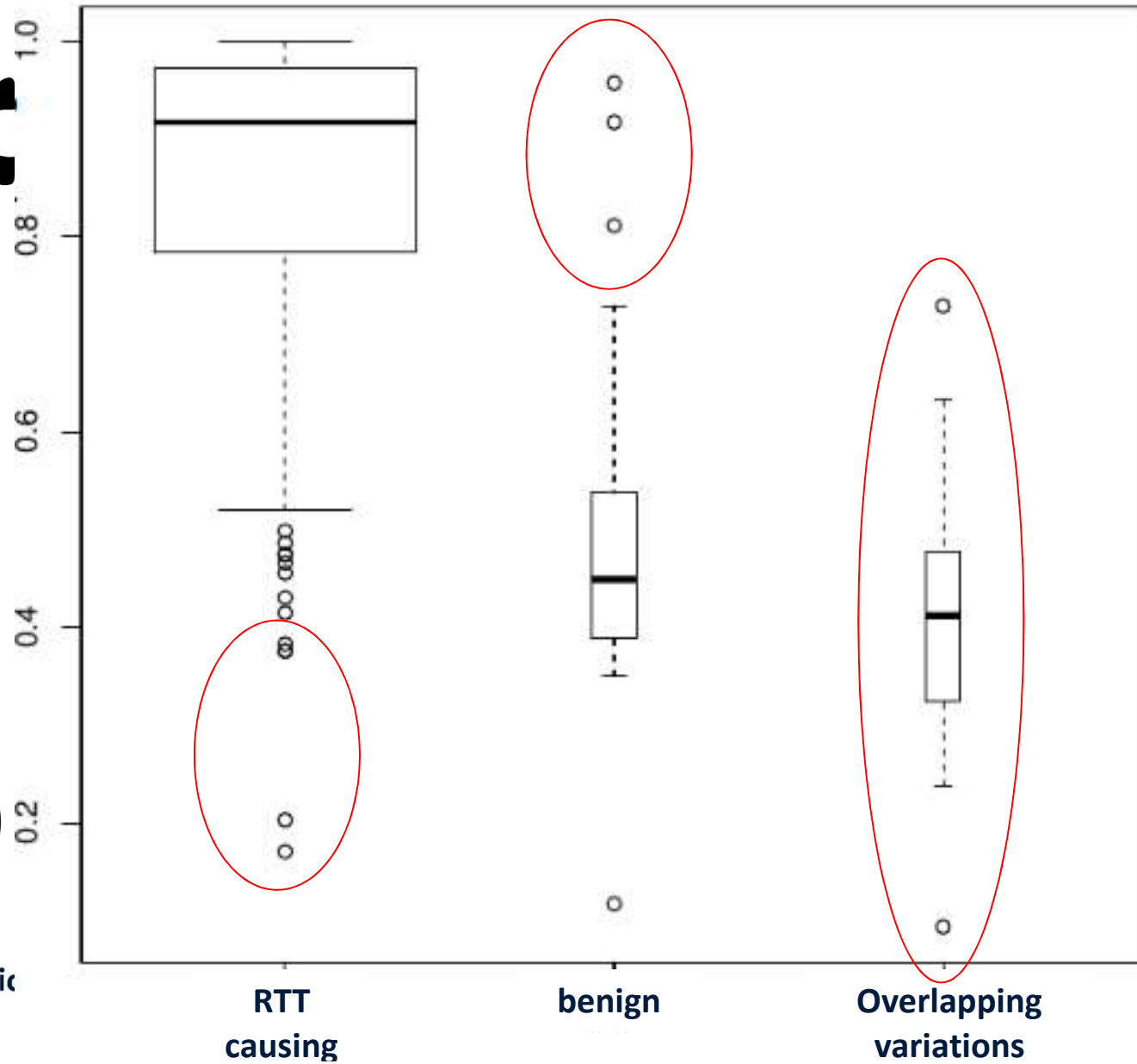


Coding consequences



## 400 benign variants

# MetaLR scores





# TL:DL (too long, did not listen)

- **What did we do?**
  - **FAIRified** original clinical dataset of rare disease genotype / phenotype data
    - Genotype – HGVS standard
    - Semantic model and data in RDF format
    - Disease specific scoring (Kerr score) – mapped to HPO
    - Clinicians annotation – mapped to ontologies
- **Identified problems**
  - Different **minimum information standards** or different **(levels of) information** recorded
  - Data **not yet mapped to ontologies** for phenotypic data used for diagnosis and inconsistent description of diagnosis itself
    - Historic data can be curated manually but only to a certain degree
    - New recorded data should be collected and properly annotated using the agreed community standard
  - **Broad variety of formats** for genetic information used – current and historic (e.g. HGVS, RS, reference sequences) - **Mutalyzer** can often fix that.
  - **Consequent application** of nomenclature in these formats (5-12% not correct) – education required!

### 3. Rare disease documentation - from anecdotes to medical catalogues and databases

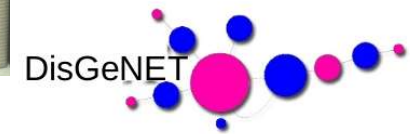
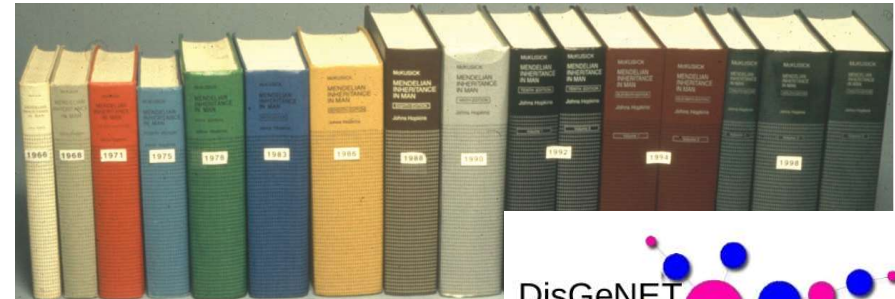
***THE BRITISH MEDICAL JOURNAL.***

**[June 30, 1888.]**

ON RARE DISEASES AND EXCEPTIONAL  
SYMPTOMS.

By JONATHAN HUTCHINSON, F.R.C.S., F.R.S., LL.D.,  
Emeritus Professor of Surgery at the London Hospital.

*“Mendelian Inheritance in Man (MIM)”* Dr. Victor A. McKusick



Since 1987 *“Online Mendelian Inheritance in Man (OMIM)”*

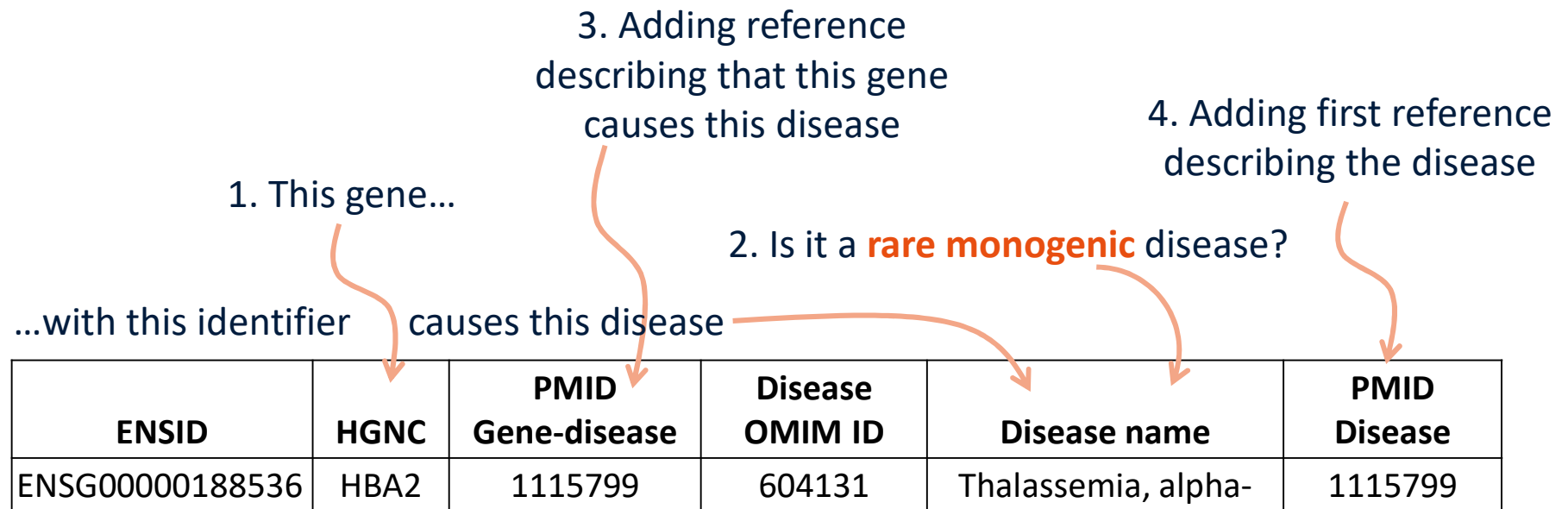


**Whonamedit?**

orphanet

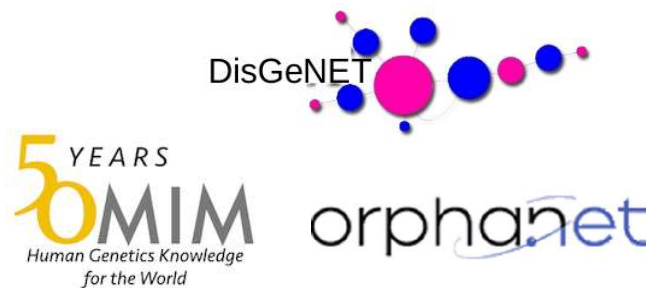


# Approach



Gene

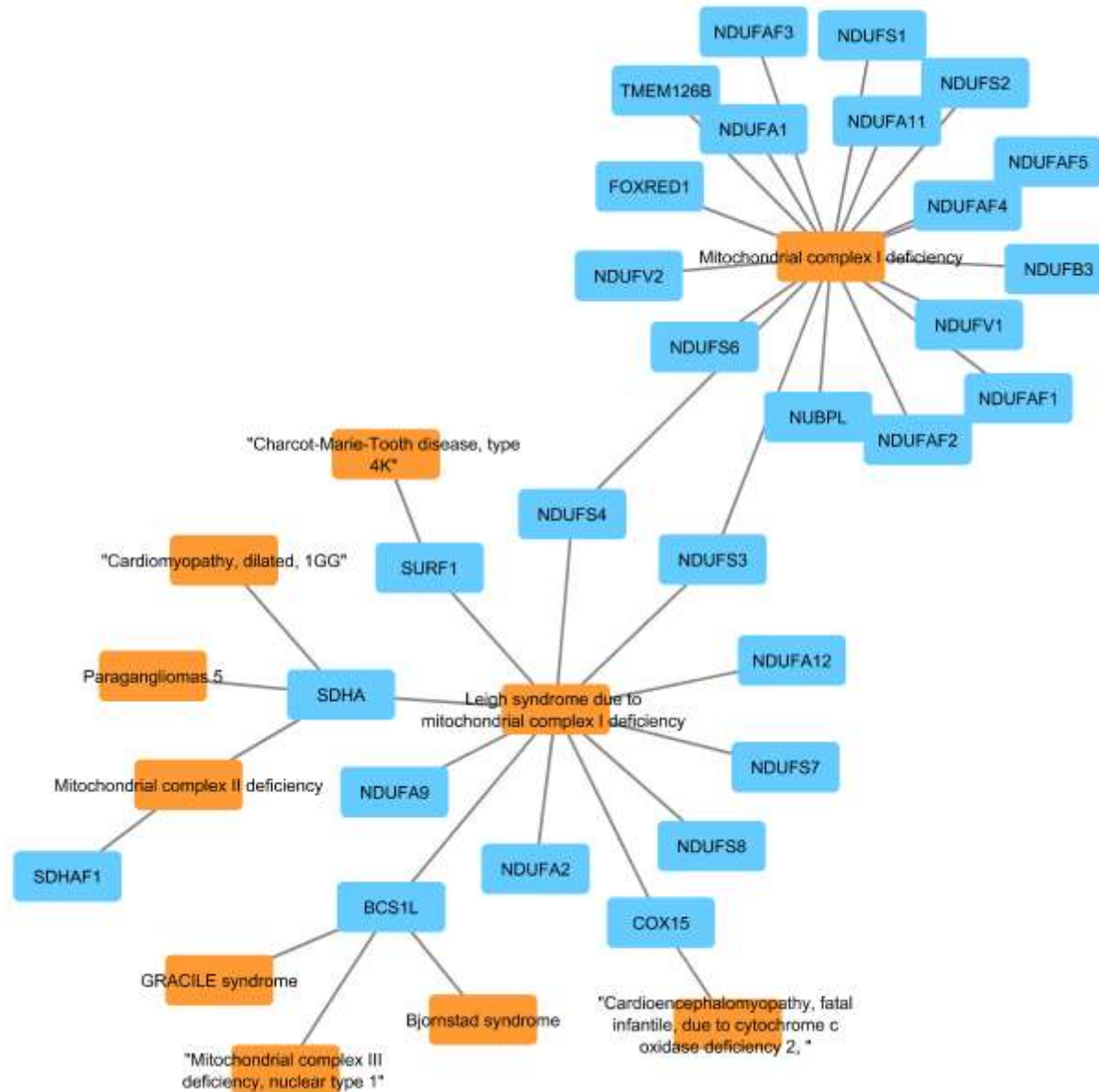
Disease



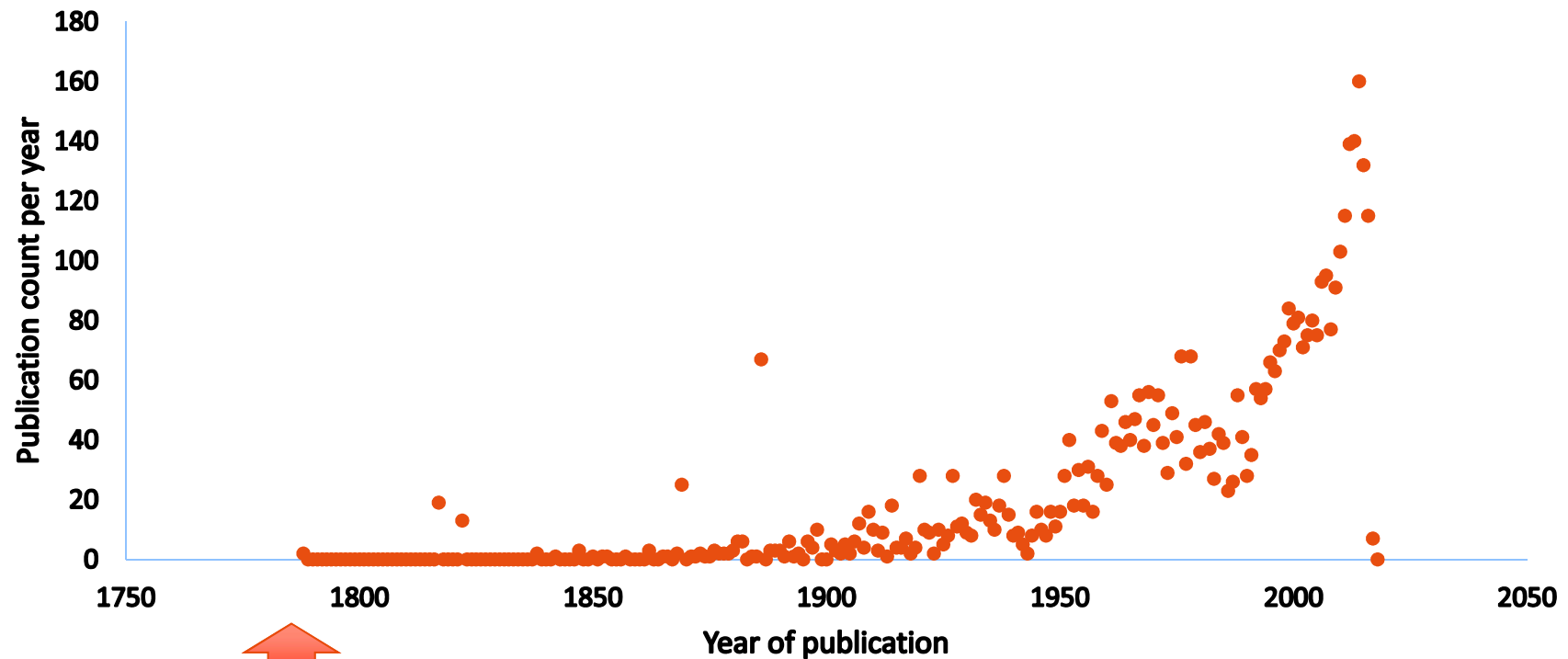
# Gene – Disease networks

E  
ENSG00

	PMID Disease
la-	1115799



# Time line of rare disease description

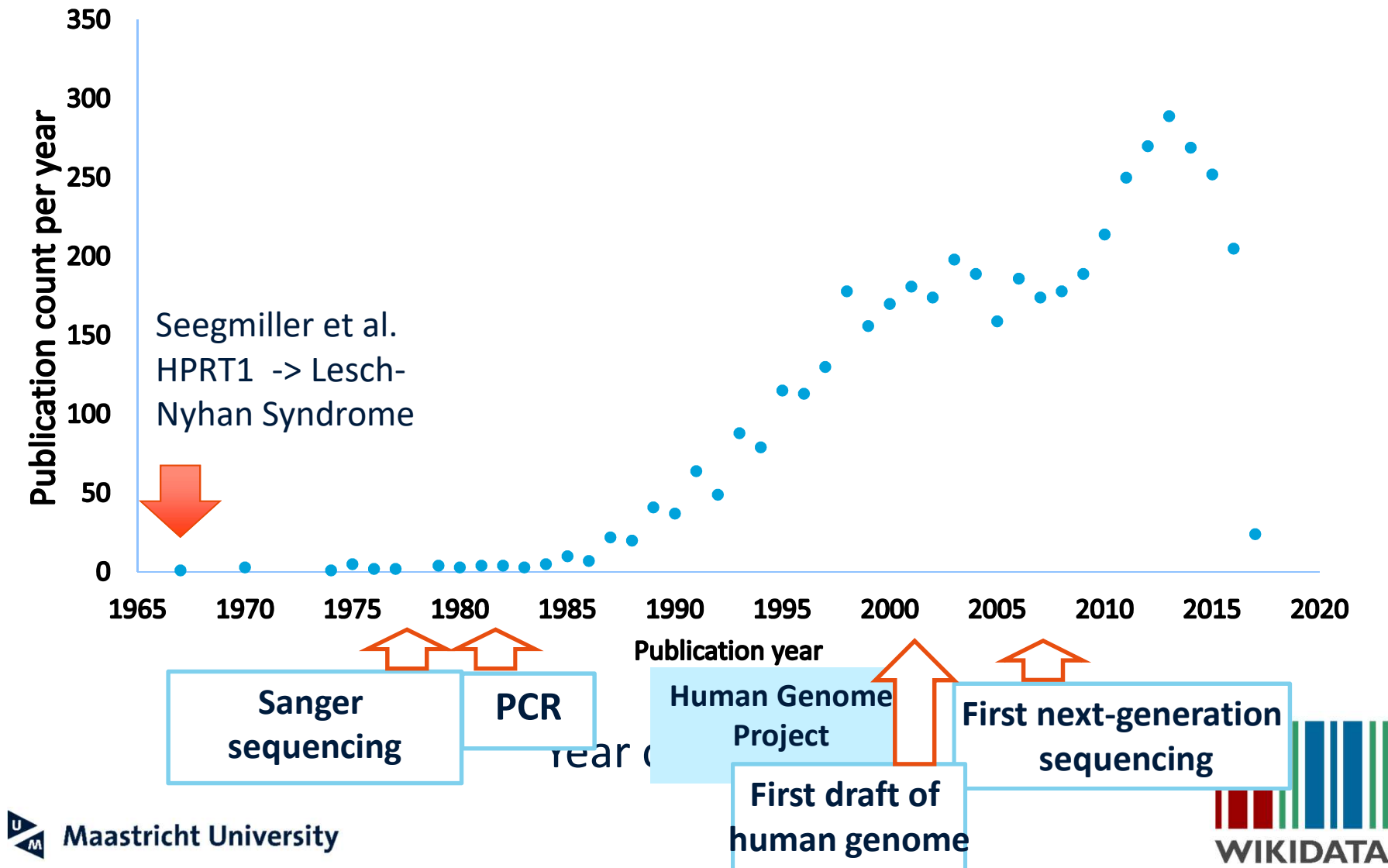


1788 Olof Ekman  
osteogenesis imperfecta

PMID

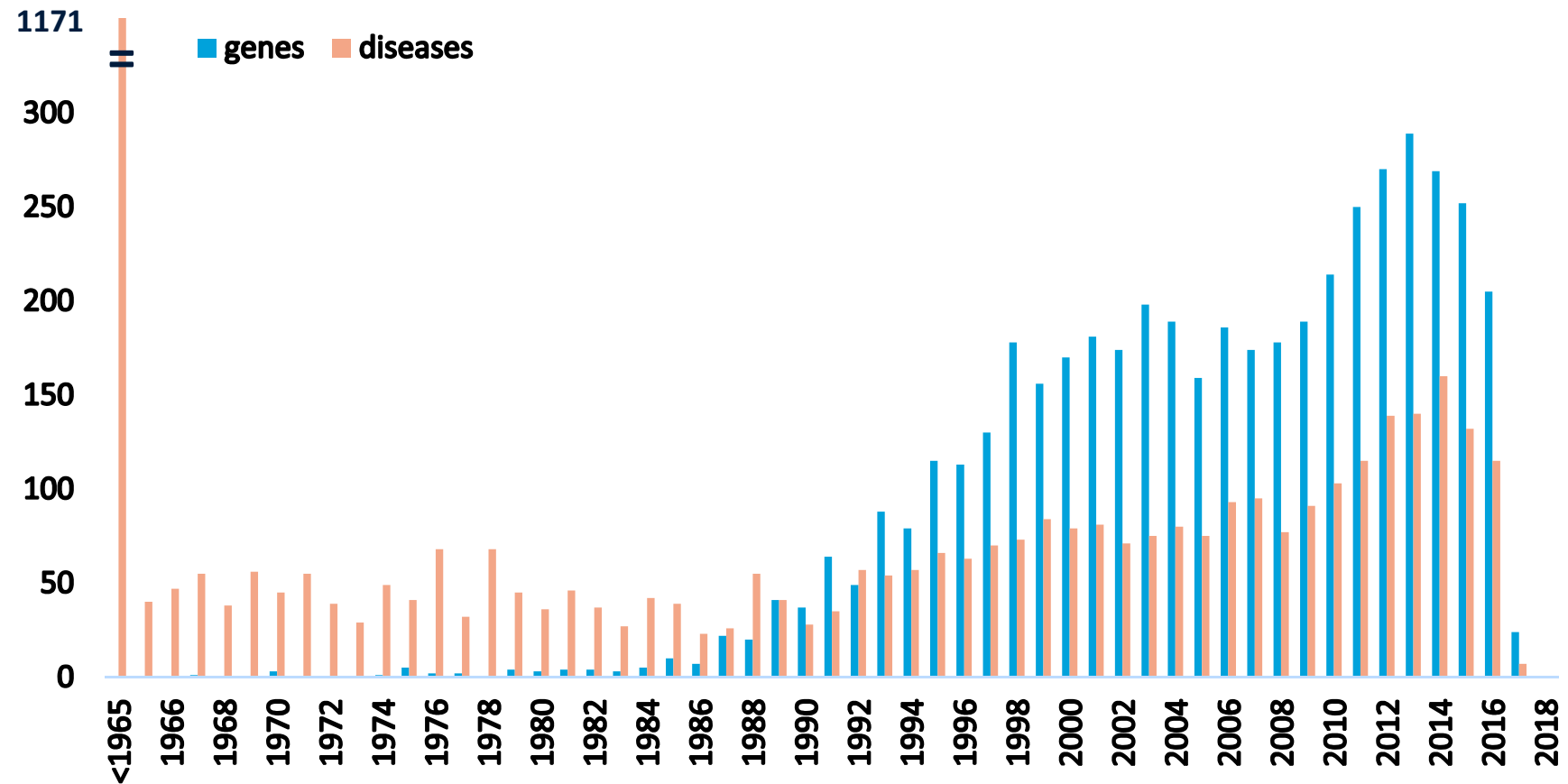


# Timeline of discovery of underlying genetic causes

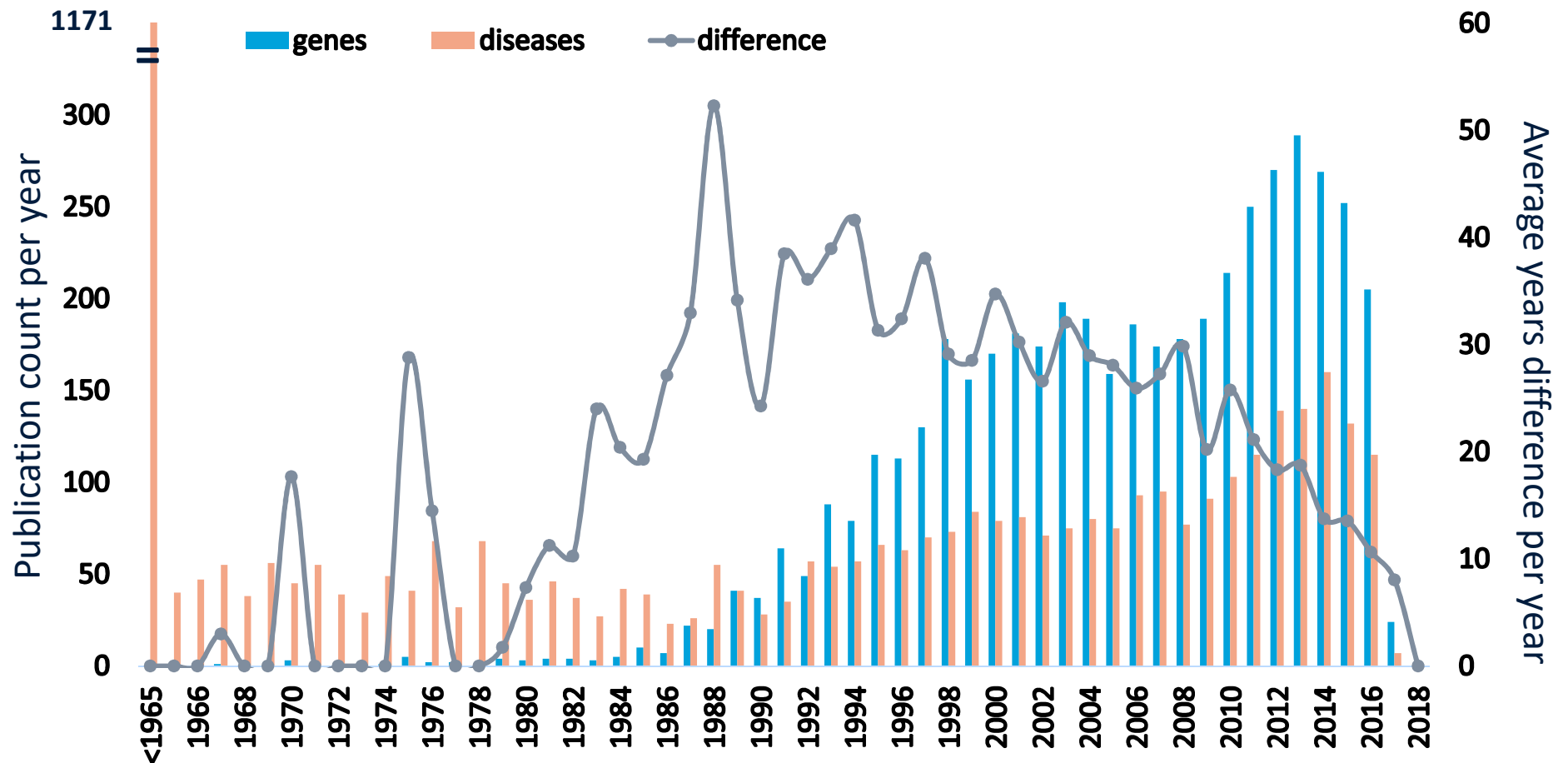




# Genes vs. Diseases



# Difference in **gene** vs. **disease** discovery



# Journals

**Disease**

**Gene**

journal	count	journal	count
<b>First description of a new disease (100% = 3144)</b>		<b>First description of new genes causing a rare disease (100% = 4263)</b>	
American Journal of Human Genetics	457	American Journal of Human Genetics	1137
Nature Genetics	145	Nature Genetics	786
American Journal of Medical Genetics Part A	137	Human Molecular Genetics	266
Journal of Medical Genetics	136	Journal of Medical Genetics	175
Human Molecular Genetics	122	The New England Journal of Medicine	148
The New England Journal of Medicine	121	Journal of Clinical Investigation	136
Journal of Clinical Investigation	78	Proceedings of the National Academy of Sciences of the United States of America	129
Neurology	77	Science	121
The Journal of Pediatrics	68	Nature	99
Proceedings of the National Academy of Sciences of the United States of America	59	Human Mutation	94

# Authors

- 2641 authors of **gene** discovery papers have an ORCID which is about 15% of all authors
- 1637 of authors listed are male (62.4%) and 986 female (37.6%) (18 no gender listed)
- 993 of first rare **disease** description papers (19.7 %)

Top 10 of **gene** discoverers:

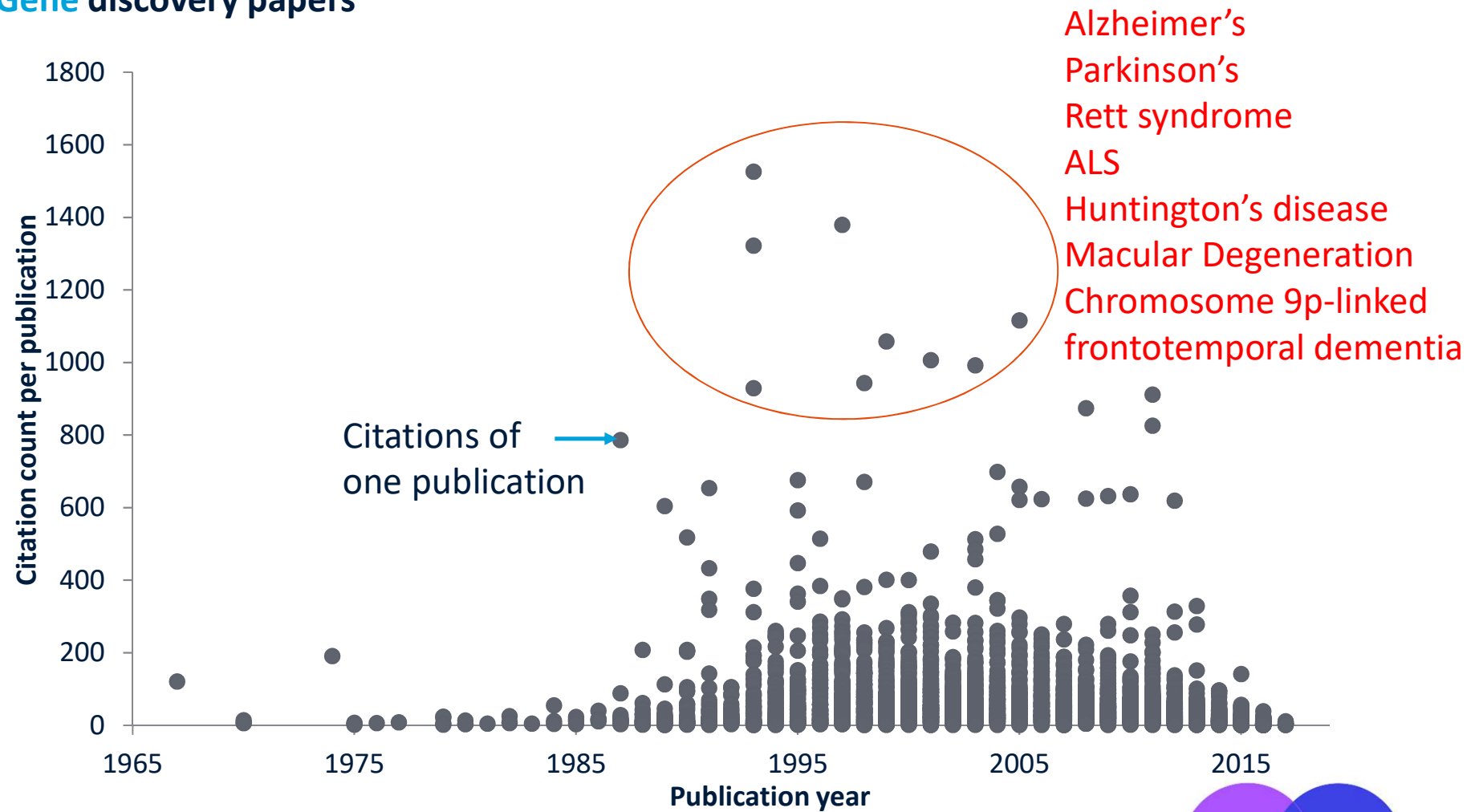
Author name	count	gender	nationality
Arnold Munnich	58	m	French
Gudrun Nürnberg	44	f	German
Peter Nürnberg	38	m	German
Thomas Meitinger	35	m	German
Nicholas Katsanis	32	m	British
Friedhelm Hildebrandt	31	m	USA
Jean-Laurent Casanova	31	m	French
Alexis Brice	30	m	French
Edgar A Otto	19	m	USA
Bruno Dallapiccola	28	m	Italian

ORCID



# Citation scores – rare or truly neglected diseases?

Gene discovery papers



# Harvesting information from other resources

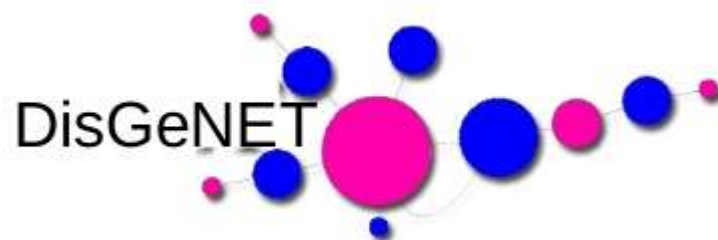
1. Identifier & entity mapping
2. Harvesting interesting information

Disease OMIM ID	CUI	Disease SemanticType	MeSH subclass	ORPHA	Epidemiology
604131	C0002312	Disease or Syndrome	Nervous System Diseases [mesh:C10]	846	>1 / 1000
609597	C1865044	Disease or Syndrome	#N/A	#N/A	#N/A
300624	C0016667	Disease or Syndrome	#N/A	908	1-5 / 10 000
604290	C1858583	#N/A	Hemic and Lymphatic Diseases [mesh:C15]	48818	1-9 / 1 000 000



Concept Unique Identifiers (CUI)

ORPHAnet ID



# Inheritance mode



OMIM ID	Inheritance	count	%
100000–299999 AND 600000 and above	autosomal	4263	93.4
300000–399999	X	299	6.5
400000–499999	Y	3	0.1

## MIM numbers

- 100000–299999: Autosomal loci or phenotypes (created before May 15, 1994)
- 300000–399999: X-linked loci or phenotypes
- 400000–499999: Y-linked loci or phenotypes
- 500000–599999: Mitochondrial loci or phenotypes
- 600000 and above: Autosomal loci or phenotypes (created after May 15, 1994)



# Epidemiology



Rare genetic causes of more common phenotypes/diseases



Epidemiology	count
>1 / 1000	855
6-9 / 10 000	5
1-5 / 10 000	72
1-9 / 100 000	208
1-9 / 1 000 000	215
<1 / 1 000 000	158
single cases/families	735
Unknown	354
#N/A	1963

No mapping from OMIM/CUI to ORPHA available



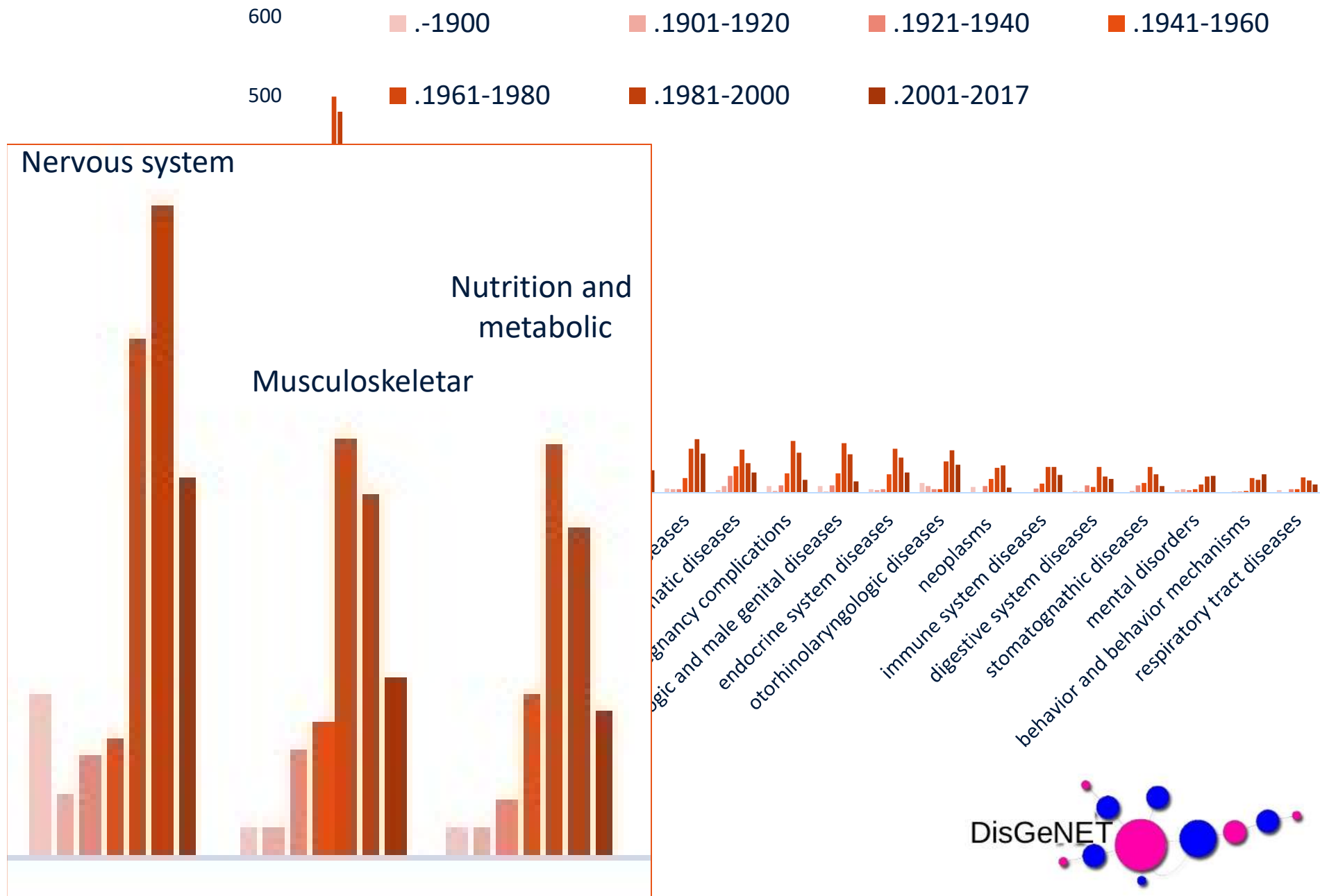
# Disease superclasses



Top 10 MeSH terms	Count
congenital, hereditary, and neonatal diseases and abnormalities	1705
nervous system disease	945
musculoskeletal diseases	595
nutritional and metabolic diseases	549
pathological conditions, signs and symptoms	415
eye diseases	328
skin and connective tissue diseases	296
cardiovascular diseases	203
hemic and lymphatic diseases	182
female genital diseases and pregnancy complications	174

58.7% of diseases are annotated

# Disease class timeline



# Thanks for attention – questions?



“Ooo, two strays to add to the database.”



Annika Jacobsen  
Marco Roos  
Rajaram Kaliyaperumal



Salvador Capella  
Maria Rigau  
Mattia Bosio



Chris Evelo  
Egon Willighagen  
Jonathan Melius

Special thanks to: Henk van Kranen, Jeroen Laros, David van Enckevort, Mark Wilkinson, Leopold Curfs

Cartoon: <https://www.tibco.com/blog/2013/03/24/here-comes-the-chief-information-integration-officer/>