

BLAH4, 10 Jan 2018

# Identifying disease-phenotype associations using PubAnnotation and PubDictionaries for supporting diagnosis of intractable diseases

Toyofumi Fujiwara  
Database Center for Life Science



# Background

---

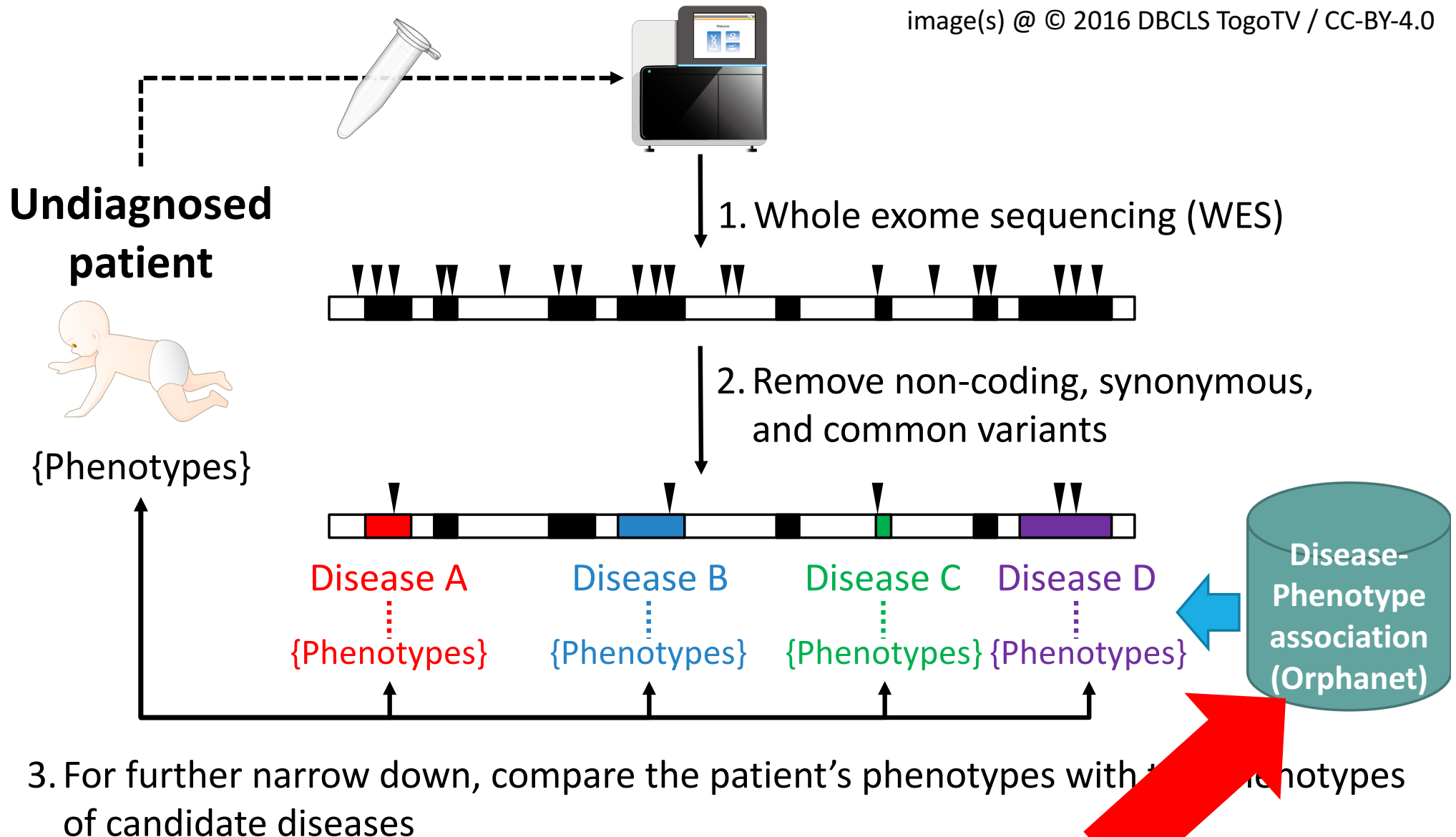
- ❑ Japan has a long history of research for diagnosis of “rare and intractable diseases” (called “Nan-Byo”)
- The Ministry of Health, Labour and Welfare of Japan (MHLW) designated 330 diseases as Nan-Byo
- MHLW made MHLW Guidelines (330 documents) for diagnosis of Nan-Byo
- MHLW supports Nan-Byo patients with medical expenses



**There are many undiagnosed patients**

# For undiagnosed patients

image(s) @ © 2016 DBCLS TogoTV / CC-BY-4.0



Disease-phenotype associations are important for this process

# Challenge

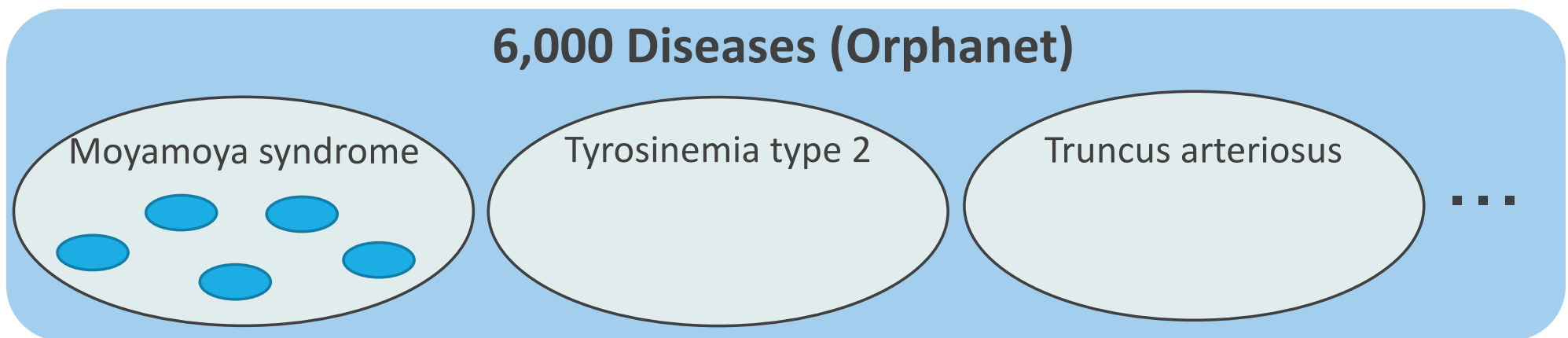
---

❑ But...

- Manually curated databases like Orphanet inherently have a limited coverage



For example, **only 1/3** of rare diseases in Orphanet are associated with phenotypes



# Prospective solution

---

- ❑ Text mining-based approach is effective to improve the coverage of disease-phenotype associations
- To automatically extract the associations from more than one million **case reports** in PubMed



As a result, **2/3** of rare diseases in Orphanet are associated with phenotypes

## 6,000 Diseases (Orphanet)



# PubCaseFinder

□ A diagnosis support system for rare diseases

**PubCaseFinder** Home About API

Query phenotype(s) + Upload File (HPO List):

HP:0004444 Spherocytosis × HP:0001744 Splenomegaly × HP:0001903 Anemia × HP:0000952 Jaundice × HP:0001297 Stroke × HP:0002721 Immunodeficiency ×

Narrow down the diseases + Upload File (Entrez Gene ID List):

ENT:286 ANK1 (SPH1) × ENT:2038 EPB42 (MGC116735 | MGC116737 | PA) × ENT:6521 SLC4A1 (CD233 | FR | RTA1A | SW | WR) × ENT:6708 SPTA1 (EL2) ×

Re-search Clear

Total: 5 20 (per page)

**Specify causative disease genes to narrow down candidate diseases**

**A ranked list of rare diseases based on phenotypic similarity**

Similarity	Disease Name	Phenotype	Causative Gene
100.0%	Hereditary spherocytosis (ORDO:822)	Hemolytic anemia   Immunodeficiency   Jaundice   Spherocytosis   Splenomegaly   Stroke	ANK1   EPB42   SLC4A1   SPTA1   SPTB
81.71%	8p11.2 deletion syndrome (ORDO:251066)	Hemolytic anemia   Microcephaly   Sacral dimple   Spherocytosis   Splenomegaly	ANK1

PubCases: A diagnosis assistant tool for rare diseases based on disease-phenotype associations extracted from published case reports.,

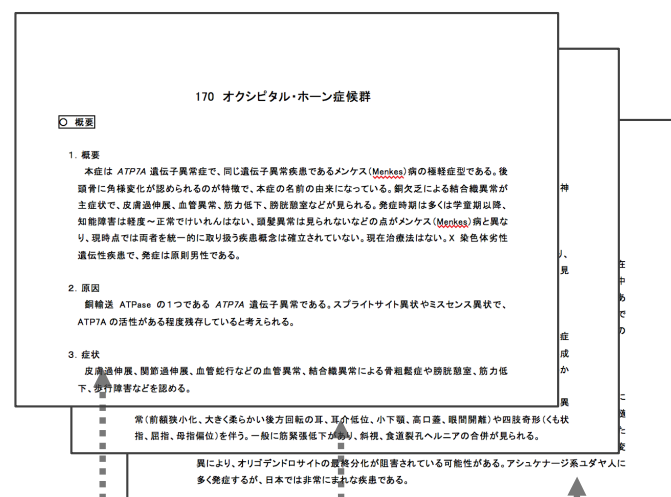
T. Fujiwara; Y. Yamamoto; J.D. Kim; T. Takagi, ASHG2017

<https://ep70.eventpilotadmin.com/web/page.php?page=IntHtml&project=ASHG17&id=170121395>

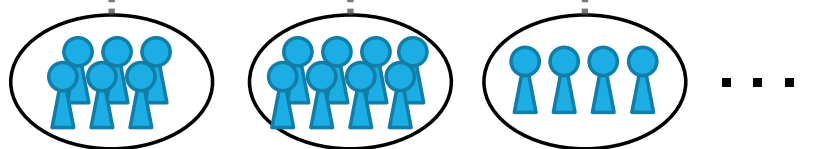
# Goal (in this hackathon)

- ❑ For getting more accurate disease-phenotype associations
  - Extracting disease-phenotype associations from MHLW Guidelines that include diagnostic criteria for Nan-Byo

330 MHLW Guidelines



Each team made each Guideline



330 research teams (composed of physicians)

HPO  
Japanese  
translations



disease-phenotype  
associations  
(HPO)

Nan-Byo 1

▪ HP:0001203,...

Nan-Byo 2

▪ HP:0023091,...

Nan-Byo 3

▪ HP:0009808,...

⋮

Nan-Byo 330

▪ HP:0012632,...

# Preliminary study

- To extract disease-phenotype associations from MHLW Guidelines for 10 Nan-Byo using PubDictionaries and PubAnnotation

## 1. To upload HPO Japanese terms to PubDictionaries

HPO  
Japanese  
translations



**PUBDICTIONARIES** [Dictionaries](#) [Find IDs](#) [Text Annotation](#) [Documentation](#) [fujitoyo](#) [Logout](#)

**HPO-japanese-alpha-18Oct2017** (10,448)

<b>creator</b>	fujiiwara@dbcls.rois.ac.jp	A Japanese version of Human Phenotype Ontology
<b>created at</b>	2017-11-27 05:49:28 UTC	
<b>updated at</b>	2017-11-28 05:09:13 UTC	

Label	Search	Id	Search
"46,XY核型での女性外性器"		HP:0008730	
"デンスデポジット糸球体腎炎, 密沈積症"		HP:0004746	
"萎縮性, 斑状禿頭"		HP:0004529	
"巨大精巣, 思春期後"		HP:0002050	
"筋波動症, ミオキミア"		HP:0002411	
"血清 calcitriol (1,25-dihydroxycholecalciferol)低値"		HP:0012052	
"血清1,25-dihydroxyビタミンD3増加"		HP:0003152	
"原始反射 (掌頭, 口とがらせ, 眉間)"		HP:0002476	
"自然切断, 自己切断 (足)"		HP:0001868	
"自然切断, 自己切断"		HP:0001218	
"女性概観をもつ性腺異発生, 男性"		HP:0008723	
"小耳, 1度"		HP:0011266	
"小耳, 2度"		HP:0008569	
"小耳, 3度"		HP:0011267	
"性腺異発生, 男性"		HP:0008668	

1 2 3 4 5 ... Next > Last >

Dictionary =  
a collection of pairs,  
(Label, ID)

Anyone can upload a  
Dic. in TSV

PubDictionaries  
immediately make it  
ready for annotation

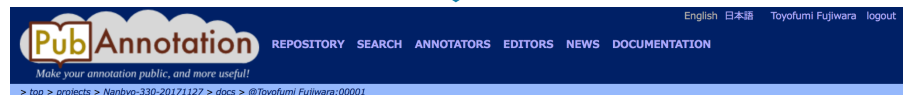


# Preliminary study

2. To upload 330 documents to PubAnnotation
3. To get dictionary-based text annotation using a REST service of PubDictionaries

MHLW guideline documents

User can upload any documents



Nanbyo-330-20171127

@Toyofumi Fujiwara:00001

JSON TEXT delete

## 1. 球脊髄性筋萎縮症

### ○ 概要

### 1. 概要

通常成人男性に発症する、遺伝性下位運動ニューロン疾患である。四肢の筋力低下及び筋萎縮、球麻痺を主症状とし、女性化乳房など軽度のアンドロゲン不全症や耐糖能異常、脂質異常症などを合併する。筋力低下の発症は通常30～60歳頃で、経過は緩徐進行性である。国際名称は Spinal and Bulbar Muscular Atrophy (SBMA) であるが、Kennedy diseaseとも呼ばれる。

### 2. 原因

X染色体長腕近位部に位置する、アンドロゲン受容体遺伝子第1エクソン内にあるCAGの繰り返し、38以上に異常延長していることが本症の原因である（正常では36以下）。CAGの繰り返し数と発症年齢との間に逆相関がみられる。男性ホルモンが、神経障害の発症・進展に深く関与していると考えられている。

Select a part of text above to get its span-url.

Annotations [VIEW] [JSON] [TextAE] delete all

- Denotations: 32
- Relations: 0
- Modifications: 0

### Automatic Annotation

Choose a predefined annotator

choose one

Prefix

a short unique character string

URL

http://pubdictionaries.org/text\_annotation

Method

☐ GET ☒ POST

Annotation can be obtained from a REST service of PubDictionaries

## 3. 症状

神経症候としては、下位運動ニューロンである顔面、舌及び四肢近位部優位の筋萎縮及び筋

力低下と筋収縮時の著明な筋線維束性収縮が主症状である。四肢腱反射は全般に低下し、上位

運動ニューロン徴候はみられない。手指の振戦や筋痙攣が、筋力低下の発症に先行することが

ある。喉頭痙攣による短時間の呼吸困難を伴うこともある。深部感覚優位の軽微

な感覚障害が、特に下肢遠位部でみら

どが見られ、呼吸器感染を繰り返すようになる。睾丸萎縮、女性化乳房、女性化乳房変化など

Automatic annotation with HPO terms

# Preliminary study

4. To manually correct dictionary-based text annotation (using TextAE, <http://textae.pubannotation.org>)
5. To collect a set of phenotype-associations for a Nan-Byo

3. 症状

神経症候としては、下位運動ニューロンである顔面、手及び四肢近位部優位の筋萎縮及び筋力低下と角反射は全般に低下し、上位運動ニューロン徴候はみられない。手指の振戦や筋痙攣が、筋力低下の発症に先行することがある。深部感覚優位の軽微

HP:000

HP:0

HP:00020  
HP:00028

HP:00020

HP:00020

どが見られ、呼吸器感染を繰り返すようになる。睪丸萎縮、女性化乳房、女性様皮膚変化など

Experts manually look up HPO terms in text

Experts confirm whether each annotation is correct or not



Corresponding HPO				
Original text	Automatic annotation	ID	Label (Japanese)	Label (English)
筋力低下	×	HP:0001324	筋虚弱	Muscle weakness
筋萎縮	○	HP:0003202	筋萎縮	Skeletal muscle atrophy
振戦	○	HP:0001337	振戦	Tremor
嚥下障害	○	HP:0002015	嚥下障害	Dysphagia

# Preliminary result

---

□ For example;

“moyamoya disease”-phenotype associations

○ Orphanet (by manual curation)

→ 5 HPO terms

Highly reliable

○ Published case reports (by automatic annotation)

→ 129 HPO terms

Less reliable

○ MHLW Guideline (by automatic annotation and manual correction)

→ 21 HPO terms

Highly reliable

# Summary & Plan in this hackathon

---

## □ Summary

- Extracting disease-phenotype associations from MHLW Guidelines that include diagnostic criteria for Nan-Byo
- PubAnnotation and PubDictionaries facilitated getting automatic annotation and manual correction by experts

## □ Plan

- The performance of automatic annotation can be improved by adding the newly identified terms to PubDictionaries

# Acknowledgements

---

## DBCLS

- Yasunori Yamamoto
- Kim Jin-Dong

## University of Tokyo

- Toshihisa Takagi

## Experts

- Chiharu Ohta
- Haruka Yoshii
- Hiromi Yamamura
- Noriko Nakabayashi

## Funding Source

MEXT KAKENHI  
(No. 221S0002)



