EDUCATION REPORT | Genetics inMedicine

Phenotate: crowdsourcing phenotype annotations as exercises in undergraduate classes

Willie H. Chang, MSc[1,2], Pouria Mashouri, MSc[1], Alexander X. Lozano, MSc[1,3,4], Brittney Johnstone, MSc, CGC[1,5], Mia Husić, MSc[1], Annie Olry, PhD[6], Sylvie Maiella, PhD[6], Tugce B. Balci, MD[7], Sarah L. Sawyer, PhD, MD[8], Peter N. Robinson, MD, MSc[9,10], Ana Rath, MD[6] and Michael Brudno, PhD[1,11,12,13]

Purpose: Computational documentation of genetic disorders is highly reliant on structured data for differential diagnosis, pathogenic variant identification, and patient matchmaking. However, most information on rare diseases (RDs) exists in freeform text, such as academic literature. To increase availability of structured RD data, we developed a crowdsourcing approach for collecting phenotype information using student assignments.

Methods: We developed Phenotate, a web application for crowdsourcing disease phenotype annotations through assignments for undergraduate genetics students. Using student-collected data, we generated composite annotations for each disease through a machine learning approach. These annotations were compared with those from clinical practitioners and gold standard curated data.

Results: Deploying Phenotate in five undergraduate genetics courses, we collected annotations for 22 diseases. Student-sourced annotations showed strong similarity to gold standards, with F-measures ranging from 0.584 to 0.868. Furthermore, clinicians used Phenotate annotations to identify diseases with comparable accuracy to other annotation sources and gold standards. For six disorders, no gold standards were available, allowing us to create some of the first structured annotations for them, while students demonstrated ability to research RDs.

Conclusion: Phenotate enables crowdsourcing RD phenotypic annotations through educational assignments. Presented as an intuitive web-based tool, it offers pedagogical benefits and augments the computable RD knowledgebase.

Genetics in Medicine (2020) 22:1391–1400; https://doi.org/10.1038/s41436-020-0812-7

Keywords: rare diseases; phenotype; crowdsourcing; medical education; machine learning

INTRODUCTION

Background

There are an estimated 6172 rare diseases (RDs) and approximately 262.9 to 446.2 million RD patients in the world, yet the paucity of individual diseases and their phenotypic variability across patients make characterizing RDs extremely difficult.[1] This makes identifying and treating RD patients a unique challenge, leaving many without accurate diagnoses and care for extended periods of time. Recent advancements in computational approaches to documenting and analyzing genetic disorders have begun addressing this problem, greatly contributing to the care of RD patients. Tools such as PhenoTips allow clinicians to capture structured data about their patients, which can then be used by Exomiser, Genomiser, and other related tools to identify genomic variants likely to cause disease.[2–4] Other applications allow users to search diseases associated with entered phenotypes for various purposes.[2,5–7] The Matchmaker Exchange, for example, is used to connect multiple RD patients and their care providers across the globe by matching patients' phenotypic and genomic profiles. This is helping to identify dozens of novel disease genes, contributing to the diagnosis of hundreds, if not thousands of patients.[6,9–13]

These tools, however, rely on thorough and accurate structured annotations of human diseases and their clinical representations (phenotypic profiles). Unfortunately, much of the available information about RD phenotypes currently exists in freeform text, such as academic literature, and there is a need to collect more data that characterizes RD phenotypic profiles using standardized, computable terms. Building a library of accurate, up-to-date, and standardized annotations—or associations between specific OMIM/Orphanet Rare Disease Ontology (ORDO) diseases and collections

[1]Centre for Computational Medicine, The Hospital For Sick Children, Toronto, ON, Canada; [2]Department of Computer Science, Princeton University, Princeton, NJ, USA; [3]Faculty of Medicine, University of Toronto, Toronto, ON, Canada; [4]Department of Materials Science & Engineering, Stanford University, Stanford, CA, USA; [5]Sunnybrook Health Sciences Centre, Toronto, ON, Canada; [6]Orphanet, Institut national de la santé et de la recherche médicale, Paris, France; [7]Medical Genetics Program of Southwestern Ontario, London Health Sciences Centre, London, ON, Canada; [8]Department of Genetics, Children's Hospital of Eastern Ontario and Children's Hospital of Eastern Ontario Research Institute, University of Ottawa, Ottawa, ON, Canada; [9]The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA; [10]Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA; [11]Department of Computer Science, University of Toronto, Toronto, ON, Canada; [12]Genetics and Genome Biology Program, The Hospital for Sick Children, Toronto, ON, Canada; [13]University Health Network, Toronto, ON, Canada. Correspondence: Michael Brudno (brudno@cs.toronto.edu)

Submitted 21 November 2019; revised 9 April 2020; accepted: 10 April 2020
Published online: 5 May 2020

of HPO phenotypes[11]—for rare and other genetic disorders will therefore augment computational knowledge of RDs. This, in turn, will enhance the potential of automating tasks such as differential diagnosis, patient matching, and more to ultimately improve the care of RD patients.

The RD community has converged on the Human Phenotype Ontology (HPO), a structured and controlled vocabulary of phenotypes, as the key resource for describing symptoms of RDs.[8] The HPO has a directed acyclic graph structure, meaning that phenotypes closer to the root of the HPO are more general (e.g., *Abnormality of the nervous system*), while more specific phenotypes (e.g., *Absence seizures*) are located further from the root. Synonymous terms are merged into single concepts. The phenotypes are hierarchically related based on shared features, such as the affected body system (e.g., nervous system) or disease type (e.g., cancer), in parent–child relationships. These relationships are determined mainly through human curation. Furthermore, each phenotype can be associated with diseases in OMIM and ORDO catalogs.[9,10] Both of these maintain annotations of rare disorders with HPO terms and, for some disorders, also record their respective frequencies in the patient population. Although many of these disease annotations are accurate and comprehensive, some are too broad or incomplete. For example, *Abnormality of the skeletal system* and *Abnormal joint morphology* are the only two listed phenotypes for anomalous coracoclavicular joint (OMIM 121350; accessed October 2019). These two phenotypes are not useful to physicians when diagnosing patients. Rather, phenotypes such as *Shoulder pain* and *Limited shoulder movement*, labeled with appropriate frequencies, would be much more helpful.[12]

Data collection and analysis have been successfully accomplished in computational medicine and bioinformatics through crowdsourcing methods in the past. For example, Phylo is a DNA sequence alignment tool presented as an online game that over 12,000 players had contributed to at the time of publication.[14] CrowdMed is an online service where undiagnosed patients can submit clinical information and test results to be examined by physicians, medical students, and laypeople around the world. Users can make and receive diagnostic suggestions, which are rated on their likelihood of being accurate. An initial study showed that CrowdMed helped 233 of 391 patients receive a correct diagnosis.[15] Some crowdsourcing projects involve students in the process of data collection, providing them with pedagogical benefits while rapidly gathering data. One such project is MetaSUB, a crowdsourcing initiative for the mapping of urban environment metagenomes, particularly in mass transit vehicles and facilities.[16,17] MetaSUB incorporates educational outreach by involving students in collecting samples, allowing them to learn about the microbiome of their city's public transit system.

The success of such crowdsourcing projects motivated our use of similar techniques for improving RD annotations. Methods such as automated text mining of disease annotations from medical literature can be error prone, while manual curation by experts is expensive and time consuming. In this project, we implement crowdsourcing in a classroom setting as a method of collecting disease annotations. By analyzing annotations contributed by nonexperts (specifically, students enrolled in undergraduate genetics courses) with a machine learning (ML) approach, we show that it is possible to construct composite annotations for genetic diseases that are comparable, both quantitatively and qualitatively, with those produced by experts such as clinical geneticists, genetic counselors, and RD researchers.

MATERIALS AND METHODS

Phenotate web application

We built Phenotate (phenotate.org), a web application for the collection and curation of disease annotations. Users with various levels of medical expertise can submit annotations through a simple user interface. While it can be used by individuals such as RD patients and citizen scientists, we designed the application primarily for deployment in classroom settings. Course instructors can create annotation exercises for students to complete, then review, grade, and comment on the students' annotations. Students use the feedback to augment their knowledge in medical genetics.

Phenotate users first create an account designated as either expert (e.g., genetics clinicians, researchers, and course instructors) or nonexpert (e.g., students and laypeople). A user can receive an expert account upon sign-up by entering a numeric code distributed to verified individuals, or after their account is made by requesting an upgrade via their Dashboard. The Phenotate user interface varies depending on the user's account type.

Once they have obtained an account, users can create and submit annotations using the annotator tool (Fig. **1a**). To do so, users select a disease from the OMIM or ORDO catalog using a search bar on their Dashboard or in My Annotations (accessed from a menu in the left sidebar). Student users can access their annotation assignments by selecting "Join a Class" and entering a class code provided by their instructor. Once a disease has been selected for annotation, users can add, remove, and modify phenotypes. When adding phenotypes, the annotator interface provides a dynamic phenotype list that continuously updates as the user types. It is powered by the PhenoTips[3] search engine and allows users with various levels of medical expertise to enter annotations without knowing the precise name of the phenotype in the HPO. The annotator interface also provides a phenotype browser that allows users to select an ancestor or child of any given HPO phenotype (Supplementary Fig. 1). In an annotation, each phenotype is categorized by system, and has modifiable attributes, including whether it is observed or absent, its frequency, age of onset, pace of progression, severity, temporal pattern, spatial pattern, and laterality. Users can also link references, such as journal articles, to the phenotypes. Students annotating diseases as part of a course

exercise are required to associate each phenotype with one or more references.

Expert users can view annotations submitted by other experts, create course assignments, access annotations submitted by their students, compare them against standard annotations, and assign scores using the interface shown in Fig. 1b.

Data collection and annotation scoring

The vast majority of annotations on Phenotate to date were collected through course assignments given by instructors who agreed to use Phenotate in their genetics courses as extra credit activities. Students were asked to use a combination of their prior knowledge and literature searches to complete

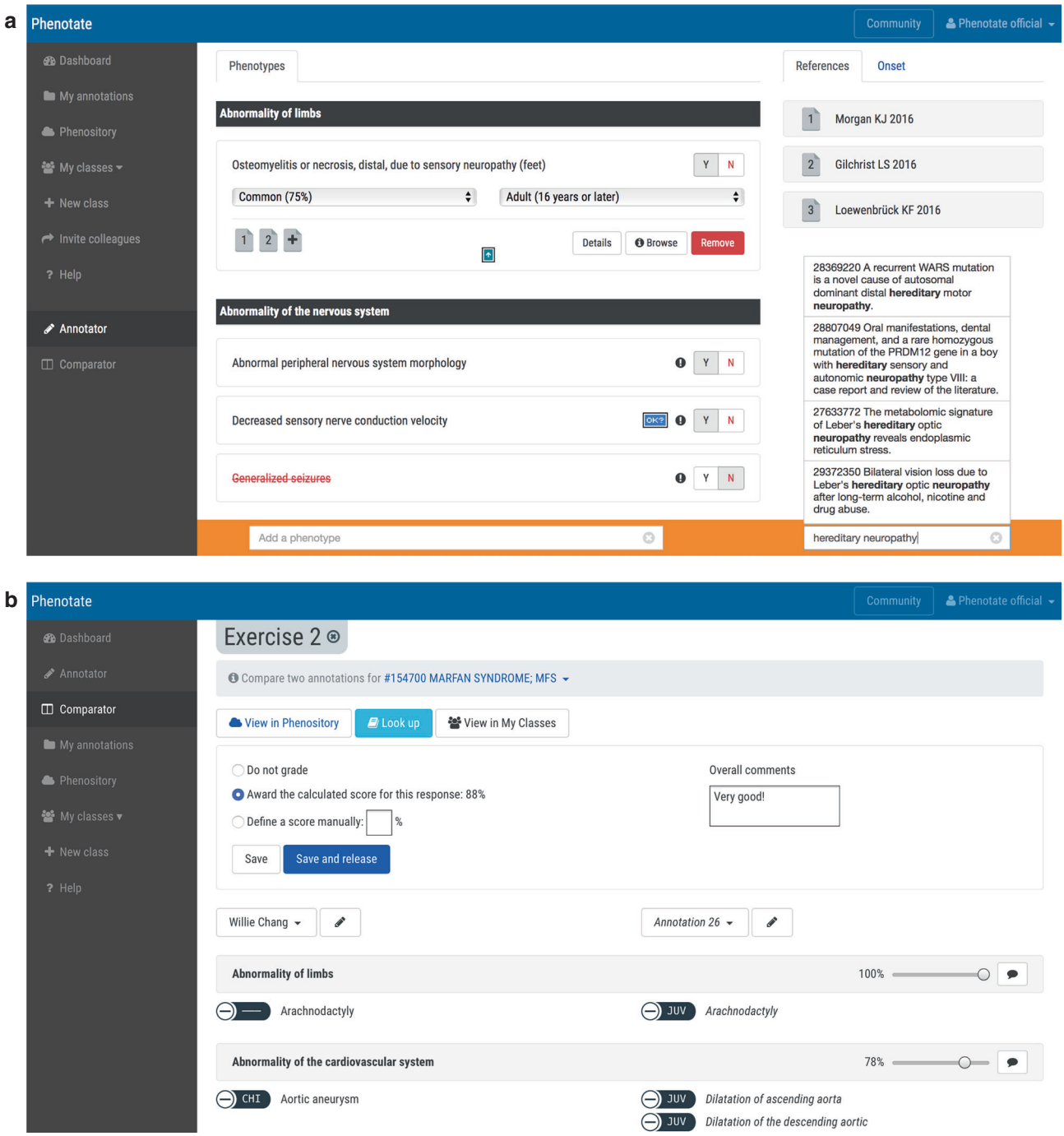

Fig. 1 The Phenotate annotator and comparator user interfaces. (**a**) The Phenotate annotator user interface is available to all users. The user adds phenotypes (left) and references (right) using the orange bar at the bottom of the interface. Clicking on a phenotype allows the user to change its age of onset and frequency. (**b**) Course instructors can compare phenotypes in a student annotation (left) against a standard annotation (right) and grade the student annotation via a comparator tab on the expert user interface. Each phenotype has a label indicating the entered frequency and age of onset. Scoring options and a comment box are available for feedback to the student.

their assigned annotations. We did not inform students about publicly available HPO annotations in OMIM or ORDO; rather, we emphasized the importance of using and citing the medical literature. We also gave them the option to annotate pertinent negative phenotypes, or those that are distinctively absent in a given disease.

Instructors were given the option to grade student annotations by comparing them against pre-existing high-quality HPO annotations or curated annotations (hereinafter termed "gold standard annotations," see next section) if they were available. Phenotate features automatic grading using the Jaccard index, which is defined as the size of intersection of two sets divided by the size of the union. This index is applied between the list of phenotypes selected by the student and their ancestors (R), and the list of phenotypes in the standard annotation (Q).

$$g = |R \cap Q| / |R \cup Q|$$

We chose to include ancestors because HPO concepts have *is-a* relationships between child and parent, meaning that a child implies all of its ancestors. This Jaccard measure was selected based on its success in previous work.[5] Although instructors can adjust the students' scores manually for grading purposes, the automatically generated score was used for data analysis in this paper.

To evaluate student annotations we relied, in part, on a set of gold standard annotations created by a genetic counselor (B.J.). These comprehensive annotations for the diseases were created under consideration from both existing databases and medical literature.

Data analysis

Generating composite annotations

First, n students each annotated two diseases: one scored against standard annotations (known) and one unscored (target). Based on the accuracy of the students' known diseases annotations, we generated composite annotations for the target diseases. First, we removed any pertinent negative phenotypes from all disease annotations. For each remaining phenotype, we added its ancestor phenotypes from the HPO, up to but excluding *Phenotypic abnormality*. We represented the n annotations of a given disease as a binary $p \times n$ matrix $M_{(disease)}$, where p is the number of all unique phenotypes selected across all students, including aforementioned ancestors, that appear in any of the n annotations. Each matrix element contained either a $(+1)$ or a (-1), indicating that the student did or did not include that phenotype in their annotation, respectively.

As part of the model learning process, we define the following four equations:

$$S_{(i,j,k)} = 1/(1 + e^{-(j+k\times i)})$$
$$G_{(scored\_disease)} = |R \cap Q| / |R \cup Q|$$
$$C_{(disease)} = M_{(disease)} \times G_{(scored\_disease)}$$
$$Y_{(disease)} = S(C_{(disease)}, \theta_0, \theta_1)$$

where S() is the sigmoid transformation applied to each element in the vector input, G is a vector of length n containing the Jaccard score for each student's known disease annotation, M() is the binary $p \times n$ matrix of a given disease, C() is a length-p composition annotation vector, and Y() is the predicted binary annotation. The sigma variables, θ_0 and θ_1, are weight parameters that are learned throughout the model training process. It is important to note that G is the Jaccard score for the students' known disease annotations in both training and test phases. For all other variables, the known disease values are used for training, while target disease values are used for generating target annotations.

We trained a linear classification model on the scored disease that takes $C_{(scored\_disease)}$ as input, and learns to predict the binary annotation $Y_{(scored\_disease)}$ by learning a slope (θ_0) and bias vector (θ_1). The model parameters outputted after training may not be optimal due to potentially converging to a local optimum. We thus trained the model five times for each scored disease, and chose the parameters that yielded the highest training F1 score. We then applied the vector $G_{(scored\_disease)}$ as well as the learned weight parameters θ_0 and θ_1, to generate the composite annotation $C_{(target\_disease)}$ and predict a binary annotation set $Y_{(target\_disease)}$.

We trained the model on the scored disease with the Adam optimizer using TensorFlow,[18] version 1.9.0-rc0 (Python 2.7.15rc1 on Ubuntu 18.04 LTS), for 100,000 epochs. A sigmoid cross-entropy loss was used as the training cost function between the predicted annotation Y and the standard annotation. This process is summarized in Supplementary Fig. 2.

Evaluating composite annotations

We generated a composite annotation for each target disease from a given known disease annotated by the same students. If multiple known diseases were available for a target disease, the known-target disease pair with the highest number of student annotations was selected. In the event of a tie, the known disease with the higher quality standard annotation source was selected. A composite annotation for a disease is composed of the disease's phenotypes and their ancestors (see "Generating Composite Annotations"). We evaluated each composite annotation by taking the F1 score of the annotation when compared against the standard annotation, including all of the selected phenotypes' ancestors in the HPO.

RESULTS

Data collection

The Phenotate application has been deployed in a total of five classes, with annotations ranging from 11 to 87 submissions per class. Across all five classes, we collected student annotations for 22 diseases. These data are summarized in Table 1.

The largest deployment of Phenotate to date was in a course assignment for the second-year undergraduate molecular genetics class MGY200 (Current Topics in Molecular Genetics and Microbiology) at the University of Toronto during spring semester 2017. We focused our analysis on this class, as it provided the most stable machine learning (ML)

Table 1 Disease annotations generated by a set of courses coordinated through Phenotate.

Course	Disease	ID	Number of annotations	Gold standard	Allotted time
MGY200 (1)	CMS6: presynaptic congenital myasthenic syndrome 6	OMIM 254210	74	Expert clinicians, genetic counselor	Several weeks
	FRDA: Friedreich ataxia 1	OMIM 229300	77	Expert clinicians, genetic counselor	
	MFS: Marfan syndrome	OMIM 154700	87	Expert clinicians, genetic counselor	
BIO476 (2)	AFAP: attenuated familial adenomatous polyposis	ORPHA 220460	24	ORDO or OMIM	1 week
	ALS4: juvenile amyotrophic lateral sclerosis 4	OMIM 602433	22	Student volunteer	
	CLASSIC: classic homocystinuria	ORPHA 394	13	ORDO or OMIM	
	CMS6: presynaptic congenital myasthenic syndrome 6	OMIM 254210	24	Expert clinicians, genetic counselor	
	DMP: distal myotilinopathy	ORPHA 98911	22	N/A	
	DVA: chronic diarrhea with villous atrophy	ORPHA 1670	23	N/A	
	FRDA: Friedreich ataxia 1	OMIM 229300	22	Expert clinicians, genetic counselor	
	FTLD: frontotemporal lobar degeneration with TDP43 inclusions	OMIM 607485	18	Expert clinicians	
	HCU-MTHFR: homocystinuria due to methylene tetrahydrofolate reductase deficiency	ORPHA 395	13	N/A	
	HSAN IE: hereditary sensory neuropathy type IE	OMIM 614116	11	Expert clinicians	
	JPS: juvenile polyposis syndrome	ORPHA 2929	11	N/A	
	LGMD2B: limb-girdle muscular dystrophy	OMIM 253601	23	Expert clinicians	
	LODM-MG: late-onset distal myopathy, Markesbery–Griggs type	ORPHA 98912	21	N/A	
	MFS: Marfan syndrome	OMIM 154700	21	Expert clinicians, genetic counselor	
	MTPD: mitochondrial trifunctional protein deficiency	ORPHA 746	23	ORDO or OMIM	
	NPD-B: Niemann–Pick disease type B	ORPHA 77293	18	ORDO or OMIM	
	SMARD1: spinal muscular atrophy with respiratory distress type 1	ORPHA 98920	22	ORDO or OMIM	
	SPG2: X-linked spastic paraplegia	OMIM 312920	22	Expert clinicians	
	TMD: tibial muscular dystrophy	ORPHA 609	22	ORDO or OMIM	
	WILSON: Wilson disease	ORPHA 905	23	ORDO or OMIM	
BIOL434 (3)	ALS: amyotrophic lateral sclerosis	ORPHA 803	23	Student volunteer	Several weeks
	WILSON: Wilson disease	ORPHA 905	23	ORDO or OMIM	
HMB311 (4)	ALS: amyotrophic lateral sclerosis	ORPHA 803	14	Student volunteer	Several weeks
	CLASSIC: classic homocystinuria	ORPHA 394	14	ORDO or OMIM	
	MD: muscular dystrophy	ORPHA 98473	14	N/A	
	WILSON: Wilson disease	ORPHA 905	14	ORDO or OMIM	
LMP408 (5)	ALS4: juvenile amyotrophic lateral sclerosis 4	OMIM 602433	11	Student volunteer	50 minutes
	FTLD: frontotemporal lobar degeneration with TDP43 inclusions	OMIM 607485	11	Expert clinicians	

Courses included University of Toronto MGY200 (Current Topics in Molecular Genetics and Microbiology), LMP408 (Genetic Modeling of Human Development and Disease), and HMB311 (Laboratory in Fundamental Genetics and its Applications), as well as University of Waterloo BIOL434 (Human Molecular Genetics). Gold standard annotations came from a number of sources, including expert clinicians, genetic counselor (B.J.), student volunteer (outside of course), and ORDO or OMIM. Allotted time indicates the length of time students were given to complete two disease annotations.

model. As a bonus exercise, 87 students annotated three diseases: Marfan syndrome (MFS, OMIM 154700), Friedreich ataxia 1 (FRDA, OMIM 229300), and presynaptic congenital myasthenic syndrome 6 (CMS6, OMIM 254210). Each student was assigned one disease as the known and one as the target. As these were bonus assignments, the students were given several weeks to complete their two assigned annotations. For each disease pair, students' annotations of the known disease were first scored against the standard annotation of that disease using the Jaccard index. These scores determined the weighting of the students' annotations for the target disease when generating the composite

annotation. Fifty percent of the bonus grade received by students for this exercise was based on the completion of the assignment, and 50% on the accuracy of the annotations, as evaluated by the Jaccard coefficient, with a small linear correction.

We likewise collected a number of annotations from clinical geneticists as part of a neuromuscular disease workshop held in December 2014 in Newcastle, UK. Geneticists were given approximately two hours to complete annotations for several disorders. We were able to collect one, two, and three annotations from clinical geneticists for MFS, FRDA, and CMS6, respectively.

Evaluating composite annotations for CMS6, FRDA, and MFS

With the data submitted by all students in MGY200, we generated composite annotations for CMS6, FRDA, and MFS. The F1 scores of every composite annotation at 50% probability threshold, along with other metrics including areas under the receiver operating characteristic curves (AUROCs), are listed in Table 2. Receiver operating characteristics (ROC) curves are shown in Supplementary Fig. 3. Composite annotations are in Supplementary File 1.

We then compared the Jaccard similarity score of each target disease for our model's predicted composite annotations with those created by a genetic counselor (B.J.), and the clinical geneticists from the neuromuscular disease workshop. The Jaccard score between the model's predictions and the genetic counselor's annotations of CMS6 was 0.430. Between the model's predictions and the workshop annotations, the score was 0.381, and for the genetic counselor's annotations against the clinical geneticists' annotations, it was 0.299. For FRDA, the Jaccard scores for these same pairings were 0.611, 0.571, and 0.305, respectively. Finally, for MFS, the Jaccard scores were 0.652, 0.374, and 0.332, respectively. For all three diseases, our model's predicted annotations had a much stronger Jaccard score against the genetic counselor's annotations than against those collected at the workshop.

A closer examination of the composite annotations revealed that they included a number of phenotypes not listed in some or all of the clinical geneticists' annotations. *Ectopia lentis* and *Dysarthria*—frequent and clinically important phenotypes of MFS and FRDA, respectively[19,20]—are in many students' annotations as well as the composite annotations, but not in those by clinical geneticists. Furthermore, the FRDA composite annotation is more specific and correct regarding the symptom *Gait ataxia*, which is listed in the geneticists' annotations as *Sensory ataxia* or simply *Ataxia*. *Apnea* and *Bulbar palsy*, possible symptoms of CMS6,[21] are listed in the composite annotations but only occur in one of the three clinical geneticists' annotations each.

Discrepancies between student and professional annotations may be accounted for by several factors. *Ectopia lentis* may have been excluded as the clinical geneticists' subspecialty was neuromuscular disorders as opposed to those of the connective tissue. *Dysarthria* is relatively nonspecific in ataxia patients, and so may not have been suggested for this reason. Furthermore, clinical geneticists were asked to work from memory, while students had access to additional resources and were required to include citations to medical literature for their assignments. Students were also not under time constraint, and had several weeks to complete the annotations.

Subsampling student annotations

To test the reproducibility of the ML model with different sample sizes, we performed subsampling analysis on disease annotations from the MGY200 class. From the original 73 usable annotations collected in this class, we trained the model and tested it using randomly selected subsamples of annotations. Ten subsample sizes were attempted, ranging from 7 annotations (10%) to all 73 annotations (100%), in increments of 7 (10%). We ran the experiment ten times per subsample size, and took an average of the results to account for different biases that each group of annotations might contain.

Higher subsample sizes performed slightly better than lower sizes. Nevertheless, the F1 scores of our composite annotations only dropped by approximately 10–15% when using 10% of the data set compared with the full data set. Therefore, although larger annotation samples are preferred, our model can still perform relatively well with a limited number of annotations. Furthermore, the model scales well and continues to improve performance as the sample size increases. Further data collection and testing will need to be done to establish convergence and saturation points of the model performance, which will be the focus of the next stage of this study. Average F1 scores of each subsample size for each scored-target disease pair are shown in Fig. 2.

Student grade distributions

One goal of having students and nonexperts annotate diseases on Phenotate is to provide a learning experience to them, helping expand their research skills and genetics knowledge. The students were not expected to have any familiarity with the specific RDs before the class. To illustrate the knowledge each student gained throughout the annotation process we generated histogram plots of the sensitivity and specificity of student annotations of target diseases against the gold standard annotation that exists within Phenotate. In the original class of MGY200, the students rarely labeled irrelevant concepts, with 100% of students achieving a specificity score above 50% for every disease. However, there was some variability in the completeness of the annotations, with 49–94% of students achieving a sensitivity score above 50% for the three diseases (Fig. 3). Specifically, for MFS only 5% of students submitted nearly complete annotations (80–100% specificity), while for CMS6 all students (100%) had nearly complete annotations.

Annotation of additional diseases

We also deployed Phenotate in four additional classes: three at the University of Toronto and one at the University of

Table 2 Results for all classes where Phenotate was deployed.

Class	Target disease	Scored disease	Number of annotations	Precision	Recall	F1 score	Accuracy	AUROC
MGY200 (1)	CMS6	MFS	73	51.65%	67.14%	58.39%	84.77%	86.20%
	FRDA	MFS	73	78.30%	68.60%	73.13%	85.68%	90.63%
	MFS	FRDA	73	94.61%	80.20%	86.81%	89.09%	95.34%
BIO476 (2)	AFAP	LGMD2B	23	91.49%	39.09%	54.78%	68.86%	84.53%
	ALS4	CMS6	22	90.91%	54.05%	67.80%	88.27%	86.63%
	CLASSIC	CMS6	13	95.65%	31.21%	47.06%	69.44%	78.35%
	CMS6	AFAP	24	53.85%	76.71%	63.28%	81.69%	88.74%
	DMP	SPG2	22	-	-	-	-	-
	DVA	CMS6	23	-	-	-	-	-
	FRDA	CMS6	22	94.64%	46.49%	62.35%	83.20%	92.52%
	FTLD	CMS6	18	65.22%	60.00%	62.50%	89.47%	89.22%
	HCU-MTHFR	CMS6	13	-	-	-	-	-
	HSAN IE	CMS6	11	50.00%	44.00%	46.81%	79.76%	76.78%
	JPS	HSAN 1E	11	-	-	-	-	-
	LGMD2B	CMS6	23	51.85%	40.00%	45.16%	86.97%	81.69%
	LODM-MG	MFS	21	-	-	-	-	-
	MFS	CMS6	21	100.00%	40.45%	57.60%	67.88%	87.92%
	MTPD	CMS6	23	84.44%	54.29%	66.09%	88.60%	94.06%
	NPD-B	FTLD	18	69.57%	55.17%	61.54%	82.22%	83.85%
	SMARD1	FRDA	22	68.99%	78.07%	73.25%	83.33%	90.85%
	SPG2	CMS6	22	80.00%	46.38%	58.72%	79.07%	79.21%
	TMD	ALS4	22	36.84%	50.00%	42.42%	86.13%	86.21%
	WILSON	CMS6	23	63.64%	40.78%	49.70%	82.40%	77.92%
BIOL434 (3)	ALS	WILSON	23	37.70%	62.16%	46.94%	79.03%	78.16%
	WILSON	ALS	23	55.88%	60.64%	58.16%	77.72%	80.15%
HMB311 (4)	ALS	WILSON	14	37.74%	62.50%	47.06%	71.52%	73.16%
	CLASSIC	ALS	14	94.29%	52.38%	67.35%	76.30%	72.50%
	MD	ALS	14	-	-	-	-	-
	WILSON	ALS	14	67.65%	58.97%	63.01%	80.29%	75.00%
LMP408 (5)	ALS4	FTLD	11	71.88%	62.16%	66.67%	81.75%	85.21%
	FTLD	ALS4	11	50.00%	36.00%	41.86%	79.34%	86.65%

The precision, recall, F1 score, accuracy (all at 50% probability thresholds), and AUROCs of composite annotations of the target disease are indicated. Composite annotations were generated with sigmoid parameters and Jaccard index scores obtained from annotations of the corresponding scored disease. Dashes indicate a gold standard annotation was not available at the time of analysis.Courses included University of Toronto MGY200 (Current Topics in Molecular Genetics and Microbiology), LMP408 (Genetic Modeling of Human Development and Disease), and HMB311 (Laboratory in Fundamental Genetics and its Applications), as well as University of Waterloo BIOL434 (Human Molecular Genetics).

AFAP attenuated familial adenomatous polyposis, *ALS* amyotrophic lateral sclerosis, *ALS4* juvenile amyotrophic lateral sclerosis 4, *AUROC* area under the receiver operating characteristic curve, *CLASSIC* classic homocystinuria, *CMS6* presynaptic congenital myasthenic syndrome 6, *DMP* distal myotilinopathy, *DVA* chronic diarrhea with villous atrophy, *FRDA* Friedreich ataxia 1, *FTLD* frontotemporal lobar degeneration with TDP43 inclusions, *HCU-MTHFR* homocystinuria due to methylene tetrahydrofolate reductase deficiency, *HSAN IE* hereditary sensory neuropathy type IE, *JPS* juvenile polyposis syndrome, *LGMD2B* limb-girdle muscular dystrophy, *LODM-MG* late-onset distal myopathy, Markesbery–Griggs type, *MD* muscular dystrophy, *MFS* Marfan syndrome, *MTPD* mitochondrial trifunctional protein deficiency, *NPD-B* Niemann–Pick disease type B, *SMARD1* spinal muscular atrophy with respiratory distress type 1, *SPG2* X-linked spastic paraplegia, *TMD* tibial muscular dystrophy, *WILSON* Wilson disease.

Waterloo. Through these classes, we collected student annotations for 19 additional diseases and generated composite annotations for all 19. The assignments were typically done as homework, with the exception of one class where assignments were done in class. For the classes in which this was given as homework, students had one week to complete two disease annotations. For in-class assignments, students had 50 minutes to complete two disease annotations. All work was graded using the automated scoring mechanism.

All metrics are shown in Table 2, and ROC curves are presented in Supplementary Figs. 4–7. Grade distribution histograms for four classes can be found in Supplementary Figs. 8–11. In general, we obtain high AUROCs (0.8–0.9) across most disease combinations. We present F1 scores and other metrics at 50% probability threshold for each phenotype. This threshold can be tweaked to obtain better F1 scores or to limit the composite annotation to only those phenotypes that are selected most often, depending on the specific downstream application.

Within the 22 total diseases that students across all five classes annotated, we created composite annotations for six diseases with no gold standard annotations at the time of analysis: distal myotilinopathy (DMP), muscular dystrophy (MD), chronic diarrhea with villous atrophy (DVA),

homocystinuria due to methylene tetrahydrofolate reductase deficiency (HCU-MTHFR), juvenile polyposis syndrome (JPS) and late-onset distal myopathy, Markesbery–Griggs type (LODM-MG). Since then, annotations for HCU-MTHFR, DMP, JPS, and LODM-MG were made available at ORDO from consultation with RD experts and literature searches. We compared our annotations with those curated by ORDO and found that our composite annotations often included either the most frequent phenotypes in the ORDO annotation or their parents. For DMP, we successfully identified *Peripheral neuropathy*, as well as the parents or siblings of 5 of the 12 phenotypes classified as either very frequent or frequent. We also identified an additional five very frequent or frequent ORDO phenotypes for DMP; however, the weighted scores of these phenotypes were below our F1 cutoff of 0.500. Our composite annotation for JPS contained direct matches to four of the eight most frequent phenotypes listed on ORDO, as well as the parents of the remaining four. For LODM-MG, our annotations included the parents of two of three frequent phenotypes, but we did not have a match for *Fatigable weakness of distal limb muscle* or its parent terms. All of our annotations also included several phenotypes labeled as "occasional" or "rare" in the ORDO annotations of these three diseases.

Overall, our predictions had more annotated terms than ORDO (46 versus 29, on average over these three diseases). However, these include some high-level terms (e.g., abnormality of the nervous system), which may not be explicitly reported in ORDO. The overall high precision of Phenotate annotations for these disorders (0.81, 0.98, and 0.65, respectively, see Supplementary Table 1) illustrates that in general Phenotate does not significantly predict extraneous phenotypes.

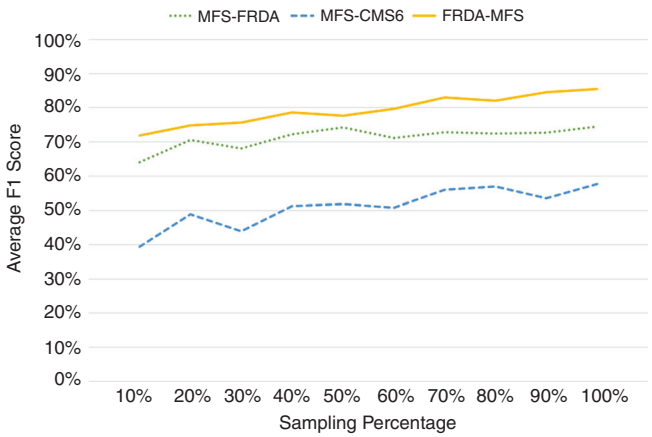

Fig. 2 Average F1 scores of composite annotations increase as the subsample percentage increases. Each data point is averaged over F1 scores from ten experiments of random samples. Percentages are out of 70 annotations (e.g., 10% = 7 annotations). In the legend, the label of each line is written in *scored disease—target disease* format. *CMS6* presynaptic congenital myasthenic syndrome 6, *FRDA* Friedreich ataxia 1, *MFS* Marfan syndrome.

Evaluation of Phenotate for clinical applications

To further evaluate the accuracy and completeness of composite annotations generated using Phenotate, we designed an experiment to compare composite annotations of 20 diseases to their OMIM/ORDO counterparts. Within these 20, we included the 4 diseases for which ORDO/OMIM

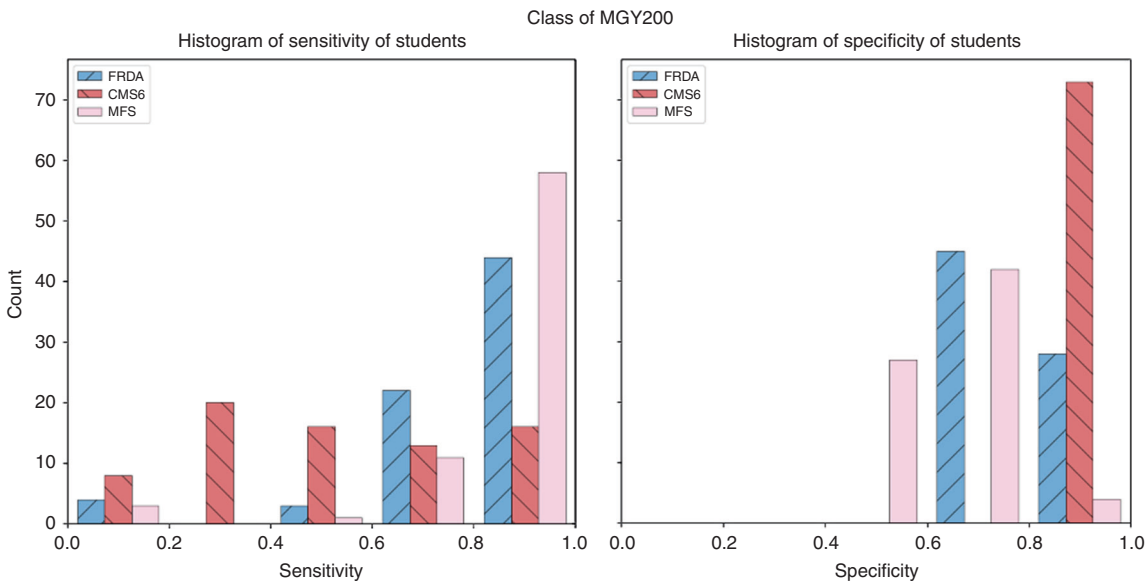

Fig. 3 Sensitivity and specificity of student disease annotations versus the gold standard annotations for Marfan syndrome (MFS), Friedreich ataxia 1 (FRDA), and presynaptic congenital myasthenic syndrome 6 (CMS6) in University of Toronto course MGY200 (Current Topics in Molecular Genetics and Microbiology). The histogram plot contains five equidistant bins between 0.0 and 1.0. The *x*-axis of each graph shows the sensitivity/specificity of the students' scores, while the *y*-axis shows a count of how many students fall within each bin. All students showed high specificity (>0.5) across all disorders; however, the sensitivity varied by disorder.

annotations were made available only after the composite annotations were generated (HCU-MTHFR, DMP, JPS, and LODM-MG). We then asked two clinical geneticists (T.B.B. and S.L.S.) to identify the diseases, while blinding them to the annotation sources. Each clinician was given ten annotations from both sources, and for any given disease one clinician received the composite annotation, while the other received an OMIM/ORDO annotation. The clinicians were asked to do this without referencing HPO, OMIM, or ORDO databases. For readability, we omitted ancestor phenotypes from each annotation. We also asked clinicians to indicate, on a scale of 1–5, how certain they were of their identifications.

The clinicians were able to identify 13 diseases using Phenotate composite annotations, and 15 diseases using OMIM/ORDO annotations. The clinicians were able to more accurately diagnose four diseases using Phenotate composite annotations, and six using OMIM/ORDO annotations. On the remaining ten diseases, they performed equally. Importantly, the clinicians used the composite annotations to either precisely diagnose or identify the correct subgroup for three of four diseases that did not previously have OMIM/ORDO annotations (JPS, LODM-MG, and HCU-MTHFR). For LODM-MG and HCU-MTHFR, clinicians performed equally well when using either the composite or OMIM/ORDO annotations, while they were only able to successfully diagnose JPS using its composite annotation. When asked about the confidence of their diagnoses, the clinicians had higher overall certainty when using OMIM/ORDO across the entire set of diseases (Phenotate average certainty: 4.05; OMIM/ORDO average certainty: 4.55; $p = 0.045$, Student's t test). Nonetheless, it should be noted that some diseases were particularly difficult to identify regardless of the annotation source, including DMP and mitochondrial trifunctional protein deficiency (MTPD). These diseases are ultrarare and primarily seen by subspecialist geneticists, contributing to the difficulty of their identification. All results are summarized in Supplementary Table 2.

To gain insight into how Phenotate may be improved for clinical use, we also asked clinical geneticists to directly compare annotations from Phenotate and OMIM/ORDO for six diseases (attenuated familial adenomatous polyposis [AFAP], amyotrophic lateral sclerosis [ALS], FRDA, JPS, MFS, and Wilson disease [WILSON]). Clinicians were given two annotations for the same disease and, without knowing the sources of the annotations, were asked to select the one that more accurately described the disease. Each clinician was asked to do this for three different diseases. Overall, clinicians showed equal preference for Phenotate and OMIM/ORDO (three disorders each). They cited preferring shorter annotations with more specific and accurate descriptions, particularly for phenotypes that help differentiate one disease from others similar to it. For example, in the case of JPS, the clinician felt that the ORDO annotation was inaccurate and presented far too many phenotypes that were either extremely rare or erroneous. This made the ORDO annotation difficult to use, particularly compared with the more concise and specific composite annotation from Phenotate.

DISCUSSION

Phenotate allows for collecting annotations of genetic diseases with HPO phenotypes. We successfully implemented Phenotate in five classes with 11–87 students each, and demonstrated that, by using a large number of annotations from the same individuals for two diseases, it is possible to generate a composite annotation for one disease given an existing standard annotation for the other. We showed that, for MFS and FRDA, the composite annotations we generated are higher in quality than individual annotations created by expert clinicians. This comparison pits data generated by undergraduate students with varied levels of genetics knowledge against those of geneticists with extensive medical training and clinical experience (albeit with limited time constraints). For no disease under consideration did the students collectively perform worse than the geneticists. We anticipate that future uses in courses and training programs will involve students annotating progressively rarer diseases for which we do not have sufficient computational annotations. Additional avenues that can be explored in future work include scaling up and integration of Phenotate into general purpose crowdsourcing using means such as Amazon Mechanical Turk.

The process of generating composite annotations depends on the ML method we developed that weighs students' annotations for one disease based on their scores from another disease with a standard known annotation. The method comprises training two parameters for a sigmoid that determines the weighting of scores. While this implementation performed well, a clear limitation is that we are using a linear classifier only, which would fail to learn any nonlinear relationships within the data set. A more complex model can also be implemented in the future to allow for various types of linear and nonlinear relationships to exist within the analysis, allowing for a more fine-tuned learning approach.

We designed Phenotate as a crowdsourcing annotation tool that has a strong educational component, with its primary deployment setting being university genetics classrooms. Phenotate gives genetics students an opportunity to learn about rare disease phenotypes associated with rare genetic diseases. Phenotate also encourages its student users to explore the relevant medical and scientific literature, allowing them to examine these diseases and their associated phenotypes in various clinical and research-based contexts. This, in turn, may help them understand how specific phenotypes relate to molecular and genetic components of particular diseases. The high sensitivity and specificity of student annotations show that they are successfully using various sources to research assigned diseases, and correctly applying the knowledge they obtained to create accurate annotations.

Our work shows that Phenotate is an effective platform for crowdsourced curation of structured RD annotations that are comparable with those created by medical professionals. We show that clinicians can use the composite annotations

generated via Phenotate to arrive at a patient diagnosis. They do so with accuracy comparable with that of annotations from sources such as OMIM and ORDO. Composite annotations also allowed for a diagnosis for several diseases that were recently unannotated, suggesting that Phenotate can be used to generate novel annotations for RDs. Clinicians did, however, have more certainty in their diagnoses when using OMIM/ORDO annotations due to higher specificity of the annotations included. We will use this feedback to refine Phenotate in future courses.

Structured data can be applied in computational methods related to diagnostics, patient matching, and more to improve RD patient care, yet such data are not often available for many RDs. The annotations compiled through Phenotate will allow for such computational approaches to be used for the documentation and analysis of various RDs. This could be done through integration with the HPO, which has high interoperability with other ontological tools and annotation databases. It also takes a collaborative approach to increasing access to disease ontology and phenotype data. Incorporating Phenotate annotations into the HPO will increase the availability of complete sets of disease phenotype annotations for RDs. The accuracy and robustness of such annotations will help refine the characterization of RDs and guide patient diagnostics.

SUPPLEMENTARY INFORMATION

The online version of this article (https://doi.org/10.1038/s41436-020-0812-7) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

We thank Orion Buske, Marta Gîrdea, and other members of the Centre for Computational Medicine for their guidance during the development stage of this project. Furthermore, we thank the clinical geneticists who have submitted annotations to the project. We also thank Peter Roy, Karim Mekhail, Alistair Dias, Bernard Duncker, and Nagham Abdalahad for integrating Phenotate into their classes. We thank their students, as well as Chloe Ng, for contributing annotations. We also thank Sana Tonekaboni for her advice on ML methods, Andrei Turinsky for his advice on statistics, and Jixuan Wang for his assistance in integrating Phenotate into LMP408. We use web-based calculators on Social Science Statistics to compute *P* values. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the ERA-NET Cofund action number 643578, E-Rare3; the Canadian component of the work was supported by the Canadian Institutes of Health Research (CIHR); and Genome Canada. A.X.L. received funding to work on Phenotate from a University of Toronto Faculty of Medicine Comprehensive Research for Medical Students (CREMS) Scholarship.

DISCLOSURE

The authors declare no conflicts of interest.

REFERENCES

1. Wakap SN, Lambert DM, Olry A, et al. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. Eur J Hum Genet. 2020;28:165–173.
2. Girdea M, Dumitriu S, Fiume M, et al. PhenoTips: patient phenotyping software for clinical and research use. Hum Mutat. 2013;34:1057–1065.
3. Smedley D, Jacobsen JOB, Jäger M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10:2004–2015.
4. Smedley D, Schubach M, Jacobsen JOB, et al. A whole-genome analysis framework for effective identification of pathogenic regulatory variants in Mendelian disease. Am J Hum Genet. 2016;99:595–606.
5. Buske OJ, Girdea M, Dumitriu S, et al. PhenomeCentral: a portal for phenotypic and genotypic matchmaking of patients with rare genetic diseases. Hum Mutat. 2015;36:931–940.
6. Masino AJ, Dechene ET, Dulik MC, et al. Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. BMC Bioinformatics. 2014;15:248.
7. Foong J, Girdea M, Stavropoulos J, Brudno M. Prioritizing clinically relevant copy number variation from genetic interactions and gene function data. PLoS One. 2015;10:e0139656.
8. Köhler S, Carmody L, Vasilevsky N, et al. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Res. 2019;47(D1):D1018–D1027.
9. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Database issue):D514–517.
10. Rath A, Olry A, Dhombres F, Brandt MM, Urbero B, Ayme S. Representation of rare diseases in health information systems: the Orphanet approach to serve a wide range of end users. Hum Mutat. 2012;33:803–808.
11. Maiella S, Olry A, Hanauer M, et al. Harmonising phenomics information for a better interoperability in the rare disease field. Eur J Med Genet. 2018;61:706–714.
12. Orphanet. Orphadata. http://www.orphadata.org/cgi-bin/index.php#phenotypesmodal. Accessed 15 November 2019.
13. Orphanet. What is HOOM (the HPO-ORDO Ontological Module)? http://www.orphadata.org/cgi-bin/img/PDF/WhatIsHOOM.pdf. Accessed 15 November 2019.
14. Kawrykow A, Roumanis G, Kam A, et al. Phylo: a citizen science approach for improving multiple sequence alignment. PLoS One. 2012;7:e31362.
15. Meyer AND, Longhurst CA, Singh H. Crowdsourcing diagnosis for patients with undiagnosed illnesses: an evaluation of CrowdMed. J Med Internet Res. 2016;18:e12.
16. MetaSUB International Consortium. The Metagenomics and Metadesign of the Subways and Urban Biomes (MetaSUB) International Consortium inaugural meeting report. Microbiome. 2016;4:24.
17. Afshinnekoo E, Ahsanuddin S, Mason CE. Globalizing and crowdsourcing biomedical research. Br Med Bull. 2016;120:27–33.
18. Abadi M, Barham P, Chen J et al. TensorFlow: A System for Large-Scale Machine Learning. Paper presented at the Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, 2–6 November 2016.
19. De Paepe A, Devereux RB, Dietz HC, Hennekam RC, Pyeritz RE. Revised diagnostic criteria for the Marfan syndrome. Am J Med Genet. 1996;62:417–426.
20. National Institute of Neurological Disorders and Stroke. Friedreich ataxia fact sheet. https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Friedreichs-Ataxia-Fact-Sheet. Accessed 23 October 2019.
21. Abicht A, Müller J, Lochmüller H. Congenital myasthenic syndromes. In: Adam MP, Ardinger HH, Pagon RA, et al., editors. GeneReviews®. Seattle, WA: University of Washington, Seattle; 1993. http://www.ncbi.nlm.nih.gov/books/NBK1168/. Accessed 23 October 2019.