Learning phenotypic patterns in genetic disease by symptom interaction modeling

Kevin Yauy [1,2,], Nicolas Duforet-Frebourg [2], Quentin Testard [1], Sacha Beaumeunier [2], Jerome Audoux [2], Benoit Simard [3], Dimitri Larue [2], Michael Blum [2], Virginie Bernard [4], David Genevieve [5], Denis Bertrand [2], PhenoGenius consortium, Nicolas Philippe [2], Julien Thevenon [1,4,*],*

[1] Institute of Advanced Biosciences, Centre de recherche UGA, Inserm U 1209, CNRS UMR 5309, Grenoble, France.

[2] SeqOne Genomics, Montpellier, France.

[3] OuestWare, Nantes, France

[4] Reference center for congenital anomalies, Department of Genetics, Genomics and ART, Grenoble-Alpes University-Hospital, Grenoble, France.

[5] Montpellier University, Inserm U1183, IRMB, Reference center for congenital anomalies, Clinical Genetic Unit, Montpellier University Hospital Center, Montpellier, France

* Corresponding authors. Email: kevin.yauy@univ-grenoble-alpes.fr and JThevenon@chu-grenoble.fr

Abstract (125 words max)

Precision medicine relies on recognizing clinically relevant groups of patients but is currently weakened by the lack of consistency in phenotyping. Inspired by medical inductive reasoning, we developed a machine-learning system which models symptom interactions in genetic diseases to overcome this phenotyping heterogeneity. Based on the Human Phenotype Ontology and using medical data from raw text, we determine 1,1 million associated pairs of symptoms in diseases. From 16,600 phenotypic traits, we defined 390 groups of interacting symptoms summarizing the overall clinical spectrum. The symptom interaction model allowed to cover 99.8% of given symptom-gene relationships from a cohort of 1,686 diagnosed patients. This method should enable new discoveries in precision medicine by standardizing clinical descriptions.

One sentence summary

The symptom interaction model standardizes clinical descriptions to handle physicians' heterogeneous phenotyping to overcome the clinical bottleneck in precision medicine.

.

Precision medicine relies on patient stratification and recognition of clinically relevant groups to improve diagnosis, prognosis, and medical treatment (*1*). Phenotyping allows homogeneous groups of individuals to be constituted, where physicians report characteristics deviating from normal morphology, physiology, and behavior using standardized descriptions in the Human Phenotype Ontology (HPO) (*2*, *3*). Despite a common ontology and abundant clinical data, medical records often lack consistency and comparability between descriptions and practitioners, which is referred to as fuzzy matching in phenotype profiles (*4*). This fluctuating information is a major challenge which has to be overcome to enable us to exploit the clinical data contained in medical records.

1. Phenotyping practices in large cohorts

To our knowledge, no studies focusing on phenotyping practices in clinical sequencing analysis have been previously undertaken in the literature. Through four international studies, including 1,686 patients in total, we collected 2501 different symptoms in HPO format and 849 different disease-causing genes (*5*–*7*) (Table S1). Nearly half of the individuals presented with symptoms belonging to the *Abnormality of the nervous system* (HP:0000707) and *Abnormality of the musculoskeletal system* (HP:0033127) classes, illustrating the current focus on rare disorders in clinical practice and confirming a recent report (*8*) (Figure 1A). Reflecting the genetic heterogeneity of rare diseases, 538 genes were declared only once across our cohort, and the most frequently mutated gene in cohort occurred in less than 2% of cases (*ABCC6,* n=21, Table S2).

We observed heterogeneity in HPO selection terms, as 47% of terms were only used once in our cohort (Figure 1B, Table S3). The median number of HPO terms per clinical description varied across observations, ranging from three (Peng *et al.* (*7*)) to seven (PhenoGenius

consortium, Seo *et al.* and Trujillano *et al.* (*5*, *6*)) (Figure 1C). The heterogeneity of physicians' clinical descriptions was also observed for patients with identical genetic diagnoses. As an example, for the 14 genes with more than ten clinical observations, 67 % of symptoms were declared in a single clinical case.

To exclude the implication of clinical examination variability in this described heterogeneity, we next investigated whether heterogeneity in clinical descriptions was reported if physicians phenotyped the same clinical observations. We settled on a prospective experiment where 12 clinical geneticists with various levels of expertise (Table S4) were asked to phenotype three independent clinical reports, from free plain text to HPO format. We observed heterogeneity in terms of the number and diversity of symptoms declared per clinical observation (Fig 1D). Two to nine symptoms were declared in clinical descriptions of the Kleefstra syndrome observation with the EHMT1 pathogenic variant. A total of 29 different terms were provided; 17 of these terms were used by two or more physicians, and none of the terms were mentioned by all 12 physicians.

2. Quantifying the overlap of symptoms-gene associations between the retrospective cohort and the medical literature

To assess if the clinical descriptions of our cohort matched available knowledge in the medical literature, we mapped the cohort's 11,526 unique symptom-gene associations to the 734,931 associations available in HPO-structured databases (Orphanet, DDG2P (*9*), and the Monarch Initiative or MI (*3*)). From these databases, only 4,913 associations (43%) matched, meaning that 57% were missing (Figure 2A).

As the medical literature is mainly available in free-text format, we developed a text-mining algorithm based on Elasticsearch to extract symptom-gene associations from free-text data in HPO format. Applied to OMIM, MedGen (*10*), and abstracts from PubMed, this text-mining algorithm identified 1,784,453 different symptom-gene associations. This approach resulted in a 3,2-fold increase in HPO-structured database associations (Figure 2B).

The text-mining algorithm provided symptom-gene associations that were significantly deeper in the ontology compared to the HPO-structured databases (median depth 6.7 and 5.2 respectively, Kolmogorov-Smirnov test p-value $< 10^{-215}$, Figure 2C). This underlines the complementarity of these approaches, as illustrated in Figure 2D. Reflecting the variability across individuals in selecting an HPO term to summarize a clinical observation, 76% of associations were exclusive to one database. We hypothesized that text-mined symptom-gene associations in the literature were related to associations available in HPO-structured databases. This hypothesis embodies the fuzzy phenotyping concept, providing human-determined alternative wordings of the same information.

To evaluate this hypothesis, for each gene we compared the average distance in the ontology of exclusive symptom-gene associations to the MI database and the text-mined OMIM database, respectively the largest database of each type (Figure 2B). Compared with a random choice of an HPO term, the average distance of the exclusive symptom-gene associations was significantly lower, suggesting these associations are related (Kolmogorov-Smirnov test p-value $< 10^{-215}$, Figure 2E). Although in this exercise the number of symptom-gene associations increased from 734,931 to 1,784,453, a match was only available for 6,226 of 11,526 associations, meaning that 43% of matches were still missing (Figure 2A).

3. From symptom-gene to symptom-symptom associations modeling

We wondered whether modeling associations between symptoms of the same genetic disorder could improve matches. The Human Phenotype Ontology is ordered according to human development, so it may not represent the interaction of symptoms in disease (Figure 3A). We explored an alternative approach to measuring symptom-symptom associations in genetic disease using Neo4j's node similarity algorithm. This method produced approximately 48 million unique symptom-symptom associations (Figure 3B).

We noted a high correlation between symptom-symptom similarity pair scores and their joint frequency occurrence in clinical observations in the cohort (Spearman correlation coefficient: 0.99). No correlation was observed between symptom-symptom similarity pair scores and the distance between symptoms in the HPO (Spearman correlation coefficient: -0,02, Figure S1), reflecting that symptom-symptom associations cannot be solely derived from the ontology architecture.

Based on similarity score distributions, we suspected that similarities above 80% were potential synonyms or highly related symptoms in diseases (Figure S2). Based on this observation, we selected 1,131,886 pairs with similarity pair scores >80%, corresponding to the 10% highest symptom-symptom association scores.

We illustrate these similarities in Figure 3C, using the symptom *Hypotonia* (HP:0001290) reported by six of the 12 practitioners in our exercise on the Kleefstra syndrome with the EHMT1 pathogenic variant. In the symptom-symptom association graph, the closest term to

Hypotonia is *Neurodevelopmental delay* (HP:0012758), with a symptom-symptom similarity pair score measuring 86%. In the HPO, these symptoms are separated by ten nodes.

Adding symptom-symptom associations with similarity pair scores >80% improved the coverage of symptoms-gene association in the cohort to 8,350 of 11,526 associations, leaving only 27% of matches missing.

4. From symptom-gene associations to groups of symptoms modeling

In contrast to node similarity methods, where pairs of symptoms are evaluated independently, we completed this method by dimensional reduction algorithms with collaborative filtering non-negative matrix factorization (NMF) (*11*).

Using the topic coherence measure (*12*), we determined that the 16,660 HPO terms could optimally be reduced to 390 groups of interacting symptoms (Figure 4A). Each symptom was positioned in the graph with a group weight determined by the algorithm, which defined the 390 groups (Figure 4B). To evaluate if these 390 groups could picture the clinical spectrum of genetic diseases in the cohort, we projected each clinical description into the group of symptoms dimension, applied agglomerative clustering, and performed UMAP (*13*) visualization. In contrast to the 16,600 symptoms dimension, we distinguished clusters of clinical descriptions using 390 groups of symptoms (Figure S3-S5, Figure 4C). 29 clusters were retained according to the distance between clusters in the dendrogram and to avoid multiple single observation clusters (Figure S6). In total, 185 different interacting groups were significantly enriched in 24 different clusters of clinical descriptions (Fisher exact test

with p-value < 0.05 with Bonferroni correction), reflecting phenotypic patterns and the spread of the clinical spectrum among the 390 groups of interacting symptoms dimension.

To compare the accuracy of this approach with the symptom-symptom associations model, we selected the top 10% of the 390 groups of interacting symptom associations to evaluate their coverage of symptoms-gene association in the cohort (Figure S7). Overall in this selection, there were 67,136 symptom-group associations, a median of 29 symptoms were included per group, and each HPO term appeared in seven groups. Each of the 390 groups of interacting symptoms included symptoms from multiple ontology classes, with a median of four out of 22 classes per group.

As an illustration, a total of 49 groups were involved in our Kleefstra syndrome exercise (Figure 4D). The most reported symptom was *Sparse hair* (HP:0008070), belonging to 13 groups (Figure S8). This symptom was strongly associated with three groups with the main symptoms *Sparse eyelashes* (HP:0000653, 21 symptoms), *Fibroma* (HP:0010614, 315 symptoms), and *Megalencephaly* (HP:0001355, 1459 symptoms). These groups respectively encompassed disease groups such as chromatinopathies (including *EHMT1*), disorders of the mTOR pathway (*PTEN, TSC1/2*), and disorders of the RAS-MAP kinase pathway (*NF1, HRAS*).

Using the top 10% of the 390 groups of interacting symptom associations enhanced the coverage of symptom-gene associations to 11,344 of the 11,526 associations from the cohort, leaving less than 2% of matches missing.

Defining a modeling knowledge approach to symptom-gene associations through the measurement of symptom-symptom association pair scores and the delineation of 390-groups of interacting symptoms enabled an increase in coverage of the available knowledge on genetic disorders. Next, we used this symptom interaction model to develop a phenotype matching system and to evaluate its potential impact on candidate gene prioritization.

5. Symptom interaction models as an efficient system for phenotype matching

We progressively explored a phenotype matching system using HPO-structured databases, text-mined knowledge, symptom-symptom association pair scores > 80%, and the global 390 groups of interacting symptoms (Figure S9). These experiments were performed on the clinical observations of our cohort of 1,686 patients using the 16,600 symptoms standard from HPO or projecting them into the dimension of interacting symptoms (Figure S10).

A phenotype match occurred if at least one symptom in the clinical description was related to the diagnostic gene in one database. According to the count of matches per gene, a personalized ranked list of genes is provided (Figure 5A, Figure S9). Using the HPO-structured databases, we obtained a phenotype match for 1566 clinical observations with an average diagnostic gene rank of 852. Applying text-mined associations led to a match for 1628 clinical observations with an average rank of 437 (Figure 5B). Using symptom-symptom similarity pair score > 80%, we matched 1663 clinical observations with an average rank of 313. Projecting clinical observations to the 390 groups of interacting symptoms dimension advanced to 1682 matches with an average rank of 263. In the latter cases, the median rank rose from 37 (symptom-symptom similarity pair score > 80%) to 57.

To illustrate this phenotype-matching system, we considered a clinical description containing two symptoms of the Kleefstra syndrome observation with the *EHMT1* pathogenic variant: *Sparse hair* (HP:0008070) and *Moderate global developmental delay* (HP:0011343). There is no match between these terms and Kleefstra syndrome in HPO-structured databases. Hence, the phenotype match score for *EHMT1* using HPO-structured databases is nil. No match is identified from text-mined symptom-gene associations either. Using symptom-symptom interaction with similarity pair scores >80%, five highly similar symptoms were available for *Sparse Hair* and two for *Moderate global developmental delay*. Among them, highly related symptoms such as *Abnormal hair quantity* (HP:0011362) allowed a phenotype match for *EHMT1*, ranking 1244 out of 5235 (top 25% of genes). *EHMT1* was associated with 55 of the 390 groups of interacting symptoms and related to these two symptoms by a single group: group 56. This was the 23rd *EHMT1*'s associated group, whose main symptom out of a total of 53 symptoms is *Delayed speech and language development* (HP:0000750) (Figure S8). Using only the two above-mentioned symptoms, the 390 groups of interacting symptoms provided a ranking of 851 out of 5235 (top 17% of genes).

This section suggests a complementarity between symptom-symptom similarity pair score > 80% and the 390 groups of interacting symptoms for phenotype matching. We questioned the variable efficiency of each system in the context of heterogeneous phenotyping across clinical descriptions.

6. Conditional algorithm of symptom interaction modeling for phenotype matching

We measured the phenotype matching performances of the symptoms interaction model according to the number of HPO terms in clinical descriptions.

Overall, the 390 groups of interacting symptoms had a better average rank with clinical descriptions with five terms or fewer. Symptom-symptom similarity pair score > 80% had a better average rank with clinical descriptions with six or more terms (Figure S11). As each symptom interaction model provides a different level of induced knowledge, we conditionally applied a model according to the number of symptoms in the clinical description. The combined system, which we called PhenoGenius provided the best performance and reached a nearly full match (99.8%, 1682/1686) for all clinical descriptions.

We compared PhenoGenius to four recently published algorithms for phenotype-driven gene prioritization: PhenoApt, Phen2Gene, CADA, and LIRICAL (*7*, *14–16*). Despite using different prioritization methodologies, these four programs demonstrated similar performances in phenotype matching (Figure 5C).

PhenoGenius ranked 63% of clinical observations in the top 100, with others ranking 53-54%. This is a relative increase of 17%. The overall performance of PhenoGenius reports a median diagnosis rank of 41, while the others rank 71-80, giving PhenoGenius a 42% better performance than the best competitor, Phen2Gene. Overall, PhenoGenius provides an average diagnostic rank of 262, corresponding to a 21% increase compared to its best competitor CADA (334). This performance was replicated across each of the four groups in the cohort (Figure S12).

Although we observed a 99.8% phenotype match with PhenoGenius, the accuracy of the result depends on the chosen symptoms. We assessed the robustness of a disease-causing gene ranking based on the number of symptoms provided. We chose 842 clinical descriptions where the disease-causing gene ranking was lower than 42. This cut-off corresponds to the median result of PhenoGenius' ranking performance. We removed 145 observations with a single symptom. The remaining 697 clinical descriptions included a total of 6,345 symptoms. Then, we randomly removed each symptom from each clinical description (6,345 times) and measured the consequence on the disease-causing gene ranking.

In most cases, phenotype matching remained robust with symptom removal (Figure 5D). Disease-causing gene ranking was identical in 34% of cases (2,188/6,345) and 15 ranks lower in 57% of the experiments (3,633/6,345). The ranking of the disease-causing gene was downgraded by at least 100 in slightly more than 1% of experiments (84/6,345). As expected, the impact of removing a symptom on the disease-causing gene ranking was greater than 100 in observations with three symptoms or fewer (9% of failed experiments, 39/415). Observations with more than four symptoms account for 1% of the experiments with a ranking difference higher than 100 (45/5930).

Discussion

This study aimed to use symptom interaction modeling to learn phenotyping patterns in genetic disease. This modeling method adds to the precision medicine toolbox, a technological paradigm that connects symptoms according to their association with genetic disease. Our approach provides a way of matching physicians' phenotyping to the medical knowledge of symptom-gene associations, helping to lead to standardized clinical descriptions.

As current knowledge overwhelms human learning abilities, an overarching goal in precision medicine is to overcome digital bottlenecks to succeed in deep phenotyping and identification of clinically relevant groups of patients. Progressive adoption of the Monarch Initiative's HPO in clinical symptoms description, the development of automatic extraction of symptoms in HPO format from electronic medical records (*17*), and the definition of the Phenopackets standard file format by GAG4H (*18*) bring the community one step forward.

This study provides an in-depth analysis of phenotyping clinical practice by bringing together four international groups of phenotyped patients with a definitive genetic diagnosis (*5–7*) and clinical reports phenotyped by 12 physicians. Complementary to recent reports (*8*, *19*), this study provides original insights on heterogenous patient phenotyping, both in the cohort's clinical descriptions and the medical literature. A striking observation is that only half of the symptom descriptions were exploited through the matches in the HPO-structured databases. More exhaustive exploitation of the symptom description should encourage more accurate and complete clinical descriptions. As well as encouraging richness of clinical description, tools must address the medical reality of summarized clinical information.

Current practice is responsible for a relative scarcity of symptom-gene matches. Lacking time in their clinical routine, physicians provide scanty phenotyping, with chosen symptoms based on strong *a priori* or phenotypic patterns. Medical inductive knowledge often proposes patterns or groups of hypotheses based on recurrently associated symptoms in their own experience and in the literature. Defining groups of symptoms represent a natural behavior of medical inductive reasoning (*20*, *21*). This could explain the heterogeneity of phenotyping across clinical observations, independently from the innate clinical heterogeneity of a disorder. Current algorithms address this issue by exploring the ontology for synonyms or related symptoms to obtain a match.

While most AI systems currently use a deep learning approach, this original study exploited the alternative computing methodology of graph algorithms to obtain a fuller picture. It allowed us to uncover the missing pieces in clinical descriptions through symptom similarity modeling and collaborative filtering using NMF methods (*11*). As such, projection into the symptoms interaction model dimension could provide a path to standardizing clinical descriptions. Graph algorithms are more accurate in this context as clinical data is limited and sparse. Moreover, graph algorithms are reproducible and interpretable, and these features are fundamental in a medical context (*22*). In addition, in contrast to the clustering approach, graphs and NMF allow free association of symptoms, which is important since the same symptom may belong to different disease groups.

For precision medicine, we need to provide an efficient phenotype matching system to expand the availability of clinical expertise (*23*). Based on symptom interaction models adapted to the number of symptoms, our system improved performance phenotype matching

performance. This system provides a proof-of-concept, and forthcoming studies are likely to optimize its matching and ranking performance. Its simplicity in scoring allows a complete understanding of the phenotype matching, thus providing an interpretable measure of potential genotype-phenotype correlation. Our approach contrasts with state-of-the-art phenotype-driven gene prioritization software, which mostly relies on complex scoring and symptom relationships based on HPO architecture. We recommend providing at least three symptoms per clinical description if possible to lower the risk of missing a phenotype match because of a fuzzy description. We observed better performance with more detailed clinical descriptions.

The objective of the current approach is to integrate multiple data sources from Electronic Health Records for deep phenotyping (*24*). Long-standing aspirations are to be able to answer the question, "Have I seen a case like that before?" among extensive clinical data, and to identify undescribed symptom-gene associations (*25*). We also seek to homogenize and improve the exploitation of the clinical data already available in clinical practices. If we can successfully deal with fuzzy phenotypic profiles and inductive medical reasoning in rare diseases, clinical data can be used for computational phenotype analysis, to improve the feasibility of precision medicine, and to support the adoption of genomic medicine.

Code availability

The PhenoGenius source code is available for resource generation and scientific experiments in Apache License 2.0 on GitHub (https://github.com/SeqOne).

Declaration of Interest

This study was been jointly funded by Association Nationale de la recherche et de la Technologie (ANRT) and SeqOne Genomics. K.Y., N.D-F., S.B., J.A., D.L., M.B., D.B., and N.P. are partially or fully employed by SeqOne Genomics; D.L, S.B, J.A., and N.P. hold shares in SeqOne Genomics. K.Y., N.D-F, S.B., D.L., J.A., N.P., and J.T. has filed two patent applications based on this work.

Acknowledgments

We sincerely thank all patients, clinicians, biologists, and bioinformaticians involved in this project. We acknowledge our scientific mentors for useful advice: Stanislas Lyonnet and Jean-Louis Mandel. Warm thanks to Jennifer Butt, which provides an extensive English language review of this manuscript. Special thanks are addressed to the Association Francophone de Genetique Clinique (https://af-gc.fr/) and to the physicians who performed phenotyping on the three clinical observations: Roseline Caumes, Mélanie Fradin, Anne-Marie Guerrot, Valentin Ruault, Godelieve Morel, Aurelia Jacquette, Gwenaël Le Guyader, Benjamin Dauriat, Geoffroy Delplancq, Bertrand Chesneau, Annick Toutain and Sarah Cluzel. This work has been partially supported by MIAI@Grenoble Alpes (ANR-19-P3IA-0003).

PhenoGenius consortium : *Yannis Duffourd [6], Nicolas Chatron [7], Cedric Le Maréchal [8], Jean-François Taly [9], Wilfrid Carre [10], Claire Bardel [11], Frederic Tran Mau-Them [6], Marc Planes [12], Marie-Pierre Audrezet [12], Laure Raymond [9] , Charles Coutton [13], Pierre Ray [13], Veronique Satre [13], Klaus Dietrich [13] , Isabelle Marey [13] , Françoise Devillard [13] , Radu Harbuz [13] , Florence Amblard [13] , Pauline Le Tanno [13] , Mouna Barat-Houari [14] , Marjolaine Willems [14], Thomas Guignard [14] , Sylvie Odent [15], Marie de Tayrac [10], Damien Sanlaville [7] , Laurence Faivre [6] , Laurent Mesnard [16]*

[6] Inserm UMR 1231 GAD, Genetics of Developmental disorders and Centre de Référence Maladies Rares Anomalies du Développement et syndromes malformatifs FHU TRANSLAD, Université de Bourgogne-Franche Comté, Dijon, France.

[7] Department of Medical Genetics, Lyon University Hospital, 69677 Lyon, France; CNRS UMR 5292, INSERM U1028, CNRL, 69500 Lyon, France; Université Claude Bernard Lyon 1, GHE, 69100 Lyon, France.

[8] Laboratoire de Génétique, UMR 1078 Génétique, Génomique fonctionnelle et Biotechnologies, Inserm, Université de Brest, EFS, CHU Brest, Brest, France.

[9] Service de Génétique, Eurofins Biomnis, Lyon, France

[10] Service de Génétique Moléculaire et Génomique, CHU Rennes, Rennes, France.

[11] Univ Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR5558, Villeurbanne, France. Service de Biostatistique-bioinformatique et plateforme NGS-CHU Lyon, Hospices Civils de Lyon, Lyon, France.

[12] Laboratoire de Génétique, CHU Brest, Brest, France.

[13] Genetic Epigenetic and Therapies of Infertility, Institute for Advanced Biosciences, Inserm U1209, CNRS UMR 5309, Université Grenoble Alpes, 38000, Grenoble, France. CHU de Grenoble, UM de Génétique Chromosomique, 38000, Grenoble, France.

[14] Département de Génétique Médicale, Maladies Rares et Médecine Personnalisée, Univ Montpellier, CHU de Montpellier, CLAD ASOOR Montpellier, France.

[15] Service de Génétique Clinique, Centre Référence "Déficiences Intellectuelles de causes rares" (CRDI), Centre de référence anomalies du développement CLAD-Ouest, CHU Rennes, 35203 Rennes, France; CNRS UMR 6290, Université de Rennes, 2 Avenue du Professeur Léon Bernard, 35043 Rennes, France.

[16] Soins Intensifs Nephrologiques et Rein Aigu, Hopital Tenon, Assistance Publique - Hopitaux de Paris-Sorbonne Universite, Paris, France.

Figures legends

Fig 1. The landscape of phenotyping practices using the Human Phenotype Ontology from a retrospective multicenter cohort of 1,686 patients and a prospective experiment on three clinical reports phenotyped by 12 physicians. A. Treemap chart of the HPO terms in the cohort per main class ontology. B. Treemap chart of the HPO terms in the cohort detailed per term. C. Violin plot of retrospective cohorts showing HPO term counts per clinical case. D. Violin plot of clinical reports phenotyped by twelve physicians, reporting HPO term counts per clinical description.

Fig 2. Quantifying the overlap of symptoms-gene associations between the retrospective multicenter cohort of 1,686 patients and the medical literature. A. Venn diagram of cohort symptom-gene association overlaps with public databases in HPO format and text-mined associations in free-text databases. B. Count distribution of symptoms-gene association exclusive to each database. C. Violin plot of the mean depth of HPO terms from root ontology according to each database. D. Illustration of exclusive symptom-gene associations found in Monarch Initiative database (blue) and text-mined OMIM database (green), using *KMT2D* as an example. E. Distribution of the mean lowest distance in the ontology between exclusive terms in the Monarch Initiative database and our text-mined OMIM database, compared to a random choice of HPO terms.

Fig 3. Modeling symptom-symptom interaction in rare diseases using node similarity algorithms on collected symptoms-gene associations. Node color represents the main HPO class. A. Graph visualization of current symptom relationship based on the human development architecture of HPO. B. Graph visualization of symptoms relationship using neo4j node similarity algorithms. C. Illustration of symptom relationships with a single observation of Kleefstra syndrome with *EHMT1* variant, phenotyped by 12 geneticists (the gray arrows represent the different clinical symptoms provided by a single geneticist, the blue arrows the closest symptom in HPO, the red arrows the closest interaction in the symptom-to-symptom interaction model, and the count number corresponds to the occurrence of symptom across all descriptions).

Fig 4. Modeling symptom-symptom interactions in disease using non-negative matrix factorization reduces symptoms from 16,600 to 390 groups. A. Topic coherence evaluation to determine the optimal number of groups. Colors represent iterative training experiments according to random state initialization. B. Visualization of symptom relationship based on 390 groups of interacting symptoms from medical literature C. UMAP visualization of cohort's clinical descriptions projected into the group of symptom dimension, colored and annotated by Agglomerative clusters. D. Illustration of these groups using Kleefstra syndrome observations with *EHMT1* variant. Node color represents the main HPO class, the count number corresponds to the occurrence of symptoms across all descriptions. Groups are in light brown, and size corresponds to group weight. The thickness of lines is proportional to the symptoms-group relative weight.

Fig 5. Modeling symptom interaction as an efficient system for phenotype matching. A. Illustration of the principle of phenotype matching, looking for the most connected genes

(purple) to the clinical description containing two symptoms of the Kleefstra syndrome observation with the EHMT1 variant: *Sparse hair* (HP:0008070, orange) and *Moderate global developmental delay* (HP:0011343, green). The thickness of lines is proportional to the probability score of symptom-gene associations available with joint HPO-structured and text-mined databases. B. Performance benchmark metrics of phenotype matching and gene prioritization systems according to a maximum number of symptoms in clinical descriptions of the cohort. C. Benchmark of a selection of state-of-the-art gene prioritization programs. The fraction of cases correctly diagnosed (y-axis) is plotted against a cumulative causal gene rank. D. Measuring the minimum number of HPO terms necessary to provide robust phenotype matching using symptom interaction models by randomly removing one symptom and monitoring ranking differences.

References

1. F. E. Dewey, M. F. Murray, J. D. Overton, L. Habegger, J. B. Leader, S. N. Fetterolf, C. O'Dushlaine, C. V. Van Hout, J. Staples, C. Gonzaga-Jauregui, R. Metpally, S. A. Pendergrass, M. A. Giovanni, H. L. Kirchner, S. Balasubramanian, N. S. Abul-Husn, D. N. Hartzel, D. R. Lavage, K. A. Kost, J. S. Packer, A. E. Lopez, J. Penn, S. Mukherjee, N. Gosalia, M. Kanagaraj, A. H. Li, L. J. Mitnaul, L. J. Adams, T. N. Person, K. Praveen, A. Marcketta, M. S. Lebo, C. A. Austin-Tse, H. M. Mason-Suares, S. Bruse, S. Mellis, R. Phillips, N. Stahl, A. Murphy, A. Economides, K. A. Skelding, C. D. Still, J. R. Elmore, I. B. Borecki, G. D. Yancopoulos, F. D. Davis, W. A. Faucett, O. Gottesman, M. D. Ritchie, A. R. Shuldiner, J. G. Reid, D. H. Ledbetter, A. Baras, D. J. Carey, Distribution and clinical impact of functional variants in 50,726 whole-exome sequences from the DiscovEHR study. *Science*. **354** (2016), doi:10.1126/science.aaf6814.

2. P. N. Robinson, Deep phenotyping for precision medicine. *Human Mutation*. **33** (2012), pp. 777–780.

3. S. Köhler, G. Kindle, P. N. Robinson, *The Human Phenotype Ontology in 2021* (2021).

4. M. A. Haendel, C. G. Chute, P. N. Robinson, Classification, Ontology, and Precision Medicine. *N. Engl. J. Med.* **379**, 1452–1462 (2018).

5. G. H. Seo, T. Kim, I. H. Choi, J.-Y. Park, J. Lee, S. Kim, D.-G. Won, A. Oh, Y. Lee, J. Choi, H. Lee, H. G. Kang, H. Y. Cho, M. H. Cho, Y. J. Kim, Y. H. Yoon, B.-L. Eun, R. J.

Desnick, C. Keum, B. H. Lee, Diagnostic yield and clinical utility of whole exome sequencing using an automated variant prioritization system, EVIDENCE. *Clin. Genet.* **98**, 562–570 (2020).

6. D. Trujillano, A. M. Bertoli-Avella, K. Kumar Kandaswamy, M. E. Weiss, J. Köster, A. Marais, O. Paknia, R. Schröder, J. M. Garcia-Aznar, M. Werber, O. Brandau, M. Calvo Del Castillo, C. Baldi, K. Wessel, S. Kishore, N. Nahavandi, W. Eyaid, M. T. Al Rifai, A. Al-Rumayyan, W. Al-Twaijri, A. Alothaim, A. Alhashem, N. Al-Sannaa, M. Al-Balwi, M. Alfadhel, A. Rolfs, R. Abou Jamra, Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2017).

7. C. Peng, S. Dieck, A. Schmid, A. Ahmad, A. Knaus, M. Wenzel, L. Mehnert, B. Zirn, T. Haack, S. Ossowski, M. Wagner, T. Brunet, N. Ehmke, M. Danyel, S. Rosnev, T. Kamphans, G. Nadav, N. Fleischer, H. Fröhlich, P. Krawitz, CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genom Bioinform*. **3**, lqab078 (2021).

8. M. Osmond, T. Hartley, B. Johnstone, S. Andjic, M. Girdea, M. Gillespie, O. Buske, S. Dumitriu, V. Koltunova, A. Ramani, K. M. Boycott, M. Brudno, PhenomeCentral: 7 years of rare disease matchmaking. *Hum. Mutat.* **43**, 674–681 (2022).

9. A. Thormann, M. Halachev, W. McLaren, D. J. Moore, V. Svinti, A. Campbell, S. M. Kerr, M. Tischkowitz, S. E. Hunt, M. G. Dunlop, M. E. Hurles, C. F. Wright, H. V. Firth, F. Cunningham, D. R. FitzPatrick, Flexible and scalable diagnostic filtering of genomic variants using G2P with Ensembl VEP. *Nat. Commun.* **10**, 2373 (2019).

10. D. N. Louden, MedGen: NCBI's Portal to Information on Medical Conditions with a Genetic Component. *Med. Ref. Serv. Q.* **39**, 183–191 (2020).

11. D. D. Lee, H. S. Seung, Learning the parts of objects by non-negative matrix factorization. *Nature*. **401**, 788–791 (1999).

12. M. Röder, A. Both, A. Hinneburg, in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15* (ACM Press, New York, New York, USA, 2015; https://dl.acm.org/doi/10.1145/2684822.2685324).

13. L. McInnes, J. Healy, N. Saul, L. Großberger, UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. **3** (2018), p. 861.

14. Z. Chen, Y. Zheng, Y. Yang, Y. Huang, S. Zhao, H. Zhao, C. Yu, X. Dong, Y. Zhang, L. Wang, Z. Zhao, S. Wang, Y. Yang, Y. Ming, J. Su, G. Qiu, Z. Wu, T. J. Zhang, N. Wu, PhenoApt leverages clinical expertise to prioritize candidate genes via machine learning. *Am. J. Hum. Genet.* **109**, 270–281 (2022).

15. M. Zhao, J. M. Havrilla, L. Fang, Y. Chen, J. Peng, C. Liu, C. Wu, M. Sarmady, P. Botas, J. Isla, G. J. Lyon, C. Weng, K. Wang, Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform*. **2**, lqaa032 (2020).

16. P. N. Robinson, V. Ravanmehr, J. O. B. Jacobsen, D. Danis, X. A. Zhang, L. C. Carmody, M. A. Gargano, C. L. Thaxton, UNC Biocuration Core, G. Karlebach, J. Reese, M. Holtgrewe, S. Köhler, J. A. McMurry, M. A. Haendel, D. Smedley, Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *Am. J. Hum. Genet.*

107, 403–417 (2020).

17. Y. Zhang, T. Cai, S. Yu, K. Cho, C. Hong, J. Sun, J. Huang, Y.-L. Ho, A. N. Ananthakrishnan, Z. Xia, S. Y. Shaw, V. Gainer, V. Castro, N. Link, J. Honerlaw, S. Huang, D. Gagnon, E. W. Karlson, R. M. Plenge, P. Szolovits, G. Savova, S. Churchill, C. O'Donnell, S. N. Murphy, J. M. Gaziano, I. Kohane, T. Cai, K. P. Liao, High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat. Protoc.* **14**, 3426–3444 (2019).

18. J. O. B. Jacobsen, M. Baudis, G. S. Baynam, J. S. Beckmann, S. Beltran, O. J. Buske, T. J. Callahan, C. G. Chute, M. Courtot, D. Danis, O. Elemento, A. Essenwanger, R. R. Freimuth, M. A. Gargano, T. Groza, A. Hamosh, N. L. Harris, R. Kaliyaperumal, K. C. K. Lloyd, A. Khalifa, P. M. Krawitz, S. Köhler, B. J. Laraway, H. Lehväslaiho, L. Matalonga, J. A. McMurry, A. Metke-Jimenez, C. J. Mungall, M. C. Munoz-Torres, S. Ogishima, A. Papakonstantinou, D. Piscia, N. Pontikos, N. Queralt-Rosinach, M. Roos, J. Sass, P. N. Schofield, D. Seelow, A. Siapos, D. Smedley, L. D. Smith, R. Steinhaus, J. C. Sundaramurthi, E. M. Swietlik, S. Thun, N. A. Vasilevsky, A. H. Wagner, J. L. Warner, C. Weiland, M. A. Haendel, P. N. Robinson, The GA4GH Phenopacket schema defines a computable representation of clinical data. *Nat. Biotechnol.* **40**, 817–820 (2022).

19. 100,000 Genomes Project Pilot Investigators, D. Smedley, K. R. Smith, A. Martin, E. A. Thomas, E. M. McDonagh, V. Cipriani, J. M. Ellingford, G. Arno, A. Tucci, J. Vandrovcova, G. Chan, H. J. Williams, T. Ratnaike, W. Wei, K. Stirrups, K. Ibanez, L. Moutsianas, M. Wielscher, A. Need, M. R. Barnes, L. Vestito, J. Buchanan, S. Wordsworth, S. Ashford, K. Rehmström, E. Li, G. Fuller, P. Twiss, O. Spasic-Boskovic, S. Halsall, R. A. Floto, K. Poole, A. Wagner, S. G. Mehta, M. Gurnell, N. Burrows, R. James, C. Penkett, E. Dewhurst, S. Gräf, R. Mapeta, M. Kasanicki, A. Haworth, H. Savage, M. Babcock, M. G. Reese, M. Bale, E. Baple, C. Boustred, H. Brittain, A. de Burca, M. Bleda, A. Devereau, D. Halai, E. Haraldsdottir, Z. Hyder, D. Kasperaviciute, C. Patch, D. Polychronopoulos, A. Matchan, R. Sultana, M. Ryten, A. L. T. Tavares, C. Tregidgo, C. Turnbull, M. Welland, S. Wood, C. Snow, E. Williams, S. Leigh, R. E. Foulger, L. C. Daugherty, O. Niblock, I. U. S. Leong, C. F. Wright, J. Davies, C. Crichton, J. Welch, K. Woods, L. Abulhoul, P. Aurora, D. Bockenhauer, A. Broomfield, M. A. Cleary, T. Lam, M. Dattani, E. Footitt, V. Ganesan, S. Grunewald, S. Compeyrot-Lacassagne, F. Muntoni, C. Pilkington, R. Quinlivan, N. Thapar, C. Wallis, L. R. Wedderburn, A. Worth, T. Bueser, C. Compton, C. Deshpande, H. Fassihi, E. Haque, L. Izatt, D. Josifova, S. Mohammed, L. Robert, S. Rose, D. Ruddy, R. Sarkany, G. Say, A. C. Shaw, A. Wolejko, B. Habib, G. Burns, S. Hunter, R. J. Grocock, S. J. Humphray, P. N. Robinson, M. Haendel, M. A. Simpson, S. Banka, J. Clayton-Smith, S. Douzgou, G. Hall, H. B. Thomas, R. T. O'Keefe, M. Michaelides, A. T. Moore, S. Malka, N. Pontikos, A. C. Browning, V. Straub, G. S. Gorman, R. Horvath, R. Quinton, A. M. Schaefer, P. Yu-Wai-Man, D. M. Turnbull, R. McFarland, R. W. Taylor, E. O'Connor, J. Yip, K. Newland, H. R. Morris, J. Polke, N. W. Wood, C. Campbell, C. Camps, K. Gibson, N. Koelling, T. Lester, A. H. Németh, C. Palles, S. Patel, N. B. A. Roy, A. Sen, J. Taylor, P. Cacheiro, J. O. Jacobsen, E. G. Seaby, V. Davison, L. Chitty, A. Douglas, K. Naresh, D. McMullan, S. Ellard, I. K. Temple, A. D. Mumford, G. Wilson, P. Beales, M. Bitner-Glindzicz, G. Black, J. R. Bradley, P. Brennan, J. Burn, P. F. Chinnery, P. Elliott, F. Flinter, H. Houlden, M. Irving, W. Newman, S. Rahman, J. A. Sayer, J. C. Taylor, A. R. Webster, A. O. M. Wilkie, W. H. Ouwehand, F. L. Raymond, J.

Chisholm, S. Hill, D. Bentley, R. H. Scott, T. Fowler, A. Rendon, M. Caulfield, 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N. Engl. J. Med.* **385**, 1868–1880 (2021).

20. J. Larkin, J. McDermott, D. P. Simon, H. A. Simon, Expert and novice performance in solving physics problems. *Science*. **208**, 1335–1342 (1980).

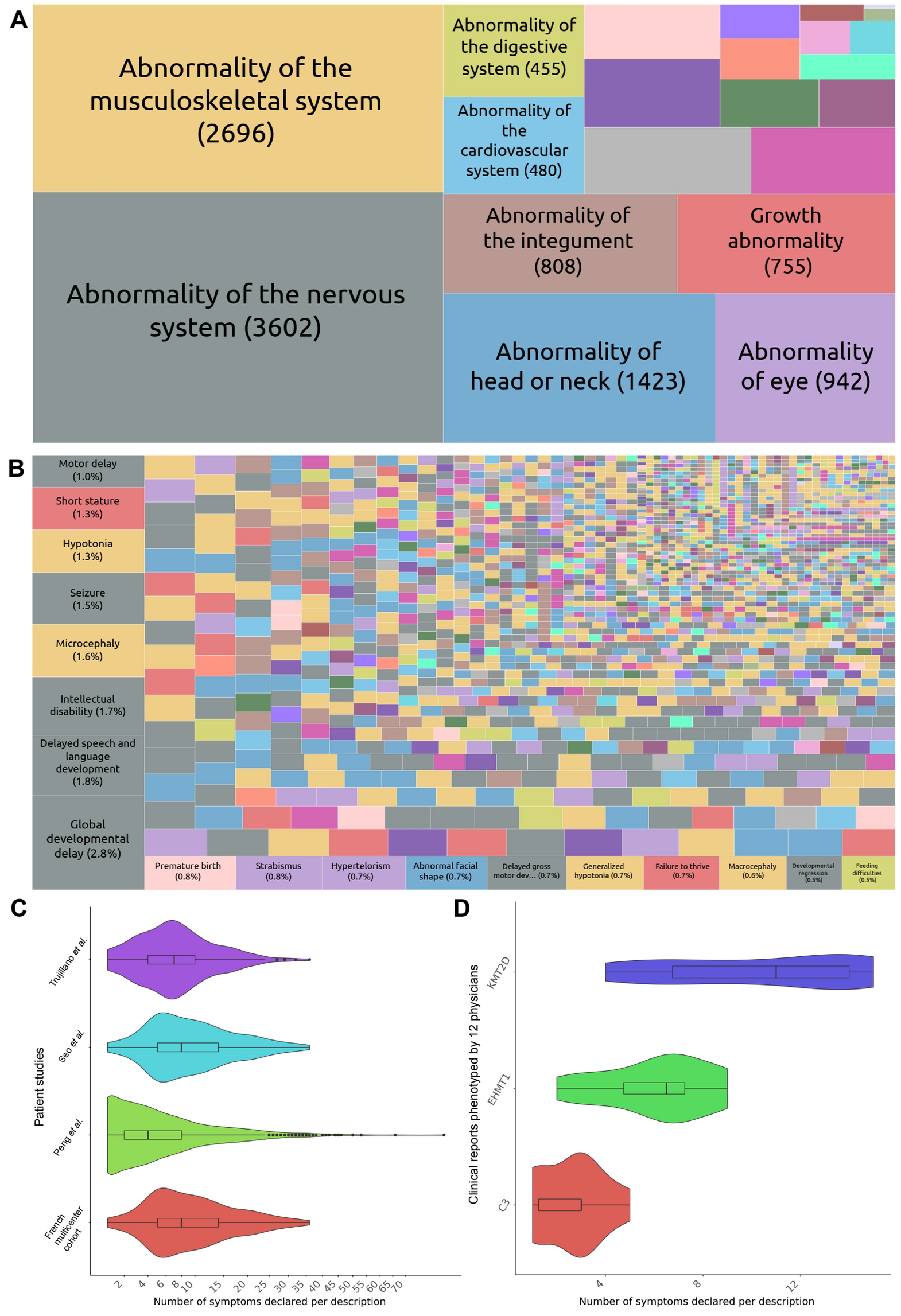
21. S. Coderre, H. Mandin, P. H. Harasym, G. H. Fick, Diagnostic reasoning strategies and diagnostic success. *Med. Educ.* **37**, 695–703 (2003).

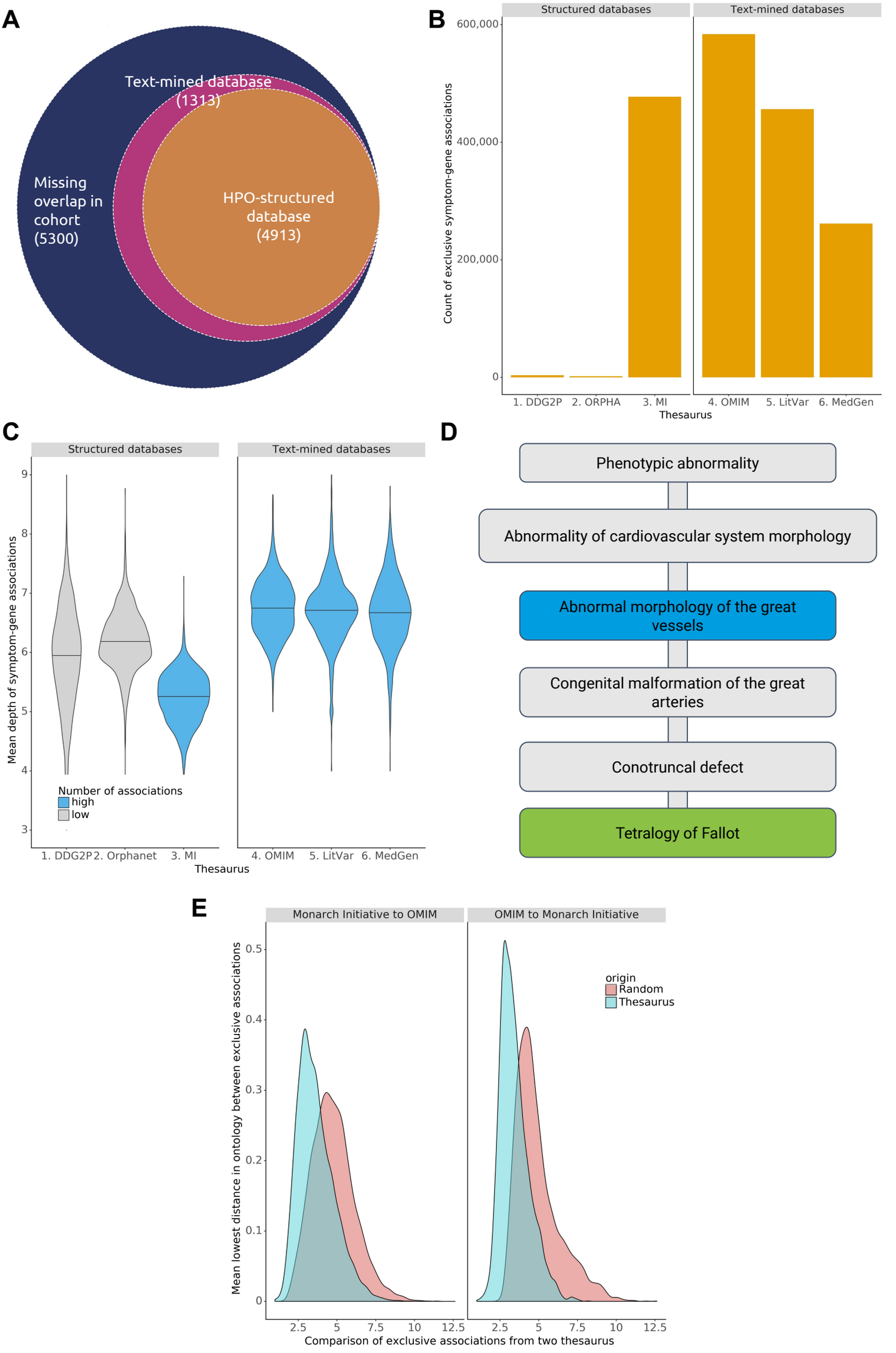
22. I. Tiddi, S. Schlobach, Knowledge graphs as tools for explainable machine learning: A survey. *Artif. Intell.* **302**, 103627 (2022).

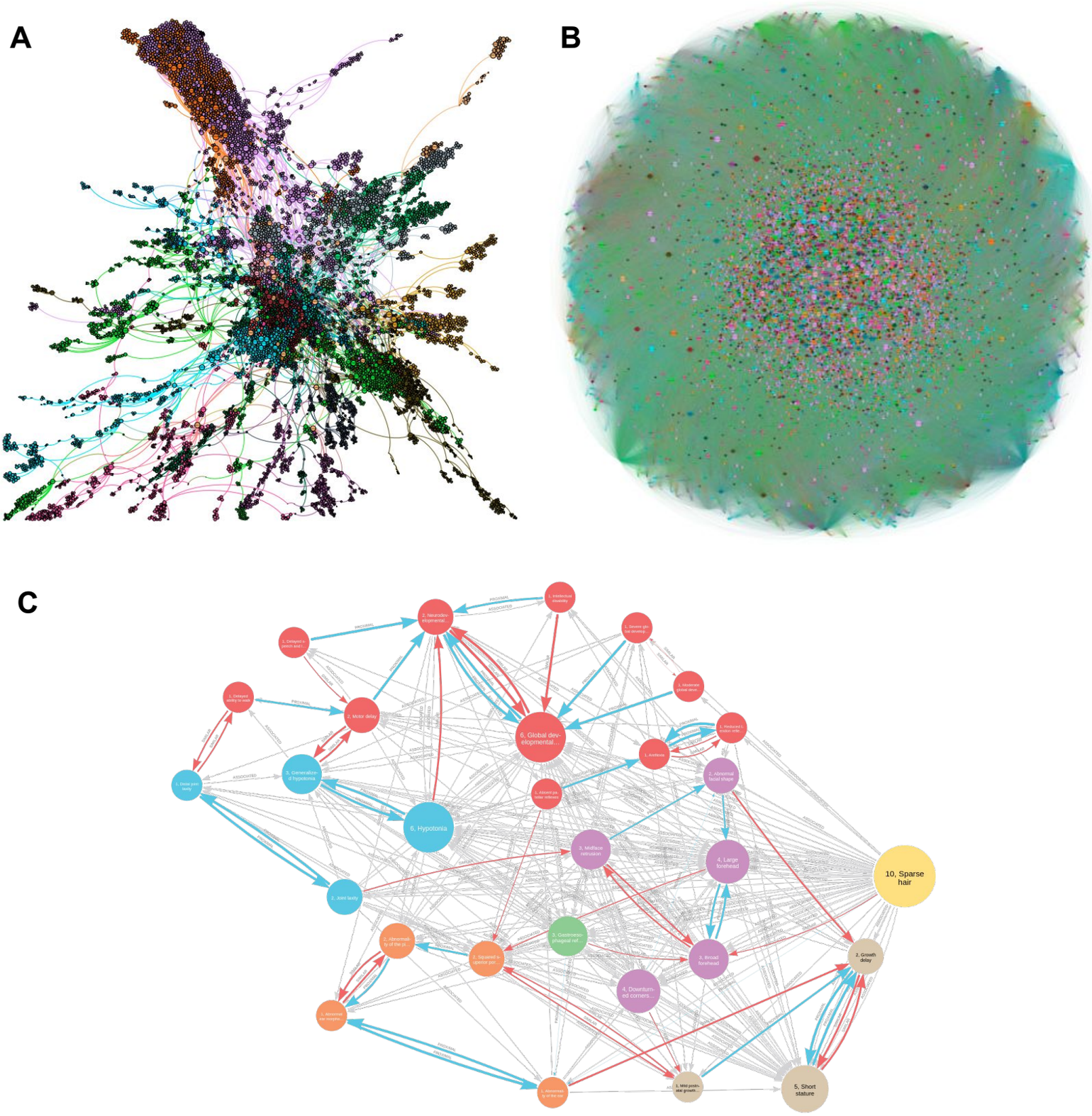
23. A. Rajkomar, J. Dean, I. Kohane, Machine Learning in Medicine. *N. Engl. J. Med.* **380**, 1347–1358 (2019).

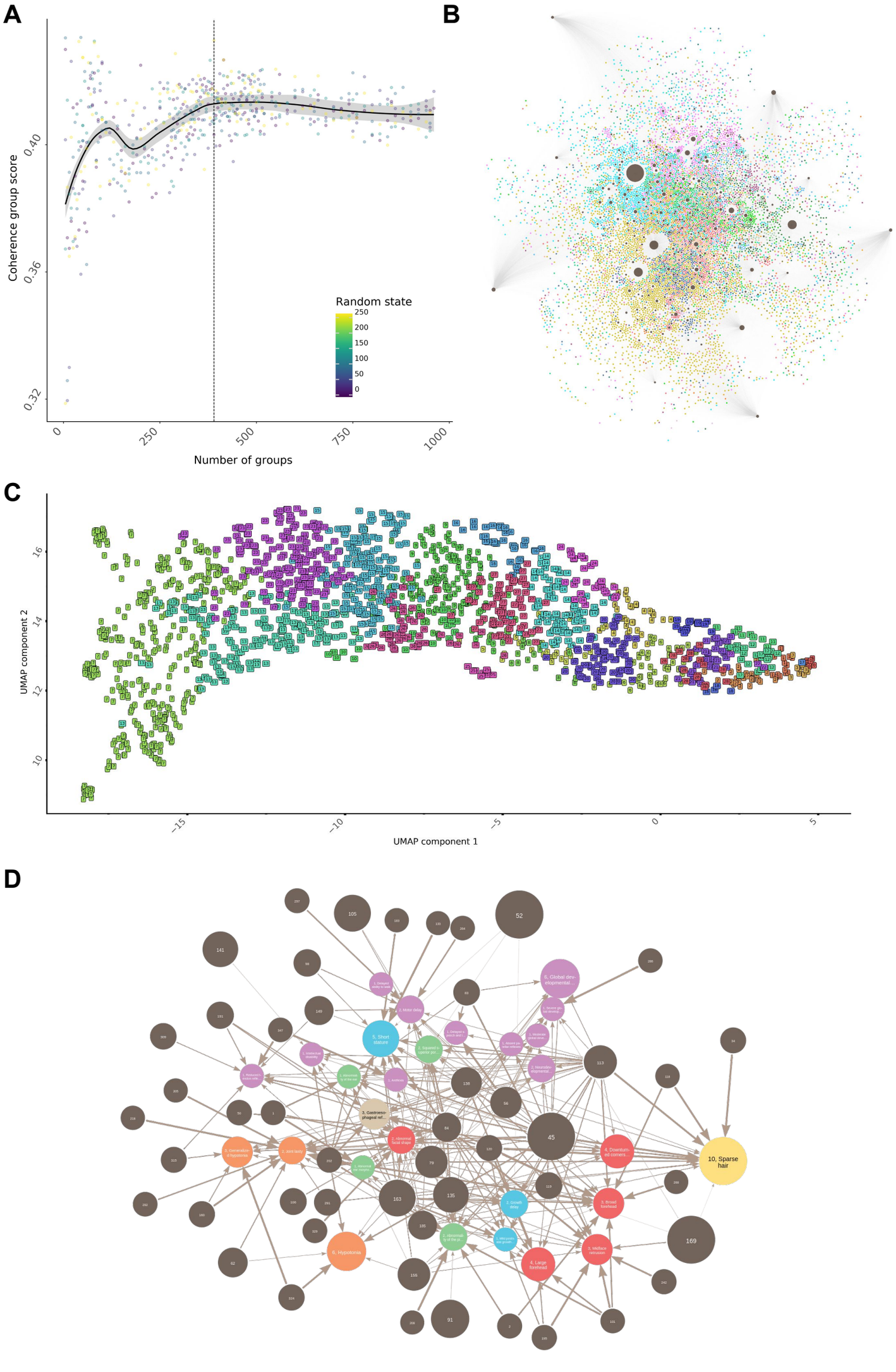
24. J. T. Yurkovich, Q. Tian, N. D. Price, L. Hood, A systems approach to clinical oncology uses deep phenotyping to deliver personalized care. *Nat. Rev. Clin. Oncol.* **17**, 183–194 (2020).

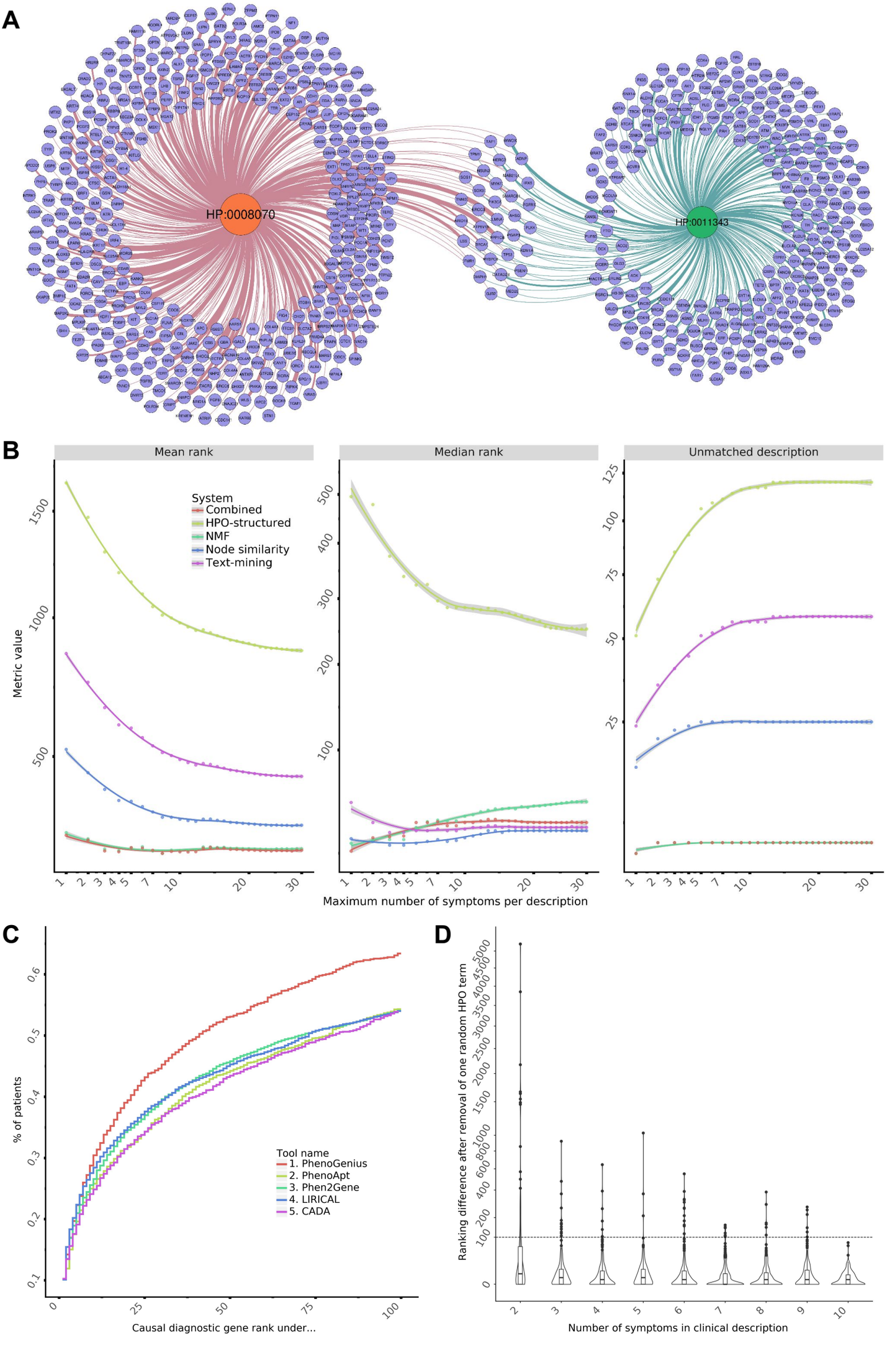
25. C. G. Chute, Clinical Data Retrieval and Analysis. I've Seen a Case Like That Before. *Annals of the New York Academy of Sciences*. **670** (1992), pp. 133–140.


A
Abnormality of the musculoskeletal system (2696)
Abnormality of the digestive system (455)
Abnormality of the cardiovascular system (480)
Abnormality of the nervous system (3602)
Abnormality of the integument (808)
Growth abnormality (755)
Abnormality of head or neck (1423)
Abnormality of eye (942)
B
Motor delay (1.0%)
Short stature (1.3%)
Hypotonia (1.3%)
Seizure (1.5%)
Microcephaly (1.6%)
Intellectual disability (1.7%)
Delayed speech and language development (1.8%)
Global developmental delay (2.8%)
Premature birth (0.8%)
Strabismus (0.8%)
Hypertelorism (0.7%)
Abnormal facial shape (0.7%)
Delayed gross motor dev... (0.7%)
Generalized hypotonia (0.7%)
Failure to thrive (0.7%)
Macrocephaly (0.6%)
Developmental regression (0.5%)
Feeding difficulties (0.5%)
C
Patient studies
Trujillano et al.
Seo et al.
Peng et al.
French multicenter cohort
2 4 6 8 10 15 20 25 30 35 40 45 50 55 60 65 70
Number of symptoms declared per description
D
Clinical reports phenotyped by 12 physicians
KMT2D
EHMT1
C3
4 8 12
Number of symptoms declared per description


A
Text-mined database
(1313)
Missing
overlap in
cohort
(5300)
HPO-structured
database
(4913)
B
Structured databases
Text-mined databases
Count of exclusive symptom-gene associations
600,000
400,000
200,000
0
1. DDG2P
2. ORPHA
3. MI
4. OMIM
5. LitVar
6. MedGen
Thesaurus
C
Structured databases
Text-mined databases
Mean depth of symptom-gene associations
Number of associations
high
low
1. DDG2P
2. Orphanet
3. MI
4. OMIM
5. LitVar
6. MedGen
Thesaurus
D
Phenotypic abnormality
Abnormality of cardiovascular system morphology
Abnormal morphology of the great vessels
Congenital malformation of the great arteries
Conotruncal defect
Tetralogy of Fallot
E
Monarch Initiative to OMIM
OMIM to Monarch Initiative
Mean lowest distance in ontology between exclusive associations
origin
Random
Thesaurus
Comparison of exclusive associations from two thesaurus


A
B
C
2, Neurodev-elopmental...
1, Intellectual disability
1, Delayed s-peech and l...
1, Severe glo-bal develop...
1, Delayed ability to walk
2, Motor delay
1, Moderate global deve...
6, Global dev-elopmental...
1, Reduced t-endon rette...
1, Anthesis
3, Generalize-d hypotonia
1, Distal joint laxity
2, Abnormal facial shape
1, Absent pa-tellar reflexes
6, Hypotonia
3, Midface retrusion
4, Large forehead
10, Sparse hair
2, Joint laxity
2, Abnormal-ity of the pi...
3, Gastroeso-phageal ref...
2, Squared s-uperior por...
3, Broad forehead
2, Growth delay
4, Downturn-ed corners...
1, Abnormal ear morpho...
1, Abnormal-ity of the ear
1, Mild postn-atal growth...
5, Short stature
PROXIMAL
ASSOCIATED
SIMILAR


A
B
C
D
Coherence group score
0.32
0.36
0.40
Number of groups
0
250
500
750
1000
Random state
250
200
150
100
50
0
UMAP component 1
UMAP component 2
−15
−10
−5
0
5
10
12
14
16
5, Short stature
6, Global dev-elopmental...
10, Sparse hair
6, Hypotonia
4, Large forehead
3, Broad forehead
4, Downturn-ed corners...
3, Gastroeso-phageal ref...
2, Abnormal facial shape
2, Joint laxity
3, Generalize-d hypotonia
2, Growth delay
3, Mid-face retrusion
2, Motor delay
1, Delayed ability to walk
1, Intellectual disability
1, Abnormal-ity of the ear
1, Aretinia
1, Abnormal ear morphy...
1, Delayed s-peech and l...
1, Absent pa-tellar reflexes
1, Moderate global deve...
1, Severe glo-bal develop...
2, Neurodev-elopmental...
2, Abnormal-ity of the pi...
1, Mild poren-tial growth...
1, Abnormal-ity of the ear
1, Reduced i-mitation rate
2, Squared s-uperior por...
1, Delayed ability to walk
52
105
141
149
113
45
169
135
163
91
56
84
79
120
119
185
155
208
202
218
315
205
300
191
347
98
183
133
204
295
94
118
298
242
101
2
195
324
62
329
291
106
160
292
1
50
202
1


A
HP:0008070
HP:0011343
B
Mean rank
Median rank
Unmatched description
System
Combined
HPO-structured
NMF
Node similarity
Text-mining
Metric value
Maximum number of symptoms per description
C
% of patients
Causal diagnostic gene rank under...
Tool name
1. PhenoGenius
2. PhenoApt
3. Phen2Gene
4. LIRICAL
5. CADA
D
Ranking difference after removal of one random HPO term
Number of symptoms in clinical description