Supplementary Methods

ClinPhen uses a phenotype ontology and thesauruses to recognize phenotypes

ClinPhen uses the Directed Acyclic Graph (DAG) of phenotypic abnormalities provided by the Human Phenotype Ontology (HPO)[1]. The HPO DAG is a large collection of phenotypes, where the more-general "parent" phenotypes are linked to their more-specific subcategories, or "child" phenotypes. "Generalized tonic-clonic seizures", for instance, is a child of "Generalized seizures", which is a child of "Seizures". HPO also has a list of synonyms for every phenotype. "Seizures", "Seizure", and "Epilepsy", for instance, all correspond to the same phenotype, represented by the ID HP:0001250. ClinPhen looks for these synonyms in the clinical notes to determine if the phenotype is mentioned.

All phenotypes descending from the node "Phenotypic Abnormality" (HP:000018) are considered. ClinPhen supplements HPO's thesaurus using the metathesauruses provided by the Monarch Initiative[2] and the Unified Medical Language System (UMLS 2017AB)[3], which match HPO IDs to a wider variety of synonyms. Together, these three databases provide 28,217 synonyms for the 13,182 HPO phenotypes (from the March 2018 release).

ClinPhen splits clinical notes into fragments and identifies negations

To extract phenotypes from clinical notes, ClinPhen splits the notes into sentences using a set of sentence delimiters. Each sentence is split into a list of subsentences using a set of subsentence delimiters. ClinPhen additionally records each sentence's "flags": words that indicate that a phenotype mention may not apply to the patient. For phenotypes such as "Negative chorea", ClinPhen will count the phenotype as validly mentioned, even though the sentence contains the flag word "negative".

Training flags and delimiters used by ClinPhen

We used the Training set to manually determine which words and characters are best used as flags or delimiters. Phenotypes from the clinical notes of these patients were extracted, once manually, and once by ClinPhen. The flags and delimiters used by ClinPhen were optimized so that ClinPhen's phenotypes would be as similar as possible to those found manually. Feature development ended when the addition of novel cases resulted in little to no further rule changes.

The set of sentence delimiters after training consisted of periods, bullet points, tabs, semicolons, newlines (ClinPhen makes two passes through the notes—once with this delimiter, once without), and the words "but", "except", "however", and "though". The set of subsentence delimiters after training consisted of commas, colons, and the word "and". The set of flags included words that indicate that the mentioned phenotype applies to a family member, not the patient (cousin, parent, mom, mother, dad, father, grandmother, grandfather, grandparent, family, brother, sister, sibling, uncle, aunt, nephew, niece, son, daughter, grandchild); words that directly negate the mentioned phenotype (no, not, none, negative, non, never, normal); and words that indicate that the phenotypes are mentioned as part of a differential diagnosis (associated, gene, recessive, dominant, variant, cause, literature, individuals).

Training additional synonyms and lemmas used by ClinPhen

To further increase sensitivity, we used the Training set to identify commonly interchanged words, and identified the following groups: the "low" group (low, decreased, deficient, deficit, reduced, lacking, insufficient, impaired, difficulty, trouble), the "high" group (high, increased, elevated, elevation), and the "abnormal" group (abnormal, unusual, atypical, abnormality, anomaly, problem). If ClinPhen finds a word in one of these groups, it will register the entire group as having appeared in the record. That way, e.g., “Decreased blood sugar” will still be recognized as a mention of “Low blood sugar”. We further used the Training set to manually augment lemmatization rules used by NLTK.

Filtering phenotypes by their frequency in the population

The STARR set consisted of the clinical notes of 5,000 randomly selected patients under the age of 18, with at least 5 recorded encounters with a physician, from Stanford’s STARR database. ClinPhen optionally ignores phenotypes that are found frequently, because frequently mentioned phenotypes are not likely helpful for rare disease diagnosis. To estimate the phenotype frequencies in a large patient population, we first detected phenotypes in the STARR patient set using ClinPhen’s phenotype-matching mechanism (described above). Phenotypes that were vague (such as “Abnormality of the nervous system”) or common (such as “Pain”, “Fever”, or “Cough”) appeared in more than 15% of these patients. By default, ClinPhen does not output detected phenotypes that occur in more than 15% of STARR patients (a user-adjustable parameter in our offered implementation).

Prioritizing phenotypes by information content

We tried prioritizing phenotypes using the information content of each phenotype as an alternative metric to number of occurrences in the notes. We calculated information content of each phenotype as described in Jagadeesh et al[4]. The information content of a phenotype estimates how indicative a phenotype is of a specific genetic disease using the number of genes that are known to cause the phenotype compared to the number of genes known to cause any phenotype. For example, the phenotype node “Neurodevelopmental delay” in the HPO DAG has an information content of 3.2 bits, because there are many kinds of neurodevelopmental delays, associated with mutations in many different genes; whereas “Expressive language delay” has a higher information content (15.3 bits), because it is more indicative of specific genetic diseases.

Variant filtering to a list of candidate causative genes

We produced a fixed list of candidate causative genes that we used to compare all gene-ranking methods described here. To produce the candidate causative gene list, we first filtered patient variants to a list of possibly pathogenic variants. Per convention[5], all missense, stop-gain, stop-loss, frameshift indel, nonframeshift indel, and splice-site variants in protein-coding regions with an allele frequency below 0.5% in all sub-populations of the Exome Aggregation Consortium (ExAC)[6] and the 1000 Genomes Project[7] were considered to be possibly pathogenic. Genes containing any of these variants comprised the list of candidate causative genes for each patient. Output from all gene-ranking tools was limited to the list of candidate causative genes for each patient.

Automatic gene ranking using Phrank

We ran Phrank[4] with the Python commands:

```
from phrank import Phrank
p = Phrank(DAG, DISEASE_TO_PHENO, DISEASE_TO_GENE)
p.rank_genes(GENES, PHENOTYPES)
```

where DAG is replaced with the path to a file containing the child-to-parent map of the HPO DAG, DISEASE_TO_PHENO with the path to a file containing an OMIM-disease-to-HPO-phenotype map, and DISEASE_TO_GENE with the path to a file containing an OMIM-disease-to-gene-symbol map, PHENOTYPES with the patient’s HPO-encoded phenotype list, and GENELIST with the patient’s candidate gene list.

Automatic gene ranking with Exomiser (PhenIX, Phive, HiPhive)

Exomiser[8] tools cannot be given custom candidate gene lists. Instead, Exomiser takes as input a Variant Call Format (VCF) file containing the patient's genetic variants, and filters it to form its own candidate causative gene list, which it subsequently ranks and returns as output. We subset the ranked genes in the Exomiser output to the same list of candidate causative genes used by Phrank to ensure a fair comparison.

We called Exomiser using the command:

```
java -Xms2g -Xmx4g -jar exomiser-cli-7.2.1.jar -f TAB-GENE --prioritiser=ALGORITHM -F 1
-hpo-ids HPOIDS -v VCF
```

where we replaced ALGORITHM with the gene-ranking algorithm we were testing (hiPhive[8], Phive[9], or PhenIX[10]), HPOIDS with a comma-separated list of HPO phenotypes to be used in the diagnosis, and VCF with the path to the same VCF file used to compile the candidate gene list used by Phrank.

For one Stanford Test patient, Exomiser filtered out the causative gene. Since 107 candidate genes were in Exomiser's output list, and 143—including the causative gene—were not, Exomiser's algorithms were all assumed to rank the excluded candidate genes at the bottom, and the causative gene at the median of these bottom ranks: $107 + \frac{143}{2} = 178.5$. This mimics the average outcome of a clinician going through the Exomiser genes, not finding the causative gene, and then going through the unranked filtered genes in random order.

References

1. Groza T, Köhler S, Moldenhauer D, et al. The Human Phenotype Ontology: Semantic Unification of Common and Rare Disease. *Am J Hum Genet*. 2015;97(1):111-124. doi:10.1016/j.ajhg.2015.05.020

2. Mungall CJ, McMurry JA, Köhler S, et al. The Monarch Initiative: an integrative data and analytic platform connecting phenotypes to genotypes across species. *Nucleic Acids Res*. 2017;45(Database issue):D712-D722. doi:10.1093/nar/gkw1128

3. Unified Medical Language System (UMLS). Unified Medical Language System (UMLS). https://www.nlm.nih.gov/research/umls/. Accessed September 27, 2017.

4. Jagadeesh KA, Birgmeier J, Guturu H, et al. Phrank measures phenotype sets similarity to greatly improve Mendelian diagnostic disease prioritization. *Genet Med*. July 2018. doi:10.1038/s41436-018-0072-y

5. Wenger AM, Guturu H, Bernstein JA, Bejerano G. Systematic reanalysis of clinical exome data yields additional diagnoses: implications for providers. *Genet Med*. 2017;19(2):209-214.

6. Exome Aggregation Consortium, Lek M, Karczewski KJ, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536(7616):285-291. doi:10.1038/nature19057

7. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

8. Smedley D, Jacobsen JOB, Jager M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*. 2015;10(12):2004-2015. doi:10.1038/nprot.2015.124

9. Robinson PN, Köhler S, Oellrich A, et al. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*. 2014;24(2):340-348. doi:10.1101/gr.160325.113

10. Zemojtel T, Köhler S, Mackenroth L, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med*. 2014;6(252):252ra123-252ra123. doi:10.1126/scitranslmed.3009262

Supplementary Figures

Supplementary Figure 1. Performance replication on patient data from an independent clinical center

The experiments performed on Stanford data to generate Figure 3 were repeated on Manton Center patients, and show similar results:

(a) ClinPhen extracts phenotypes with higher precision and sensitivity than both cTAKES and MetaMap.

(b) Limiting to the highest-priority phenotypes improves gene-ranking performance.

(c) ClinPhen outperforms other automatic and human phenotype extractors when used as input to automatic gene-ranking algorithms.

(d) ClinPhen is much faster than all other (human and automated) alternatives for phenotype extraction, taking less than 5 seconds for the task across all medical records. The alternatives are all more than 20x slower on average.

Supplementary Figure 2. Performance of Phrank when phenotypes are prioritized by information content

The same experiment used to generate Figure 3b was done here: taking the *n* highest-priority phenotypes (for every n from 1 to 100), using Phrank to automatically rank genes for each patient, and taking the average Phrank rank of the causative gene across the full cohort of patients (**a**: the Stanford Test set, **b**: the Manton Test set). Here we compare two phenotype prioritization schemes: prioritization by number of occurrences in the clinical notes (implemented in ClinPhen), and prioritization by information content associated with each phenotype (phenotypes with higher information content are prioritized; see Supplementary Methods). When limiting to the highest-priority phenotypes, prioritizing by information content degrades automatic gene ranking performance compared to prioritizing by number of occurrences in the medical record.

Members of the Undiagnosed Diseases Network

David R. Adams
Aaron Aday
Mercedes E. Alejandro
Patrick Allard
Euan A. Ashley
Mahshid S. Azamian
Carlos A. Bacino
Eva Baker
Ashok Balasubramanyam
Hayk Barseghyan
Gabriel F. Batzli
Alan H. Beggs
Babak Behnam
Hugo J. Bellen
Jonathan A. Bernstein
Gerard T. Berry
Anna Bican
David P. Bick
Camille L. Birch
Devon Bonner
Braden E. Boone
Bret L. Bostwick
Lauren C. Briere
Elly Brokamp
Donna M. Brown
Matthew Brush
Elizabeth A. Burke
Lindsay C. Burrage
Manish J. Butte
Hsiao-Tuan Chao
Shan Chen
Gary D. Clark
Terra R. Coakley
Joy D. Cogan
Heather A. Colley
Cynthia M. Cooper
Heidi Cope
William J. Craigen

Precilla D'Souza
Mariska Davids
Jean M. Davidson
Jyoti G. Dayal
Esteban C. Dell'Angelica
Shweta U. Dhar
Katrina M. Dipple
Laurel A. Donnell-Fink
Naghmeh Dorrani
Daniel C. Dorset
Emilie D. Douine
David D. Draper
Annika M. Dries
Laura Duncan
David J. Eckstein
Lisa T. Emrick
Christine M. Eng
Gregory M. Enns
Ascia Eskin
Cecilia Esteves
Tyra Estwick
Liliana Fernandez
Carlos Ferreira
Elizabeth L. Fieg
Paul G. Fisher
Brent L. Fogel
Noah D. Friedman
William A. Gahl
Rena A. Godfrey
Alica M. Goldman
David B. Goldstein
Sarah E. Gould
Jean-Philippe F. Gourdine
Catherine A. Groden
Andrea L. Gropman
Melissa Haendel
Rizwan Hamid
Neil A. Hanchard
Frances High
Ingrid A. Holm

Jason Hom
Ellen M. Howerton
Yong Huang
Fariha Jamal
Yong-hui Jiang
Jean M. Johnston
Angela L. Jones
Lefkothea Karaviti
Emily G. Kelley
David M. Koeller
Isaac S. Kohane
Jennefer N. Kohler
Donna M. Krasnewich
Susan Korrick
Mary Koziura
Joel B. Krier
Jennifer E. Kyle
Seema R. Lalani
C. Christopher Lau
Jozef Lazar
Kimberly LeBlanc
Brendan H. Lee
Hane Lee
Shawn E. Levy
Richard A. Lewis
Sharyn A. Lincoln
Sandra K. Loo
Joseph Loscalzo
Richard L. Maas
Ellen F. Macnamara
Calum A. MacRae
Valerie V. Maduro
Marta M. Majcherska
May Christine V. Malicdan
Laura A. Mamounas
Teri A. Manolio
Thomas C. Markello
Ronit Marom
Martin G. Martin
Julian A. Martínez-Agosto

Shruti Marwaha
Thomas May
Allyn McConkie-Rosell
Colleen E. McCormack
Alexa T. McCray
Jason D. Merker
Thomas O. Metz
Matthew Might
Paolo M. Moretti
Marie Morimoto
John J. Mulvihill
David R. Murdock
Jennifer L. Murphy
Donna M. Muzny
Michele E. Nehrebecky
Stan F. Nelson
J. Scott Newberry
John H. Newman
Sarah K. Nicholas
Donna Novacic
James P. Orengo
J. Carl Pallais
Christina GS. Palmer
Jeanette C. Papp
Neil H. Parker
Loren DM. Pena
John A. Phillips III
Jennifer E. Posey
John H. Postlethwait
Lorraine Potocki
Barbara N. Pusey
Genecee Renteria
Chloe M. Reuter
Lynette Rives
Amy K. Robertson
Lance H. Rodan
Jill A. Rosenfeld
Jacinda B. Sampson
Susan L. Samson
Kelly Schoch

Daryl A. Scott
Lisa Shakachite
Prashant Sharma
Vandana Shashi
Rebecca Signer
Edwin K. Silverman
Janet S. Sinsheimer
Kevin S. Smith
Rebecca C. Spillmann
Joan M. Stoler
Nicholas Stong
Jennifer A. Sullivan
David A. Sweetser
Queenie K.-G. Tan
Cynthia J. Tifft
Camilo Toro
Alyssa A. Tran
Tiina K. Urv
Eric Vilain
Tiphanie P. Vogel
Daryl M. Waggott
Colleen E. Wahl
Nicole M. Walley
Chris A. Walsh
Melissa Walker
Jijun Wan
Michael F. Wangler
Patricia A. Ward
Katrina M. Waters
Bobbie-Jo M. Webb-Robertson
Monte Westerfield
Matthew T. Wheeler
Anastasia L. Wise
Lynne A. Wolfe
Elizabeth A. Worthey
Shinya Yamamoto
John Yang
Yaping Yang
Amanda J. Yoon
Guoyun Yu

Diane B. Zastrow
Chunli Zhao
Allison Zheng