IMPROVE-DD: Integrating multiple phenotype resources optimizes variant evaluation in genetically determined developmental disorders

Stuart Aitken,1,6 Helen V. Firth,2,3 Caroline F. Wright,4 Matthew E. Hurles,2 David R. FitzPatrick,1,5 and Colin A. Semple1,5,*

Summary

Diagnosing rare developmental disorders using genome-wide sequencing data commonly necessitates review of multiple plausible candidate variants, often using ontologies of categorical clinical terms. We show that Integrating Multiple Phenotype Resources Optimizes Variant Evaluation in Developmental Disorders (IMPROVE-DD) by incorporating additional classes of data commonly available to clinicians and recorded in health records. In doing so, we quantify the distinct contributions of sex, growth, and development in addition to Human Phenotype Ontology (HPO) terms and demonstrate added value from these readily available information sources. We use likelihood ratios for nominal and quantitative data and propose a classifier for HPO terms in this framework. This Bayesian framework results in more robust diagnoses. Using data systematically collected in the Deciphering Developmental Disorders study, we considered 77 genes with pathogenic/likely pathogenic variants in ≥10 individuals. All genes showed at least a satisfactory prediction by receiver operating characteristic when testing on training data (AUC ≥ 0.6), and HPO terms were the best predictor for the majority of genes, though a minority (13/77) of genes were better predicted by other phenotypic data types. Overall, classifiers based upon multiple integrated phenotypic data sources performed better than those based upon any individual source, and importantly, integrated models produced notably fewer false positives. Finally, we show that IMPROVE-DD models with good predictive performance on cross-validation can be constructed from relatively few individuals. This suggests new strategies for candidate gene prioritization and highlights the value of systematic clinical data collection to support diagnostic programs.

The importance of phenotype to ranking candidate disease-causing genes is established in research and increasingly so in clinical practice. The primary data resource used in computational phenotype analyses is the Human Phenotype Ontology (HPO).1 Despite promising results,2,3 the exploitation of other information readily available to clinicians, including quantitative anatomic measurements and patient images, is less prevalent.4 The HPO resource, HPO-encoded disease models, and patient’s disease descriptions in HPO terms support diverse tasks: protocols start with a systematic description of an individual’s phenotype and may progress to suggested diagnoses.5 While numerous computational and statistical approaches have been proposed, the advantages of likelihood ratios for the interpretation of genomic and phenomic data in rare disease have been demonstrated.6

Probabilistic methods for combining HPO terms with genetic data in Mendelian disease have been proposed,7,9 as have statistical criteria10–12 and deep learning,13,14 commonly as components of a variant prioritization workflow. Others have aimed to support users through ontology-assisted visualization and ranking.15–18 Here, we show that integrating multiple phenotype resource optimizes variant evaluation in developmental disorders (IMPROVE-DD) by utilizing a range of clinical datasets coupled with gold-standard diagnoses confirmed by clinical evaluation.

Probabilistic models are often compared through the likelihood ratio that uses Bayes rule to decompose the joint probability of the models under consideration (M1 and M2) and the data to the conditional probability of the data given the model and the prior:

\[ LR(M_1, M_2) = \frac{P(D|M_1)P(M_1)}{P(D|M_2)P(M_2)} \] (Equation 1)

This formulation avoids calculating the probability of observing the data P(D), which can be difficult to evaluate.

The Deciphering Developmental Disorders (DDD) study recruited individuals with severe or extreme developmental disorders (DDs) in whom clinical assessment and baseline genetic investigation were unable to establish a molecular diagnosis.19–22 Whole exome sequencing (WES) was performed in >13,500 unrelated individuals with 85% analyzed as nuclear trios (affected child with both parents) and the remainder as singleton WES. Detailed phenotypic information (see below) was recorded by clinicians using the secure portal within the DECIPHER system23 (deciphergenomics.org). A combination of rational filtering

---

1MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh EH4 2JU, UK; 2Wellcome Sanger Institute, Hinxton, Cambridgeshire CB10 1SA, UK; 3Clinical Genetics Department, Addenbrooke’s Hospital Cambridge University Hospitals, Cambridge CB2 0QQ, UK; 4University of Exeter Medical School, Royal Devon & Exeter Hospital, Barrack Road, Exeter EX2 5DW, UK
5These authors contributed equally
6Lead contact
*Correspondence: colin.semple@ed.ac.uk
https://doi.org/10.1016/j.xhgg.2022.100162.
© 2022 The Authors. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
for impactful genotypes at known DD loci and statistical
genomic approaches to the discovery of novel loci and ge-
etic mechanisms has proven to be successful in diag-
nosing in >33% of the cohort.

At the time of access, the DDD dataset included informa-
on 13,439 individuals. In addition to age, sex, and
HPO terms, the available clinical information included
the following: five growth attributes, which were gestation
(in months), birthweight, height, weight, and head
circumference (expressed as Z scores); plus four develop-
mental milestones (in months) marking when the child
first walked independently, spoke their first words, ex-
pressed a social smile, and sat independently.

We used genetic diagnoses assigned by referring clinical
centers as the ground truth and only considered individu-
als with a single diagnosis. Diagnoses in one of 856 genes
were recorded for 4,112 individuals. However, we required
at least 10 individuals per genetic disorder to build gene-
disease models and a relatively complete record of quanti-
tative data in order to be able to model a gene. This reduced
the number of genes we could consider to 77 in 1,730 in-
dividuals. The median number of individuals per gene
was 17 (maximum 81).

In IMPROVE-DD (web resources), we apply Bayesian
methods to integrate diverse quantitative phenotypic
data types and measure their contribution to decision-
making. Here, decision-making is formalized as classifying
data types and measure their contribution to decision-
methods to integrate diverse quantitative phenotypic
each data type (Figures 1A and 1B) using the R package
implementing a naive Bayes classifier for each gene from
(sex, growth, development, and HPO annotations) by im-
nosis of each of the available data types available in DDD
approach to explore the contributions to making a diag-
the phenotype under consideration is both consistently
value when its distribution for
width) to model the data. A feature will have diagnostic
a smoothed kernel (the nrd0 kernel with increased band-

ontology-encoded phenotypes to aid prediction.8,20

When considering the entire corpus of annotations (to
all 13,439 individuals whether diagnosed or not), the least
frequently used terms are most informative but describe


Measures from information retrieval including term fre-
frequency (TF) and inverse document frequency (IDF) have
previously been adopted for the selection of relevant
HPO terms.25 To compare our set of IPTs with those from
an information theoretic (TF IDF IC) approach, we
computed these measures for our dataset and examined
the position of our IPTs in a ranking of terms per gene
(Figure 2A). IPTs seldom ranked highest by TF IDF IC and
can rank rather lowly, so they would not likely be selected
by such a method. Terms ranking highly by TF IDF IC had
considerably higher information content than IPTs, but

When considering the entire corpus of annotations (to
all 13,439 individuals whether diagnosed or not), the least
frequently used terms are most informative but describe


We next sought methods for selecting HPO terms and
deriving useful probabilistic models from them in the like-
lihood framework. Information content (IC), defined as –
log(probability of term),24 has been combined with
genomic frequency and has been used to compare
ontology-encoded phenotypes to aid prediction.8,20

Measures from information retrieval including term fre-
frequency (TF) and inverse document frequency (IDF) have
previously been adopted for the selection of relevant
HPO terms.25 To compare our set of IPTs with those from
an information theoretic (TF IDF IC) approach, we
computed these measures for our dataset and examined
the position of our IPTs in a ranking of terms per gene
(Figure 2A). IPTs seldom ranked highest by TF IDF IC and
can rank rather lowly, so they would not likely be selected
by such a method. Terms ranking highly by TF IDF IC had
considerably higher information content than IPTs, but
considering the top 10 such terms, the number of terms
per individual was low (Figure 2B). Applying the same
ranking procedure to the terms that define an HPO gene
model1,25(see web resources), we observed a more uniform
ranking of HPO terms, where some terms ranked highly
and others very lowly (Figure S1). We conclude that our
approach uncovers a set of terms that are unlikely to
have been selected by conventional metrics.
As summary statistics, we report the F1 score, being the harmonic mean of precision and recall, as a measure of the accuracy of decision-making, and area under the curve (AUC) as a measure of the rank of true positives irrespective of decision threshold across the entire dataset. The individual classifier outputs were combined by a set of weights, found per gene, that optimized F1 from the four input likelihoods plus an additional constant representing the prior. We discuss F1 and AUC from testing on the training data, as indicative of the potential of phenotype modeling, and results from a leave-one-out cross-validation that better controls for model overfitting.

Beginning with the results from testing on the training data, the IPT-based HPO classifier had the best performance in decision-making for most genes (Figure 3A) with F1 scores of up to 0.46 when testing on training data. Exceptions were apparent; for example, growth was a better predictor for NSD1, while development was a better predictor for GRIN2B (Sotos syndrome MIM: 117550; GRIN2B MIM: 138252). Genes with larger numbers of individuals tended to score well when testing on training data: Pearson correlation between F1 from the HPO classifier and the number of individuals per gene was 0.61 (0.51 in cross-validation; p < 1e-5) indicating that performance...
was positively influenced by the number of individuals. In the interpretation of F1 scores, it should be noted that the prior probabilities $P(M_i)$ derived from the number of individuals range from $1/150$ to $1/19$, whereas the alternative hypotheses are much more probable ($95/100$ to $99/100$). Consequently, the evidence from the data commonly failed to outweigh the prior, so precision or recall was $0$, and F1 could not be calculated. We found the classification results were insensitive to the annotation frequency criteria used to select IPTs (supplemental information).

The best F1 scores from growth data alone were from NSD1, PTEN, and DNMT3A (Figures 3A, S2, and S3) with $20$, $12$, and $14$ individuals, respectively (top three on cross-validation and top five when testing on training data). Turning to developmental milestones, the best F1 scores were from SCN8A, FOXG1, and GRIN2B (Figures 3A, S2, and S3) with $14$, $21$, and $25$ individuals, respectively (top three by cross-validation and in the top eight testing on training data).

The best predicted genes from HPO annotations were PTPN11, KMT2A, and ARID1B with $25$, $69$, and $71$ individuals, respectively (top three by leave-one-out cross-validation, top five when testing on training data). The likelihoods of individual IPTs can be examined for each individual and provide potentially useful diagnostic information to a clinician. We list the three most likely and three least likely KMT2A diagnoses according to phenotype modeling to show the balance of HPO terms for and against and, in the final case, the positive contribution from growth outweighed by the negative contribution.

| Table 1. Examples of informative phenotypic terms |
|-----------------------------------------------|
| Abnormality of the nervous system (HP:0000707) | Abnormality of the cardiovascular system (HP:0001626) | Abnormality of limbs (HP:0000664) |
| Mild global developmental delay (HP:0001342) | Abnormal heart morphology (HP:0001627) | Abnormal 5th finger morphology (HP:0004207) |
| Moderate global developmental delay (HP:0001343) | Abnormality of the vasculature (HP:0002997) | Abnormal thumb morphology (HP:0001172) |
| Severe global developmental delay (HP:0001344) | Abnormal cardiovascular system physiology (HP:0011025) | Abnormal fingertip morphology (HP:0001211) |

Figure 2. Informative phenotypic terms differ from the highest scoring terms by information theoretic criterion

(A) Heatmap showing for each gene (row) the occurrence of an IPT in a ranking of HPO terms by TF IDF IC. HPO terms are ordered left to right by decreasing TF IDF IC. Top panel shows the number of diseases for which an informative term is found in rank $i$ (from 1 to 800), and colors indicate scale, where each covers 100 positions. Although occurring toward the top of the 2,634 terms for which this measure can be calculated, IPTs are seldom in the top 50, and genes such as ANKRD11, ARID1B, and KMT2A in the bottom rows rank 400 and below.

(B) Histograms of mean term IC (left) and mean number of individuals per term (right) when selecting the top 10 terms scored by TF IDF IC per gene.

(C) Histograms of term IC (left) and number of individuals per term (right) for the 157 IPTs.
from HPO terms (Table 2). Of note, the top three cases are pathogenic frameshifts, whereas the bottom three are likely pathogenic missense variants.

As would be expected, sex alone was a poor predictor, but this attribute has information by AUC for three genes: SMC1A, MECP2, and DDX3X (Figure 3B). A strong female bias is known for these X-linked syndromes (OMIM MIM: 300040; MIM: 300005; DDX3X MIM: 300160).

Performance measures from testing on the training data (the apparent error) are expected to be optimistic estimates of the true rates. In contrast, the estimates from leave-one-out cross-validation are believed to have less bias and more variance. To further explore issues of cross-validation, we performed a .632 bootstrap cross-validation that combined the apparent error with that from held out samples in a bootstrap procedure. Due to the imbalance of the classes, the method could only be applied to the recall of the class of interest, which was shown to be intermediate between the apparent and leave-one-out values as would be expected (supplemental information).

We next optimized the F1 score for each gene by combining the likelihood ratios for each data type using five weights \( w_0 - w_4 \) (Equation 2). For a given set of weights, a value from Equation 2 greater than 0 is a classification to the gene, from which true and false positives can be determined and F1 calculated.

\[
F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}
\]

\[
w_0 (Ps(M_1) / Ps(M_1))^{w_1} \cdot (Pd(M_1) / Pd(M_1))^{w_2} \cdot Ph(M_1) / Ph(M_1))^{w_3}
\]

Figure 3. Integrating multiple phenotypic models improves classification

(A) F1 per gene from sex, growth, development, and HPO classifiers (all values plotted at the same x coordinate). Vertical bars and filled symbols highlight the HPO classifier performance. Number of individuals per gene (upper).

(B) AUC per gene highlighting sex.

(C) F1 per gene highlighting the optimized score.

(D) F1 per gene highlighting the optimized score when combining likelihoods from cross-validation.

(E) Violin plots of likelihood ratios from HPO terms for selected genes (symbols are individuals). Likelihoods are shown for a prior of 0.5 to give a common zero reference line for all models. The per-gene prior is shown by the horizontal bar. The HPO model may represent a consensus for a majority of individuals, yet the model may not be sufficient to give a diagnosis for some individuals (KMT2A, ARID1B).

Alternatively, the phenotypic spectrum may be broad (NSD1, PPP2R5D) and correctly diagnose all but a few cases.

(F) Violin plots for the gene models in (E) applied to individuals whose diagnosis differs. A log likelihood above the per-gene prior is a false positive.
In Equation 2, $P(D|M)$ is abbreviated to $P_s(M)$ for sex, and similarly for growth ($P_d$), development ($P_d$), and HPO ($P_h$). Simulated annealing implemented in the R package GenSA was used to find the optimal weights (web resources).

For all genes, optimization improved F1 over any individual data source (Figure 3C), achieving F1 scores greater than 0.7 for 12 genes when testing on training data. Genes with lower numbers of individuals were high in the ranking, indicating that good models can be found for them, but overfitting in the original model training may be at play. The F1 score was generally reduced in leave-one-out cross-validation where 14 genes had F1 0.3–0.5, namely, ARID1B, CHD7, DYSK1A, EFTUD2, EP300, FOXP1, ITPR1, KIF1A, KMT2A, NDIS1, PPP2RS5, PTEN, PTPN11, and SATB2 (Figures 3D and S3). Genes with larger numbers of individuals tended to have higher F1 (Pearson correlation 0.27; $p = 0.015$).

We also examined the distribution of likelihoods of a diagnosis (Figure 3E) versus all others (Figure 3F) from HPO gene models. This analysis highlighted a number of genes, including KMT2A, where the evidence from the fit to the HPO model did not outweigh the prior for many individuals and hence gave false negatives.

We then asked if an optimized HPO classifier would rival the combined data classifier. Optimal values were found for $w_0$ and $w_4$ (for the prior and HPO likelihood) using the same procedure. Using only HPO terms gave 147 fewer true positives (8.5% of the 1,730 individuals) and 2,041 more false positives in total (summing false positives over 77 genes). Per gene, median recall was reduced from 0.57 to 0.5, and median precision from 0.64 to 0.21. The benefits of additional phenotypic information, specifically growth and development data, are clear from these results. As an additional comparison, classifiers based on top terms by TF IDF IC and by disease model were also assessed and found to not perform as well as IPT-based classifiers (supplemental information).

To further investigate the generality of each model in each data type, we assessed growth, development, and HPO models through their contribution to the optimized log likelihood for all individuals for each gene (Figure 4A). This revealed models that worked well in decision-making did not necessarily capture all diagnosed individuals; for example, growth in NDIS1 and development in GRIN2B captured distinctive subsets of individuals. To visualize this across all 77 genes, we selected individuals at quartiles 1, 2, and 3 representing poor, typical, and good fits to the gene models (Figure 4B). Where the scaled values were negative, the model contradicted the assigned diagnosis. We found HPO models agreed with the correct

### Table 2. Clinical evidence supporting or opposing the diagnosis of KMT2A in individuals with a pathogenic de novo variant

| DECIPHER ID | Growth | Dev. | HPO | Total | Informative phenotypic terms supporting KMT2A | Informative phenotypic terms opposing KMT2A |
|-------------|--------|------|-----|-------|-----------------------------------------------|---------------------------------------------|
| 295774      | 2.5    | 2.0  | 18.2| 12.9  | Abnormal size of the palpebral fissures (1.4) Short stature (1) Intellectual disability mild (0.45) Abnormality of the endocrine system (0.44) | — |
| 258419      | 5.0    | 0.8  | 16.2| 12.2  | Short stature (1) Abnormality of the endocrine system (0.44) Cognitive impairment (0.39) | — |
| 294226      | 2.3    | 0.9  | 18.2| 11.7  | Abnormal size of the palpebral fissures (1.4) Short stature (1) Abnormal hair quantity (0.95) Abnormality of upper lip (0.41) | Abnormality of the nervous system (−0.011) |
| 304702      | −1.5   | 1.4  | −1.3| −11.1 | Epicanthus (0.88) Syndactyly (0.64) Abnormality of the genital system (0.23) Abnormality of prenatal development or birth (0.22) | Abnormality of the musculoskeletal system (−0.61) Abnormality of the forehead (−0.41) Abnormality of lower lip (−0.39) Abnormality of the fontanelles or cranial suture (−0.37) Thick vermilion border (−0.37) Moderate global developmental delay (−0.34) Abnormality of toe (−0.017) |
| 273901      | −0.6   | −0.4 | −7.9| −19.3 | Abnormal hair quantity (0.95) Abdominal symptom (0.45) Abnormality of eye movement (0.038) | Involuntary movements (−1.4) Generalized-onset seizure (−1.3) Abnormality of coordination (−0.82) Gait disturbance (−0.64) Dialectic seizure (−0.46) Non-motor seizure (−0.28) Hypotonia (−0.22) Sleep disturbance (−0.2) Abnormality of the respiratory system (−0.041) Abnormality of the immune system (−0.02) |
| 305957      | 5.7    | −0.4 | −17.6| −22.1 | — | Abnormal ear physiology (−1.2) Abnormality of calvarial morphology (−1) Abnormality of skin pigmentation (−0.91) Abnormal emotion/affect behavior (−0.62) Abnormality of the musculoskeletal system (−0.61) Abnormality of the middle ear (−0.28) Localized skin lesion (−0.075) |

*Values are log likelihood ratios. Note that the total includes values not listed here.*
diagnosis at each quartile. Defining models to generalize well as those in which a positive likelihood ratio is found for the median individual, all but four growth models generalized well. However, development models for 22 genes (29%) failed to generalize by this criterion.

In summary, we found that HPO terms are the best predictor of a correct diagnosis for most genes, but 17% (13/77 genes) were better predicted by growth or development, and prediction from combined data performed better than any individual source. Predictions from combined data also gave notably fewer false positives than prediction from HPO terms alone, as median precision increased to 0.64 from 0.21 when only HPO terms were used. While more individuals are to be preferred when building models, we found gene models with good predictive performance on cross-validation could be built from a relatively small number of individuals (n ≥ 10).

The derivation of likelihood ratios from individual annotations using HPO-encoded disease models of LIRICAL is closest to our approach. However, rather than define HPO models per gene, we began with the extensive database of individual annotations of the DDD study from which we are able to assess HPO term usage in clinical practice irrespective of diagnosis: our disease models are then computed from the observed annotations for each gene.

Bayesian methods are recognized as adding quantitative rigor to the combination of evidence for and against variant pathogenicity in rare disease while making explicit any assumptions regarding strength of evidence and disease prevalence. In addition, they are responsive to changes in the evidence base as new observations are made. We extend the application of this paradigm to phenotypic data making use of the extensive acquisition of growth measures and developmental milestones in addition to HPO terms in the DDD study. This approach could be extended by the incorporation of additional phenotypic data (e.g., images, epigenomic profiles, biochemical assays, etc.) to further improve gene-disease models and make them more applicable to other rare disorders. Although phenotypic models are unlikely to be sufficiently predictive by themselves, particularly for genetically heterogeneous disorders such as DD, they can be used to update posterior probabilities found from genomic analyses of

Figure 4. The contribution of data type to diagnosis varies by gene
(A) Heatmaps of log likelihood per individual (column) for each data type (row) for selected genes. Values are scaled by optimization weight and columns ordered right to left from highest to lowest total likelihood (negative values in blue, positive in green). Upper bar shows true positives in black and false negatives in red.
(B) Heatmaps of log likelihood multiplied by optimization weight per gene (column) for each data type (row). Heatmaps show values for individuals at quartiles 1, 2, and 3 successively for each data type. The hierarchical clustering reflects groupings at quartile 1. For example, for GRIN2B, the individual with the first quartile score fits the development model poorly (indicated in blue), and the median GRIN2B individual has a small negative contribution from development, and the contribution from development is positive by the third quartile. Only positive contributions are found from quartile 3, and HPO has a positive contribution at each quartile. The color scale truncates absolute values above 3 in order to focus on the range –3 to 3.
variant pathogenicity and thus have an important role in increasing the robustness of a diagnosis. This is illustrated for an individual with a missense variant in NSD1 with weak genetic evidence that could be strengthened to likely pathogenic through the likelihood ratio from HPO terms by the methods reported here.

Quantitative patient data is of importance to clinical interpretation, and we have proved its value in computational disease modeling; however, systematic collection of and computational access to such data is often lacking in health data systems, potentially hindering diagnostic and biological insights.

Data and code availability

The code used to identify informative HPO terms and to run the classification and optimization procedures can be found in the IMPROVE-DD github repository. Sequence and variant-level data and phenotypic data for the DDD study data are available from the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/) with study ID EGA000001000775. Clinically interpreted variants and associated phenotypes from the DDD study are available through DECIPHER (https://www.deciphergenomics.org/).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100162.

Acknowledgments

The DDD study presents independent research commissioned by the Health Innovation Challenge Fund (grant number HICF-1009-003), a parallel funding partnership between Wellcome and the Department of Health, and the Wellcome Sanger Institute (grant number WT098051). The views expressed in this publication are those of the authors and not necessarily those of Wellcome or the Department of Health. The study has UK Research Ethics Committee approval (10/H0305/83, granted by the Cambridge South REC, and GEN/284/12 granted by the Republic of Ireland REC). The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. D.R.F. is funded as part of the National Institute for Health Research, through the Comprehensive Clinical Research Network. This study makes use of DECIPHER (https://www.deciphergenomics.org/), which is funded by the Wellcome. H.V.F. is supported by the Wellcome Trust (award 200990/Z/16/Z) “Designing, developing and delivering integrated foundations for genomic medicine.” The research team acknowledges the support of the National Institute for Health Research, through the Comprehensive Clinical Research Network. D.R.F. is funded as part of the MRC Human Genetics Unit grant to the University of Edinburgh. C.A.S. and S.A. are supported by MRC Core funding to the MRC Human Genetics Unit (MRC grant MC_UU_000716). For the purpose of open access, the authors have applied a CC-BY public copyright license to any author-accepted manuscript version arising from this submission.

Declaration of interests

M.E.H. is a co-founder, consultant, and non-executive director of Congenica Ltd.

Received: May 25, 2022
Accepted: November 22, 2022

Web resources

DECIPHER, https://www.deciphergenomics.org/
EGA, https://www.ebi.ac.uk/ega/
genSA, https://CRAN.R-project.org/package=GenSA.
HPO https://hpo.jax.org/app/
HPO, Gene Models http://purl.obolibrary.org/obo/hpoa/genes_to_phenotype.txt.
IMPROVE-DD scripts, https://github.com/Stuart-Aitken/IMPROVE-DD,
Naivebayes, https://CRAN.R-project.org/package=naivebayes

References

1. Köhler, S., Gargano, M., Matentzoglu, N., Carmody, L.C., Lewis-Smith, D., Vasilevsky, N.A., Danis, D., Balagura, G., Baynam, G., Brower, A.M., et al. (2021). The human phenotype ontology in 2021. Nucleic Acids Res. 49, D1207–D1217. https://doi.org/10.1093/nar/gkaa1043.
2. Aitken, S., Firth, H.V., McRae, J., Halachev, M., Kini, U., Parker, M.J., Lees, M.M., Lachlan, K., Sarkar, A., Joss, S., et al. (2019). Finding diagnostically useful patterns in quantitative phenotypic data. Am. J. Hum. Genet. 105, 933–946. https://doi.org/10.1016/j.ajhg.2019.09.015.
3. Ferry, Q., Steinberg, J., Webber, C., FitzPatrick, D.R., Ponting, C.P., Zisserman, A., and Nellåker, C. (2014). Diagnostically relevant facial gestalt information from ordinary photos. Elife 3, e02020. https://doi.org/10.7554/eLife.02020.
4. FitzPatrick, D.R., and Firth, H.V. (2020). Genomically aided diagnosis of severe developmental disorders. Annu. Rev. Genomics Hum. Genet. 21, 327–349. https://doi.org/10.1146/annurev-genom-120919-082329.
5. Köhler, S., Olen, N.C., Buske, O.J., Groza, T., Jacobsen, J.O.B., McNamara, C., Vasilevsky, N., Carmody, L.C., Goudrine, J.P., Gargano, M., et al. (2019). Encoding clinical data with the human phenotype ontology for computational differential diagnostics. Curr. Protoc. Hum. Genet. 103, e92. https://doi.org/10.1002/cphg.92.
6. Robinson, P.N., Ravanmehr, V., Jacobsen, J.O.B., Zhang, X.A., Carmody, L.C., Gargano, M.A., Thaxton, C.L., UNC Biocuration Core, and Karlebach, G., et al. (2020). Interpretational clinical genomics with a likelihood ratio paradigm. Am. J. Hum. Genet. 107, 403–417. https://doi.org/10.1016/j.ajhg.2020.06.021.
7. Li, Z., Zhang, F., Wang, Y., Qiu, Y., Wu, Y., Lu, Y., Yang, L., Qu, W.J., Wang, H., Zhou, W., and Tian, W. (2019). PhenoPro: a novel toolkit for assisting in the diagnosis of Mendelian disease. Bioinformatics 35, 3559–3566. https://doi.org/10.1093/bioinformatics/btz100.
8. Li, Q., Zhao, K., Bustamante, C.D., Ma, X., and Wong, W.H. (2019). Xrare: a machine learning method jointly modeling phenotypes and genetic evidence for rare disease diagnosis. Genet. Med. 21, 2126–2134. https://doi.org/10.1038/s41436-019-0439-8.
9. Xue, H., Peng, J., and Shang, X. (2019). Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO. BMC Syst. Biol. 13, 34. https://doi.org/10.1186/s12918-019-0697-8.
10. Gao, J., Liu, L., Yao, S., Huang, X., Mamitsuka, H., and Zhu, S. (2019). HPOAnnotator: improving large-scale prediction of HPO annotations by low-rank approximation with HPO semantic similarities and multiple PPI networks. BMC Med. Genomics 12, 187. https://doi.org/10.1186/s12920-019-0625-1.

11. Akawi, N., McRae, J., Ansari, M., Balasubramanian, M., Blyth, M., Brady, A.E., Clayton, S., Cole, T., Deshpande, C., Fitzgerald, T.W., et al. (2015). Discovery of four recessive developmental disorders using probabilistic genotype and phenotype matching among 4, 125 families. Nat. Genet. 47, 1363–1369. https://doi.org/10.1038/ng.3410.

12. Javed, A., Agrawal, S., and Ng, P.C. (2014). Phen-Gen: combining phenotype and genotype to analyze rare disorders. Nat. Methods 11, 935–937. https://doi.org/10.1038/nmeth.3046.

13. Kulmanov, M., and Hoehndorf, R. (2020). DeepPheno: predicting single gene loss-of-function phenotypes using an ontology-aware hierarchical classifier. PLoS Comput. Biol. 16, e1008453. https://doi.org/10.1371/journal.pcbi.1008453.

14. Boudellioua, I., Kulmanov, M., Schofield, P.N., Gkoutos, G.V., and Hoehndorf, R. (2019). DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC Bioinf. 20, 65. https://doi.org/10.1186/s12859-019-2633-8.

15. Pontikos, N., Yu, J., Moghul, I., Withington, L., Blanco-Kelly, F., Vulliamy, T., Wong, T.L.E., Murphy, C., Cipriani, V., Fiorentino, A., et al. (2017). Phenopolis: an open platform for harmonization and analysis of genetic and phenotypic data. Bioinformatics 33, 2421–2423. https://doi.org/10.1093/bioinformatics/btx147.

16. Köhler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dölken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am. J. Hum. Genet. 85, 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003.

17. Bone, W.P., Washington, N.L., Buske, O.J., Adams, D.R., Davis, J., Draper, D., Flynn, E.D., Girdea, M., Godfrey, R., Golas, G., et al. (2016). Computational evaluation of exome sequence data using human and model organism phenotypes improves diagnostic efficacy. Genet. Med. 18, 608–617. https://doi.org/10.1038/gim.2015.137.

18. Yang, J., Dong, C., Duan, H., Shu, Q., and Li, H. (2021). RDmap: a map for exploring rare diseases. Orphanet J. Rare Dis. 16, 101. https://doi.org/10.1186/s13023-021-01741-4.

19. Deciphering Developmental Disorders Study, Gerety, S.S., Jones, W.D., van Kogelenberg, M., King, D.A., McRae, J., Morley, K.I., Parthiban, V., Al-Turki, S., Ambridge, K., et al. (2015). Large-scale discovery of novel genetic causes of developmental disorders. Nature 519, 223–228. https://doi.org/10.1038/nature14135.

20. Deciphering Developmental Disorders Study, Clayton, S., Fitzgerald, T.W., Kaplanis, J., Prigmore, E., Rajan, D., Sifrim, A., Aitken, S., Akawi, N., Alvi, M., et al. (2017). Prevalence and architecture of de novo mutations in developmental disorders. Nature 542, 433–438. https://doi.org/10.1038/nature21062.

21. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet 385, 1305–1314.

22. Firth, H.V., Wright, C.F.; and DDD Study (2011). The deciphering developmental disorders (DDD) study. Dev. Med. Child Neurol. 53, 702–703. https://doi.org/10.1111/j.1469-8749.2011.04032.x.

23. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M., and Carter, N.P. (2009). DECIPHER: database of chromosomal imbalance and phenotype in humans using ensemble resources. Am. J. Hum. Genet. 84, 524–533. https://doi.org/10.1016/j.ajhg.2009.03.010.

24. Resnik, P. (1999). Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. 11, 95–130.

25. Groza, T., Köhler, S., Moldenhauer, D., Vasilyevsky, N., Baynam, G., Zemotel, T., Schriml, L.M., Kibbe, W.A., Schofield, P.N., Beck, T., et al. (2015). The human phenotype ontology: semantic unification of common and rare disease. Am. J. Hum. Genet. 97, 111–124. https://doi.org/10.1016/j.ajhg.2015.05.020.

26. Köhler, S., Carmody, L., Vasilyevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.-P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al. (2019). Expansion of the human phenotype ontology (HPO) knowledge base and resources. Nucleic Acids Res. 47, D1018–D1027. https://doi.org/10.1093/nar/gky1105.

27. Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. J. Am. Stat. Assoc. 78, 316–331. https://doi.org/10.2307/2288636.

28. Tavtigian, S.V., Greenblatt, M.S., Harrison, S.M., Nussbaum, R.L., Prabhu, S.A., Boucher, K.M., Biesecker, L.G.; and ClinGen Sequence Variant Interpretation Working Group ClinGen SVI (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet. Med. 20, 1054–1060. https://doi.org/10.1038/s41436-017-210.

29. Wright, C.F., Campbell, P., Eberhardt, R.Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E.J., Chundra, V.K., Lindsay, S.J., et al. (2022). Optimising diagnostic yield in highly penetrant genomic disease. Preprint at medRxiv. https://doi.org/10.1101/2022.07.25.22278008.