GenomeDiver: A platform for phenotype-guided medical genomic diagnosis.

Nathaniel Pearson PhD^{a,1}, Christian Stolte^{a,2}, Kevin Shi PhD^{a,2}, Faygel Beren BA³, Noura S. Abul-Husn MD PhD^{4,5,6}, Gabrielle Bertier PhD⁷, Kaitlyn Brown MS⁸, Jacqueline A. Odgis MS⁴, Sabrina A. Suckiel MS⁶, Carol R. Horowitz MD⁷, Melissa Wasserstein MD^{8,9}, Bruce D. Gelb MD⁷, Eimear E. Kenny PhD^{4,5,6}, Charles Gagnon BEng², Vaidehi Jobanputra PhD², Toby Bloom PhD², John M. Greally DMed PhD^{8,9,10}

^aCo-first authors

Institutional addresses:

¹Root Deep Insight, Inc. 76 Easton St Unit 5, Boston, MA 02134, USA.

²New York Genome Center, 101 Avenue of the Americas, New York, NY 10013, USA.

³Columbia University, Graduate School of Arts and Sciences, 535 W 116th St #109, New York, NY 10027, USA.

⁴Institute for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

⁵Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029

⁶Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA.

⁷Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA.

⁸Division of Genetics, Department of Pediatrics, Children's Hospital at Montefiore, 3415 Bainbridge Ave, Bronx, NY 10467, USA.

⁹Department of Genetics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA.

¹⁰Center for Epigenomics, Albert Einstein College of Medicine, 1301 Morris Park Avenue, Bronx, NY 10461, USA.

Current addresses:

CS: Stolte Design, 358 Main Road, Islesboro, ME 04848, USA

TB: eGenesis, Inc, 300 Tech Sq, Cambridge, MA 02139, USA

Corresponding author

John M. Greally, telephone 718 678 1234, email john.greally@einsteinmed.org

CONFLICTS OF INTEREST

NSAH was previously employed at Regeneron Pharmaceuticals and has received an honorarium from Genentech. EEK has received speaker honorariums from Regeneron Pharmaceuticals and Illumina, Inc.

ABSTRACT

Purpose: Making a diagnosis from clinical genomic sequencing requires well-structured phenotypic data to guide genotype interpretation. A patient's phenotypic features can be documented using the Human Phenotype Ontology (HPO), generating terms used to prioritize genes potentially causing the patient's disease. We have developed GenomeDiver to provide a user interface for clinicians that allows more effective collaboration with the clinical diagnostic laboratory, with the goal of improving the success of the diagnostic process.

Methods: GenomeDiver is designed to prompt reverse phenotyping of patients undergoing genetic testing, enriching the amount and quality of structured phenotype data for the diagnostic laboratory, and helping clinicians to explore and flag diseases potentially causing their patient's presentation.

Results: We show how GenomeDiver communicates the clinician's informed insights to the diagnostic lab in the form of HPO terms for interpretation of genomic sequencing data. We describe our user-driven design process, the engineering of the software for efficiency, security and portability, and an example of the performance of GenomeDiver using simulated genomic testing data.

Conclusions: GenomeDiver is a first step in a new approach to genomic diagnostics that enhances laboratory-clinician interactions, with the goal of directly engaging clinicians to improve the outcome of genomic diagnostic testing.

INTRODUCTION

Compelling economic evidence now supports how genomic sequencing saves both money and time^{1,2} and improves quality-adjusted life years³ in the diagnosis of the 3.5–5.9% of the population with rare diseases.⁴ To make this genomic diagnostic process more efficient, substantial attention has been appropriately paid to detecting and understanding the effects of sequence variants in the human genome.⁵ Linking genetic variants to diseases is challenging and is most successful in case-control studies or when comparing multiple affected and unaffected family members, for which excellent analytical approaches have been developed.^{6,7} In a clinical context, however, testing can only be ordered on the proband and one or both parents (if available), with results needed sufficiently rapidly to influence care. The genomic diagnostic laboratory relies on linking the variants present in the patient's genome with the patient's phenotypic features. When a pathogenic genetic variant has previously been found to be the cause of a disease, that associated disease is characterized by a set of Human Phenotype Ontology (HPO) terms.⁸ If the patient has a variant that appears potentially pathogenic, causality is supported if the list of HPO terms in the disease associated with a pathogenic variant of that gene significantly overlaps the HPO terms used to describe the patient. A number of phenotype-based variant prediction tools exist to perform this kind of variant prioritization.⁹

If the diagnostic laboratory is provided with a greater number of HPO terms describing the patient on whom sequencing was performed, the diagnostic yield is increased.^{10,11} The comprehensive collection of these HPO terms is very difficult to scale effectively. Ordering genomic tests for rare diseases involves providing the lab with a primary indication for the testing (e.g. hearing loss, seizures), but also requires communicating additional findings in the patient, sometimes through checklists completed as part of the test order, also potentially involving the diagnostic laboratory reviewing notes from the patient's health record. The diagnostic laboratory staff then extract information from these sources and can make decisions about how the phenotype can be represented as HPO terms. In critically ill

children, an approach has been described that involves automated extraction of HPO terms from the electronic health record (EHR) using natural language processing,¹² helping with the speed of generation of HPO terms.

We have developed GenomeDiver to help clinicians contribute more effectively to the diagnostic process. GenomeDiver facilitates the curation by the clinician of case-relevant HPO terms quickly, in greater numbers and more accurately than is currently typical, with the goal of improving our ability to make diagnoses using genomic information.

MATERIALS AND METHODS

The user-centered design process for GenomeDiver

We developed GenomeDiver using evidence-based participatory methods from design thinking.¹³ User research began with initial interviews with physicians and genetic testing laboratory staff, followed by a one-day stakeholder workshop and a design sprint to understand user needs, to define requirements, to sketch candidate workflows, to choose among them and to prototype. We included physicians (with and without specialist genetics training), genetic counselors, genetic testing laboratory personnel, research scientists, software tool builders, bioinformaticians and biologists, a designer, and a software engineer. The resulting concepts, sketches, and paper prototypes informed the designs for user experience, workflow, and user interfaces.

The GenomeDiver workflow: overview of a 'dive'

The workflow for a case (**Figure S1**) begins with the upload by the genetic testing lab of the patient's genomic information as a variant call format (VCF) file, along with the HPO terms that they derived from the test requisition documents. The clinician then starts a 'dive' by selecting the patient from the GenomeDiver interface. The next step, refining the patient's phenotype, requires the clinician categorize HPO terms generated by GenomeDiver by dragging each into one of three screen areas: 'Present' in the patient, 'Absent', or 'Unknown'. Using this updated and enriched phenotypic information, GenomeDiver re-analyzes the genomic data and presents candidate diseases for the clinician to review and flag, adding any comments that they would like to communicate to the diagnostic laboratory. To conclude the workflow, the updated HPO term categorizations, flagged diseases, clinician comments and the re-ranked shortlist of variants are returned to the diagnostic laboratory for interpretation.

Variant prioritization and presentation of HPO terms

We illustrate the variant prioritization steps in **Figure S2**. The VCF and starting HPO information are used for an Exomiser¹⁴ analysis. Exomiser is a gene prioritization tool that combines a score quantifying the likely pathogenicity of a variant associated with a gene (variant score) with a second score that measures the similarity of the phenotypic features of the patient with those associated with a pathogenic variant of the gene (phenotype score). Both values are used to generate the combined score for the variant.

We describe the prioritization and filtering steps in detail in the **Supplementary Information**. An Exomiser run is initiated based on the patient's VCF and the HPO terms from the test requisition. This generates a list of variants, each of which is associated with a gene and one or more diseases, which in turn are associated with HPO terms present in individuals who have pathogenic variants of that gene, all of which are collected at this point. We then filter HPO terms for redundancy, focusing on those that are descendants of 'Phenotypic abnormality' (HP:0000118), and those associated with the specific disease prioritized by Exomiser. This leaves a subset of HPO terms that go through multiple rounds of selection, with the goal to present to the clinician ≤ 5 non-redundant HPO terms associated with each candidate gene, generating a maximum of 25 new HPO terms for categorization.

Re-analysis based on enriched HPO information and disease exploration

A further potential input by the clinician is made possible by re-running Exomiser, now updated with the newly categorized HPO terms. The modified list of HPO terms updates the Exomiser phenotype to generate revised combined scores. These allow the refinement of the findings for clinician review, listed by gene, presenting the new list of genes ranked by the absolute combined score, displaying the magnitude and direction of change of the scores, and linking to associated candidate diseases. The hyperlink embedded with the names of each candidate disease brings the clinician to a description of the disease, letting

them judge whether the syndrome of features plausibly fits their patient. Any disease of interest can be flagged, with the final step of the dive involving the clinician returning the list of updated HPO terms, any flagged diseases and free text comments as a file for the diagnostic laboratory.

Software system architecture

We provide a detailed description of the GenomeDiver software in the **Supplementary Information** section.

RESULTS

GenomeDiver performance with simulated data

A video of the diagnostic laboratory interaction with the interface to set up a patient in the GenomeDiver system appears as **Supplementary Video 1**. Once set up, this allows a clinician to access the separate interface shown in **Figure 1** for categorization of HPO terms. The upper section 'Add Phenotype Feature' allows text to be entered, prompting HPO terms as a dropdown list for selection by clicking. The design is intended to focus the clinician on the middle section to 'Classify All Phenotype Features', dragging and dropping individual HPO terms into categories of 'Present', 'Absent' or 'Unknown' (not known whether present in the patient). After all terms have been categorized, the 'Submit' button on the bottom right activates and the updated information is sent back to GenomeDiver. This interaction appears in **Supplementary Video 2**.

We use the VCF for the publicly-available NA12878 genome,¹⁵ adding a variant of uncertain significance in the *FBN1* gene (ClinVar accession VCV000200085.4, NM_000138.4(*FBN1*): c.6449G>T(p.Arg2150Leu)). To simulate a request for analysis of a patient presenting with a clinical suspicion of Marfan syndrome (OMIM 154700), we used HPO terms for ascending tubular aorta aneurysm (HP:0004970), scoliosis (HP:0002650) and arachnodactyly (HP:0001166) as those potentially used in a test requisition in a clinical scenario.

We detail each step of the process of variant and gene filtering and prioritization in the **Supplementary Information** section (**Figure S2**), as well as the selection of HPO terms for presentation to the clinician (**Figures S3-S4**). For this particular combination of genomic and phenotypic information, two genes were selected, *FBN1* and *SKI*, the latter implicated in Shprintzen-Goldberg syndrome (OMIM 182212), whose phenotypic features likewise include aortic aneurysm, scoliosis and arachnodactyly.¹⁶

The HPO terms presented to the clinician are listed in **Supplementary Table 1**. In our experience, the process of categorization of ≤25 HPO terms in practice takes the user no

more than ~3 minutes, once reasonably familiar with the interface. We show how each term in our sample case was categorized in **Supplementary Table 1** and illustrate how a second Exomiser run presents candidate diseases in **Figure 2** (using steps illustrated in **Figure S5**). Of note, GenomeDiver never exposes information about variants to the clinician, preventing the clinician from bypassing the formal laboratory diagnostic reporting process.

DISCUSSION

In this initial application of GenomeDiver, we focused on improving the diagnostic process when performing genome-wide (exome, genome) sequencing in the diagnosis of rare diseases. We recognize that we can potentially improve both the rate of successful diagnosis and decrease the time spent by diagnostic laboratory personnel if we can facilitate the provision of the HPO terms that are most likely to discriminate the highest ranked gene candidates for causing the patient's disease. With the appreciation that clinician time is limited, the interface is designed to maximize intuitiveness and ease of use, limiting the number of candidate HPO terms so that the entire categorization session should last no more than several minutes per patient. The second stage of input, facilitating the exploration of diseases that could be affecting the patient, permits further clinician insights to be provided, this time studying the syndrome of phenotypic features in candidate diseases. These steps of identifying individual phenotypic features and then considering how they might aggregate in diseases and syndromes reflect reasonably accurately the typical clinical genetics evaluation, but with the added intent of contributing to the laboratory diagnostic process. While we present certain choices for variant and HPO prioritization in the current version, these are intended to serve as the basis for an initial, functional version of GenomeDiver, but with the goal of incorporating community input for algorithmic improvement over time.

It is likely that HPO term harvesting from sources like EHRs¹⁷ and image analysis¹⁸ represents an area of innovation that will expand over time. A value of the GenomeDiver interface is that it allows a clinician to assess whether candidate terms gleaned from diverse, automated sources are actually present in a patient, creating the potential that GenomeDiver can act as a clearinghouse for downstream curation of candidate phenotypic features. The categorization of HPO terms as 'absent' is not yet used by Exomiser, but the updated LIRICAL approach, based on a likelihood ratio framework that includes information about

HPO prevalence data,¹⁹ can exploit such information and will be served by this functionality within GenomeDiver.

There are more potential uses for GenomeDiver than for the initial diagnosis of a patient with a rare disease. GenomeDiver can facilitate re-analysis of patients with initially uninformative results whose phenotypes may have changed over time, and whose genomic variants may have undergone re-classification. GenomeDiver is fundamentally a tool to prompt ‘reverse phenotyping’ – the evaluation whether specific phenotypic features are present in a patient based on their genotypic information. As such, GenomeDiver has potential to serve clinicians not trained in clinical genetics and dysmorphology. With the development of HPO terms that now also encompass common diseases,²⁰ GenomeDiver’s potential value as a reverse phenotyping tool that generates a genotype-based differential diagnosis could accordingly extend to much of medical genomics.

DATA AVAILABILITY

The GenomeDiver software is available at <https://github.com/GenomeDiver/>,

ACKNOWLEDGMENTS

Research reported in this publication was part of the NYCKidSeq project, supported by the National Human Genome Research Institute and National Institute for Minority Health and Health Disparities of the National Institutes of Health under Award Number 1U01HG0096108.

AUTHOR INFORMATION

Conceptualization NP CS KS FB NAH KB JAO SAS BDG EEK VJ TB JMG, Data curation VJ, Formal analysis NP FB CG TB JMG, Funding acquisition CRH MW BDG EEK,

Investigation NP CS KS FB TB JMG, Methodology NP CS KS FB CG VJ TB JMG, Project administration NP FB GB EEK CG VJ TB JMG, Resources CRH MW BDG EEK CG VJ TB JMG, Software NP CS KS FB JMG, Supervision GB EEK CG JMG, Validation NP CS KS FB, Visualization NP CS KS FB JMG, Writing – original draft NP CS KS CG JMG, Writing – Review & Editing NP CS KS FB NAH GB KB JAO SAS CRH MW BDG EEK CG TB JMG.

REFERENCES

1. Splinter, K. *et al.* Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
2. Chung, C. C. Y. *et al.* Rapid whole-exome sequencing facilitates precision medicine in paediatric rare disease patients and reduces healthcare costs. *The Lancet Regional Health - Western Pacific* **1**, 100001 (2020).
3. Schofield, D., Rynehart, L., Shrestha, R., White, S. M. & Stark, Z. Long-term economic impacts of exome sequencing for suspected monogenic disorders: diagnosis, management, and reproductive outcomes. *Genet. Med.* **21**, 2586–2593 (2019).
4. Nguengang Wakap, S. *et al.* Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur. J. Hum. Genet.* **28**, 165–173 (2020).
5. Rehm, H. L. & Fowler, D. M. Keeping up with the genomes: scaling genomic variant interpretation. *Genome Med.* **12**, 5 (2019).
6. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
7. Hu, H. *et al.* VAAST 2.0: improved variant classification and disease-gene identification using a conservation-controlled amino acid substitution matrix. *Genet. Epidemiol.* **37**, 622–634 (2013).
8. Köhler, S. *et al.* Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* **47**, D1018–D1027 (2019).
9. Pengelly, R. J. *et al.* Evaluating phenotype-driven approaches for genetic diagnoses from exomes in a clinical setting. *Sci. Rep.* **7**, 13509 (2017).
10. Trujillano, D. *et al.* Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur. J. Hum. Genet.* **25**, 176–182 (2017).
11. Thompson, R. *et al.* Increasing phenotypic annotation improves the diagnostic rate of

- exome sequencing in a rare neuromuscular disorder. *Hum. Mutat.* **40**, 1797–1812 (2019).
12. Clark, M. M. *et al.* Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci. Transl. Med.* **11**, (2019).
 13. Brown, T. Design thinking. *Harv. Bus. Rev.* **86**, 84–92, 141 (2008).
 14. Smedley, D. *et al.* Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat. Protoc.* **10**, 2004–2015 (2015).
 15. Zook, J. M. *et al.* Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* **3**, 160025 (2016).
 16. Greally, M. T. in *GeneReviews®* (eds. Adam, M. P. *et al.*) (University of Washington, Seattle, 1993).
 17. Deisseroth, C. A. *et al.* ClinPhen extracts and prioritizes patient phenotypes directly from medical records to expedite genetic disease diagnosis. *Genet. Med.* **21**, 1585–1593 (2019).
 18. Hsieh, T.-C. *et al.* PEDIA: prioritization of exome data by image analysis. *Genet. Med.* **21**, 2807–2814 (2019).
 19. Robinson, P. N. *et al.* Interpretable Clinical Genomics with a Likelihood Ratio Paradigm. *medRxiv* (2020). doi:10.1101/2020.01.25.19014803
 20. Groza, T. *et al.* The human phenotype ontology: semantic unification of common and rare disease. *Am. J. Hum. Genet.* **97**, 111–124 (2015).

FIGURES

FIGURE 1: The clinician interface for GenomeDiver. HPO terms can be added in the 'Add Phenotype Feature' field by entering text which will bring up a list of HPO terms from which the appropriate term can be chosen. The HPO terms generated by GenomeDiver are shown in the 'Classify all phenotype features' section. These can be dragged into each of the categories underneath, 'Present', 'Absent' or 'Unknown'. Three terms have been moved into the 'Present' category as an example.

FIGURE 2: Following the categorization of the HPO terms, a second Exomiser run re-prioritizes variants, genes and associated diseases. The clinician is presented with a shortlist of genes, ranked by the Exomiser combined_score which is shown, as well as the change in score (green positive, red negative) resulting from the HPO term categorization. We list the disease and a link that allows the clinician to explore whether a disease description resembles their patient, allowing them to flag one or more as being candidates for causing the patient's presentation. Free text can be used in the 'Add comment' box, finalizing the process by clicking 'Confirm analysis' to return the information to the diagnostic laboratory.

Refine phenotype

> Emily Dickinson

ADD PHENOTYPE FEATURE:
i.e. Intellectual Disability

CLASSIFY ALL PHENOTYPE FEATURES: (10)

Mitral annular calcification
HP:0005136

Striae distensae
HP:0001065

Patient identifier

Manual addition of HPO terms

GenomeDiver-prompted HPO terms

GenomeDiver clinician interface

medRxiv preprint doi: <https://doi.org/10.1101/2020.11.22.20236430>; this version posted November 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

PRESENT (3)

Scoliosis
★ HP:0002650

Ascending tubular aorta aneurysm
★ HP:0004970

Arachnodactyly
★ HP:0001166

ABSENT (0)

UNKNOWN (0)

Drag/drop categorization of HPO terms

Gene/disease exploration interface

Flag any disease(s) likely fitting your patient

Gene	Score (Change)	Flag?	Disease	Explore Disease
FBN1	0.9965 (+0.0032)		Marfan Syndrome; MFS	OMIM:154700
TGFB2	0.9149 (-0.0669)		Loeys-Dietz Syndrome 4; LDS4	OMIM:614816
COL5A1	0.8874 (-0.0723)		Vascular Ehlers-Danlos Syndrome	ORPHA:286

[Download CSV](#)

Add Comment

UPDATE COMMENT

CONFIRM ANALYSIS