

ARTICLE PREVIEW

view full access options

NATURE METHODS | BRIEF COMMUNICATION

Phen-Gen: combining phenotype and genotype to analyze rare disorders

Asif Javed, Saloni Agrawal & Pauline C Ng

Nature Methods **11**, 935–937 (2014) doi:10.1038/nmeth.3046

Received 01 April 2014 Accepted 16 June 2014 Published online 03 August 2014

We introduce Phen-Gen, a method that combines patients' disease symptoms and sequencing data with prior domain knowledge to identify the causative genes for rare disorders. Simulations revealed that the causal variant was ranked first in 88% of cases when it was a coding variant—a 52% advantage over a genotype-only approach—and Phen-Gen outperformed other existing prediction methods by 13–58%. If disease etiology was unknown, the causal variant was assigned the top rank in 71% of simulations. Phen-Gen is available at <http://phen-gen.org/>.

Print

Subject terms: Software Genomics Translational research Genetics research

READ THE FULL ARTICLE

Subscribe to
Nature Methods
for full access:
\$199

Subscribe

ReadCube
Access*:
\$3.99 rent
\$9.99 buy

*printing and sharing restrictions apply

Buy/Rent now

Purchase article
full text and PDF:
\$32

Buy now

Already a subscriber? **Log in now** or **Register** for online access.

Additional access options:

Use a document delivery service | Login via Athens | Purchase a site license | Institutional access

References

1. de Ligt, J. *et al.* *N. Engl. J. Med.* **367**, 1921–1929 (2012).
2. Yang, Y. *et al.* *N. Engl. J. Med.* **369**, 1502–1511 (2013).
3. Cordero, J.F. *N. Engl. J. Med.* **352**, 2032 (2005).
4. Amberger, J., Bocchini, C.A., Scott, A.F. & Hamosh, A. *Nucleic Acids Res.* **37**, D793–D796 (2009).
5. Chakravarti, A. *Genome Res.* **21**, 643–644 (2011).

6. Köhler, S. *et al. Am. J. Hum. Genet.* **85**, 457–464 (2009).
7. Sifrim, A. *et al. Nat. Methods* **10**, 1083–1084 (2013).
8. Yandell, M. *et al. Genome Res.* **21**, 1529–1542 (2011).
9. Singleton, M.V. *et al. Am. J. Hum. Genet.* **94**, 599–610 (2014).
10. Robinson, P.N. *et al. Genome Res.* **24**, 340–348 (2014).
11. Stenson, P.D. *et al. Hum. Genet.* **133**, 1–9 (2014).
12. Fu, W. *et al. Nature* **493**, 216–220 (2013).
13. Visel, A. *et al. Nature* **464**, 409–412 (2010).
14. Khurana, E. *et al. Science* **342**, 1235587 (2013).
15. Pruitt, K.D. *et al. Genome Res.* **19**, 1316–1323 (2009).
16. Sim, N.-L. *et al. Nucleic Acids Res.* **40**, W452–W457 (2012).
17. Adzhubei, I.A. *et al. Nat. Methods* **7**, 248–249 (2010).
18. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. *Proc. Natl. Acad. Sci. USA* **106**, 3871–3876 (2009).
19. Schwarz, J.M., Rödelberger, C., Schuelke, M. & Seelow, D. *Nat. Methods* **7**, 575–576 (2010).
20. Lewin, B. *Genes VIII* (Benjamin Cummings, 2004).
21. Price, A.L. *et al. Am. J. Hum. Genet.* **86**, 832–838 (2010).
22. Davydov, E.V. *et al. PLoS Comput. Biol.* **6**, e1001025 (2010).
23. Cooper, G.M. *et al. Genome Res.* **15**, 901–913 (2005).
24. Prabhakar, S. *et al. Genome Res.* **16**, 855–863 (2006).
25. Derrien, T. *et al. Genome Res.* **22**, 1775–1789 (2012).
26. Kozomara, A. & Griffiths-Jones, S. *Nucleic Acids Res.* **39**, D152–D157 (2011).
27. Smith, N.G.C., Webster, M.T. & Ellegren, H. *Genome Res.* **12**, 1350–1356 (2002).
28. He, L. & Hannon, G.J. *Nat. Rev. Genet.* **5**, 522–531 (2004).
29. Esteller, M. *Nat. Rev. Genet.* **12**, 861–874 (2011).
30. McLean, C.Y. *et al. Nat. Biotechnol.* **28**, 495–501 (2010).
31. The 1000 Genomes Project Consortium. *Nature* **467**, 1061–1073 (2010).
32. Sherry, S.T. *et al. Nucleic Acids Res.* **29**, 308–311 (2001).
33. MacArthur, D.G. *et al. Science* **335**, 823–828 (2012).
34. Wu, G., Feng, X. & Stein, L. *Genome Biol.* **11**, R53 (2010).
35. Matthews, L. *et al. Nucleic Acids Res.* **37**, D619–D622 (2009).
36. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. *Nucleic Acids Res.* **40**, D109–D114 (2012).
37. Schaefer, C.F. *et al. Nucleic Acids Res.* **37**, D674–D679 (2009).
38. Stark, C. *et al. Nucleic Acids Res.* **34**, D535–D539 (2006).
39. Franceschini, A. *et al. Nucleic Acids Res.* **41**, D808–D815 (2013).
40. Ashburner, M. *et al. Nat. Genet.* **25**, 25–29 (2000).
41. Obayashi, T. *et al. Nucleic Acids Res.* **41**, D1014–D1020 (2013).

42. The 1000 Genomes Project Consortium. *Nature* **491**, 56–65 (2012).

[Download references](#)

Author information

Affiliations

Computational and Systems Biology Group, Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore.

Asif Javed, Saloni Agrawal & Pauline C Ng

Contributions

A.J. conceived of and designed the project, designed and implemented the analysis framework, implemented methods, conducted experiments, interpreted results, wrote the initial manuscript and revised and proofread the paper. S.A. implemented methods, conducted experiments, set up the web server and revised and proofread the paper. P.C.N. conceived of and designed the project, revised and proofread the paper and supervised the project.

Competing financial interests

The authors declare no competing financial interests.

Corresponding authors

Correspondence to: Asif Javed or Pauline C Ng

Supplementary information

Supplementary Figures

1. Supplementary Figure 1: Overall workflow. (141 KB)
Patient disease symptoms are matched against known disorders and the probability of a symptomatic match is assigned to genes implicated for the respective disorder. These probabilities are permeated to known gene associates using a random walk with restart on the interaction network. In parallel the patient's sequencing data is analyzed and the damaging impact of each variant estimated and pooled within genes. These two predictions are combined to implicate the gene(s) involved.
2. Supplementary Figure 2: Distribution of SIFT and PolyPhen-2 scores for damaging and benign nonsynonymous mutations. (231 KB)
The distribution of SIFT and PolyPhen-2 scores for HGMD-reported damaging nonsynonymous mutations and neutral nonsynonymous fixed substitutions inferred from human-chimp alignment are shown. The plots indicate general agreement between the two methods.
3. Supplementary Figure 3: Deleteriousness predictions around splice site. (77 KB)
The figure depicts the probability of deleteriousness around donor and acceptor sites for splice site mutations.
4. Supplementary Figure 4: Probability of deleteriousness using the genomic predictor. (154 KB)
The figure illustrates the predicted deleteriousness of different combination of five annotations: GERP++ (G), PhyloP (P), near-genic (N), transcription factor binding sites (T), and DNase hypersensitive sites (D). The predictions are binned according to the number of annotations (shown on the x-axis). Each bin is further canonically sorted based on the fore mentioned order of annotations.
5. Supplementary Figure 5: Confidence intervals for positive and benign mutation set combinations. (283 KB)
The 90% confidence intervals of different combination of genomic annotations are shown. The order from Supplementary Figure 4 is maintained. With the four sub-figures representing combinations of the two positive sets (HGMD and GWAS) and the two neutral sets (common variation in dbSNP and Complete Genomics MAF>0.30), respectively.
6. Supplementary Figure 6: Histograms of the null distribution of deleteriousness of genes. (201 KB)
The top 1 percentile of damaging variants in each gene is shown. The histogram of this null distribution cutoff for all genes under dominant and recessive inheritance pattern for coding and genomic predictors is shown. Most genes do not harbor any putative damaging variants and hence the distributions are dominated by the left most bar; which has been truncated for better visual representation.

7. Supplementary Figure 7: Performance of variant predictors. (101 KB)

The distribution of damaging probabilities assigned to different classes of HGMD variants is shown. The top three panels employ the coding predictor. A genomic predictor was used for the bottom panel and applied to noncoding regulatory variants. The histograms depict the distribution of the scored variants. The pie charts on the right explicate the distribution of omitted and predicted variants in each category. Common variants (white) were observed in 1000 Genomes, ESP, or dbSNP with MAF 0.01. Commonly mutated genes indicate that the variants failed to exceed the null distribution of the respective gene (light green). Missed indicates that the variant eluded our regions of interest (dark blue).

8. Supplementary Figure 8: Prediction of heterozygous variants. (143 KB)

The figure depicts how compound heterozygous variants are evaluated. When both damaging variants reside within the coding region, the coding predictor is used to estimate the damaging impact of these variants. In cases when one or both variants lay outside the exon boundaries, both variants are evaluated using the genomic predictor.

9. Supplementary Figure 9: Phen-Gen and VAAST comparison for phenotypically heterogeneous disorders. (109 KB)

The comparison of Phen-Gen and VAAST in simulations using 44 phenotypically heterogeneous disorders and nonsynonymous mutations in HGMD is shown. In both panels the ability of both methods to narrow down the true gene search within 1, 5 and 10 genes is depicted. For Phen-Gen, the bar is split into the predictive power based on genotypic prediction and the added advantage gained from disease symptoms. VAAST only uses the genomic data and assign multiple genes the same rank at the top of the order. For a fair comparison, the true gene was assigned the worst, average and best rank among similarly ranked peers. The three components of the bar reflect the performance across these scenarios.

PDF files

1. Supplementary Text and Figures (5,267 KB)

Supplementary Figures 1–9, Supplementary Tables 1–7 and Supplementary Note

Nature Methods ISSN 1548-7091 EISSN 1548-7105

© 2014 Macmillan Publishers Limited. All Rights Reserved.
partner of AGORA, HINARI, OARE, INASP, ORCID, CrossRef and COUNTER