COLD SPRING HARBOR
Molecular Case Studies

PRECISION MEDICINE IN PRACTICE

When "*N* of 2" is not enough: integrating statistical and functional data in gene discovery

Christopher A. Cassa,[1,2] Sebastian Akle,[3] Daniel M. Jordan,[4] and Jill A. Rosenfeld[5]

[1]Brigham and Women's Hospital, Division of Genetics, Boston, Massachusetts 02115, USA; [2]Harvard Medical School, Department of Medicine, Boston, Massachusetts 02115, USA; [3]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA; [4]Department of Genetic and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA; [5]Baylor Genetics, Houston, Texas 77030, USA

Abstract The expanding use of genomic sequencing promises to improve clinical diagnostics and to drive the discovery of new disease genes. Candidate genes are increasingly being identified through recurrent cases (e.g., two or more independent cases ["*N* of 2"] in which variants are present in the same gene). These second case hits provide statistical evidence of an association, which may then be combined with functional validation or familial segregation studies to bolster the evidence that a gene is truly causal. Here, we discuss how to integrate different forms of functional evidence with human genetics case and segregation data to improve the significance of new disease–gene associations.

It is estimated that there may be thousands of Mendelian disease genes yet to be discovered (Boycott et al. 2013), which are likely to be identified through a long tail of observations in rare disease cases (Krawitz et al. 2015). There are many ways in which recurrent cases can help identify Mendelian disease genes. These cases may include single families that are studied in isolation and aggregated through the Matchmaker Exchange or collaborations (Philippakis et al. 2015), case series in specialized clinics or Centers for Mendelian Genetics, data from large clinical sequencing laboratories (Lee et al. 2014; Yang et al. 2014), or large-scale population screens (Table 1; Saleheen et al. 2015). This issue is now acutely in focus as *Molecular Case Studies* (DeBerardinis and Mardis 2015) and systems such as the Matchmaker Exchange expand potential discovery cohorts (Philippakis et al. 2015). As case volumes grow, increasing numbers of cases will have strong candidate variants for which there is insufficient evidence of disease–gene association without a separately ascertained case (Regalado 2014).

For example, in a large clinical exome-sequencing cohort, there is great potential for the identification of new disease–gene associations through recurrent observations. In a set of 500 solved clinical exome cases from Baylor College of Medicine (Yang et al. 2014), 291 of those cases involved repeated genes that were responsible for a molecular diagnosis. More than 90% of those genes related to diagnoses did not have a sufficient knowledge base to be readily available on clinical genetic testing panels (National Center for Biotechnology Information 2012), which demonstrates that gene recurrence in clinical cases is uncovering or helping to establish the evidence base for a substantial number of new disease genes.

Such recurrent observations can help to identify potential new disease–gene associations; however, alone they are not necessarily sufficient to prove disease causality.

Corresponding author: cassa@alum.mit.edu

Published by Cold Spring Harbor Laboratory Press

doi: 10.1101/mcs.a001099

Table 1. Modes of new gene discovery that use clinical sequencing data from multiple cases to identify potential gene–phenotype associations

Mode of discovery	Description
Clinical case series (*N* > 1 individuals studied with the same phenotype)	In large case series, several patients with similar phenotypes may have similar variants in the same gene, allowing the identification of possible new disease–gene associations. These associations should be statistically and functionally validated before assigning causality, including consideration of variant type (e.g., repeated rare variants vs. repeated de novo variants).
Matches identified through the Matchmaker Exchange and/or clinical collaborations	Cases identified across institutions through matchmaking services or collaborations highlight potential disease–gene associations. These too must be checked for potential false-positive associations, given the small sample size, and functionally validated before assigning causality.
Large-scale population studies	These studies identify variants in large population cohorts where phenotypic data are available. These cohorts may include populations at large medical centers or consanguineous families that are enriched in identity by descent. Consanguineous population studies are particularly enriched for rare variant types (e.g., nonsense or canonical splicing variants), which may help generate new phenotypic associations at higher rates.

Recurrent case observations (e.g., matches among two or more cases of the same gene and phenotype) provide variable levels of evidence depending on the gene in question, the variant's functional consequence, and whether the variant is de novo or segregating.

WEAK STATISTICAL EVIDENCE FROM GENE MATCHES IN CLINICAL EXOMES

When matches in the same gene are identified in separately ascertained cases, there are three possibilities: A potential novel disease gene has been identified, a phenotypic expansion has been found in a previously identified gene, or the new cases have produced a false-positive disease–gene association. The statistical evidence for disease–gene association that may be gleaned from two matching cases depends on a number of factors based on the gene and phenotype(s) in question.

Evidence from a Matching Gene

How informative is a gene match in two unrelated individuals with the same phenotype? In large genes (e.g., *TTN*), it is not unusual to find matching variants in two cases by chance alone. The likelihood that two individuals each carry variants in that gene may be directly estimated. We have developed a tool (RD-Match; http://genetics.bwh.harvard.edu/rdmatch/) that assesses the likelihood that two individuals would carry variants in a specified gene that are unrelated to the phenotype by chance alone using specific variant, gene, and case parameters (Akle et al. 2015). We estimate this probability using data from the Exome Aggregation Consortium, a set of more than 60,000 exomes from individuals without severe disease (Lek et al. 2016). This tool is freely available and open-source, and the use of such calculations should be encouraged whenever evidence for causality is provided from recurrent case data (Box 1).

▶ **Box 1. Application note**

When using RD-Match, we suggest starting with a conservatively large *N* if there is no better way of independently estimating this parameter. If the patient population is broadly representative of cases that might be available for matching, then the total number of individuals in a cohort (e.g., Matchmaker Exchange) could be used to estimate the cohort size for each phenotype. We have updated the online tool to provide estimates of these cohort sizes for each phenotype based on clinical exome case phenotypic frequencies.

Of the recurrent matches identified in the Baylor clinical exome-sequencing program, which genes might have sufficient statistical evidence to support a potential association from matching frequency alone? For each implicated gene, we statistically analyzed the number of matched cases, the variant's functional consequence, and the mode of inheritance, and we found that merely having a match in two patients with the same phenotype is not very strong evidence of association (Akle et al. 2015). Overall, 91% of the 291 cases from the Baylor cohort would not have statistical significance from matches identified at the gene level given the number of recurrent observations in the affected gene and the number of individuals in the disease group.

The evidence from matching cases is not uniform; it very much depends on the specific gene and other case parameters. For example, when considering only cases with an autosomal recessive mode of inheritance, in which we expect matches at the gene level to be more statistically informative, we find that 68% of these cases would reach statistical significance given the specific gene and the variant's functional consequence. Both of these analyses conservatively assume that any match within the same broad phenotypic category (e.g., neurologic disease phenotype) would be considered for analysis.

This demonstrates that the statistical evidence from identifying two separate cases with the same phenotype with variants in the same gene is not very strong. Where statistical information is insufficient, investigators must make use of detailed phenotypic data from matching cases, information from historical cases, and other resources to make assessments of causality. For this reason, matches within clinical case series or through the Matchmaker Exchange must be further investigated and supplemented with functional validation studies at the variant level.

INTEGRATING FUNCTIONAL DATA TO BUILD EVIDENCE FOR CAUSALITY

How can information from recurrent observations be supplemented with other sources of available evidence, such as functional validation data, to demonstrate causality? The community continues to establish standards for the clinical interpretation of variants (Richards et al. 2015), which often must rely on integrating multiple, disparate lines of evidence to ascribe causality and to prevent false-positive associations from entering into the published literature (Cassa et al. 2013).

There are many forms of functional evidence that can be useful in supporting a disease–gene association, but integrating these diverse types of evidence adds statistical complexity. Furthermore, some functional validation assays may not be suitable for specific genes or phenotypes. So, instead, we evaluate a small set of functional data sources for which we would expect a broad phenotypic impact in order to determine which of these sources might be used to supplement Mendelian gene discovery efforts. We specifically focus on high-throughput, neutrally ascertained model data, including two lines of evidence from in vitro

assays of cell essentiality (Wang et al. 2015) and in vivo data from embryonic lethal mouse knockout data (International Mouse Phenotyping Consortium; http://www.mousephenotype.org/data/embryo), as measures of gene essentiality that we expect to have broad phenotypic impact.

To measure the utility of these functional data from clinical exome sequences for the purpose of gene discovery, we annotated the recurrent genes that appeared in more than one case in the aforementioned Baylor cohort that were responsible for a molecular diagnosis ($N = 192$ genes). Each gene in this set was annotated to indicate whether it was found to be essential for cellular or embryonic development, and the genes were separated by clinical mode of inheritance. We then compared the number of genes for each functional assay that were found to be essential for development with a similarly sized set of genes without any clinical annotations, again separated by mode of inheritance from the clinical diagnosis (Table 2). We found that certain assays, including systematic clustered regularly interspersed short palindromic repeats (CRISPR)-based KBM7 human tumor cell line inactivation and yeast gene traps, are statistically predictive in the identification of autosomal dominant disease genes from clinical exome cases, whereas lethal mouse knockouts are predictive of autosomal recessive disease genes.

Although these functional annotations in aggregate are informative for causality assessment in novel disease gene discovery, they do not guarantee that a given gene with these

Table 2. Evidence for causality from aggregate functional validation assay data in a set of recurrent genes identified in a clinical exome-sequencing program

Number of genes with importance in functional model	Essential in KBM7 human cell assay	Essential in gene trap assay	Lethal in IMPC mouse knockout
Autosomal dominant (AD) disorders (N = 118)			
Recurrent genes associated with AD disorders from clinical exome-sequencing case data found to be essential for cellular or embryonic development	27	15	4
Expected number of genes in a similarly sized set of unannotated genes found to be essential for cellular or embryonic development	12.35	7.83	2.58
χ^2 *P*-value	**1.62×10^{-5}**	**5.82×10^{-3}**	0.303
Autosomal recessive (AR) disorders (N = 74)			
Recurrent genes associated with AR disorders from clinical exome sequencing case data found to be essential for cellular or embryonic development	11	7	6
Expected number of genes in a similarly sized set of unannotated genes found to be essential for cellular or embryonic development	7.75	4.91	1.62
χ^2 *P*-value	0.172	0.271	**3.01×10^{-4}**

Statistically significant results are in bold. The unannotated gene set included any gene without a ClinVar or Human Gene Mutation Database (HGMD) annotation ($N = 10{,}719$) and was adjusted in size for each gene group. It represents a candidate set of novel genes that might be associated with disease in the future. If a gene is required for cell essentiality, it is significantly more likely to be associated with new autosomal dominant disease genes than a gene with no disease annotations. Conversely, if a gene is required for mouse embryonic development, it is significantly more likely to be lethal in a mouse knockout than any unannotated gene.
IMPC, International Mouse Phenotyping Consortium.

annotations will be causal. But clinical laboratories have long integrated this type of uncertain information into variant causality assessments. For example, the use of parametric linkage analysis (e.g., logarithm of odds [LOD] scores) in individual pedigrees (Morton 1955) has long been used to provide evidence for association at the human genetics level, and χ^2 *P*-values have been used to describe the statistical certainty of association at the population level. Similarly, it is time to begin using the statistical certainty that recurrent case matches and functional validation assays provide at the gene level in novel disease gene assessment.

As recurrent gene observations begin to represent a larger source of gene discovery, this complementary form of evidence also requires community standards for statistical certainty. Likewise, evidence from in vivo and in vitro assays must appropriately weigh the potential for false-positive associations by considering the worldwide frequency of phenotypic observations and model outcomes. One path forward would be to develop a composite probability of association given the independent probabilities derived from orthogonal lines of evidence. This poses challenges, though, as it is difficult to be certain that two lines of evidence are truly independent of one other. Using a Bayesian methodology, one can integrate these separately ascertained lines of evidence into a posterior probability. This can be done by assuming a uniform prior probability that any gene in the genome is causal for an observed phenotype and then considering the likelihood of the observed match, used together with other lines of evidence such as those presented above.

Appropriately weighing each line of evidence may be challenging when there are biases in the ascertainment and generation of data sets (e.g., mouse knockouts that have largely been targeted for orthologs of suspected disease genes). While efforts are ongoing, resources such as ClinGen are cataloging and analyzing annotation data that can be used to generate broader statistical evidence for different sources of data (Rehm et al. 2015).

ADDRESSING STATISTICAL UNCERTAINTY FROM FAMILIAL DATA AND PHENOTYPES

Evidence from Segregation Studies within Pedigrees

When there is additional evidence from a pedigree, the different numbers of segregation events can be integrated into novel gene assessment using tools like SORVA (Significance of Rare Variants; https://sorva.genome.ucla.edu). SORVA allows users to measure the significance of matches across cases and to integrate evidence from segregation within families. Various case parameters may be used to measure the significance from within single families (e.g., the coefficient of relationship for individuals who share the variant, how many observations were made for individuals of the same level of relatedness, and whether the variants are de novo), and similar parameters are used for findings that span different families, as in RD-Match.

Prevalence of Phenotype in Matching

Another consideration that may influence the statistical evidence for disease–gene association is the total number of patients in a cohort with the same phenotype. For example, if a new patient with congenital hearing loss is added to the Matchmaker Exchange, the probability of a false-positive match depends on the total number of patients in the database who also have that phenotype (*N*). In practice, the number of individuals in a cohort with the same particular phenotype is difficult to estimate given that in many instances either the information is not structured using a standardized phenotype ontology such as the Human Phenotype Ontology (HPO) or is unavailable. The degree of relatedness of any given set

of phenotypes may also be considered, and phenotypes can be used in aggregate to identify syndromes, which can then be used as the basis for a "phenotypic match."

Multiple Candidate Variants or Phenotypes in Each Case

Another factor that can contribute to false-positive associations through recurrent observations is the number of candidate variants or associated phenotypes in each case. Although the protocol used to interpret clinical sequencing data can dramatically alter the number of candidate variants (Brownstein et al. 2014), many services review variants in disease-associated genes and inspect rare coding variation that is predicted to be deleterious (Adzhubei et al. 2013). Depending on the population and consanguinity in a case, it is reasonable to expect multiple promising candidate variants. For example, in a case where the putative mode of inheritance is recessive, we may find several rare homozygous alternative coding variants per genome (MacArthur et al. 2012; Francioli et al. 2015).

When there are multiple variants, each candidate variant is less likely to be causal, which must be taken into account when analyzing matches in case data. Although some fraction of cases may have multiple segregating variants that are linked to multiple disorders (Yang et al. 2014), the majority of solved Mendelian cases are linked with a single locus (particularly if the phenotype is not diffuse). RD-Match assumes that all variants are reported and thus gives a conservative multiple test correction for all the genes in the genome.

Similarly, when testing multiple phenotypes or syndromes (e.g., a single clinical exome case may have two separately segregating Mendelian disorders or several HPO terms; see Box 2), the *P*-value must be adjusted by the number of associated phenotypes. Clusters of

▶ Box 2. Many adult clinical exome sequencing patients have several HPO terms, many of which are common

In a sample of 85 cases from an adult clinical exome service, we find that, on average, each patient is associated with 11 HPO terms (Fig. 1, left). The likelihood that two randomly selected patients will have a matching HPO term depends on the frequency of each category. In our sample, roughly one-third of the HPO categories are unique (i.e., belonging to a single individual), whereas the other categories are very common. The four most common categories are present in more than 25 of the 85 cases (Fig. 1, right).

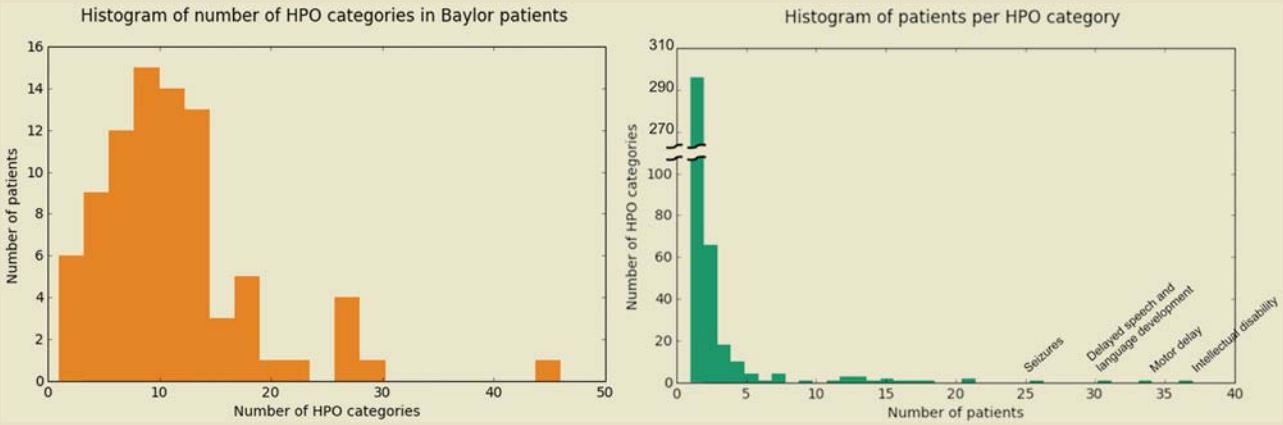

Figure 1. (*Left*) Histogram showing the number of Human Phenotype Ontology (HPO) categories associated with each patient in the Baylor adult exome-sequencing cohort (Posey et al. 2015). (*Right*) Histogram showing the number of patients associated with each HPO category in the Baylor adult exome-sequencing cohort.

HPO terms that form distinct syndromes should not be treated as separate phenotypic matches, as doing so might reduce the statistical certainty of a match (Greene et al. 2016).

Although the vast majority of clinical exome cases are indicated for neurologic disorders, there are now significant numbers of cases from adult populations involving a broader set of disorders and related genes (Posey et al. 2015; Retterer et al. 2015). These populations may be enriched for recurrent cases, as they are less likely to be solved during initial sequence analysis (with a 17.5% diagnosis rate in adult exome cases vs. 25%–30% in pediatric cases); however, these cohorts are only modestly sized.

This can be addressed conservatively using a standard Bonferroni correction (i.e., multiplying the *P*-value by the number of HPO categories or phenotypic clusters present in the patient showing the match). This correction is unlikely to erode the data because most of the HPO categories are shared by only a few patients, resulting in small values of *N*.

CONCLUSION

The broader availability of clinical sequencing data allows for the identification of novel disease genes using recurrent observations. Clinical sequencing cases from this journal and from matchmaking services provide rich data for recurrent gene case matching, and they may also provide additional phenotypic or human genetics data (e.g., segregation analysis). These repeated observations can facilitate the identification of new genes in rare disorders, but investigators must carefully consider the appropriate statistical evidence derived from repeated observations along with other sources of evidence of association.

Competing Interest Statement

The authors have declared no competing interest.

Received March 24, 2016; accepted in revised form January 25, 2017.

ADDITIONAL INFORMATION

Acknowledgments

This research was supported by the National Institutes of Health (NIH)/National Human Genome Research Institute (NHGRI) grant R00-HG007229. J.A.R. receives salary support from Baylor Genetics, a clinical genetic testing laboratory.

REFERENCES

Adzhubei I, Jordan DM, Sunyaev SR. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7:** Unit7 20.

Akle S, Chun S, Jordan DM, Cassa CA. 2015. Mitigating false-positive associations in rare disease gene discovery. *Hum Mutat* **36:** 998–1003.

Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet* **14:** 681–691.

Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, Dechene ET, Towne MC, Savage SK, Price EN, et al. 2014. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. *Genome Biol* **15:** R53.

Cassa CA, Tong MY, Jordan DM. 2013. Large numbers of genetic variants considered to be pathogenic are common in asymptomatic individuals. *Hum Mutat* **34:** 1216–1220.

DeBerardinis RJ, Mardis ER. 2015. From " *N* of 1" to *N* of more. *Mol Case Stud* **1:** a000521.

Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I; Genome of the Netherlands Consortium, , van Duijn CM, Swertz M, Wijmenga C, et al. 2015. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* **47:** 822–826.

Greene D, NIHR BioResource, Richardson S, Turro E. 2016. Phenotype similarity regression for identifying the genetic determinants of rare diseases. *Am J Hum Genet* **98:** 490–499.

Krawitz P, Buske O, Zhu N, Brudno M, Robinson PN. 2015. The genomic birthday paradox: how much is enough? *Hum Mutat* **36:** 989–997.

Lee H, Deignan JL, Dorrani N, Strom SP, Kantarci S, Quintero-Rivera F, Das K, Toy T, Harry B, Yourshaw M, et al. 2014. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA* **312:** 1880–1887.

Lek M, Karczewski K, Minikel E, Samocha K, Banks E, Fennell T, O'Donnell-Luria A, Ware J, Hill A, Cummings B, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536:** 285–291.

MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, Jostins L, Habegger L, Pickrell JK, Montgomery SB, et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335:** 823–828.

Morton NE. 1955. Sequential tests for the detection of linkage. *Am J Hum Genet* **7:** 277–318.

National Center for Biotechnology Information. 2012. Genetests.org. http://www.ncbi.nlm.nih.gov/sites/GeneTests/.

Philippakis AA, Azzariti DR, Beltran S, Brookes AJ, Brownstein CA, Brudno M, Brunner HG, Buske OJ, Carey K, Doll C, et al. 2015. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum Mutat* **36:** 915–921.

Posey JE, Rosenfeld JA, James RA, Bainbridge M, Niu Z, Wang X, Dhar S, Wiszniewski W, Akdemir ZHC, Gambin T, et al. 2015. Molecular diagnostic experience of whole-exome sequencing in adult patients. *Genet Med* **18:** 678–685.

Regalado A. 2014. EmTech: Illumina Says 228,000 Human genomes will be sequenced this year. *Technol Rev.* http://www.technologyreview.com/news/531091/emtech-illumina-says-228000-human-genomes-will-be-sequenced-this-year/ (Accessed January 11, 2014).

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, et al. 2015. ClinGen—The Clinical Genome Resource. *N Engl J Med* **372:** 2235–2242.

Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, Vertino-Bell A, Smaoui N, Neidich J, Monaghan KG, et al. 2015. Clinical application of whole-exome sequencing across clinical indications. *Genet Med* **18:** 696–704.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* **17:** 405–423.

Saleheen D, Natarajan P, Zhao W, Rasheed A, Khetarpal S, Won H-H, Karczewski KJ, O'Donnell-Luria AH, Samocha KE, Gupta N, et al. 2015. *Human knockouts in a cohort with a high rate of consanguinity.* http://biorxiv.org/lookup/doi/10.1101/031518.

Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, Wei JJ, Lander ES, Sabatini DM. 2015. Identification and characterization of essential genes in the human genome. *Science* **350:** 1096–1101.

Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, Ward P, Braxton A, Wang M, Buhay C, et al. 2014. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA* **312:** 1870–1879.