Genome Medicine

COMMENTARY

Likelihood ratios for genome medicine

Alexander A Morgan[1,2], Rong Chen[1] and Atul J Butte*[1,3]

Abstract

Patients are beginning to present to healthcare providers with the results of high-throughput individualized genotyping, and interpreting these results in the context of the explosive growth of literature linking individual variants with disease may seem daunting. However, we suggest that results of a personal genomic analysis may be viewed as a panel of many tests for multiple diseases. By using well-established methods of evidence based medicine, these very many parallel tests may be combined using likelihood ratios to report a post-test probability of disease for use in patient assessment.

Introduction

Although there has been continuing discussion and debate over the ethical implications and clinical utility of a large-scale genotyping for an individual patient [1-3], the issue is somewhat moot. Patients are now being genotyped using either (i) measurement platforms run by several different direct-to-consumer companies that sequence nearly a million single nucleotide polymorphisms (SNPs) [4], or (ii) whole genome sequencing, which is beginning to be offered to selected individuals [5-8]. Patients are beginning to present to their healthcare provider before or during an evaluation, including an extensive genotyping scan [9]. It may appear overwhelming and a nearly impossible task to take the complexity of genetic variation and interpret it in the context of the enormous amount of literature on human genetics [10], some of which seems mercurial and contradictory. However daunting, it is incumbent upon a healthcare provider to try to help patients make informed decisions in light of the information available, and to not ignore this genetic information.

Discussion

Although DNA variants unique to an individual, or at least extremely rare in the general population, may have major impact on personal phenotypes and may explain much of the 'missing heritability' [11,12] of common variants, we currently have very little power to interpret the impact or predictive power of these rare variants. Additionally, individual sequence data, which are able to probe for more rare variants, are not yet as common as parallel genotyping assays, which primarily probe common variants. There is a large body of published research associating common variants with disease [13]. Admittedly, those relationships are through association, which does not necessarily indicate a direct functional relationship for the outcome or phenotype being studied. However, having a direct model of mechanism has never been a requirement for the value of a medical test. Many features used in physical examinations or laboratory tests have an indirect relationship with the clinical phenotype (typically disease state) being measured. For instance, the well-known relationship between clubbing and impaired lung function is through association, not mechanism, but that does not reduce the predictive value. Association of a genotype with clinical phenotype has value as a predictive tool independent of mechanism.

We envision that patients may present to a healthcare provider with a large panel of genotyping studies or a whole genome sequence (both of these are referred to here as DNA analysis) generally for three reasons. The first might be to seek reproductive counseling, and there is already extensive existing methodology in this area, including professional certification for counselors in the USA and Canada by the American Board of Genetic Counseling. The second might be for an individual with clinical complaints, and the genotyping analysis might have been performed with the hope of providing assistance in the refinement of a diagnosis or an improved, personalized treatment plan. The third might be for a healthy patient looking for suggestions into lifestyle modifications or information on long-term prognosis and early identification of potential problems; this situation is not unique to a genetic screen and is typically the goal with a well physical. Here, we are addressing patients presenting for the latter two reasons.

By viewing a DNA analysis as a series of multiple laboratory tests that each have predictive power for different phenotypes, it becomes clear how these fit into the well-established methods of evidence based medicine [14-16]. The measurement of each DNA variant turns

*Correspondence: abutte@stanford.edu
[1]Department of Pediatrics and the Department of Medicine, Stanford University School of Medicine, 251 Campus Drive, MS-5415, Stanford, CA 94305-5479, USA
Full list of author information is available at the end of the article

Table 1. Example calculations of post-test probabilities

Type of disease and associated variants probability of disease (%)	Pre-test probability of disease (%)	Likelihood ratio	Post-test
Common disease, weakly associated variant	15.0	1.1	16.256
Common disease, several weakly associated variants	15.0	1.1 × 1.1 × 1.1 × 1.1 = 1.46	20.486
Rare disease, weakly associated variant	0.01	1.1	0.011
Rare disease, strongly associated variant	0.01	5	0.050
Rare disease, several weakly associated variants	0.01	1.1 × 1.1 × 1.1 × 1.1 = 1.46	0.015
Rare disease, several moderately associated variants	0.01	2 × 2 × 2 × 2 = 16	0.160

Post-test probabilities may be calculated for common or rare diseases with weakly and strongly associated variants using example values for likelihood ratios and pre-test probabilities. The definition of strongly versus weakly associated is in the context of genetic associations, where likelihood ratios from large-scale studies rarely reach higher than 3. Many clinical laboratory tests have likelihood ratios of 10 or more.

into an individual test. That test provides a likelihood ratio for phenotype (we will focus primarily on current or future disease state as the phenotype of interest) based on the result of that test.

Armed with a reasonable assessment of pre-test odds, the framework of evidence based medicine, which has been taught in medical schools and in residency programs for decades, simply multiplies the likelihood ratios of disease state, given the results of the tests, to produce a post-test odds of disease. The fact that the results of genotype analysis of any individual variant are extremely precise should not be confused with the fact that individual tests for disease need not be exceptionally accurate to have value. The DNA analysis is just a very large panel of such tests.

Calculation of likelihood ratios, and pre- and post-test probabilities

A likelihood ratio is the ratio of the probability of a positive test, in this case a particular genotype, in a diseased person to that in a non-diseased person:

Likelihood ratio = Probability of genotype in diseased person/Probability of genotype in non-diseased person = LR_i

Likelihood ratios multiplied by the pre-test odds of disease give the post-test odds of disease (Table 1), and these likelihood ratios may be chained together (Figure 1):

Pre-test odds = Probability of disease/1 - Probability of disease

Pre-test odds × LR_1 × LR_2 ×...× LR_n = Post-test odds

Post-test probability = Post-test odds/Post-test odds + 1

The assumption of independence made here is that each test is independent of one another. Note that assuming independence of tests is actually a different assumption than assuming that each variant contributes independently to risk. The independence of risk contributions may be an accurate model if each genetic variant measured does causally contribute independently to risk, but there is only very little indication [17] that this is broadly the case for most genetic associations, and there are difficulties with many models that do assume independent risk contributions [18]. If we view each measured variant as an independent test probing disease state, this is arguably closer to our understanding of their use as markers associated with disease instead of actual causal variants. In this case, assuming independence as tests of disease is a more appropriate approximation.

A key advantage of considering genotyping assays by likelihood ratios is that this methodology directly takes the prior probabilities into account. Genetic features suggesting relatively dramatic increase in associated risk may still only suggest modest post-test probabilities of rare diseases. Variants that do not contribute dramatically to risk will leave common diseases as being common (that is, having a high post-test probability) and should not substantially change most current guidelines for preventative screening. In addition, the specific pre-test probabilities are also adjustable in the context of a patient with other clinical findings. The calculation of post-test probabilities in this manner will allow the results of genetic screens to more easily fit into discussions of the numbers needed to treat, numbers needed to harm, and many issues in cost-benefit analysis.

Considering genotyping assays by likelihood ratios and post-test probabilities [16] also addresses previously suggested 'incidentalome' issues [19], where incidental findings, even many of them, that weakly suggest increased likelihood of rare diseases will be largely irrelevant in a patient free from clinical complaints and with correspondingly low post-test probabilities of these diseases. Physicians have been taught to consider threshold post-test probabilities for continuing testing or initiating therapy, with thresholds set based on careful consideration of the risks and benefits of continued testing or initiation of therapy. If physicians are presented with panels of post-test probabilities, instead of being presented with genotypes or odds ratios, we suggest they

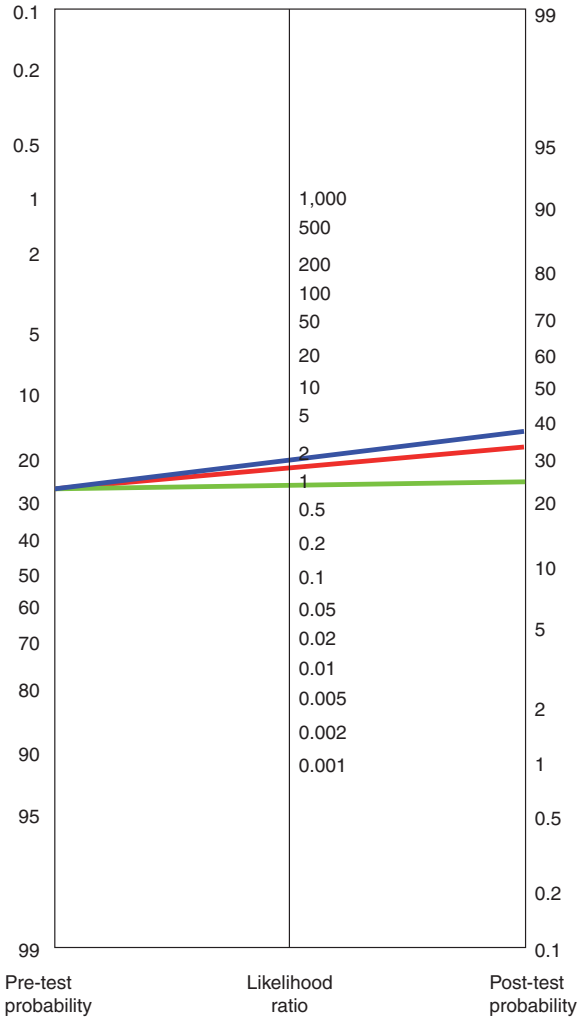

Figure 1. Nomogram for likelihood ratios. The pre-test and post-test probabilities and likelihood ratios of any diagnostic test, including a genetic test, can be visualized using a nomogram familiar to most physicians and medical students. The nomogram shown is derived from the Fagan nomogram [14], and modified from one generated using a web-based tool [28]. The left side of the figure indicates a hypothetical pre-test probability of disease of 27%. Three lines represent the three possible genotypes, from top to bottom: homozygous risk alleles with a likelihood ratio of 1.61, heterozygous alleles with a likelihood ratio of 1.26, and homozygous protective alleles with a likelihood ratio of 0.83. The right side of the figure indicates three possible post-test probabilities resulting from the three genotypes. Multiple such tests can be 'chained' together serially, if they describe independent risks and cover the same pre-test assumptions.

have the training to make the determination of future courses based on post-test probabilities.

Challenges

Unfortunately, much of the information necessary to support this method of using likelihood ratios is not being published in the primary publications associating genotypes with disease. Although many studies have been performed examining the association between common variants and disease, many of these reports still do not provide enough information to calculate a likelihood ratio from a specific genotype, do not characterize the sample population and the prior probability of disease in this population, or do not make clear what other variants were measured to help adjust for multiple hypothesis testing and other biases.

Traditionally, the published literature on genetic associations has focused on suggesting interesting variants with possible mechanistic involvement in the disease of study. Hence, authors may only report an odds ratio as a measure of effect size, and a *P* value to show that the variant is significantly associated with the disease. Many such studies do not even report the risk genotype at the site of the SNP; this is a particular problem because the relationship of the common allele in the population under study to a reference genome is unknown, and the reference genome may actually contain the risk-associated allele. For example, a study that reports that having a variation at an identified location in the genome doubles the risk for a disease, without reporting which variant (A, C, T or G) is actually associated with the increase of risk, is failing to report essential information.

We recently curated 2,174 articles reporting primary data on gene-disease associations of variants in the National Center for Biotechnology Information (NCBI) SNP database (dbSNP) [20]. Of these publications, only 46% contained information on actual genotype-associated risk, enabling the calculation of a likelihood ratio yielding a total of 2,092 disease-variant associations. Although any particular genetic association study may not be intended for use in informing a clinical diagnostic test or interpretation, information on the actual proportion/frequency of subjects with each associated genotypic variant in the relevant phenotype categories (such as with and without disease) should be made available for use in further studies and meta-analyses. This information aids in attempts at replication of results and in calculating overall estimates of the power of a particular genotype to predict disease state. The prostate cancer study by Duggan and colleagues [21] contains a particularly illuminating example of this kind of detailed reporting in Table 2 of the article. At a bare minimum, the actual risk allele should be reported; this is something not explicitly required by current guidelines [22].

One reason that additional data specifying the exact proportion of individuals of each genotype in each disease category is not given in publications is possibly due to the concern in being able to identify a patient's disease class if detailed data from the study are made available [3]. However, such re-identification of disease state does still require that one has the patient's genotype.

Having an individual's genotype at thousands of phenotype-associated loci by itself enables you to know a considerable amount about that individual, independent of their involvement in any association studies. As knowledge of human genetics increases, possession of an individual's genetic sequence will continue to be the level at which invasion of individual rights and privacy must be protected. Thus, the potential re-identification of a patient into a study group should not dissuade researchers from reporting detailed information in genome-wide association studies.

Many genetic association studies still do not report information about the characteristics of the population studied, such as age, gender and ethnicity. This information would substantially increase the clinical relevance of the study, and it is a key part of using literature in evidence based medicine [23]. Analyses showing association of a single biomarker with disease typically report very detailed characteristics of the populations studied; this is radically different from typical genetic association studies, which often report almost nothing about the subjects.

Another challenge in applying likelihood ratios from genetic tests is that there are very few sources available that provide enough information to calculate the pre-test probabilities of disease states, particularly in the same populations under genetic study or populations resembling many presenting patients. A concerted effort to calculate prevalence and incidence statistics, and report them both in genetic association studies and as general epidemiological features, will improve the quality of the clinical interpretation of genotyping dramatically.

Finally, there are many established techniques for addressing many of the biases in reporting results of many statistical tests, and the 'winner's curse' is a well-known phenomenon [24,25]. Genetic studies that combine a discovery for a significant association with disease with an estimate of associated risk are strongly biased to overestimate the level of risk [26]. However, if it is clear which associations are measured and what the overall results are, we can attempt to address these biases and apply the appropriate correction to the estimated effect size, in this case predicted risk with a confidence estimate [27].

Conclusions

In summary, we suggest that the methods for using a personal genotype to improve clinical evaluation already exist. For many diseases, actual genotypes and their associated risks are currently being collected in high volumes, and as more of these data are presented in publications, our ability to assess a patient through genotype will be greatly enhanced. If we have reasonable estimates of the pre-test probability of disease for a patient, by using careful methods of meta-analysis to combine the results of studies that report genotype level risk to compute good estimates of likelihood ratios, we can provide post-test probabilities that a physician can use in assessment and a patient could use for potential lifestyle modification.

Abbreviation

SNP, single nucleotide polymorphism.

Competing interests

AJB receives or has received consulting fees from Johnson & Johnson, Genstruct, Lilly and Tercica, and has received lecture fees from Siemens and Lilly, and equity ownership/stock from Genstruct and NuMedii.

Authors' contributions

All the authors have contributed to the conceptualization and preparation of this manuscript.

Acknowledgements

This work was supported by Lucile Packard Foundation for Children's Health, the Hewlett Packard Foundation, National Institute of General Medical Sciences (R01 GM079719), US National Library of Medicine (R01 LM009719 and T15 LM007033), and Howard Hughes Medical Institute. We thank Alex Skrenchuk and Boris Oskotsky from Stanford University for computer support.

Author details

[1]Department of Pediatrics and the Department of Medicine, Stanford University School of Medicine, 251 Campus Drive, MS-5415, Stanford, CA 94305-5479, USA. [2]Biomedical Informatics Training Program, Stanford University School of Medicine, 251 Campus Drive, Stanford, CA 94305, USA. [3]Lucile Packard Children's Hospital, 725 Welch Road, Palo Alto, CA 94304, USA.

Published: 17 May 2010

References

1. Heeney C, Hawkins N, de Vries J, Boddington P, Kaye J: **Assessing the privacy risks of data sharing in genomics.** *Public Health Genomics* 2010, in press.
2. Kaye J, Boddington P, de Vries J, Hawkins N, Melham K: **Ethical implications of the use of whole genome methods in medical research.** *Eur J Hum Genet* 2010, **18:**398-403.
3. Lumley T, Rice K: **Potential for revealing individual-level information in genome-wide association studies.** *JAMA* 2010, **303:**659-660.
4. Ng PC, Murray SS, Levy S, Venter JC: **An agenda for personalized medicine.** *Nature* 2009, **461:**724-726.
5. Kim J, Ju Y, Park H, Kim S, Lee S, Yi J, Mudge J, Miller N, Hong D, Bell C: **A highly annotated whole-genome sequence of a Korean individual.** *Nature* 2009, **460:**1011-1015.
6. Levy S, Sutton G, Ng P, Feuk L, Halpern A, Walenz B, Axelrod N, Huang J, Kirkness E, Denisov G: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5:**e254.
7. Pushkarev D, Neff N, Quake S: **Single-molecule sequencing of an individual human genome.** *Nat Biotechnol* 2009, **27:**847-850.
8. Wheeler D, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen Y, Makhijani V, Roth G: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452:**872-876.
9. Lupski J, Reid J, Gonzaga-Jauregui C, Rio Deiros D, Chen D, Nazareth L, Bainbridge M, Dinh H, Jing C, Wheeler D: **Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy.** *N Engl J Med* 2010, **362:**1181-1191.
10. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury M: **A navigator for human genome epidemiology.** *Nat Genet* 2008, **40:**124-125.
11. Goldstein DB: **Common genetic variation and human traits.** *N Engl J Med* 2009, **360:**1696-1698.
12. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.
13. Frazer K, Murray S, Schork N, Topol E: **Human genetic variation and its**

contribution to complex traits. *Nat Rev Genet* 2009, **10:**241-251.
14. Fagan T: **Nomogram for Bayes theorem.** *N Engl J Med* 1975, **293:**257.
15. Kassirer J, Kopelman R: *Learning Clinical Reasoning.* Baltimore: Williams & Wilkins; 1991.
16. Stern S, Cifu A, Altkorn D: *Symptom to Diagnosis: An Evidence-Based Guide.* 2nd edn. San Francisco: Lange Medical; 2010.
17. Orozco G, Hinks A, Eyre S, Ke X, Gibbons L, Bowes J, Flynn E, Martin P: **Combined effects of three independent SNPs greatly increase the risk estimate for RA at 6q23.** *Hum Mol Genet* 2009, **18:**2693.
18. Wray N, Goddard M, Larizza L, Roversi G, Volpi L, Boles R, Lovett-Barr M, Preston A, Li B, Adams K: **Multi-locus models of genetic risk of disease.** *Genome Med*, **2:**10.
19. Kohane I, Masys D, Altman R: **The incidentalome: a threat to genomic medicine.** *JAMA* 2006, **296:**212.
20. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, Pushkarev D, Neff NF, Hudgins L, Gong L, Hodges LM, Berlin DS, Thorn CF, Sangkuhl K, Hebert JM, Woon M, Sagreiya H, Whaley R, Knowles JW, Chou MF, Thakuria JV, Rosenbaum AM, Zaranek AW, Church GM, Greely HT, Quake SR, *et al.*: **Clinical assessment incorporating a personal genome.** *Lancet* 2010, **375:**1525-1535.
21. Duggan D, Zheng S, Knowlton M, Benitez D, Dimitrov L, Wiklund F, Robbins C, Isaacs S, Cheng Y, Li G: **Two genome-wide association studies of aggressive prostate cancer implicate putative prostate tumor suppressor gene DAB2IP.** *J Natl Cancer Inst* 2007, **99:**1836-1844.
22. Little J, Higgins J, Ioannidis J, Moher D, Gagnon F, Von Elm E, Khoury M, Cohen B, Davey-Smith G, Grimshaw J: **Strengthening the reporting of genetic association studies (STREGA): an extension of the STROBE statement.** *Hum Genet* 2009, **125:**131-151.
23. Richardson W, Wilson M, Guyatt G, Cook D, Nishikawa J: **Users' guides to the medical literature: XV. How to use an article about disease probability for differential diagnosis.** *JAMA* 1999, **281:**1214.
24. Kraft P: **Curses--winner's and otherwise--in genetic epidemiology.** *Epidemiology* 2008, **19:**649-651; discussion 657-648.
25. Zollner S, Pritchard JK: **Overcoming the winner's curse: estimating penetrance parameters from case-control data.** *Am J Hum Genet* 2007, **80:**605-615.
26. Ioannidis JP: **Why most discovered true associations are inflated.** *Epidemiology* 2008, **19:**640-648.
27. Zhong H, Prentice RL: **Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies.** *Biostatistics* 2008, **9:**621-634.
28. **Diagnostic Test Calculator** [http://araw.mede.uic.edu/cgi-bin/testcalc.pl]

doi:10.1186/gm151
Cite this article as: Morgan AA, *et al.*: Likelihood ratios for genome medicine. *Genome Medicine* 2010, **2:**30.