# Personal Genomic Measurements: The Opportunity for Information Integration

**RB Altman**[1]

[1]Departments of Bioengineering, Genetics, and Medicine, Stanford University, Stanford, California, USA.

## Abstract

High-throughput genomic measurements initially emerged for research purposes but are now entering the clinic. The challenge for clinicians is to integrate imperfect genomic measurements with other information sources so as to estimate as closely as possible the probabilities of clinical events (diagnoses, treatment responses, prognoses). Population-based data provide *a priori* probabilities that can be combined with individual measurements to compute *a posteriori* estimates using Bayes' rule. Thus, the integration of population science with individual genomic measurements will enable the practice of personalized medicine.

## High-throughput genomics technologies are moving to the clinic

Genomics technologies have enabled inexpensive, accurate, and high-throughput measurement of the genome and transcriptome, and other technologies for measuring the metabolome, proteome, microbiome, and immune system are emerging. In the research arena, we have seen provocative analyses of high-throughput data suggesting potential clinical applicability. For example, a clinical annotation of a complete human genome sequence looked for risk of common disease, rare disease, and predicted drug response.[1] Investigators assessed the genetic risk for hundreds of common diseases and rare diseases, and generated clinically relevant advice for nearly 100 drugs. A similar analysis of a family quartet (mom, dad, daughter, son) showed increased accuracy with a better reference genome and an ability to improve error detection with multiple related genomes.[2] Impressive studies have uncovered the genetics of rare disorders using genome sequencing,[3–5] and applications in cancer therapy are accelerating.[6] There has been progress in understanding the bacteria colonizing the gut and constituting the microbiome— undoubtedly important for drug response as well as immune response. In terms of integrating several modalities, a single individual was followed over 14 months with an analysis of the genome, transcriptome, proteome (including autoantibody panels), and metabolome.[7] In the context of a viral infection, the investigators observed the putative onset of type 2 diabetes mellitus, integrating millions of measurements and suggesting that we may one day watch the evolution of molecular responses and rapidly intervene to improve health. Thus, it appears likely that personalized medicine will ultimately involve several "omic" measurement technologies. I argue that it must also involve a broader set of population-based data sources to be successfully integrated into the clinic.

Correspondence: RB Altman (russ.altman@stanford.edu).

There is already an emerging industry for health-related genomics. This industry currently focuses on DNA genotyping and sequencing, as well as transcriptomic measurements, primarily in the context of cancer.[8] Many hospitals and companies are pursuing College of American Pathologist accreditation and Clinical Laboratory Improvement Amendments–approved DNA genotyping and sequencing facilities and adopting standardized practices to demonstrate reliability, reproducibility, and quality control. The US Food and Drug Administration (FDA) has shown interest in this topic and has held several workshops to follow progress in measurement quality and reliability, focusing primarily on DNA sequencing technologies.[9] Indeed, there are interesting questions about whether a comprehensive genomics measurement should be considered a diagnostic test (similar to a lab test such as a complete blood count or comprehensive metabolic panel) in which the interpretation is primarily left to the ordering physician, or whether they should be considered an expert consultation (as in a radiologist interpreting an X-ray film or a pathologist examining a histology sample) in which the primary data are interpreted for the ordering physician by another physician who presents the findings and their potential clinical significance.

In any case, it is clear that many "omics" measurements are rapidly becoming available to clinicians. The quality of the information is not perfect but is improving; there are pressures on companies to provide high-quality data. As we consider how to integrate these new data, we must consider their intrinsic time scales. For example, genome information is generally static (except for cancer, somatic mutations, and other important exceptions), and so the relevant time scale is the lifetime of the organism. The genetic variants that are seen in the genome have long-term relevance. Thus, the genome is useful for assessing disease risk, risk of inherited disease (particularly relevant in the context of reproduction), and the modulation of drug response. On the other hand, other emerging measurements are dynamic and may be reassessed periodically. Cellular transcription responses may change in minutes to hours, in response to stimuli such as infections, medications, diet, and other systemic exposures. We may have to interrogate the transcriptome frequently as patient status evolves. It is exciting to imagine the eventual integration of transcriptomic data (showing potential changes in messenger RNA levels) with proteomic data (showing the actual levels of protein) and metabolomic data (showing the products of enzymatic and signaling cascades) to completely understand the current patient state. We have only begun to assess the value of this information and the ways it could be presented to the clinician in support of decision making.

## Integrating genomic information with traditional health information

A primary challenge for genomics measurements is their integration with more traditional sources of health-care information. Perhaps the most important of these information sources is the electronic health record or the electronic medical record (EMR). The EMR contains key information that is critical for interpreting genomics measurements, including the history of diagnoses, drug and other substance exposures, social history of employment, travel and lifestyle choices, as well as the physical exam. Family history is particularly useful as a complex combination of genetic factors and environmental and lifestyle exposures. The EMR also contains the results of laboratory tests, radiological images, pathology reports—all of which can be compared, contrasted, and combined with genomics measurements. In fact, genomic information should ultimately improve the precision of diagnosis and treatment, but we will require robust and general methods for combining multiple sources of information. The National Institutes of Health's Electronic Medical Records and Genomics (eMERGE) network has been studying the issues in integrating genomic information with medical records and has published its preliminary experience in defining clean phenotypes and replicating known associations. [10] On a larger scale,

population databases looking at particular types of health information are also very valuable for genomics. For example, the FDA Adverse Events Reporting System collects almost 500,000 reports of adverse reactions each year and makes these available for research. The Framingham longitudinal study of cardiovascular health has created risk scores based on epidemiological data, as have several other large population- based efforts. Emerging sources of health data are also appearing, such as the use of search engine data to track flu outbreaks and online communities of patients. Although some might argue that the information gleaned from all these sources (genomics, EMR, population databases) is too noisy and incomplete to be useful, the theory of medical decision making provides us with an information integration solution: Bayesian reasoning.

Clinical medicine is concerned with estimating probabilities, for example, what is the probability of a disease or diagnosis? What is the probability that a particular treatment will work? Clinicians make decisions by assessing possible actions and their potential outcomes — both the probability and the desirability (utility) of these outcomes. In the end, clinicians make decisions that they believe optimize the expected outcome for the patient. (Admittedly, these calculations are often not quantitative but instead based on a qualitative assessment.) Bayes' rule for estimating probabilities is a general method for information integration. It states that the *a posteriori* probability of a clinical event (e.g., a diagnosis, whether a drug will work) is a function of both the *a priori* probability of that event as well as the information contained in any newly measured data. Box 1 presents an example of Bayes' rule in a hypothetical drug response scenario for pharmacogenomics. It illustrates how general population-based data can be combined with individual genomic data to create an updated estimate of the relevant clinical probabilities. It is meant as a simple illustration, and the value of and issues with Bayesian reasoning have been extensively documented.[4]

## The importance of population data for personalized medicine

Recently, there has been excitement about the opportunities for using "big data" in medicine. Our ability to collect, organize, and store biomedical information has increased markedly over the past several years, and data-mining methods are increasingly discovering new patterns in the data. For clinical medicine, "big data" come in roughly two flavors: (i) big data about the individual, including all of his or her high-throughput genomic measurements, and (ii) big data about populations, providing very useful statistics about the overall probability of diseases, drug responses, surgical outcomes, and environmental exposures. It seems that the individual "big data" are about personalized medicine and the population "big data" are about "one size treats all" medicine. However, Bayes' rule uses the population data as critical prior probabilities and uses the individual data to update these for an individual. In a very real sense, the population data are just as important for estimating the *a posteriori* probabilities as the genomic data, and so both are critical for the implementation of personalized medicine. Thus, we need not despair about our lack of perfect knowledge about how to interpret genomic measurements. Instead, we combine prior estimates with our best understanding of the new information to estimate the probabilities that clinicians require to make good decisions.

In the context of drugs and therapeutic response, the data submitted to the FDA, the European Medicines Agency, and other regulatory agencies provide the population expectations for drug response phenotypes. The measurement of genomic variants important for drug response provides the individual data. In some cases the variants are common, and their effect on the drug response is known and can be integrated with population expectations directly. In other cases, the variants may be rare and not previously studied; we must estimate their impact using predictive methods. For example, if a patient has stop

codons in a critical metabolizing enzyme, we may infer that he or she has substantial loss of function and update our probabilities based on this inference.

The introduction of high-throughput genomic information into the clinic provides an exciting opportunity to create clinical decision-support systems that combine population-level information with individual-level measurements so as to implement personalized medicine. We can use Bayesian reasoning to integrate prior knowledge and new sources of data to estimate key clinical probabilities. We then evaluate these in the context of the anticipated value of each outcome to make the best clinical decisions possible. If we appropriately instrument the health-care system with an infrastructure that allows us to track outcomes, then we can make each individual patient a small experiment that contributes to better estimates of these probabilities—a rapid-learning paradigm.[10] We will thus continuously improve our understanding of how to use these powerful new measurements for precise and accurate clinical decisions. In this way, the new technologies can begin to influence care soon, while we refine our understanding of the information that they contain.

## Acknowledgments

## References

1. Ashley EA, et al. Clinical assessment incorporating a personal genome. Lancet. 2010; 375:1525–1535. [PubMed: 20435227]

2. Dewey FE, et al. Phased whole-genome genetic risk in a family quartet using a major allele reference sequence. PLoS Genet. 2011; 7:e1002280. [PubMed: 21935354]

3. Worthey EA, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet. Med. 2011; 13:255–262. [PubMed: 21173700]

4. Mardis ER. Applying next-generation sequencing to pancreatic cancer treatment. Nat. Rev. Gastroenterol. Hepatol. 2012; 9:477–486. [PubMed: 22751458]

5. Chen R, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell. 2012; 148:1293–1307. [PubMed: 22424236]

6. US Food and Drug Administration. Ultra High Throughput Sequencing for Clinical Diagnostic Applications-Approaches to Assess Analytical Validity. <http://www.fda.gov/MedicalDevices/NewsEvents/WorkshopsConferences/ucm255327.htm> (23 June 2011).

7. Cronin M, et al. Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. Clin. Chem. 2007; 53:1084–1091. [PubMed: 17463177]

8. McCarty CA, et al. The eMERGE Network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Med. Genomics. 2011; 4:13. [PubMed: 21269473]

9. Robert, C. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. New York: Springer; 2007.

10. Etheredge LM. A rapid-learning health system. Health Aff. (Millwood). 2007; 26:w107–w118. [PubMed: 17259191]

## Bayes' rule example for pharmacogenomics

Scenario: we are considering the probability that a drug response will be favorable for a particular patient using both population-level information and individual genetic data (e.g., a particular SNP). We can use Bayes' rule as follows:

**P(D)** = P(drug works) = the overall probability in the general population that the drug will be effective

**P(¬D)** = P(drug doesn't work) = 1 − P(D) = the overall probability in the general population that the drug is not effective

**P(G)** = P(genetic data) = the probability of observing the genetic data in the general population

**P(G|D)** = P(genetic data | drug works) = the probability of observing the genetic markers in an individual patient, knowing that the drug works on that patient

**P(G|¬D)** = P(genetic data | drug doesn't work) = the probability of observing the genetic markers in an individual patient, knowing that the drug doesn't work on that patient

**P(D|G)** = P(drug works | genetic data) = the probability that the drug will be effective for an individual patient given that he or she has the genetic markers of interest

**Bayes' rule tells us:**

**P(D|G) = P(G|D) × P(D) / P(G)**

**with P(G) = P(G|D) × P(D) + P(G|¬D) × P(¬D)**

So, if P(drug works) = 80% = 0.8 = P(D)

(it works on most people…)

P(drug doesn't work) = 20% = 0.2 = P(¬D)

(…but doesn't work on some people)

P(genetic data | drug works) = 30% = 0.3 = P(G|D)

(30% of people for whom drug works have the SNP)

P(genetic data | drug doesn't work) = 0.05 = P(G|¬D)

(5% of people for whom drug doesn't work have the SNP)

then

P(D|G) = P(drug works | genetic data) = $0.3 \times 0.8/(0.3 \times 0.8 + 0.05 \times 0.2) = 0.96$

A patient with this SNP has a 96% chance of responding well to the medication. The genetic data do not have perfect predictive power, but they increase our estimate of the probability that the drug will work from 80% (the population expectation) to 96% (the individualized estimate).