Computational and Statistical Issues in Personalized Medicine

Computational methods can be used to find associations between our genome and our traits, and new optimizations to these computations promise to do it much faster.

By *Christoph Lippert and David Heckerman*

DOI: 10.1145/2788502

There is a genomics revolution underway. Fourteen years ago, first drafts of the human genome were constructed at the staggering cost of three billion dollars. Today, the human genome can be sequenced for less than \$1,000. If you plot cost as a function of time, you see a curve that beats Moore's Law (see Figure 1) [1]. The time it takes to sequence a genome has seen a similar dramatic improvement, and can be done in about a day. As a result of these improvements, there has been a flood of genomics data leading to important findings and applications.

One important finding is, remarkably, the differences between two individuals' DNA are extremely small—about 0.1 percent. That is, if you look at the three billion *As*, *Cs*, *Ts*, and *Gs*

in the human genome (times two, because we each have two copies of DNA, one from mom and one from dad), you'll only find millions of differences—so-called single, nucleotide polymorphisms or SNPs. (There are also more complex differences such as insertions, deletions, and multiple copies of small regions of DNA, but we won't get into that here.) These “small” differences are a big plus for computer scientists and computational biologists who are working with and analyzing genomic data, as it drastically reduces the scale of their computational problems. Another interesting finding is that almost all SNPs take on only two nucleotide values (*A*, *T*, *C* or *G*), which

makes it possible to talk about a major and minor nucleotide or “allele.”

Perhaps the most compelling application of the genomics revolution is personalized or precision medicine, where medical care is customized to the patient based on their genome. As an example, it turns out if you happen to have a particular set of SNPs in the region of your DNA that codes for immune-system-related genes known as Human Leukocyte Antigens (HLAs), and you take the seizure-suppressing drug carbamazepine, you will very likely develop Stevens-Johnson syndrome. This horrible disease causes layers of your skin to separate from one another. One idea behind personalized medicine is to avoid

such fates by checking your genome before administering drugs. More generally, with personalized medicine, your genome could be used to identify drugs that would work well for you or to warn you of diseases for which you are at high risk, such as diabetes and heart disease, so you can take measures to avoid them.

This application brings us to the question: How do we identify associations between your genome and various traits such as whether you will get a disease or whether a particular drug will harm you? Although older techniques have been used for decades, the genomics revolution has led to use of a new technique known as Genome-Wide Association Studies or GWAS.

With GWAS you identify a set of individuals (now typically numbering in the thousands), measure many or all of their SNPs, and then identify correlations between those SNPs with the trait. With exceptions to be noted, the genomics community typically looks for associations between a single trait and a single SNP. Also typical is the use of a “frequentist” test yielding a p -value indicating how likely a putative correlation is due to chance. This approach has proved to be quite successful, yielding ever-increasing numbers of significant associations (see Figure 2).

One surprising realization from these studies is the Stevens-Johnson syndrome example—where only a few SNPs are associated with the trait—is the exception rather than the rule. Namely, it is now recognized that, typically, many SNPs (hundreds and more) are causally related to traits. Furthermore, it turns out that whether an association can be detected depends on the frequency of the minor allele and the strength of the SNP’s effect on the trait. SNPs with larger effect tend to have very rare minor alleles due to natural selection. In contrast, SNPs with more common mi-

nor alleles tend to have little effect on the trait. Thus, extremely large sample sizes—on the order of 10^5 individuals—are required to create a comprehensive picture of the relationships among SNPs and traits. Many groups, including the American and British governments, are working to generate such data.

Unfortunately large datasets often lead to the introduction of confounders—unobserved factors or variables that introduce spurious associations, those that are not truly causal. As an example, a data set might include individuals from two different populations (say, Europe and Africa). If one population is correlated with both the trait and a SNP, then the trait and SNP will be spuriously correlated. Confounding of this sort is known as “population structure.” In large data sets, we also often see individuals who are related to each other, either closely or distantly. These relationships, known as “family relatedness,” can also introduce confounders and spurious correlations. This problem raises the question: How can we deal with confounders and identify only the truly causal associations?

To help answer this question, we

turn to graphical models. The problem of confounders can be represented with the graphical model shown in Figure 3. Here, the variables c_i are SNPs that causally influence the trait y , the variables s_j are SNPs that are not causally related to the trait, and h is an unobserved confounder. This model can capture population structure by having h encode the population identity of an individual. The model can capture family relatedness by having h represent (in some sense) the family identity of an individual. Note that this model is somewhat idealized. Recombination during gamete (sperm and egg) production leads to correlations among adjacent SNPs. In addition, the confounder is sometimes directly correlated with the trait, for example when h is related to causal environmental factors. Nonetheless, this idealized graphical model is sufficient for purposes of our discussion.

The graphical model readily explains how spurious associations can arise. (If you are unfamiliar with graphical models, please see the discussions about d-separation in *Probabilistic Reasoning in Intelligent Systems* [2].) In particular, suppose we test for

Figure 1. The cost of sequencing a human genome is dropping faster than Moore's Law. [Used with permission from Hindorff [1].]

the presence of a correlation between s_1 and y . Here, because h and the causal SNPs are unobserved, there will be an open path between y and s_1 corresponding to a non-causal association.

This explanation suggests an immediate solution: When testing for a correlation between a SNP and the trait, do so by taking into account the observations of the causal SNPs. That is, in the language of probability, evaluate the correlation conditional on the causal SNPs. This approach blocks all paths between s_1

and y and eliminates the non-causal association. Unfortunately, the identity of the causal SNPs is unknown. One way around this problem would be to condition on all SNPs, except the one being tested or (when taking into account relatedness among adjunct SNPs due to recombination) those SNPs near the one being tested. But this approach then leads to a second problem: How do we condition on so many SNPs without drastically increasing the complexity of the model? Let us look at one answer

Figure 2: The number of published genome-wide association reports is increasing rapidly. [Used with permission from Hindorff [1].]

that is proving successful in the genomics community.

Suppose we want to determine whether SNP s is correlated with trait y , conditioned on a set of SNPs $X=(x_1, \dots, x_K)$. One approach used by the community is to compute a p -value for the possible association based on a likelihood ratio test (LRT). To perform this test, two probabilities are computed: (1) the probability of observations of the trait (across N individuals) given s and X , and (2) the probability of the observations of the trait given only X . It turns out that twice the logarithm of the ratio of these probabilities follows a chi-squared distribution. We can use this fact to determine the p -value. Here, let us consider the computation of the second probability. The computation of the first probability is similar, but involves more complex notation.

To represent these observations over N individuals, we will use bold-face versions of the variable names. In particular, we will use $y=(y_1, \dots, y_N)^T$ and $X=(X_1, \dots, X_N)^T$ to denote the observations of y and X , respectively, across the individuals. Note that X is thus an $N \times K$ matrix, where the ij^{th} element corresponds to the j^{th} SNP of the i^{th} individual. In the genomics community, a common assumption is that y is a multiple linear regression on X :

$$y \sim \mathcal{N}(\mu + X\beta; \sigma_e^2 I)$$

where μ is a constant offset, $\beta_T=(\beta_1, \dots, \beta_K)$ are the weights relating corresponding SNPs ($x_1, \dots, x_K$) to the trait, and σ_e^2 is the variance of the multivariate normal distribution denoted by $\mathcal{N}(\cdot; \cdot)$. In practice, the use of linear regression often yields good results even when the phenotype takes on discrete values 0 (absent) and 1 (present). In addition, we assume the β s are mutually independent, each having a normal distribution

$$\beta_i \sim \mathcal{N}(0; \sigma_g^2), i=1, \dots, K$$

Averaging over the distributions of the β_i s, we obtain

$$y \sim \mathcal{N}(\mu; \sigma_e^2 I + \sigma_g^2 X X^T) \quad (1)$$

where X^T is the transpose of X . Model (1) has been derived independently by mul-

tiple researchers over the years and goes by different names including Bayesian or regularized linear regression. Model (1) is also known as a linear mixed model with random effect having the covariance matrix $\mathbf{XX}^T$, and a Gaussian process with a linear covariance (or kernel) function. In any case, researchers in the genomics community have investigated this approach and found it to perform quite well for GWAS, in the sense that it yields good control of type-one error, a criterion related to avoiding spurious associations that is widely adapted by the community [3, 4].

However, the use of model (1) raises computational issues. Namely, evaluation of the LRT probabilities associated with it naively require manipulations of $\mathbf{XX}^T$ that scales cubically with sample size N , yielding an overall runtime complexity of $O(MN^3)$ when testing M SNPs. Thus, the model is infeasible for GWAS with sample sizes on the order of 10^5 . Algebraic transformations offer a partial solution. If we first factor $\mathbf{XX}^T$ into the matrix product $\mathbf{USU}^T$, where $\mathbf{U}$ is an orthogonal matrix and $\mathbf{S}$ is a diagonal matrix (a procedure known as a spectral decomposition), then it can be shown that model (1) can be written as the linear regression

$$\mathbf{U}^T \mathbf{y} \sim \mathcal{N}(\mathbf{U}^T \boldsymbol{\mu}; \sigma_e^2 \mathbf{I} + \sigma_g^2 \mathbf{S}) \quad (2)$$

That is, the model takes the form of a linear regression after the data is rotated by $\mathbf{U}^T$, where the rotation scales quadratically in N . Model (2) can be used to test M SNPs with a runtime complexity of $O(MN^2)$. Unfortunately, this improvement is still unlikely to make large-scale GWAS feasible.

A path to a truly feasible large-scale GWAS may lie in the strong correlations among adjacent SNPs that arise due to recombination. Given these strong correlations, it might be possible to use a subset of SNPs (e.g., evenly spaced across the genome) for conditioning during GWAS. If the number SNPs used, K , is less than the sample size N , then it is possible to perform GWAS with an overall runtime complexity of $O(MNK)$ [5]. This result follows from additional algebraic manipulations that make use of the factored form of the covariance term $\mathbf{XX}^T$.

Model (2) and the additional al-

Figure 3: A graphical model for GWAS, showing causal SNPs c_i , non-causal SNPs s_j , trait y , and unobserved confounder h for an individual. The box around the model indicates that observations are repeated for N individuals.

gebraic improvements allowing an $O(MNK)$ runtime is known as factored spectrally transformed linear mixed model or FaST-LMM [5]. The model is being used widely by the genomics community and is available at <https://github.com/microsoftgenomics>.

The value of K needed to effectively block the open paths in the graphical model will depend on the strength of correlations among nearby SNPs, which will in turn depend on the data set under analysis. Although experiments with phenotypes generated synthetically from the real SNPs can be used to determine a minimum value for K [4], it remains an open question as to whether K will be sufficiently small to allow analysis in practice on the large datasets that are being created.

Recently, researchers have been looking at how to identify associations between a set of SNPs and a trait. The idea is that the association between a trait and any one SNP may be too weak to detect, but the association between a trait and a well-chosen set of SNPs (e.g., the set of SNPs in a gene or in a region thought to regulate a gene) can be detected [6]. Another line of work focuses on improving the analysis of binary traits by modeling them as thresholded hidden continuous traits [7]. In yet another line of work, traits are modeled jointly based on the recognition that there are substantial interactions or dependencies among them [8].

This short article just scratches the surface of ongoing efforts in GWAS for personalized medicine. Efforts to improve the computational techniques by reducing the time complexity and improving the accuracy of causal asso-

ciation discovery promise to increase the rate at which personalized medicine can become a reality. Finally, as with any field experiencing exponential growth, there are sure to be surprises.

References

- [1] Hindorf, L. et al. A Catalog of Published Genome-Wide Association Studies. National Human Genome Research Institute; <http://www.genome.gov/gwastudies/>. Accessed March 2015.
- [2] Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, San Francisco, 1988.
- [3] Kang, H. M. et al. Efficient control of population structure in model organism association mapping. *Genetics* 178, 3 (2008), 1709–1723.
- [4] Widmer, C. et al. Further improvements to linear mixed models for genome-wide association studies. *Scientific Reports* 4, 6874 (2014).
- [5] Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nature Methods* 8, 10 (2011), 833–835.
- [6] Lippert, C. et al. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* 30, 22 (2014), 1–9.
- [7] Weissbrod, D., Lippert, C., Geiger, D. and Heckerman, D. Accurate liability estimation improves power in ascertained case-control studies. *Nature Methods* 12, 4 (2015), 332–334. doi:10.1038/nmeth.3285
- [8] Lippert, C., Casale, F. P., Rakitsch, B., Casale, F. P. and Stegle, D. LIMIX: Genetic analysis of multiple traits. *bioRxiv*. 2014; <http://biorxiv.org/content/biorxiv/early/2014/05/22/003905.full.pdf>

Biographies

Christoph Lippert is a researcher at Microsoft Research. He works on machine learning, and computational and statistical methods for genetics and genomics, including Gaussian-process regression and methods for genome-wide association studies. Lippert has done research at the Max Planck Institutes in Tübingen and received his Ph.D. from the University of Tübingen in 2013.

David Heckerman is a distinguished scientist at Microsoft Research. He is known for his work in showing the importance of probability theory in artificial intelligence, for developing methods to learn graphical models from data, and for developing machine learning and statistical approaches for biological and medical applications, including the design of a vaccine for HIV and algorithms for genome-wide association studies. Heckerman received his Ph.D. [1990] and M.D. [1992] from Stanford University, and is an ACM and AAAI Fellow.

© 2015 Copyright held by Owner(s)/Author(s).