nature genetics

Predicting the influence of common variants

An ever-larger proportion of the liability to common and complex disease can be obtained by progressively larger studies. However, for most diseases, the sample sizes required to gain usable predictions will be out of reach of sequencing technologies for the foreseeable future. Array-based genotyping genome-wide association studies (GWAS) still offer a reliable harvest of biological hypotheses for many diseases, together with the secondary benefit of slowly improving prediction.

GWAS have an amazing track record in rapidly discovering the genetic contribution to over 700 common and complex diseases and phenotypes. Indeed, the technique may well have mapped out a large proportion of the regulatory variation associated with human traits. Still, by the benchmark of genetic epidemiologists, it has been slow to deliver. The set of well-replicated SNPs together do not account for the phenotypic variance that can be attributed to additive genetic variance (narrow-sense heritability). Because of this, the loci are of limited usefulness in risk prediction. We are simply not yet playing with a full deck.

Consortia of genetic epidemiologists now work on very large population samples. For example, in this issue, our Focus on cancer risk loci (p 343) reports the results of genotyping ~200,000 SNPs in a total of ~200,000 individuals. The implications of these studies are explored in two Commentaries (pp 345, 349) and in detail online in our editorial threads (http://nature.com/ng/focuses/icogs) linking the coordinated COGS publications. These roughly double the harvest of loci associated with these cancers and finally make clinical prediction a testable reality. For three common cancers, breast (p 353), ovarian (p 362) and prostate (p 385), genetic variants now explain about a third of the familial relative risk (link to Primer 1 online). With each of these studies now able to identify the individuals at the greatest genetic risk, SNP genotypes can be used in stratification approaches and tested in population screening (p 349). Variants contributing to disease can be found by considering not only the replicated variants of significant effect but also all genotyped variants using polygenic analytical methods that take into account the much larger set of contributory SNPs (*Nat. Genet.* **42**, 565–569, 2010). In this issue, Nilanjan Chatterjee and colleagues (p 400) show that the predictive accuracy attained by larger studies is limited not only by the samples available to train the polygenic model but also by the distribution of effect sizes of the genetic variants themselves. For some diseases, prediction is readily achievable, but in no case do they anticipate that common SNPs will fully account for all of the genetic variance, even when the thousands of variants with individually undetectable effect sizes are included. For many diseases, the polygenic model suggests that GWAS deliver no more prediction accuracy beyond studies of 100,000–200,000 individuals, but for a few conditions, such as coronary artery disease (p 422), it may be useful to continue studies to five times that size.

Risk prediction has many different aims. For most diseases, it should be possible to identify the individuals with the highest genetic risk. However, if the aim is to identify individuals with just twice the mean population risk, we cannot currently do that with SNPs. For most diseases, only a small proportion of individuals with twice the mean risk can currently be identified genetically. Risk locus discovery can be an iterative process, with subtypes of disease being initially lumped together or discovered in the process. Loci that have larger effect sizes in disease subtypes than in the broader condition can be more valuable in predictive classification.

Although we currently deal with diseases and loci independently, common diseases themselves may not be independently distributed in the population. For example, James Sorace and colleagues (*Popul. Health Manag.* **14**, 161–166, 2011) examined the conditions for which over 32 million US citizens billed the Medicare system in 2008, classifying the population by the combinations of two or more conditions for which they were treated in that one year. What they found was that the population divided approximately into thirds. One third of the population with no recorded treatment accounted for 6% of the expenditure. The next third accounted for 15% of the expenditure on the 100 most common disease combinations, and the final third, with larger numbers of disease combinations, accounted for 79% of the expenditure. The healthcare cross section was also very diverse. Sixty percent of the beneficiaries (and 90% of the expenditure) were accounted for by over 2 million disease combinations comprising of one of the 20 most prevalent conditions with one or more other condition.

To us, this study suggests that, if prediction is to be used in the real world, it will be interesting to examine the genetic risk profiles for common diseases mapped so far by GWAS in a cross section of health care users. ■