

- 18 Abbott, A. (2002) Betting on tomorrow's chips. *Nature* 415, 112–114
- 19 Mann, M. *et al.* (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473

- 20 Mandl, K.D. *et al.* (2002) Newborn screening program practices in the United States: notification, research, and consent. *Pediatrics* 109, 269–273
- 21 Williamson, A.R. (2001) Gene patents: socially acceptable monopolies or an unnecessary

- hindrance to research? *Trends Genet.* 17, 670–673
- 22 Woolley, A.T. *et al.* (2000) Direct haplotyping of kilobase-size DNA using carbon nanotube probes. *Nat. Biotechnol.* 18, 760–763

One gene and one outcome? No way

Stephen Chanock and Sholom Wacholder

The world of molecular diagnostics is undergoing major change because of both technical advances and the availability of rapidly expanding genetic databases generated by the study of human genomics. These resources comprise an extraordinary opportunity to decipher the biological importance of genetic aberrations, and link our understanding with clinical utility. The challenge lies in sorting through the information and developing effective strategies to advance molecular diagnostics.

Published online: 9 May 2002

Before we go overboard in predicting imminent progress or even the attainment of a medical utopia based on the recent advances in genomics, we must acknowledge that we are not yet sufficiently equipped with the knowledge to make sense of the enormous opportunities that lie ahead [1]. The sheer volume of information harvested from interrogation of still more single nucleotide polymorphisms (SNPs), transcripts or chromosomal regions, using smaller and smaller quantities of material, is enormous, and the expectations are too great to simply slow down the process [2]. Certainly, the opportunity to look at old problems using new approaches will yield unexpected findings and, perhaps, redefine the questions we ask [3,4]. But even as technical advances create more opportunities to look at larger numbers of genetic observations, which could have an impact on outcomes, we do not currently possess adequate means to examine existing data effectively [5]. The widening gap between the streams of data generated by new technologies and the underlying biological insights required to make sense of the observations will tax molecular diagnostics, which relies on a tight connection between a test and its 'clinical significance'.

Numerous technologies exist that detect genetic alterations [e.g. DNA sequencing, molecular profiling by cDNA microarray analysis, comparative genomic hybridization (CGH), fluorescent *in situ* hybridization (FISH) and spectral karyotyping (SKY) analyses] that could improve diagnostic acumen and treatment decisions. Similarly, new platforms have been developed to survey many genes at once by SNP or cDNA microarray analyses, and these can also be informative

for predicting outcome risk. Moreover, investigation of genomic and epigenetic alterations could lead to new insights into the mechanisms underlying classical monogenic disorders (e.g. hemophilia) and more complex genetic disorders (e.g. cancer, mental illnesses), which range from individual mutations to rearrangements of chromosome segments [6,7].

Assigning significance to genetic markers

The era of a one-to-one correspondence between a single gene and a single disease is at an end. Even classical gene mapping for highly penetrant monogenic disorders is being revisited in search of secondary genes that modify the effect of the primary mutation. Penetrance is a measure of the proportion of individuals with a defined genotype who manifest an expected phenotype and it is usually observed under a specific set of environmental conditions that might not be applicable in the complex, real world of clinical medicine. The ability of an individual SNP to disrupt a process might be too low to have a major effect on penetrance. Even the identification of a SNP with elevated penetrance is troublesome because of the difficulty in isolating the individual effect on phenotypic outcome. Furthermore, it is subject to the cumulative minor effects of other SNPs.

No single SNP is sufficient or necessary for predicting a clinically significant phenotype with acceptable accuracy. Even when we consider a monogenic disorder like cystic fibrosis (CF), there is a wide spectrum of clinical outcomes observed in patients bearing the same primary mutation, even within families. Surely, phenotypic differences can be ascribed partially to environmental factors, but several recent studies have provided preliminary evidence that additional genes modify the risk of severe outcomes in supposedly monogenic disorders. For example, several studies have now shown that common exon 1 variants in the mannose binding protein (encoded by the *MBL2* gene) are associated with deterioration in pulmonary function in patients with cystic fibrosis [8,9]. What is the 'penetrance' of these variants? How can one characterize the penetrance of a primary mutation when it is modified substantially by common SNPs? So far, we have not yet figured out how to communicate genetic modifiers to patients with monogenic diseases such as sickle-cell anemia or chronic granulomatous disease (CGD). The very concept of penetrance has reduced appeal outside the narrow 'one gene, one disorder' paradigm; when SNPs arrive as components in the everyday arsenal of clinical diagnostics (especially for complex diseases such as hypertension or diabetes), the message to scientists, clinicians and

Stephen Chanock
Section on Genomic
Variation, Pediatric
Oncology Branch, Center
for Cancer Research,
National Cancer Institute,
Bethesda, MD, USA.
e-mail: sc83a@nih.gov

Sholom Wacholder
Division of Cancer
Epidemiology and
Genetics, National Cancer
Institute, Bethesda,
MD, USA.

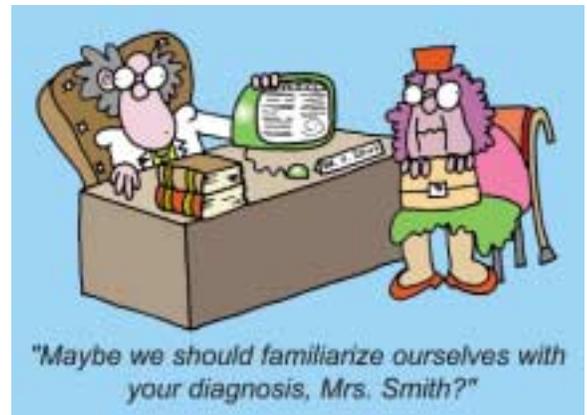
patients will have to be counched as the difference in the risk for disease outcome.

When we turn to the clinical utility of this information, we are struck by the difficulty of interpreting SNP data without consideration for the context. For example, why does the 'penetrance' of *BRCA1* mutations for breast cancer seem to be higher in studies of 'high-risk families' than in population-based studies? [10] Does the risk for breast or ovarian cancer vary according to the primary *BRCA1* or *BRCA2* mutation [11]? Does a woman's risk vary with the presence of other SNPs or by her parity or past use of oral contraceptives? How does one consider an intervention to reduce the risk of one outcome when it has the potential for an adverse effect on another? Nonetheless, the classical approach of dissecting heritable components of complex diseases persists, even though we know that no simple definition of penetrance for SNPs is adequate.

Analyzing SNPs in molecular diagnostics

SNPs do not act in isolation, but against the background of thousands of other SNPs, on top of environmental factors. SNPs are common variations defined by a frequency of the minor allele greater than 1% in one or more populations. Strewn throughout the genome, SNPs can be located within genes and in intergenic regions at a frequency far greater than previously imagined; there can be as many as 10 million SNPs per person, which is still a tiny fraction (<0.1–0.2%) of the total genome's size. The sheer number of SNPs, even if restricted to those predicted to have functional significance (i.e. those having a measurable phenotypic effect because of a change in amino-acid sequence or alteration in gene expression, accounts for roughly 50 000–250 000 overall), is staggering [12].

It is clear that identifying each of the potentially multiple genes that might affect a complex phenotype will emerge as one of the most formidable challenges in molecular diagnostics. Until recently, the paradigm for determining whether a SNP is informative has been to identify the effect of a single gene on a single outcome. When one considers the multi-dimensional implications of combinations of SNPs in a profile, instead of one SNP at a time, the level of complexity increases dramatically. For example, intense effort has been directed at identifying SNPs or variants that alter the risk for severe malarial infection [13,14]. So far, common variants in at least ten different genes have been reported to influence the risk for severe malaria. It will be a daunting task to isolate the relative contribution of any single variant. To date, no published study has the required sample size or design to isolate the effects of or determine the absolute and relative importance of distinct variants/SNPs [15,16]; the joint effect of two or three or even more variants could override the effect of any single one. Alternatively, the effect of a specific SNP might not be apparent unless in the presence or absence of others. More, informative SNPs might be infrequent, or



uncommon variants [17]. As long as we rely upon the venerable 'single gene, single outcome' paradigm for confirmation of results, we will not achieve even cursory knowledge of the effect of individual SNPs.

There is a crucial difference between looking at a SNP variant as a susceptibility factor in the general population and in a population under stress, or already under extraordinary risk. In the context of a substantial stress (i.e. an underlying 'hit', which can be either a genetic mutation, an underlying condition such as HIV or HPV infection, or a powerful environmental exposure like smoking), the effect of a minor phenotypic change in a complementary biological pathway could be magnified [18]. Therefore, it is not surprising that genetic modifiers have been reported in a range of monogenic disorders, including CF, CGD and sickle-cell anemia. Pharmacogenomics and pharmacotoxicology are examining the effect of a SNP or set of SNPs in concert with a pharmacological interventions [19]; the responsiveness (i.e. efficacy) or toxicity (i.e. side effects) of a SNP might be very different in the presence of a stressor potent enough to alter a biological process or pathway [20].

Some have argued that haplotypes, which are made up of common SNPs inherited in blocks, will be a more efficient means to screen for significant associations [21,22]. The expectation is based upon the premise of the existence of common haplotypes, which should decrease the number of variants required to be screened in the first stage of a study [23]. But this type of analysis also raises uncomfortable questions about population-specific haplotypes and new complexities in design and analysis for looking at linked markers. To understand which components within a haplotype actually confer any effect, positive or negative, haplotypes will have to be deconstructed and the individual variants examined in a process that has not been well developed. SNPs that possess no discernible effect on phenotype can create meaningless differences in risk, because they will be linked to one or more SNPs, which confer the observed biological effect. It will be difficult to distinguish definitively between statistical noise and true signal without the laboratory-based understanding of the functional implications of the components of the haplotype.

The challenge of molecular diagnostics

Most experts assume that molecular diagnostics is entering a new age, driven by the availability of new techniques coupled with bioinformatic tools to mine the rapidly expanding genetic databases. More data are becoming available, but does this lead to more information? Patients expect and deserve to know why a test is performed and what its consequences are. Are we ready to deluge patients with new tests and explain their implications? Will the medical provider or the patient accept the findings blindly, or for that matter, take the time to understand the significance of each data point? In other words, is it reasonable to perform tests when the data cannot be fully explained to the patient? This is unlikely. Indeed, the immediate challenge will be to provide guidance to care providers who face the reality of offering patients test results emanating from techniques not yet fully understood by the care providers. Education of practitioners will have to include a basic understanding of the types of genetic tests, as well as their results and implications for clinical care.

Patients expect and deserve to know why a test is performed and what its consequences are. Are we ready to deluge patients with new tests and explain their implications?

We are at a critical juncture and must choose between two mutually exclusive courses. One is to wait for the basic-scientific research process to catch up, which is highly unrealistic. It is unlikely in the near future that computational approaches will be developed with sufficient sophistication to organize and analyze the staggeringly large amount of genetic information and present us with a clear picture of the meaning of the data [5]. The second option is to begin to analyze genetic markers (e.g. SNPs, cDNA profiles) in clinical settings, but without satisfactory biological validation *a priori*. This second approach, while more practical, is also more daunting, because it will require continual reassessment of analytical tools as well as the criteria for applying new and old data to clinical utility, based upon experience and not an *a priori* model.

The tantalizing lure of examining more genes will pay off sooner for both basic science and clinical applications when studies are designed to yield biological insights, not only estimates of risk or penetrance or disease association. Instead of looking at the 'one gene, one outcome' hypothesis, we now have the means to design studies to ask how collections of genes, either alone or in combination, contribute to complex diseases like cancer or mental illness. Herein lies the opportunity to look at pathways of genes, in large association studies, which could provide insight into the balance between factors that comprise a biological pathway. For example, it is now possible to

study SNPs within the genes that encode for Th1/Th2 cytokines or the complement pathway. In each of these examples, the analysis of subtle genetic differences could be useful in investigating the relative importance of one or more genes (and their gene-gene interactions) *in vivo*. While this approach utilizes a classical pathway, defined according to a model of action to which each component contributes, it is also possible to look at motifs embedded within a collection of genes that participate in different pathways. One could ask whether variants in genes with SH2 domains or immunoglobulin-binding regions alter function, and thus result in an unexpected disease association. In other words, the search must be not only for more than genetic markers, but also for the biological significance of the markers, if we are to ask patients to consider the information ultimately pertinent to health.

Future considerations

It is necessary to address two major gaps in the implementation of molecular diagnostics: (1) the chasm between existing genetic information and its clinical utility, and (2) the translation of collected data into effective clinical practice. Exploitation of the great advances ahead will require novel study designs to look at the etiology of disease in ways that do justice to their complexity. A new paradigm will have to be forged to address ethical and economic issues; specifically, the informed consent process will have to undergo substantial transformation to account for the volume of information. In this regard, the overall probability of an event, even if incompletely defined, might have to suffice. It will not be possible to discuss each gene included in a microarray study; instead, the cumulative answer will have to be adequate. At the same time, we will need to carefully consider how insurers will incorporate the uncertainty of complex genetic risk into a fair and equitable system. We will have to develop safe systems for storing genetic information, so that we will protect the confidentiality of the individual, and at the same time, permit release of data, after informed consent is obtained, to ongoing studies in an effort to accumulate a sufficiently large sample to achieve adequate power.

Finally, we will have to re-educate the medical community and develop standards for assessing data. In turn, there will have to be a slow but steady education of the general public concerning the importance of risk, particularly as it pertains to complex genetic diseases. This could generate a substantial degree of misunderstanding, even a backlash against all of molecular diagnostics. However, with a clear program, based on high standards for study design, analytical methods and criteria for implementation, it is possible to envision the impact of the age of molecular diagnostics changing the practice of medicine. To effect this transformation and realize the promise of molecular diagnostics, we will have to look carefully at the design of studies, so that the information can be applied to

both general and selected populations. We will have to take the next steps with not only advanced computational tools in hand, but also with a measured

and useful eye towards the clinical context, one that guides us towards the introduction and utilization of genetic information to improve health worldwide.

References

- 1 Collins, F.S. (1999) Shattuck lecture – medical and societal consequences of the Human Genome Project. *N. Engl. J. Med.* 341, 28–37
- 2 Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature* 405, 847–856
- 3 Schork, N.J. *et al.* (2000) Single nucleotide polymorphisms and the future of genetic epidemiology. *Clin. Genet.* 58, 250–264
- 4 Collins, F.S. and Guttmacher, A.E. (2001) Genetics moves into the medical mainstream. *J. Am. Med. Assoc.* 286, 2322–2324
- 5 Brookes, A.J. (2001) Rethinking genetic strategies to study complex diseases. *Trends Mol. Med.* 7, 512–516
- 6 Lupski, J.R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet.* 14, 417–422
- 7 Stankiewicz, P. and Lupski, J.R. (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet.* 18, 74–82
- 8 Garred, P. *et al.* (1999) Association of mannose-binding lectin gene heterogeneity with severity of lung disease and survival in cystic fibrosis. *J. Clin. Invest.* 104, 431–437
- 9 Chanock, S.J. and Foster, C.B. (1999) SNPing away at innate immunity. *J. Clin. Invest.* 104, 369–370
- 10 Struwing, J.P. *et al.* (1997) The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N. Engl. J. Med.* 336, 1401–1408
- 11 Modan, B. *et al.* (2001) Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a BRCA1 or BRCA2 mutation. *New Engl. J. Med.* 345, 235–240
- 12 Chanock, S.J. (2001) Candidate genes and single nucleotide polymorphisms in the study of human disease. *Disease Markers*, 17, 89–98
- 13 Hill, A.V. *et al.* (1997) Genetic analysis of host-parasite coevolution in human malaria. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 352, 1317–1325
- 14 Hill, A.V. (2001) The genomics and genetics of human infectious disease susceptibility. *Annu Rev Genomics Hum Genet.* 2, 373–400
- 15 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
- 16 Dahlman, I. *et al.* (2002) Parameters for reliable results in genetic association studies in common disease. *Nat. Genet.* 30, 149–150
- 17 Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69, 124–137
- 18 Foster, C.B. *et al.* (1998) Host defense molecule polymorphisms influence the risk for immune-mediated complications in chronic granulomatous disease. *J. Clin. Invest.* 102, 2146–2155
- 19 Roses, A.D. (2002) Pharmacogenetics place in modern medical science and practice. *Life Sci.* 70, 1471–1480
- 20 Evans, W.E. and Relling, M.V. (1999) Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 286, 487–491
- 21 Collins, A. *et al.* (1999) Genetic epidemiology of single-nucleotide polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.* 96, 15173–15177
- 22 Reich, D.E. *et al.* (2001) Linkage disequilibrium in the human genome. *Nature* 411, 199–204
- 23 Fallin, D. *et al.* (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. *Genome Res.* 11, 143–151

The value of microarray techniques for quantitative gene profiling in molecular diagnostics

Stephen A. Bustin and Sina Dorudi

There has been an explosion of interest in microarray technologies that allow the quantification of whole-genome RNA expression data. The apparent correlation of expression profiles with clinically relevant parameters such as disease outcome has raised expectations with respect to the clinical usefulness of the data generated. Yet the accuracy and biological relevance of these data remain contentious, even in basic research applications. Therefore, numerous issues related to format, quality, validation and interpretation remain to be resolved before microarray profiling can become a diagnostic tool of clinical relevance for routine work.

Published online: 8 May 2002

The current fashion for functional genomics has put the spotlight on microarray technologies that are capable of comprehensive, quantitative RNA

expression analysis, with their promise of a global approach to the quantification of gene expression [1]. However, there are two conflicting opinions on the value of microarrays: they are either promoted as an exploratory-driven replacement for hypothesis-driven biology [2], or they are criticized as expensive fads [3] that cannot substitute for the old-fashioned, low-throughput approach to experimental biology [4].

Variability

Unquestionably, microarrays can reveal associations between gene-expression signatures and the biology and outcome of disease, for example by identifying clinically significant subtypes of cancers [5,6]. This has raised expectations that the expression profile of a cell can be used as a diagnostic and/or prognostic aid in cancer management. However, although microarray experiments generate long lists of genes with altered expression, the interpretation of these data depends on the judgement of the investigator performing the experiment [7]. Furthermore, a comparison of the same microarray experiment performed a few weeks apart can demonstrate a considerable lack of reproducibility [8]. This is exacerbated by an apparent lack of robustness of the microarray data, as demonstrated by the variability in the results obtained in different laboratories. Indeed, a comparison of data obtained in independent studies performed with different microarray platforms, for example in lung [9,10] or colorectal [11–13] cancers, shows both similarities and significant differences. Moreover, a comparison of 47 and 98 genes identified from independent studies to be associated with metastasis [14,15] does not reveal a