points for therapeutic intervention. The same is probably true for drug-response genes, in which single-nucleotide polymorphism (SNP) mapping is being extensively used in clinical studies to measure population response to marketed drugs[8].

As for new targets with direct utility in small-molecule drug discovery, the 30,000 genes identified in the draft sequence must encode a considerably larger number of ligand-binding domains (LBDs), many of which could be therapeutic targets in their own right. This implies that original estimates of 5,000–10,000 extra "targets" hidden within the human genome[9], while an overestimate based on gene number alone could actually be an underestimate at the LBD level.

Although it is fair to say that the bulk of the paper is not concerned with pharmaceutical research applications, examples are given of a number of potential new therapeutic targets the discovery of which has been aided by access to the genome sequence. Some of these are likely to represent interesting new candidate drug targets, such as the putative dopamine, purinergic, and insulin-like growth factor receptors quoted in the paper—possibly the tip of a target iceberg? In this sense, the fruits of the genome sequence are now available to all researchers—if any fruit is left on the tree after it has been systematically harvested in the patents of Celera (Rockville, MD), Incyte (Palo Alto, CA), Human Genome Sciences (Rockville, MD), and others!

With the way signposted by genomics, there has never been a greater choice of targets on which to speculate, but the potential for expensive failures in the clinic is vast with unvalidated targets. Their triage necessitates careful clinical trial design and highlights the urgent need for new surrogate markers of disease progression and drug efficacy. Quantitative gene and protein expression analyses, facilitated by the draft sequence, will play central roles in developing such markers.

In Figure 1, we predict a continuity of *in silico* advances that will underpin future drug discovery. Already, the human genome initiative is pointing the way to the next wave of *in silico* research, from the one-dimensional world of DNA sequence information to the three-dimensional world of structural genomics.

Major international efforts focused on determining protein structures on an industrial scale[10] are yielding essential data for the next phase of *in silico* drug discovery. Currently, bioinformatics algorithms, such as PROSITE[11], can identify around 1,400 distinct sequence patterns, based on their occurrence within linear protein sequences. Efforts will continue to focus on developing methods that expand this search to three-dimensional protein architectures, allowing the comprehensive definition of all possible ligand-binding sites. A major advance springing from comparative site analyses will be a new understanding of site selectivity and how to design drugs that exploit it. The automated definition of the ligand-binding sites themselves will be critical to progress in this area.

A key development in the computational world has been the arrival of *de novo* design algorithms that use all available spatial information to be found within the target to design novel drugs[12]. Coupling these algorithms to the rapidly growing body of information from structural genomics provides a powerful new route for exploring design to a broad spectrum of genomics targets, including more challenging examples such as protein–protein interactions.

A gauntlet has now been thrown down for the drug discovery industry. Clearly, molecular biology is capable of reduction to practice on an industrial scale. Because technology moves in waves (see "The next wave"), we believe it is only a matter of time before structural biology and structure-based drug design are also transformed by new approaches. The genome sequence is a triumph of collaborative science and computational biology, but it is merely the beginning; from it we can glimpse a new *in silico* wave of pharmaceutical discovery.

1. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
2. Data from S.G. Cowen (2001).
3. Liang, F. *et al. Nat. Genet.* **25**, 239–240 (2000).
4. Roizman, B. *Nature* **288**, 2327–2328 (2000).
5. Paarmann, I., Frermann, D., Keller, B.B.B.U. & Hollmann, M. *J. Neurochem.* **74**, 1335–1345 (2000).
6. Pandey, A. & Mann, M. *Nature* **405**, 837–846 (2000).
7. Olszewski, K.A., Yan, L., Edwards, D. & Yeh, T. *Comput. Chem.* **24**, 499–510 (2000).
8. Roses, A.D. *Nature* **405**, 857–865 (2000).
9. Drews, J. *Science* **287**, 1960–1964 (2000).
10. Service, R.F. *Science* **287**, 1954–1956 (2000).
11. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. *Nucleic Acids Res.* **27**, 215–219 (1999).
12. Leach, A.R., Bryce, R.A. & Robinson, A.J. *J. Mol. Graph. Model.* **18**, 358–367 (2000).

Mapping a role for SNPs in drug development

Publication of the human genome SNP map is an early benchmark in efforts to increase the efficiency of drug development and improve the provision of medicine.

Bonnie E. Gould Rothberg

The International SNP Map Working Group has now presented an initial map of human genome sequence variation[1]. The map identifies and localizes 1.42 million single-nucleotide polymorphisms (SNPs) throughout the genome, most of which are located in noncoding regions. Although the work is likely to have a limited impact on drug development in the near term, it provides a foundation for efforts to shift biomedical research away from candidate disease/drug response genes and toward the systematic characterization of individual gene products and the effects of sequence variants on function.

Overall, the published SNP map (a composite of all publicly available polymorphisms as of November 2000) comprises three sets of data (see Fig. 1): 1,023,950 SNPs derived from reduced representation shotgun sequence of 3–5× redundant sequence coverage (originating from a two-year-old public–private effort funded by 10 pharmaceutical companies— the SNP Consortium); 971,077 SNPs identified as sequence differences in regions of overlap between large-insert clones (originating from human genome project sequence analysis); and 71,000 SNPs originating from public gene-based studies. Independent validation of a representative subset has demonstrated over a 95% true-positive rate, with over 82% occurring at minor allele frequencies of greater than 10% (ref. 1).

Construction of this genome-wide SNP map is an important benchmark in ongoing efforts to better characterize and correlate genes with complex traits such as drug response. Indeed, one of the main aims of the SNP Consortium was to rapidly identify SNPs that are either linked or contribute to individual drug response variability across a patient population—the field of pharmacogenetics.

Insufficient therapeutic efficacy or unanticipated side effects in "outlier" individuals of drugs previously demonstrated to be both safe and efficacious in clinical trials is cur-

Bonnie E. Gould Rothberg is group leader, pharmacogenomics at CuraGen Corporation, 555 Long Wharf Drive, New Haven, CT 06511 (bgould@curagen.com).

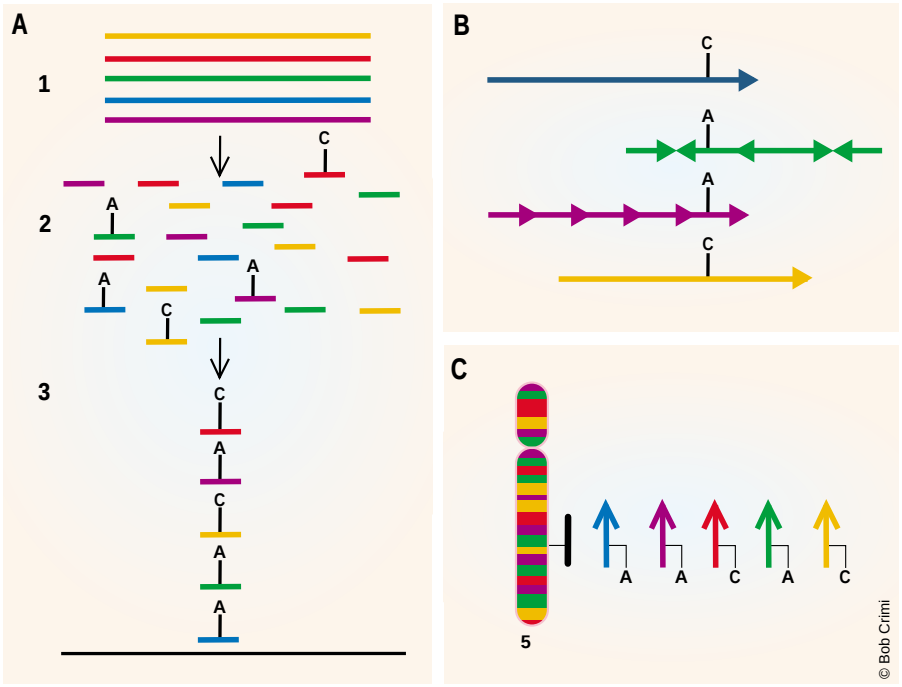

Figure 1. Methods used for International SNP Map Working Group SNP discovery. (A) Reduced representation shotgun sequencing followed by genomic alignment. Bacterial artificial chromosomes affording 3–5× coverage for selected genome regions are (1) generated and (2) fragmented using standard "shotgun" methodology. (3) The individual fragments are sequenced and readouts compared with publicly available large-insert clone sequences to ensure proper alignment. (B) Identification of SNPs through sequence differences in regions of overlap between large-insert clones sequenced by the human genome project. Ordered and unordered sequence clones are assessed using conventional alignment algorithms. This approach contributes dense clusters of SNPs spaced throughout the genome. (C) SNPs discovered through directed gene-based studies. These are identified by automated detection of single base differences in clusters of overlapping expressed sequence tags (ESTs) or by targeted resequencing efforts for candidate disease genes. In both (A) and (B), SNPs are then determined using Neighborhood Quality Standard (NQS) or Polybayes, Bayesian analysis of confidence scores performed on the raw sequence.

rently a tremendous problem for health care practitioners. A 1998 report indicated that 6.7% of hospitalized patients in the United States suffer serious adverse reactions as a result of drug administration—0.32% of these events resulting in death[2]. Extrapolating these data to include individuals treated as outpatients, individuals suffering milder adverse drug reactions, and all patients demonstrating limited drug efficacy, it is clear that a very high proportion of prescribed medicines do not result in the desired outcome.

Since the first correlation over half a century ago of adverse drug responses with amino acid variations in two drug-metabolizing enzymes (plasma cholinesterase and glucose-6-phosphate dehydrogenase), careful genetic analyses have linked sequence polymorphisms in over 35 drug metabolism enzymes, 25 drug targets, and 5 drug transporters with compromised levels of drug efficacy or safety[3]. In the clinic, such information is already being used to prevent drug toxicities; for example, patients are routinely screened for SNPs in thiopurine methyltransferase (TPMT) that cause decreased metabolism of 6-mercaptopurine or azathioprine, allowing the design of personalized regimens based upon genotype[4,5]. Yet, only a small percentage of observed drug toxicities have been adequately explained by the set of pharmacogenetic markers validated so far, and this level of personalized medication cannot be universally applied.

Pharmacogenetic candidate gene validation relies on elements of population and quantitative genetics in four main areas: first, identification of candidate pharmacogenetic target genes; second, identification of all potential alleles (and their relative frequency) for each candidate gene in the general population; third, genotyping of a clinically relevant population for the set of relevant alleles; and fourth, application of robust statistical metrics to establish linkage between any allele and a selected response/nonresponse phenotype. The recent introduction of high-throughput genomic methods has served to reduce both the time and the cost of pharmacogenetic discovery.

For the classical drugs and drug targets with a well-documented molecular pharmacology and toxicology, thorough literature surveys have proved successful in proposing candidate genes relevant to their pharmacokinetics and pharmacodynamics for pharmacogenetic follow-up. Indeed, most pharmacogenetic variations validated thus far have been identified in this fashion[3].

Literature review, however, has proved inadequate not only with drugs for which modes of action, metabolism, or disposition are poorly understood, but also with the newest generation of therapeutics designed to impact newly identified subclasses of drug targets with equally nascent receptor biology. It is for these drugs and targets that genomic approaches to pharmacogenetics will leave their mark.

Four methods can be used to identify pharmacogenetic candidate genes, two of which will be significantly facilitated by the SNP map[1] and its companion paper on the human genome sequence[6]. The strategy most relevant to the SNP data set is whole-genome linkage analysis on populations of poor responders to identify chromosomal regions likely to house candidate genes. The Working Group SNP collection places a SNP every 2–3 kilobases, with only 4% of the genome having a gap of >80 kilobases between SNPs. Whereas this fine degree of genetic mapping should ultimately pinpoint specific regions of linkage, the current cost of genotyping 100,000 unique SNPs across a reasonable clinical population continues to make such research impractical.

A second approach involves mining human sequence databases for unique paralogs of accepted pharmacokinetic and pharmacodynamic regulators. These genes may invoke previously unknown alternative pathways for drug activity that cause clinically detectable response variability. This data set is discussed among the findings published by the human genome project[6].

The last two candidate identification approaches are differential gene expression profiling in drug response models and construction of protein–protein interaction maps for drug receptors and their immediate effectors. Although these approaches (rather than genome linkage analysis and homology searches) are likely to be of immediate use in identifying new pharmacogenetic candidates, they are not dependent on the recently published data sets and thus will not be further discussed here.

In any case, it is clear that pharmacogenetics discovery will increasingly involve comprehensive candidate gene triage for all SNPs capable of modifying protein function and

the subsequent determination of each allele's frequency in appropriate responder/nonresponder clinical populations. To put some perspective on the utility of current data, of the 1.42 million SNPs presented in the *Nature* paper[1], only 10% (60,000 exonic SNPs and as many as 120,000 upstream promoter and intronic SNPs) are found within gene loci. Previous reports on sequence diversity indicate that, among coding region SNPs, 50% produce amino acid changes or premature terminations[7]. A smaller percentage, however, are expected to affect function, and it is also unclear how many nonexonic SNPs can influence gene function. Thus, of the total SNPs found so far, a staggering 1.2 million are unlikely to have any impact on drug response.

With the scale of the task now apparent, the challenge for researchers is to identify the exact SNP contributing to a functional change, and not an alternative SNP in linkage disequilibrium with the causative SNP. A related issue concerns the allele frequency for proposed candidate SNPs. In the *Nature* paper[1], 1,276 random SNPs were genotyped across three ethnic populations using pooled resequencing. Around 82% of assayed SNPs had a minor allele frequency of greater than the 10% detection limit in at least one population.

Based upon clinical data, it is questionable whether these common SNPs are the most likely pharmacogenetic candidates. For example, mibefradil, a new T-class calcium channel antagonist, was withdrawn from the market following increased mortality due to cytochrome P450 (CYP3A4)-mediated drug–drug interactions in a 2,590-patient trial[8]. A smaller trial conducted on 229 patients did not detect this toxicity[9]. Furthermore, troglitazone was withdrawn from the US market following significantly morbid outcomes in 650 individuals among over 1 million prescriptions[10]. It seems less likely that variants present in 10% of the population will be causative for clinical events observed in 1/1,000 to 1/10,000 patients. Although these SNPs should be included in any analysis, directed SNP mining in reference populations of 96 or 384 individuals is becoming standard practice for the identification of rare variants.

Appropriate statistical analyses to determine positive genotype/variable drug response correlations are also needed to exploit future diagnostics and/or improved compound screens. In this area, the development of robust mathematical models and statistical algorithms for processing genotypic data remains a priority. Because robust pharmacogenetic analysis algorithms must accommodate polygenic as well as monogenic causes of drug response variability, quantitative genetics metrics seem most likely to be useful.

It is unclear as yet whether individual SNPs or gene locus haplotypes represent most appropriate measures of genetic variability. For example, a recent evaluation of the efficacy of β_2-adrenergic receptor (β_2AR) agonists in a cohort of asthmatics uncovered a significant correlation between drug response and β_2AR gene haplotypes, but not in individual SNPs (ref. 11). Even so, statistical algorithms indicate that there is no correct empiric choice and that both single SNPs and gene haplotypes are equally suitable metrics for assessing correlations between genotype and clinical outcome. One way of solving this problem would be to perform SNP or haplotype assessment once the total number of candidate gene SNPs and their resulting haplotypes has been determined. Using this approach, the metric that produces the fewest measurable states should afford the greatest statistical power (J.S. Bader, personal communication).

Although it is still early days, it is clear that pharmacogenetics will transform future medical practice. It can already be used to increase the efficiency of the preclinical phases of drug development by enabling parallel screening of all functionally distinct drug receptor variants. Using such approaches, drug companies can select drug candidates that demonstrate equal levels of efficacy across all genotypes; similar strategies are currently being explored to eliminate variability in compound metabolism and excretion.

The application of genomic tools and analysis strategies to pharmacogenetics will undoubtedly increase the number of drugs for which preemptive diagnostics will establish correct patient dosing. Through its publication, The International SNP Map Working Group data set has established the baseline standards for all future pharmacogenetic research.

1. The International SNP Map Working Group. *Nature* **409**, 928–933 (2001).
2. Lazarou, J., Pomeranz, B.H. & Corey, P.N. *JAMA* **279**, 1200–1205 (1998).
3. Evans, W.E. & Relling, M.V. *Science* **286**, 487–491 (1999).
4. Yates, C.R. *et al. Ann. Intern. Med.* **126**, 608–614 (1997).
5. Krynetski, E.Y. & Evans, W.E. *Pharm. Res.* **16**, 342–349 (1999).
6. International Human Genome Sequencing Consortium. *Nature* **409**, 860–921 (2001).
7. Nickerson, D.A. *et al. Nat. Genet.* **19**, 233–240 (1998).
8. Levine, T.B. *et al. Circulation* **101**, 758–764 (2000).
9. Bittar, N. *Clin. Ther.* **19**, 954–962 (1997).
10. FDA Talk Paper. *Rezulin to be withdrawn from the market*. (FDA, Washington, DC, March 21, 2000).
11. Drysdale, C.M. *et al. Proc. Natl. Acad. Sci. USA* **97**, 10483–10488 (2000).