INTRODUCTION

The SNP Consortium (TSC) was established in the spring of 1999 as an innovative collaboration among major pharmaceutical companies, the world’s largest medical research charity [the Wellcome Trust], and five leading academic centers. The TSC’s focus has been to advance the field of medicine and the development of genetic-based diagnostics and therapeutics, through the creation of a high-density single nucleotide polymorphism (SNP) map of the human genome. This map will be a powerful research tool to enhance the understanding of disease processes and facilitate the discovery and development of safer and more effective medications.

SNPs are emerging as a critical tool for human genetics in general and pharmacogenomics in particular. It is becoming clear that human susceptibility to common disease and response to treatment is influenced by common DNA sequence variations. Although there is tremendous variation in the human population, it is estimated that the common SNPs which likely underlie common disease constitute a finite collection of perhaps 3–4 million total variants. The value of SNPs resides in their simplicity, frequency, and distribution throughout the human genome (the full set of genetic instructions, encoded in long strands of DNA). There are roughly 3 billion nucleotide pairs (the “building blocks” of DNA, also called “base pairs”) in the human genome, 99.9% of which are the same for everyone. A variation in a single base—that is, a SNP—occurs roughly every 1000 bases. SNPs are the most common form of genetic variation.

Prominent examples of the utility of SNPs include ApoE in Alzheimer’s disease; CKR5 in susceptibility to infection by HIV; factor V in risk of deep venous thrombosis; MTHFR in cardiovascular disease and neural tube defects; p53 in HPV infection; and, of course, various cytochrome p450s in drug metabolism and HLA in autoimmune disease. A representative SNP map should be useful in identifying human disease susceptibility either directly via “association studies” or indirectly via linkage disequilibrium studies. The former studies test for enrichment of a specific allele in susceptible individuals. Stated differently, SNP patterns from a target population—such as patients who suffer from a particular disease or who respond poorly to a particular drug—are compared with SNP patterns from unaffected populations to find genetic variations shared only by the affected group. It is from these association studies that disease-specific genes can be identified, and from which novel therapeutic avenues and even tailor-made therapies are expected to evolve. Linkage disequilibrium studies identify the presence of common ancestral chromosomes among susceptible individuals.

To summarize, SNPs should help pinpoint the subtle genetic differences that predispose some but not others to diseases such as arthritis, Alzheimer's, cancer, diabetes, and depression, and underlie variability in individual responses to a given drug. In turn, such knowledge will spur the development of new medical therapies to better treat, or even prevent, illness. Through the TSC, it was expected that a high-density, high-quality map could be created more quickly, and with shared financial risk and less duplication of effort than if each company had pursued development of a SNP map on its own. Now that it is created, the map provides important data for industry scientists searching for new ways for medicines to cure and prevent disease, and for medicines that more precisely target specific patient groups.

Moreover, the availability of the map to academic, government, and independent medical researchers worldwide should enable investigation of genes associated with rare diseases, which otherwise might not be feasible because of the significant investment that would be required.

THE SNP CONSORTIUM: A BRIEF HISTORY

The history of TSC goes back to early 1998. The Consortium emerged from an initiative by Glaxo Wellcome to establish a “private entity” to pursue the following three important genomic activities: (i) Identify and map enough SNPs to create a high-quality, genome-representative map; (ii) develop alone or in conjunction with other companies a cost-effective technology...
for “scoring” (genotyping) high numbers of SNPs in a reasonable time frame; and (iii) develop “multi-point” software programs to facilitate the analysis of pharmacogenomic studies.

After a series of meetings throughout the summer and fall of 1998 with major pharmaceutical companies and the Wellcome Trust, the group decided to establish the TSC as a nonprofit consortium. Numerous options were considered (e.g., waiting for academic developments to progress, working to accelerate other public domain efforts, and leaving the work to existing private firms), but in the final analysis, it was decided the nonprofit, consortium model was the most efficient way to proceed. Organizing in this manner would greatly facilitate the member recruiting, simplify operations, and allow the TSC to be exempt from taxation, thereby increasing the funds available for research. The central concept was to establish the TSC as an entity that would function in the public good. This was accomplished by making the data generated from the TSC’s efforts available to all parties, at the same time, and at no cost. Dues paying members of the TSC would gain access to the data under the same terms and conditions as nonmembers.

The spirit of these early deliberations and decisions also led to a narrowing of the Consortium’s activities and focus, versus those initially envisioned by Glaxo Wellcome. The focus became the identification and mapping of enough SNPs to create a high-quality, genome-wide map.

The initial TSC funding ($53 million) was provided by the Wellcome Trust ($14 million) and ten pharmaceutical companies, including Bayer Group AG, Bristol-Myers Squibb Company, Glaxo Wellcome PLC, Aventis, Monsanto Company, Novartis AG, Pfizer Inc, Roche Holding Ltd., SmithKline Beecham PLC, and Zeneca Group PLC. In addition, Motorola, Amersham Pharmacia Biotech, and IBM became members of the TSC, reflecting the commitment of the TSC mission to building a membership beyond the pharmaceutical field.

The academic centers involved in the project include the Whitehead Institute/MIT Center for Genome Research, Washington University, the Sanger Center, Stanford University, and the Cold Spring Harbor Laboratory [CSHL]. Using DNA from a diversified, representative panel of anonymous volunteers, sequence information from the publicly funded Human Genome Project, and advanced sequencing technologies, the Whitehead Institute, Washington University, the Sanger Center are identi-
to re-allocate the work across the three centers if agreed upon quarterly performance levels are not achieved by any center.

SNP mapping is being performed in silico by CSHL to the extent permitted by the draft human genome sequence. In silico mapping was expected to yield conservatively 24,000 SNPs from the 97,500 SNPs identified in year one. Additional mapping of 36,000 SNPs was conducted by the radiation hybrid method at Stanford University and the Sanger Center. SNP mapping in year two was restricted to that which was achievable in silico, with a conservative year two target of 111,000 additional mapped SNPs.

The TSC research timetable is summarized in Figure 2.

TSC SNP QUALITY MANAGEMENT PROGRAM

To deliver on these objectives, an effective quality management program was critical to ensure the identified SNPs were of sufficient quality to be useful to researchers in conducting association studies. The TSC is executing a program to generate overall quality measures for the SNPs prior to release to the public.

The TSC quality management program consists of the two core activities: Genome Sequencing Centers ["GSCs"] Quality Verification and Overall Quality Verification of TSC SNPs by an external, independent party.

Quality Management Activity I—GSCs Quality Verification

The GSCs are obligated to identify a predetermined (minimum) number of SNPs within each of the eight research production quarters. In addition to these quantity objectives, there is an agreed upon quality objective. The TSC’s SNP identification quality standard is to achieve 95% confidence or greater that all of the SNPs submitted for mapping are in fact real, legitimate SNPs. To document performance to this quality standard, the GSCs verify 100 randomly selected SNPs each quarter, by sequencing the individual DNA samples. These data are provided by the GSCs in each quarter of the research program.

It was also deemed important to have a second, independent quality program to document the quality of the “pooled” population of SNPs.

Quality Management Activity II—External Independent Verification of GSC-Identified SNPs

All three genome centers are using similar but different software tools to estimate the quality of their SNPs. The centers are submitting to the Data Coordination Center (DCC) at CSHL all of the putative SNPs identified by reduced representative sequencing that have >95% possibility of being truly polymorphic.

The TSC is working with an external genotyping firm, Orchid BioComputer, to format, assess, validate, and develop a working assay for a minimum of 4500 of the TSC’s pooled SNPs. The samples are randomly selected from the combined outputs of the genome centers, including SNPs generated by the DCC from publicly available genome sequence. The number of SNPs being validated from each GSC are proportional to the percent contributed by each center.

THE DCC AND DATA RELEASE POLICIES

The TSC’s data need to be assembled into a uniform format, checked for consistency and completeness, and maintained in confidence until its release. The DCC is responsible for these activities.

The DCC has six major areas of focus:

- **Maintain the TSC SNP database**—The SNP database is the central repository of all SNPs discovered during the course of the project. The information maintained in the database includes the nucleotide sequence upstream and downstream of the polymorphism, the polymorphism itself and its alleles, the sequencing trace files used to discover the polymorphism, PCR assays for the polymorphism, confirmation data, and mapping information derived from RH mapping and/or alignment with genomic sequence.

- **Import information from the genome centers**—Information is collected in a continuous manner from the genome centers and imported into the database. This involves file transfer and format conversion, which must occur in a timely and automatic fashion. After importation, but prior to release, the data is embargoed until the SNP can be fully mapped. During this period, no one outside the DCC has access to the data, except the mapping centers.

- **Export information to the mapping center**—Each quarter, a subset of the confirmed SNPs are exported to the mapping centers for radiation hybrid mapping.

- **Create an integrated map of SNPs, genes, and genomic sequences**—The DCC combines mapping information from the mapping centers and public sources to create an integrated map that depicts the SNPs in the context of genes and other important genomic landmarks. As the human genomic sequence neared completion, it became increasingly possible to map candidate SNPs to the genome by sequence alignment. The DCC periodically scans the public genome databases for matches against SNP sequences and incorporates the positional information gained in this way into the SNP database.

- **Release data to the TSC and the public**—The data are released simultaneously to the TSC, the genome centers, and the public at large on a quarterly basis. Data release involves export-

<table>
<thead>
<tr>
<th>Objectives and Timetable:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Human Genome Project:</strong></td>
</tr>
<tr>
<td>Working draft (90%) by Q2 2000</td>
</tr>
<tr>
<td>SNP Identification:</td>
</tr>
<tr>
<td>1Q99  2Q99</td>
</tr>
<tr>
<td>97,500</td>
</tr>
<tr>
<td>2001</td>
</tr>
<tr>
<td>2Q01</td>
</tr>
<tr>
<td>202,500</td>
</tr>
<tr>
<td>3Q01</td>
</tr>
<tr>
<td>300,000 SNPs</td>
</tr>
<tr>
<td>SNP Mapping:</td>
</tr>
<tr>
<td>1Q99  2Q99</td>
</tr>
<tr>
<td>18,000</td>
</tr>
<tr>
<td>3Q01  1Q00</td>
</tr>
<tr>
<td>18,000</td>
</tr>
<tr>
<td>2Q01  2Q00</td>
</tr>
<tr>
<td>36,000 (RH)</td>
</tr>
<tr>
<td>3Q01  1Q00</td>
</tr>
<tr>
<td>24,000</td>
</tr>
<tr>
<td>2Q01  1Q00</td>
</tr>
<tr>
<td>111,000</td>
</tr>
<tr>
<td>3Q01  1Q00</td>
</tr>
<tr>
<td>135,000 (in silico)</td>
</tr>
</tbody>
</table>

Figure 2. Objectives and timetable.
ing the data to the National Center for Biotechnology Information (NCBI), and making the SNP database accessible on the internet. The data coordination and data flow are summarized in Figure 3.

- Provide graphical and programmable interfaces to the SNP database—The SNP database is made available to the public via a browseable Web interface. This interface provide access to SNP sequences, confirmation data, trace data, and mapping data, using a combination of dynamically generated HTML pages, graphical images, and Java applets.

INTELLECTUAL PROPERTY

The overall IP objective is to maximize the number of SNPs that (i) enter the public domain at the earliest possible date, and, (ii) are free of third-party encumbrances such that the map can be used by all without financial or other IP obligations. It must be clearly emphasized that the TSC intends to hold no IP at the end of the SNP mapping program. To meet the second objective, the TSC intends to withhold public release of identified SNPs until mapping has been achieved to prevent facilitating the patenting of the same SNPs by third parties.Mapped SNPs were publicly released quarterly (in year one) and monthly (in year two). The IP plan is intended to maintain the priority dates of discovery of the unmapped SNPs during the period between identification and release, for use as “prior art”.

The tenets of the TSC’s IP Program are:
1. The IP program is only a program to protect the identified SNPs until they can be mapped and released to the public.
2. On a regular basis the TSC will file provisional patent applications on all identified SNPs to establish priority dating.
3. Once fully mapped, the TSC will convert all provisional patent applications to utility applications.
4. As SNPs are released, the utility applications are converted to invention records.

RESULTS SUMMARY

The TSC was established as one of the first significant research collaborations within the pharmaceutical industry. It has been supported by 14 members, including The Wellcome Trust. All of the data generated by the TSC are protected until they can be fully mapped and released to the public. As of November 2001, the TSC successfully completed its release of SNP data to the public.

The program’s overarching goal was to provide a high-quality, representative SNP map of the human genome. This goal has been fully achieved. The program had as its original objective to identify 300,000 common SNPs; in fact, we identified 1.7 million SNPs. Of these, to date 1.5 million have been mapped to the draft sequence of the human genome (see Figure 4); 1.3 million have been mapped to a single position on the draft. The balance of the identified SNPs map to two or more locations of the draft sequence. As the sequence of the human genome is finalized, it is expected the majority of SNPs will be mapped to a single location.

In summary, the significance of TSC’s efforts, in the final analysis, are really twofold:

1. A high-quality, ordered SNP map of the human genome in the public domain will hopefully facilitate pharmacogenetic/genomic research by accelerating the identification of specific markers involved in both common and rare diseases. It will facilitate the discovery of new ways to intervene in disease processes, the development novel diagnostic tests, and the use of existing and the development of new medicines to personalize therapies for more effective clinical management of the patient.
2. The TSC initiative demonstrates the practicality and benefits of cross-industry research collaboration. This model will become increasing relevant to the life science industry as the cost and risk of basic research continue to increase.

Address correspondence to:
Dr. Arthur L. Holden
First Genetic Trust and
The SNP Consortium
3 Parkway North Center
Suite 150 North
Deerfield, IL 60015, USA
e-mail: aholden@firstgenetic.net

Figure 3. Data coordinating summary.

Figure 4. SNP identification results.