Making the right choice of targets at the beginning of the pipeline will be the first step down the long road of creating innovative medicines

Keynote review: Disease-specific target selection: a critical first step down the right road

Allen D. Roses, Daniel K. Burns, Stephanie Chissoe, Lefkos Middleton and Pamela St. Jean

Relevance of a drug target for a disease is often inferred with strong belief but fragile evidence. Here, a program for early identification of human disease-specific drug targets using high-throughput genetic associations is described. Large numbers of well-characterized patients (>1000) and matched controls are screened for genetic associations using several thousand (>7000) single nucleotide polymorphisms from more than 1500 genes. The genes were selected because they are members of target classes for which there are precedents for high-throughput chemical screening technology. This review summarizes the methods and intensive data analyses leading to target gene identification for type 2 diabetes mellitus, including the statistical permutation methodology used to correct for many variables.

Discussions of target identification for drug discovery have become technically oriented and complicated over the past decade [1,2], a period of time that coincides with decreases in productivity across the pharmaceutical industry [3–6]. The latest technologies for selecting targets can be fascinating and imaginative, but are these targets relevant to treating human diseases [7,8]? Although methods for high-throughput chemical screening and for optimizing lead molecules have undoubtedly made major advances, target selection remains a crucial step [9]. Currently, there is no strategy of equally high-throughput for the selection of targets that are directly associated with human diseases. This need becomes even clearer when animal models are considered the ultimate target validation [7,10].

The drug discovery process should be clear-cut – identify the best molecule, for the most effective treatment, as fast and efficiently as possible. Although high-throughput molecular methods have progressed, matching the molecule to the appropriate clinical indication

Allen D. Roses*
Daniel K. Burns
Stephanie Chissoe
Pamela St. Jean
GlaxoSmithKline R&D,
Research Triangle Park,
NC 27709, USA
*e-mail: allen.d.roses@gsk.com
Lefkos Middleton
GlaxoSmithKline R&D,
London, UK

ALLEN D. ROSES

Allen D. Roses was appointed Senior VP for Genetics Research at GlaxoSmithKline (GSK) in 2000. Before joining GlaxoWellcome in 1997, Roses was at Duke University, where he was the Jefferson Pilot Professor of Neurobiology and Neurology and Director of the Center for Human Genetics. Roses led the team that identified apolipoprotein E as a major, widely confirmed susceptibility gene in common late-onset Alzheimer's disease. While at GSK, he was charged with organizing genetic strategies for susceptibility gene discovery, developing and implementing pharmacogenetics approaches and integrating genetics into medicine discovery and development; translation of genetic and genomic research into pathway analyses, drug discovery and pharmacogenetics in development is ongoing at GSK.

DANIEL K. BURNS

STEPHANIE CHISSOE

LEFKOS MIDDLETON

PAMELA ST. JEAN

1359-6446/04/$ – see front matter ©2005 Elsevier Ltd. All rights reserved. PII: S1359-6446(04)03321-5

for development is still an arduous task: disease-relevant target identification can now be an important criterion for determining the correlation between molecule and indication. The foundation of the successful reputation of the 1980s and early 1990s, when newly discovered biological pathways and receptors were selected as targets, was a rich biological literature identifying the 'low-hanging fruit' [11]. Biology has continued to direct targets to a small extent, which is probably best illustrated by biopharmaceuticals [12]. For small-molecule screening over the past decade, the selection of targets using animal models, simple organism genomics and genetically manipulated animals has a less than successful track record [13,14] – why not return to patients to select targets for their diseases?

If the speed and efficiency of identifying targets that are relevant to human diseases were improved, then the technical advances for lead validation, high-throughput chemical screening and lead optimization could be applied to molecules with a greater probability of success. For example, polymorphic variants of target genes might have different interactions with lead molecules. Studies of mechanism, chemical lead validation and on- or off-target effects can be conducted in directed mouse models [15]. Although leads from mouse model targets are frequently ineffective in human clinical trials (this is particularly true in cancer and CNS diseases [2]), for most discovery research, animal models set a gold standard for target selection [7,16,17]. A complimentary approach would be to identify first those targets that are genetically associated with human diseases and then create appropriate knockout and conditional knockin models with target gene variants. This suggestion differs significantly in the scope and depth of human phenotyping from proposals to phenotype a broad catalogue of knockout mice to suggest targets [10].

Here, early (but not preliminary) data are presented from the application of a high-throughput human disease-specific target program (called HiTDIP within GlaxoSmithKline) that focuses on the association of tractable targets with specific patient groups. The principles of complex gene association studies and disease susceptibility will be addressed, particularly with respect to matching targets with clinical indications. Several reviews have recently been published that cover genetic association studies [18–26], and these broad, complex, and sometimes esoteric, disciplines are not discussed here. Rather, the focus of this article is the strategy and process of genetic association studies to match pharmaceutical targets with clinical indications related to several human diseases.

The pharmaceutical industry and its business analysts track attrition at various stages in the drug discovery and development pipeline. For example, 95% of candidate quality leads fail to produce a medicine. Of the molecules that enter Phase I clinical trials after surviving preclinical testing, only 21.5% reach the market [27]. Furthermore, the number of new molecular entities (NMEs) – drugs with a novel chemical structure – submitted to the FDA over the past decade has decreased (Figure 1) [13].

Candidate leads evaluated at Phase IIA for efficacy represent products of target selection generated during the past ten years, generally before the possible contributive effects of the completion of the human genome sequence could be realized. The major sources of attrition after entering Phase I trials are toxicity and lack of efficacy. One method of decreasing attrition at an early clinical stage is to apply prospective efficacy pharmacogenetics (PGx) at Phase IIA [3]. Another strategy is to select the right target initially.

The pharmaceutical applications of HiTDIP methodologies were adopted as a response to the apparent increased failure rate of clinical development that has plagued the industry over the past decade. A basic hypothesis for this attrition might be stated simply as: 'A reason that many molecules are ineffective in clinical trials is that the selection of the originally screened target was based on data and rationale that are actually irrelevant to the etiology or pathogenesis of the human disease'. For a drug to be successful in treating a disease, two variables must be matched – the target and the right therapeutic indications. The success rate of discovery of molecules from screens is therefore directly related to target choice, and targets identified using animal models and sophisticated genomic analyses have so far provided fewer molecules for NME submission than anticipated [28,29].

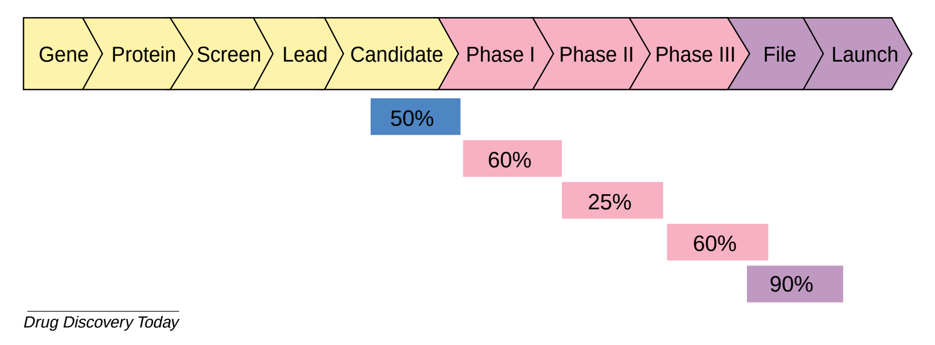


FIGURE 1

NCE success ratios: probability of progressing through each phase. Only one in 25 NCE candidate compounds is approved by the regulators (Table 1). Note that only 25% of those molecules that survive Phase II successfully pass into Phase III. Molecules derived from genetically-associated targets could increase the success rate. A small increase would have a significant effect on approvals. The addition of safety pharmacogenetics can also contribute to decreased attrition during development [60; http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid=4].

A critical view of genomic applications for target identification

Analysis of the human genome sequence promised to bring a flood of new targets,

TABLE 1

The pharmaceutical pipeline: definition by milestone

R&D stage (milestone passed)	Description
From target identified to screening hit and/or lead compound	Success[a] in identifying a compound with the desired pharmacological activity at the desired molecular, cellular or mechanistic target: this compound might not have all the characteristics required to be a viable drug
From lead compound to drug candidate	Success in identifying a viable drug candidate by optimizing the characteristics of the initial screening hit and/or lead; typically, this requires appropriate potency, selectivity and efficacy but can also involve other criteria such as bioavailability, metabolic stability and preliminary safety screening
From drug candidate to FTIH and/or Phase I	Success in progressing a drug candidate into initial studies in humans, usually (Phase I) in healthy volunteers
From Phase I to entry into Phase II	Success in progressing a clinical development candidate into small-scale exploratory studies in patients that have the targeted disease
… to proof of concept	Usually considered to be the point at which a drug candidate has demonstrated efficacy in its intended patient population, typically within Phase II: therefore, project attrition can be measured for progression to or from this milestone, as well as to or from the more traditional clinical development milestones (Phases I–III)
From Phase II to entry into Phase III	Success in progressing a medicinal candidate into large-scale (pivotal) clinical trials suitable for registration
From Phase III to regulatory filing	Success in progressing a data package into a regulatory submission
From regulatory review to approval	Success in gaining regulatory approval
From regulatory approval to launch	Success in launching an approved product, added indication and/or label-change

[a]Project progression can be quoted as attrition (failure) or success. Abbreviation: FTIH, first time in human.

and therefore increase the potential throughput of the pharmaceutical pipeline [28–31]. This scenario has also proved disappointing in not reaching the projected goals. What factors have limited target selection and drug discovery productivity? Although HTS technologies were successfully implemented and spectacular advances in mining chemical space have been made, the universe for selecting targets expanded, and in turn almost exploded with an inundation of information. Perhaps the best explanation for the initial modest success observed was the dramatic increase in the 'noise-to-signal' ratio, which led to a rise in the rate of attrition at considerable expense. The difficulty in making the translation from the identification of all genes to selecting specific disease-relevant targets for drug discovery was not realistically appreciated. There was certainly a large pharmaceutical industry investment in the intellectual property speculation surrounding the sequencing of all possible genes, with the hope that a stream of new targets for drug discovery would result. Senior R&D scientists recently recognized the need for a 'quantal step-up in discovery' [32]. To feed a high-throughput pipeline, a high volume flow of specific, disease-relevant targets is necessary. Whether or not individual researchers believe in a particular disease hypothesis, in the specific relevance of a target class to some aspect of human disease pathogenesis or in particular animal models of human disease, the evidence that 'validates' (substitute believe in, consensus view or champion, among others) the choice of a target molecule for a potential therapeutic strategy in humans is crucial to starting down the right road.

Target validation is one of those terms that scientists use in multiple ways. With respect to target identification and selection in the pharmaceutical industry, validation is interpreted as providing increased confidence to initiate expensive chemical screens and subsequent discovery programs. On occasion, support of possible relevance comes from the sheer weight of 'potentially' (substitute believed in, validated, rational or accepted, among others) relevant information. For example, a target gene could be determined to be expressed in the tissue that is affected pathologically by a particular disease, differentially expressed in disease-relevant tissues or have a visible effect in animal models when manipulated [33]. These data might provide modest human disease-specific support as the starting point for a drug development program to treat a particular disease. A putative target located on neuronal surfaces could be relevant to a human neurological or psychiatric disease – but which one? Proof of concept in humans can only occur on completion of preclinical testing and Phase I safety studies of an optimized lead candidate. A Phase II clinical trial is an extremely costly hurdle with which to justify the target choice after many years of confidence-building research. The situation could be compounded when a molecule with exceptional drug qualities, but an unclear clinical indication, is tested in several clinical trials involving multiple clinical endpoints.

The success of a target is judged after many years – usually in hindsight by counting marketed products. The success rate of efficacy (proof of concept) studies is probably a much earlier indicator of pipeline health. If the right target was selected more often and this led to the selection of effective lead candidates more frequently, then attrition would be reduced. Shots on goal are good, but center forwards who miss 99% of the time (and take all the shots) are not hired by professional teams. The pharmaceutical industry must move away from the numbers game and strive for specificity and speed at the earliest stages of the pipeline. Targets for specific diseases that are chosen based on strongly held beliefs have a significant probability of being the totally wrong target. Current 'knowledge' (substitute accepted beliefs, strongly held views or reasonably good ideas, among others) might not define disease-relevant hypotheses accurately. It is the result – an effective and safe medicine – that is of importance to the patients, physicians and industry. Indeed, the mechanisms of action of many successful medications are still unknown.

Can target molecules be directly associated with human diseases that have highly statistically significant data? Yes. Can chemical leads be produced from screening these targets that can enter the pharmaceutical pipeline? Yes. Will lead candidates produce a higher rate of future success in demonstrating efficacy compared with current metrics? A decision on this aspect has yet to be reached – more time is needed to study the flow through the pipeline into human testing. However, when a genetically associated target gene has already been screened chemically, there can be a rapid progression from identification of the target to the entry of lead molecules into clinical development.

Gene-specific target association study design

With the human genome sequenced, it is possible to define virtually all genes belonging to the known target classes using analogous sequence regions that define specific structures or functions. However, there are few real indications as to which gene might be specifically associated with a particular disease. It is possible to test each gene individually for disease relevance using genetic association studies, but only if sufficiently large and well-characterized patient and control groups are available. What of high-throughput association studies of many sequence variations within all genes of each target class? As an example hypothesis, assume there is a G-protein-coupled receptor (GPCR, sometimes referred to as a 7-transmembrane repeat) target class gene variant that is associated pathologically with Alzheimer's disease. Which GPCR is it? If there are almost 500 known GPCR genes, then is the Alzheimer's disease-specific GPCR the third on the list? Is it number 222, or maybe number 407? High-throughput, gene-based single nucleotide polymorphism (SNP) genotyping technologies provide the opportunity for the rapid testing of each of the GPCR variants for disease association. Genetic association studies provide an evidence-based opportunity to inform target choice rapidly and more specifically.

There is an implied crucial assumption when using a gene-disease association strategy: disease-specific associations might be identified for genes that are selected simply because it is known how to screen them against large chemical libraries. Seven years ago, the scientific community was reluctant to take such a risk. However, more recently, retrospective data were published that support this assumption. Goldstein *et al.* [34] examined 42 sequence variants of genes that had been associated with a drug response at least twice. These investigators found that 21 of the 42 variants were in the target or in a known pathway of the target. It is therefore reasonable to propose that the probability of identifying a disease-relevant target would be increased by screening all potential targets for genetic association with well-defined diseases. That is to say, there are now data available to support the hypothesis that candidate leads derived from genetically associated targets can increase the probability of success (decrease attrition) at Phase II or Phase III of clinical trials.

Another important scientific contribution to gene-disease association studies has come from advancements in the fields of genetic epidemiology and statistical genetics. These are specialized disciplines to most of the pharmaceutical industry, but the ability to analyze rapidly many clinical traits simultaneously in a high-throughput fashion, and with appropriate methods to define statistical significance, is probably the most useful contribution to such studies other than the sequenced genome template. By 2002, this advancement made it possible to test therapeutic class genome-wide variants for statistical associations with particular human diseases. Additional support for a disease gene-association strategy can be found in the plethora of recent studies linking specific gene variants to particular diseases [35–40]. When the effect of the variant results in expressed clinical disease, it is generally viewed as a disease mutation. Where multiple variants of several genes contribute to the expression of a disease, they are now commonly referred to as susceptibility genes. The practical problem of solely studying disease genetics to generate targets is that most susceptibility genes are not drug targets, and therefore the high-throughput methodologies currently available cannot be used to screen for the formation of chemical interactions [41].

Would the initial identification of high-throughput targets with human disease-specific associations result in a more efficient pipeline with less attrition? Because there was great confidence that the human genome projects (both public and private) would eventually provide gene sequencing, and some variant, information, it was anticipated that a large, low-throughput resource would be needed; that is, prospective, well-phenotyped patient collections and appropriate controls, each of which would have consent for commercial applications.

The patients define the relevance

To study the association of gene variants with the clinical expression of human diseases, a large number of consenting patients and controls must be carefully examined and the data placed into accessible databases; in addition, DNA must be collected and stored. Because the clinical examination of patients is performed one patient at a time, the generation of large patient collections that are

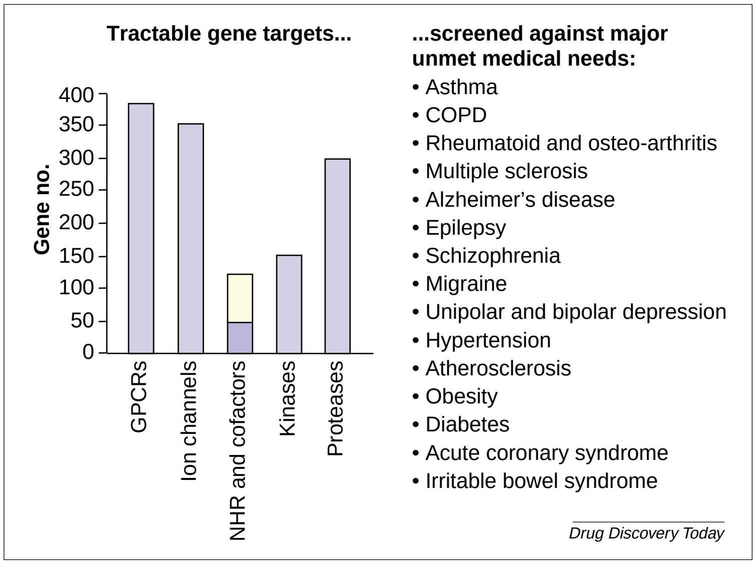


FIGURE 2

HiTDIP: genetic associations between 'tractable' targets and major diseases. High-throughput analysis of approximately 7000 polymorphisms in 1800 candidate genes (numbers of validated SNPs and candidate target genes have increased over time) are screened for association in common diseases, such as asthma, schizophrenia, depression, osteoarthritis, Alzheimer's disease, metabolic syndrome, hypertension, acute coronary syndrome and others. The approximate numbers of genes in various target-classes used in these experiments are also indicated. Genes of marketed products and other target genes such as enzymes and protein ligands are also included in the target classes. As more target groups become tractable, they are added to the screen. Between 2002 and 2004, the gene list has expanded from approximately 1450 genes to >1800. For the NHR and cofactor column, the yellow color represents the number of NHR cofactor gene targets. Abbreviations: COPD, chronic obstructive pulmonary disorder; NHR, nuclear hormone receptor.

suitable for disease association studies using multiple markers is time- and resource-intensive, and consequently is not high throughput. Anticipatory clinical research of this type requires sustained access to clinical expertise, extensive supporting resources and commitment over a period of years. Few such prospective collections exist and, in those academic environments that possess such databases, informed consent for commercial uses is usually absent.

In 2005, the genome is sequenced, target class gene variants are known, rapid genotyping technologies are available and interactive databases with analytical capabilities have been built. Since 1997, GSK has organized external clinical specialists, who are expert in more than a dozen important diseases, and has accumulated more than 80,000 patients and controls, each examined with a prospectively standardized protocol, informed consent and stored DNA samples. Several association experiments were completed that provided exciting new putative targets for pipeline consideration. The leading edge of chemical leads has begun to enter the pipeline.

As molecules resulting from insights of HiTDIP reach the published portfolio of GSK over the next few years, there will be a relatively straightforward method for the comparison of attrition with historical metrics. With respect to the gene variants identified for early drug discovery, specific disclosure of early targets and leads in the pipeline could be limited by regulatory and commercial concerns. However, the best and most rapid biological and genetic validation of gene variants associated with disease can occur where the data are available for confirmation by academia and industry. It is therefore planned that these large association experiments will be published in scientifically reviewed journals to enable the full weight of academic and industry disease-specific target validation to be focused in this area. Furthermore, pharmaceutical companies principally share many of the same targets – why not compete on the screening and lead chemistry of disease-relevant targets and well-designed drug development?

High-throughput disease-specific target discovery – the experiment

To perform this experiment for the identification of the targets associated with human disease, three major components are required: (i) selection of the gene targets to be screened; (ii) well-characterized clinical data; and (iii) genetic data generation and statistical analyses.

The targets

Before the sequencing of the human genome, the pharmaceutical industry knew of perhaps 500 targets [42]. Widely appreciated target classes, for example, nuclear receptors, kinases and GPCRs (Figure 2) now constitute ~1200 genes, and there are many additional enzymatic screens that bring one estimate of the total druggable genome to over 3000 genes [43]. One method of dealing with screenable targets is to create knockout or other modified mice and search for phenotypes. This undoubtedly provides some additional support, but the phenotypic correlation between mouse and man can be difficult to interpret [17]. It would seem that direct associations with human disease phenotypes would be a promising and efficient point to search for knockout and conditional knockin models. Kola and Landis [2] listed five places along the pharmaceutical pipeline where attrition might be managed. Their first point was that 'building the need to get very strong evidence for proof of mechanism into the discovery paradigm is critical'. In the past, this was possible because an extensive literature had developed as scientists concentrated on biochemistry, physiology, pharmacology and other disciplines before suggesting targets. If the 100 best-selling drugs are examined retrospectively, the targets were initially selected because of strong and confirmed (in the literature) biology, including studies in man [17]. The prediction that knockout models will have the same efficiency prospectively as that provided by retrospective analyses will only be possible with extensive and specific phenotyping of each knockout. Screening the phenotypes of many knockouts might be much more superficial [16]. Perhaps, the most efficient approach would

TABLE 2

Listing of family and case-control studies ongoing at GSK as of September 2004

Disease	No. of sites	Family or case-control	No. of subjects collected	Target no. of subjects
Asthma	14	Family	5909	5909
Alzheimer's disease	9	Case-control	1504	2000
COPD	11	Case-control and family	5121	5460
Schizophrenia	4	Case-control	1317	1953
Metabolic syndrome	6	Case-control and family	4842	4842
Osteoarthritis	8	Case-control and family	4137	5685
Rheumatoid arthritis	1	Case-control	1736	2600
Parkinson's disease	1	Case-control and family	3039	3039
Unipolar depression	9	Case-control and family	3781	3861
Obesity	2	Case-control	2019	2000

be to use knockout and conditional knockin mice with genetic variants as additional target validation for genes that are statistically associated with the expression of human diseases.

Industry-initiated partnerships, such as The SNP Consortium, and efforts to sequence the human genome rapidly catalyzed SNP discovery [44]; this information was used to identify validated SNP assays for the ~1800 targets that constitute, for example, the GPCRs and kinases. For large-scale, disease-association analyses, the testing of these variants requires two sets of reagents: (i) validated SNP assays of each gene; and (ii) DNA from clinically well-phenotyped patients, appropriate controls and integration with extensive bioinformatic support systems. The standard candidate gene association experiment is to test variations of a specific gene for association to a single defined human disease. For HiTDIP, it is possible to extend this experiment to hundreds of genes (~1800), using thousands of SNP variations (~7000), and to several diseases (17+) in independent test and secondary screens.

The clinical data

The rate-limiting step in the HiTDIP program is the examination, acquisition of consent and collection of DNA from large collections of well-phenotyped patients and controls. Over the past decade, the key practical problems encountered are that the majority of patient collections of DNA reside in academic laboratories or small countries, and chemical-screening libraries are located in industry. Institutional review boards require that patient populations have specific informed consent for the commercial use of DNA, thus rendering many collections of patients enrolled in academic institutions largely unusable for commercial HTS. Patient and control clinical evaluations are not high throughput, because they are performed on an individual basis. Therefore, even if the technical screening capacity were generally available, as it is now, the essential clinical populations with available, consented DNA samples are not.

Disease phenotype does not simply depend on a diagnostic label

Over the past seven years, GSK has sponsored an extensive series of disease-specific clinical collaborations. These networks eventually involved ~200 specialist physician collaborators and currently comprise at least 17 diseases in several ethnic populations. Large groups of highly phenotyped patients and controls have been collected for these associations: ten of the case-control or case-control family studies currently completed (asthma and type 2 diabetes were the first in 2003 and 2004, respectively) or scheduled to be finalized are summarized in Table 2. Productivity gains have resulted in an increase in genotyping capacity, thus six of the listed diseases are scheduled to be analyzed in 2005, with other diseases to follow on completion of clinical enrollments. The lag phase for HiTDIP analyses is the process of registering subjects (patients and controls). After that, the process is 'industrialized' with planned overlapping primary, secondary and, in some cases, tertiary screening blocks.

Unlike more common retrospective patient collections, where information is gleaned from records, the network physicians were required to agree on diagnostic criteria in advance of subject collection. Data are often missing in retrospective collections, which can give rise to speculation on the implications of a 'blank' answer in the record of an individual. GSK created a clinical database that encapsulates the complete clinical and demographic information about each patient and each control prospectively: a prospective study ensures that data on all individuals are collected, and the database is as comprehensive as possible.

Each network of clinical experts established a core database of phenotypes with defined clinical descriptions for each disease, with the added caveat that additional supplementary clinical information that any participating clinical investigator wanted to collect would also be included in the phenotypic database. This resulted in disease-specific databases encompassing clinical parameters agreed by all network physicians, in addition to the sub-sets collected with respect to the particular research interests of an investigator.

This database format provides the opportunity to test genetic associations with more granularity than simply diagnosis. The data can be analyzed using the physicians' agreed disease diagnosis, single symptoms or signs or groups of clinical findings. There can be several contributing pathogenic processes in complex diseases, any of which might not be expressed concurrently to produce active disease, but each of which might be associated at some level with similarly diagnosed patient populations.

For example, in studying type 2 diabetes mellitus, inclusion of ophthalmologic examinations, renal function indicators and a defined neurological physical examination enables specific sub-type association studies. If the association of SNP variants in those patients with peripheral neuropathy was required, then performing and specifically recording the presence or absence of ankle jerks on all patients and controls is more accurate than trying to select patients with retrospective lack of data. Similarly, if diabetic retinopathy were an interest, few retrospective-controlled studies would have this information.

Illustrative study example: asthma

Asthma was the first disease to be analyzed in HiTDIP against multiple phenotypic variations. The association of thousands of SNP variants was measured against five separate but correlated asthma-related traits in large family sets including: (i) physicians' diagnosis of asthma; (ii) atopy (positive skin prick tests); (iii) atopic asthma (physicians' diagnosis plus atopy); (iv) strict asthma (two or more classical symptoms and a positive methacholine challenge test or bronchodilator test); and (v) bronchial hyper-responsiveness (positive methacholine response at or below 10 mg/ml of methacholine) [37]. The subjects were ascertained prospectively, but retrospectively grouped for each set of clinical criteria. Association studies using high-throughput genotyping of SNPs from tractable targets to define associations with various clinical definitions have been productive.

When thinking of asthma as a complex disease and/or syndrome, it is understandable that there might be multiple genetic and environmental susceptibilities. Certainly, there should be no expectation that all or most of the phenotypic variables would be present in all cases, as might be expected for a specific genetic mutation in a highly penetrant, rare inherited disease. A comprehensive approach that improves this situation would be to test each of the clinical forms that occur within families for genetic association with particular genetic variants. Confirmation of association could be performed in a subsequent set of families as well as in series of sporadic cases.

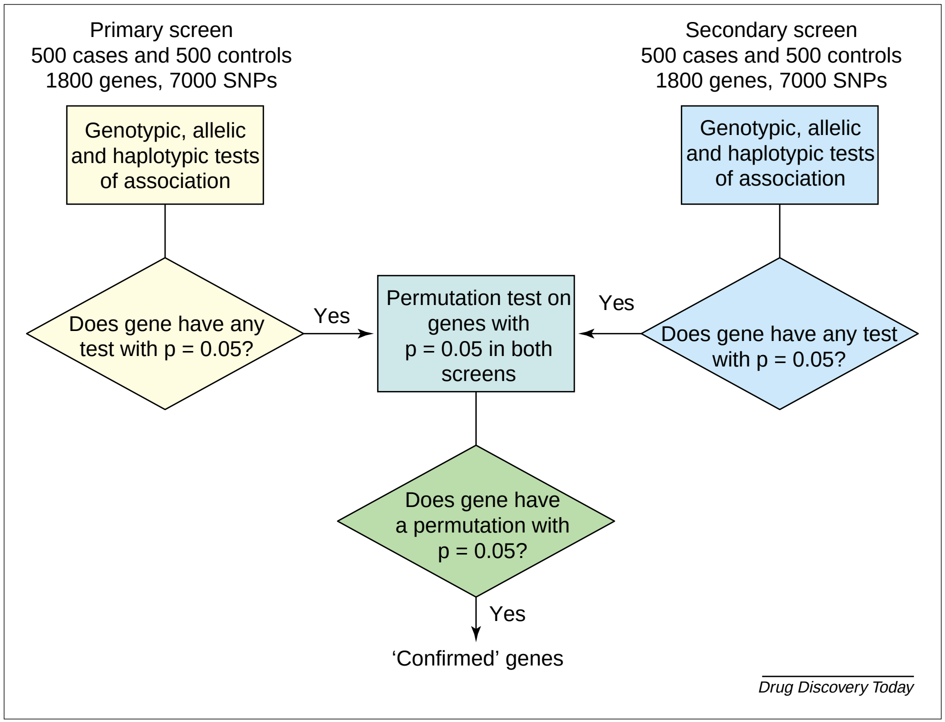
When the initial pilot asthma-screening was performed in 2001 with ~2700 SNPs from 1244 genes, interesting patterns of association were observed. Many of the associated gene variants were related to three or more of the selected clinical phenotype definitions. The association of some genes with physicians' diagnoses tended to separate to bronchial sensitivity-related signs and symptoms and atopy-related phenotypes. These early data regarding phenotypic criteria seemed to support the convergence of multiple susceptibility loci for the production of symptomatic complex disease.

In genetic linkage or association literature, there are frequently tacit assumptions, for example, that the expert physicians' diagnoses are sacrosanct, or that inclusion and/or exclusion criteria of any sort could often narrow the disease populations to be non-representative of all patients with the disease. Many of the conflicting reports of associations in the literature, where independent scientists have not 'confirmed' published linkage or association results with the same 'disease', could be the result of variations in the selected phenotypes or even the critical controls. With the same corps of physicians collecting the large prospective patient and control test series, as well as the subsequent confirmation series, variation between individual physician diagnostic skills can be minimized and the phenotype definitions can be stabilized.

Illustrative study example: metabolic syndrome

Metabolic syndrome is usually described as a combination of obesity, diabetes mellitus, hypertension and dyslipidemia – thus providing considerable opportunities for selection of who might be included in genetic studies. In 2002, The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation and Treatment of High Blood Cholesterol in Adults, Adult Treatment Panel III (ATP-III) published criteria f or a clinical diagnosis of the metabolic syndrome*. However, the GSK Genetics of Metabolic Syndrome (GEMS) patient collections for target association and susceptibility gene studies were initiated before this ATP-III report was created. By comparing the detailed phenotypic definitions for the collected patient information with those of the samples, and subsequently contrasting this with the definitions provided by a high level, independent expert panel, it was possible to analyze the relevance of our contemporary operational research definitions. In GEMS, two simple lipid-based criteria were used to define the disease-affected individuals: low high-density lipoprotein (HDL)-cholesterol and a concomitant elevation of plasma triglyceride concentrations. These criteria were selected because they are primary features of atherogenic dyslipidemia, are closely related to insulin resistance, detectable at an early stage in the development of metabolic syndrome, highly heritable and easy to measure. Wyszynski *et al.* [45] (clinical investigators supporting the GEMS Network) reported that 86% of individuals greater than 35 years of age met both the ATP-III and GEMS criteria. Conclusions based on genetic linkage, genetic association of susceptibility loci and PGx studies can thus be more accurately interpreted across studies. Accurate, interpretable, reproducible phenotypic definitions, even of complex syndromes mixing several complex diseases, can increase the value of clinical collections and facilitate analysis. Metabolic syn-

*NCEP criteria for clinical diagnosis of metabolic syndrome requires any three of the following: fasting plasma glucose of at least 110 mg/dl (6.10 mmol/l); serum triglycerides of at least 150 mg/dl (1.70 mmol/l); serum HDL cholesterol of less than 40 mg/dl (1.04 mmol/l) and 50 mg/dl (1.30 mmol/l) for males and females, respectively; and blood pressure of at least 130 mm Hg systolic and 85 mm Hg diastolic, or waist circumference (a measure of central adiposity) of more than 102 cm and 88 cm for males and females, respectively.


Primary screen
500 cases and 500 controls
1800 genes, 7000 SNPs
Secondary screen
500 cases and 500 controls
1800 genes, 7000 SNPs
Genotypic, allelic and haplotypic tests of association
Genotypic, allelic and haplotypic tests of association
Does gene have any test with p = 0.05?
Yes
Permutation test on genes with p = 0.05 in both screens
Yes
Does gene have any test with p = 0.05?
Does gene have a permutation with p = 0.05?
Yes
'Confirmed' genes

Drug Discovery Today

FIGURE 3

Experimental design algorithm for HiTDIP analyses. The basic design of the HiTDIP program was to screen approximately 500 patients and 500 controls, all prospectively examined, with full clinical information and commercial informed consent obtained. The current gene panel consists of approximately 1800 genes and 7000 validated SNP assays. Although highly significant p-values will be observed for some genes in the primary screen, the probability for occurrence of false positives is high. A secondary study of approximately 500 patients and 500 controls (obtained prospectively) is also tested. A gene with a p-value of ≤0.05 in primary and secondary screens is assessed by permutation (Box 1). Any gene with a permutation of p ≤0.05 is considered 'confirmed'. Calculations of random gene association assessed by permutation were confirmed (Figure 4).

drome is an excellent example of variable clinical definitions and hypotheses defining the disease.

Experimental data, analyses and statistical significance

During the past two years, up to ~1800 tractable genes represented by up to ~7000 SNPs were genotyped against several large patient and unrelated control groups for disease associations. The SNPs were identified predominantly from public databases, with focused SNP discovery performed as necessary. The SNPs selected were common, with 28% average minor allele frequency, and mapped typically to intergenic and promoter regions. Genotyping was performed using a multiplexed, bead-based approach published previously (Figure 2) [46,47]. A bioinformatic system, called SubjectLand, was designed and implemented to contain patient data, SNP (and other genetic variation) data, analyses programs and analytical results, with the ability to add-on data-mining capabilities. The standard experiment for each disease was designed to test ~500 well-phenotyped patients and ~500 matched controls in the initial, primary screen and follow-up with a secondary screen using an independent set of patients and controls ascertained by the same group of physicians.

The data were statistically analyzed using a gene-based approach (Figure 3). Allelic, genotypic and haplotypic tests of association were conducted in the primary and secondary screens. A 'fast fishers exact' test in SAS/BASICS® software was used for the allelic and genotypic tests [48], whereas the 'composite haplotype method' developed by Zaykin (unpublished results) was used for haplotypic tests. The use of a gene-based approach for analysis and replication, as used in HiTDIP, has recently been championed by Neale and Sham [49]. Cardon and Bell [50] stress that incorrectly adjusting for multiple testing can either unnecessarily reduce statistical power if too stringent a correction is applied or increase the false-positive rate if too weak a correction is used. As a result of the large number of tests conducted, adjustments for multiple testing were made using a gene-based permutation approach (Box 1).

The type 2 diabetes study made use of legacy (GlaxoWellcome, Burroughs-Wellcome, Glaxo, Smith Kline Beecham or Beecham) in-house collections – ~400 cases and controls were examined in the primary screen and >1100 cases and controls in the secondary screen. Among the 1405 genes examined in the primary screen, 256 genes had a p-value of ≤0.05. Of these 256 genes, 53 also had a p-value of ≤0.05 in the independent secondary screen but only 21 of the 53 genes were confirmed by passing the permutation process (Figure 4, Box 1). As well as conducting further investigations regarding the 21 confirmed genes that passed the permutation test and the 32 genes that did not, analyses of the probability that random genes would demonstrate similar statistical significance were also performed. Of the 21 permutation-confirmed genes, ten could be identified in pathways directly related to precedented mechanisms or metabolic pathways associated with disease-specific hypotheses. In the case of type 2 diabetes mellitus, four of the 21 genes had already been chemically screened in legacy companies and provided several leads. Statistical analyses of each disease will be submitted for formal peer-review in specialty journals by the appropriate network physicians.

In some diseases, such as type 2 diabetes mellitus, there are reasonable prior hypothesis concerning pathogenesis. For example, it is generally agreed that glucose metabolism and insulin sensitivity play a role. Among the HiTDIP type 2 diabetes confirmed genes, several are supported by prior published hypotheses and appear in expanded metabolic pathways. Such coincidences resulting from

BOX 1

The permutation testing process

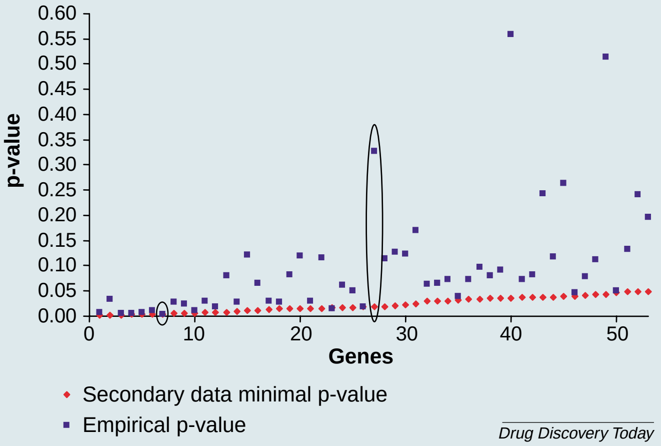
The objective of this study was to identify tractable genes associated with disease. Some small genes might have only one or two SNPs analyzed, whereas 30–40 SNPs could be assessed in the case of some of the larger genes, such as those encoding ion channels. The greater the number of SNPs and tests performed on a gene, the greater the probability that the gene will appear significant than by chance alone. To account and correct for the variable number of tests conducted across genes, a gene-based permutation test was applied. Permutation testing is a standard method used to assess significance in the statistical analysis of genetic data [57]. Any gene significant at a p-value of ≤0.05 in the primary and secondary screens was further assessed by performing this permutation process on the data from the secondary screen. For each permutation, affection status was shuffled among the cases and controls. The genetic data for all SNPs across each subject was not altered. This maintains the underlying correlation between SNPs within a gene. All the SNPs within the gene were analyzed using allelic, genotypic and haplotypic association tests via the same methods used for the observed data. The smallest p-value across all tests was recorded for each permutation. The permutations were repeated up to 5000 times per gene. The empirical p-value is the proportion of minimal p-values from the permutations that are less than the observed minimal p-value from the actual data. The empirical p-value is estimated using Equation i.

$$\frac{r+1}{n+1} \qquad \text{[Eqn i]}$$

where r is the number of permutation p-values as small as or smaller than in the actual data and n is the number of permutations [58,59].

A gene with an empirical p-value of ≤0.05 was considered to be 'confirmed' with respect to statistical association with disease. For example, if the smallest p-value of gene A in the observed data was 0.004 and among 5000 permutation (n = 5000) a p-value of ≤0.004 was observed 50 times (r = 50), then the permutation process would generate an empirical p-value for gene A of 0.01, and it would be classed as a 'confirmed' gene.

For the type 2 diabetes study assessed by permutation, the minimal p-value from the actual secondary dataset for each of the 53 genes analyzed and their empirical p-values from the permutation process are shown in Figure i (observed and gene-based permutation results). Of the 53 genes assessed, 21 genes had empirical p-values of ≤0.05 and thus were considered 'confirmed'. The observed and empirical p-values for gene 7, a confirmed gene, were highly similar, differing by only ~0.0001. For gene 27, which was not confirmed, the corresponding values showed striking differences: the permutated p-value was approximately 0.33, whereas the observed p-value was <0.02.



Drug Discovery Today

FIGURE i

Gene-based results for 53 type 2 diabetes mellitus genes assessed by permutation. The observed and gene-based permutation results for the 53 genes assessed by permutation in the type 2 diabetes data are rank ordered according to their observed p-values before permutation. Two of the 53 genes (genes 7 and 27) are circled to illustrate how the observed and empirical p-value can differ. It is useful to note that the first 12 genes were confirmed by permutation; it was expected – and quite reassuring – that the genes with consistently the lowest p-values would be confirmed. Permutation assessments confirmed four of the subsequent ten genes. Of the remaining 31 genes, five were confirmed. Hirschhorn *et al.* [18] reported that, by meta-analyses, only six of 166 genes replicated consistently across studies – again not unexpected in single-arm experiments, which are many studies with much smaller numbers of cases and controls and no opportunity to assess permutation (Box 2).

The gene-based permutation process controls for multiple correlated SNPs and multiple tests performed for each SNP in a gene in the secondary screen. A related question is – how many of the 1405 genes examined in the type 2 diabetes study would be expected to confirm under the null hypothesis of no association? That is, how many genes would have a minimal p-value of <0.05 in the primary screen and a permutation p-value of <0.05 in the secondary screen? To address the question of the permutation p-value, an additional round of 1000 permutations was conducted using the type 2 diabetes primary screen data. At each permutation, the affection status of the cases and controls was shuffled and the number of genes observed to have at least one p-value of ≤0.05 was recorded and multiplied by 0.05, the α level already enforced by the gene-based permutation applied to the secondary dataset. Of 1000 permutations performed on 1405 genes, the average number of genes that was confirmed by chance was 9.8 (95% confidence interval 8.6–11.0).

screening so many hypothesis-independent genes are, at the very least, interesting. This provides increased priority for screening specific targets in defined pathways and, subsequently, for designing relevant clinical trials. In diseases where there is a paucity of information to form the basis of supposition, for example, schizophrenia, specific genes and pathways identified in HiTDIP can provide insights to new hypotheses. Results for four other secondary screens (schizophrenia, obesity, migraine and unipolar depression) will be available in early 2005.

Additional experimental data for the association of each gene can also be generated by testing whether the target gene is within a region of extended linkage disequilibrium (LD) [51–53]. This is also important because the particular SNP used in the HiTDIP analysis might not be the variant responsible for the disease association, but has simply evolved concomitantly with the disease-variant. SNPs that are in LD with the causal variant can provide positive association data, as is the case for susceptibility genes for Alzheimer's disease, migraine, Crohn's disease and psoriasis [52–55]. Analyzing for regions of extended LD is also important to define whether the association signal is driven by the tested HiTDIP gene – and not by one of its neighboring genes [51]: documenting extended

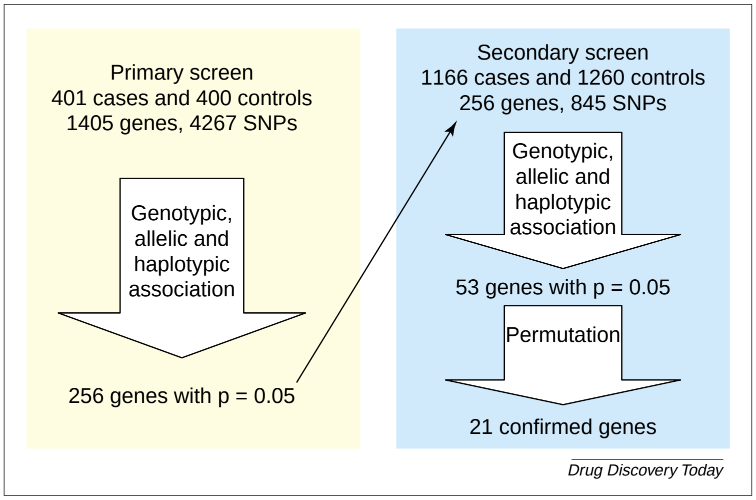


FIGURE 4

Type 2 diabetes mellitus HiTDIP results. The type 2 diabetes mellitus HiTDIP experiment was the first to be performed with two case-control screens. The prior asthma screens used a family-based series, thus the statistical analyses were significantly different. The primary screen consisted of 401 cases and 400 controls and used 4267 validated SNP assay for 1405 genes. A set of 256 genes yielded a p-value of ≤0.05 (a large number illustrating the high false-positive rate that is probable with a single screen). The secondary screen enrolled 1166 patients and 1260 controls. At the time that the secondary screen was initiated, the genotyping capacity was limited and thus only the 256 genes and their 845 SNPs were included. A set of 53 genes yielded a p-value of ≤0.05, and were assessed by permutation, with 21 genes being confirmed (Figure i). Estimation of the number of genes being confirmed by chance in this study yielded a rate of approximately ten genes per 1400 tested (Box 1).

LD around a HiTDIP gene can increase (or decrease) confidence in its validity.

It is reasonable to expect that a large screen with many tested variables could be subject to false positives. Although the same limitations also characterize susceptibility genes resulting from linkage and association studies, they are not usually tractable targets for chemical screening. Susceptibility gene studies provide new insights that are relevant to disease pathogenesis: we anticipate that the HiTDIP genes will also generate new theories and provide support for existing hypotheses. One immediate strategy could be to provide validation support by concentrating thorough phenotypic studies on knockout or otherwise manipulated mice [16]. For the type 2 diabetes example, this would mean knockouts and conditional knockins for ~21 genes and their specific variants associated with disease. Each mouse line could be examined with highly focused and complete biochemical and physiological phenotyping in parallel with high-throughput chemical screening.

There is no documented method for predicting which HiTDIP genes will point to lead candidates for medicines. The selection of genes with which to initiate high-throughput assay development is empirically based on several criteria, one of which is the relative ease of creating the chemical screening assay. A few confirmed genes for type 2 diabetes mellitus were screened previously in GSK legacy companies because of pre-existing literature rationales for potential involvement in disease pathogenesis. Enlarging or repeating the chemical screens with the current GSK compound libraries is an immediate option, as is re-evaluating earlier lead programs. In addition, there are several genes that the literature indicates are involved in interesting, potentially novel mechanisms. There are other genes, particularly those for which the design of high-throughput assays is relatively uncomplicated, that can be prioritized. In the GSK R&D structure, scientists in the Centers for Excellence in Drug Discovery (CEDDs), who are most familiar with the disease, prioritize the targets.

It is particularly pertinent to mention here that GSK is now dealing with a relatively large, staggered flow of new target genes for which there is excellent support for genetic association. If the leads from screened HiTDIP genes reduce attrition, this will be evidenced in a larger proportion of positive Phase IIA efficacy studies over the next 3–5 years. Remember, hits and leads must be optimized and undergo preclinical regulatory testing before human testing can begin.

HiTDIP provides genetic validation for target selection to drive the pipeline. The confirmed targets are genetically associated in some manner with specific human diseases or therapeutic indications that can be encompassed by the disease diagnostic label. Indeed, genes that are associated with specific clinical variables could help to define complex disease heterogeneity. A significant difference in success compared with historical target selection can be tested against benchmarks. Attrition that is the result of a lack of relevance to a particular human disease might only be appreciated when there is an absence of clinical efficacy many years down the pipeline. Reducing attrition at the proof of efficacy stage (Phase IIA) will increase the efficiency of the pipeline and could provide a sustainable stream of effective medicines related to human diseases. Furthermore, additional hypotheses might be generated at Phase IIA by the application of efficacy pharmacogenetics. These hypotheses can be tested reiteratively in Phases IIB, III and IV. Of course, the consented DNA samples remain available for testing new target classes as new chemical synthesis capabilities are developed.

HiTDIP to discovery shunt – some matches are made in heaven

Trying to decide what to wear for a particular occasion frequently involves a quick scan of your collection of clothes. Large pharmaceutical companies, particularly those with legacy compound libraries resulting from long histories of drug discovery, have a lot of unworn items in their collections. Many discovery programs are initiated and numerous targets are screened only for leads to be 'left on the shelf' (some with the price tags still on!) when times and circumstances changed. The initial assumption for HiTDIP was that by screening all the known tractable targets against several well-defined diseases, new targets would be identified. Each would then require the design of a high-throughput screen and a new screen of the chemical compound library.

However, once targets with highly statistically significant associations with a particular disease were defined, another surprisingly rich supply of existing leads, which required no novel chemical screening, was identified immediately. Indeed, several targets were already screened against the legacy company chemical libraries, with resulting leads left on the shelf because the program was dropped, or abandoned after a molecule failed for a particular therapeutic indication. Thus, immediately after completing confirmation analyses for the initial HiTDIP programs, several targets were identified for which there were lead – or better – quality molecules on the shelf. At present, several molecules have either entered preclinical testing for Phase I or have already been tested in clinical trials designed for different therapeutic indications.

BOX 2

Learning of the power and statistical significance of association studies

Hirschhorn *et al.* [18] conducted an interesting meta-analysis of 166 initial associations to determine the probability of their being reliably replicated. A particularly notable finding was that of 166 initial associations from multiple studies, only six replicated consistently across investigations (i.e. had p-values <0.05 in 75% or more of the studies identified), whereas 97 were observed to have at least one significant replication. Replication rates of 16–30%, depending on the definition of replication, have been identified in the association literature [20,21]. Replication of an association study has not usually been included in the experimental design, nor have the numbers of patients and controls been sufficient to expect replication.

The initial gene-based HiTDIP scans of type 2 diabetes mellitus, using a large cohort of patients and controls, identified five genes with a statistical significance of p ≤0.0005, 40 genes with p ≤0.0050 and 211 with p ≤0.0500 (Table i). Although these data might comprise false negatives, or undetected genes, the vital question is how to eliminate false positives. Because of the large number of tests performed, the majority of the signals identified from the initial, primary test-screen, across all levels of significance, were expected to be false positives – it is not reasonable to anticipate that the testing of less than 2000 genes will result in 45 genes with a p-value of <0.005 being real. It should be emphasized that most published association studies are performed with much smaller patient and control groups. Although meta-analyses of small and heterogeneic studies might have been the best available secondary analysis method in the past, it was never considered ideal.

The factor that distinguishes the HiTDIP study design is the availability of an equally powerful set of patients and controls, all of whom have been examined and clinically characterized by the same physicians, thereby reducing phenotypic heterogeneity. Thus, the question can be posed – how many of the genes identified in the initial association studies with highly significant p-values are actually confirmed by a second confirmation study using statistical analyses that are appropriate for complex genome-based association studies? The answer provides a fascinating insight into the poor track record of association study confirmations. A significant proportion of the genes from the large initial association study with a p-value of <0.005 were not confirmed after the secondary screen of similar size (Table i). However, it is extremely important to realize that the published association literature undergoing meta-analyses traditionally use data that are typically much less significant – and the only confirmation comes from other small studies or takes the form of the meta-analyses.

TABLE i

Comparison of numbers of confirmed genes from primary and secondary screens based on p-values (before the permutation testing process)

Primary screen		Secondary screen[a]	Permutation[a]
p-value	No. of genes	No. of genes	No. of genes
≤0.0005	5	1	0
≤0.0050	40	8	3
≤0.0500	211	44	18
Total	256	53	21

[a] Performed at a p-value of 0.05.

The match between the molecule and disease is the key data gained from confirmed genetic association. Using molecules for which considerable data and work already exists will no doubt accelerate pipeline dynamics. More shots on goal quickly – or better shots at more defined goals – results from being able to mine data generated from years of previous drug discovery programs. Occasionally, after considerable prior hard work, serendipity plays a part – which is indeed food for thought for those who wonder what the advantages of corporate size could be! Although legacy company names might disappear, contributions to capabilities, such as prior screens of targets, are maintained.

Whole-genome testing and pathway analyses

Although nothing in science is unanimous, there is a growing belief that high-density, whole-genome SNP-screening, using newer statistical methods and powerful computing capacities, can identify susceptibility genes for a disease. From a practical viewpoint, and the experience of studying almost 2000 genes selected solely because they could be drug targets, it would seem reasonable to perform additional genome-wide (all genes) SNP-screening association studies. Family linkage studies of the past two decades have defined large regions of chromosomes (1–10 Mb) using 'log of the odds' scores that successfully identified disease and susceptibility genes. By narrowing these large regions using candidate genes, many positive results were reported. Much smaller regions of extended LD (50–250 kb) containing disease susceptibility genes were identified using large patient and control collections for case-control association studies. The DNA is there, the methods are increasingly more feasible economically and there is a growing literature of statistical methods to support these large studies [56]. From an academic point of view, whole-genome screening studies might be too expensive to be readily available, but that does not mean they will not provide scientifically valid data and disease insight. Therefore, after the timetable of HiTDIP confirmation screening is completed, whole-genome SNP-screening will follow. Making the right choice of targets at the beginning of the pipeline will be the first step down the long road of creating innovative medicines.

Within three years

The first practical metric that will provide an estimate of the success of HiTDIP will be comparing Phase II attrition with historical benchmarks. If the target genes selected result in more frequent early-phase efficacy successes, then the 'quantal step-up in discovery' described by Weisberg will be apparent [32]. There will be other parallel strategies for selecting targets. However, comparisons of the success and attrition rates of drug candidates resulting from screens of genetically associated targets can be compared with benchmarks for other historical or current strategies. A retrospective view will be of interest but, for the present, the first indications will come from decreased attrition at Phase II. Along the way, there will be a stream of lead molecules with which to prime a high-throughput pipeline. Combining the choice of the right target and the application of efficacy pharmacogenetics at proof of concept, when necessary, should result in better defined medicines for patients with complex diseases [3]. The objective of the process is to reduce attrition, cycle times and expense, as well as provide safe and effective medicines.

Acknowledgements

This research was supported by GSK R&D. We wish to acknowledge the work of hundreds of GSK scientists, almost two hundred collaborating clinicians, many academic consultants and others who have contributed to the HiTDIP programs. As each set of disease-specific results is published, it is anticipated that many of the contributors will be identified and will be co-authors. Alison Gatherum and Mindy McPeak were extremely helpful in preparing the manuscript. Several important visionaries must also be acknowledged for a program that has taken seven years to begin to deliver sustained value to the pipeline: James Niedel, who was the Chairman of R&D in GlaxoWellcome and supported the initiation of the clinical networks (1997–2000); Tachi Yamada, Chairman of R&D in GSK, who backed and expanded this strategy as a core objective of the R&D organization of GSK (2000–present); Jean-Paul Garnier, CEO of GSK, whose support for the early, middle and late parts of the R&D pipeline was crucial in sustaining these efforts; and the thousands of people in the GSK organization who believe in 'Focus on the Patient' to create safe and effective medicines.

References

1 Chapman, T. (2004) Drug discovery – the leading edge. *Nature* 430, 109

2 Kola, I. and Landis, J. (2004) Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* 3, 711–715

3 Roses, A. (2002) Genome-based pharmacogenetics and the pharmaceutical industry. *Nat. Rev. Drug Discov.* 1, 541–549

4 Aono, S. *et al.* (1995) Analysis of genes for bilirubin UDP-glucuronosyltransferase in Gilbert's syndrome. *Lancet* 345, 958–959

5 Fijal, B.A. *et al.* (2000) Clinical trials in the genomic era: effects of protective genotypes on sample size and duration of trial. *Control. Clin. Trials* 21, 7–20

6 Monaghan, G. *et al.* (1996) Genetic variation in bilirubin UPD-glucuronosyltransferase gene promoter and Gilbert's syndrome. *Lancet* 347, 578–581

7 Chanda, S. and Caldwell, J. (2003) Fulfilling the promise: drug discovery in the post-genomic era. *Drug Discov. Today* 8, 168–174

8 Glassman, R. and Sun, A. (2004) Biotechnology: identifying advances from the hype. *Nat. Rev. Drug Discov.* 3, 177–183

9 Johnston, P.A. and Johnston, P.A. (2002) Cellular platforms for HTS: three case studies. *Drug Discov. Today* 7, 353–363

10 Austin, C.P. *et al.* (2004) The knockout mouse project. *Nat. Genet.* 36, 921–924

11 Drews, J. (2000) Drug discovery: a historical perspective. *Science* 287, 1960–1964

12 Marecki, S. and Kirkpatrick, P. (2004) Efalizumab. *Nat. Rev. Drug Discov.* 3, 473–474

13 Food and Drug Administration (2004) Innovation stagnation, challenge and opportunity on the critical path to new medical products (http://www.fda.gov/oc/initiatives/criticalpath/whitepaper.html)

14 Berns, A. (2001) Cancer – improved mouse models. *Nature* 410, 1043–1044

15 Liggett, S.B. (2004) Opinion: genetically modified mouse models for pharmacogenomic research. *Nat. Rev. Genet.* 5, 657–663

16 Zambrowicz, B.P. *et al.* (2003) Predicting drug efficacy: knockouts model pipeline drugs of the pharmaceutical industry. *Curr. Opin. Pharmacol.* 3, 563–570

17 Zambrowicz, B.P. and Sands, A. (2003) Knockouts model the 100 best-selling drugs – will they model the next 100? *Nat. Rev. Drug Discov.* 2, 38–51

18 Hirschhorn, J. *et al.* (2002) A comprehensive review of genetic association studies. *Genet. Med.* 4, 45–61

19 Neale, B. and Sham, P. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* 75, 353–362

20 Lohmueller, K. *et al.* (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* 33, 177–182

21 Ioannidis, J. *et al.* (2003) Genetic associations in large versus small studies: an empirical assessment. *Lancet* 361, 567–571

22 Tabor, H.K. *et al.* (2002) Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* 3, 391–397

23 Thomas, D. and Clayton, D. (2004) Betting odds and genetic associations. *J. Natl. Cancer Inst.* 96, 421–423

24 Weiss, K. and Terwilliger, J. (2000) How many diseases does it take to map a gene with SNPs. *Nat. Genet.* 26, 151–157

25 Risch, N. (2000) Searching for genetic determinants in the new millennium. *Nature* 405, 847–856

26 Colhoun, H.M. *et al.* (2003) Problems of reporting genetic associations with complex outcomes. *Lancet* 361, 865–872

27 Tufts Center for the Study of Drug Development (2003) Post-approval R&D raises total drug development costs to $897 million. *Impact Report: Analysis and Insight into Critical Drug Development Issues* 5, May–June

28 Venter, J.C. (2000) Genomic impact on pharmaceutical development. *Novartis Foundation Symposium* 229, 14-18

29 Collins, F. and McKusick, V. (2001) Implications of the Human Genome Project for medical science. *J. Am. Med. Assoc.* 285, 540–544

30 Lander, E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

31 Venter, J. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304

32 Fishman, M.C. (2004) An audience with. *Nat. Rev. Drug Discov.* 3, 292

33 Lindsay, M. (2003) Innovation – target discovery. *Nat. Rev. Drug Discov.* 2, 831–838

34 Goldstein, D. *et al.* (2003) Pharmacogenetics goes genomic. *Nat. Rev. Genet.* 4, 937–947

35 Altshuler, D. *et al.* (2000) The common PPAR gamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat. Genet.* 26, 76–80

36 Martin, E. *et al.* (2001) Association of single-nucleotide polymorphisms of the tau gene with late-onset Parkinson disease. *J. Am. Med. Assoc.* 286, 2245–2250

37 Van Eerdewegh, P. *et al.* (2002) Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* 418, 426–430

38 Funke, B. *et al.* (2004) Association of the DTNBP1 Locus with Schizophrenia in a U.S. Population. *Am. J. Hum. Genet.* 75, 891–898

39 Duan, J. *et al.* (2004) Polymorphisms in the trace amine receptor 4 (TRAR4) gene on chromosome 6q23.2 are associated with susceptibility to schizophrenia. *Am. J. Hum. Genet.* 75, 624–638

40 Giallourakis, C. *et al.* (2003) IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease

and identification of a novel association with ulcerative colitis. *Am. J. Hum. Genet.* 73, 205–211

41 Schmith, V. *et al.* (2003) Pharmacogenetics and disease genetics of complex diseases. *Cell. Mol. Life Sci.* 60, 1636–1646

42 Drews, J. (1996) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14, 1516–1518

43 Hopkins, A.L. and Groom, C. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730

44 Sachidanandam, R. *et al.* (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933

45 Wyszynski, D.F. *et al.* The relationship between atherogenic dyslipidemia and the adult treatment program-III definition of metabolic syndrome: The Genetic Epidemiology of Metabolic Syndrome Project. *Am. J. Cardiol.* (in press)

46 Taylor, J.D. *et al.* (2001) Flow cytometric platform for high-throughput single nucleotide polymorphism analysis. *Biotechniques* 30, 661–669

47 Chen, J. *et al.* (2000) A microsphere-based assay for multiplexed single nucleotide polymorphism analysis using single base chain extension. *Genome Res.* 10, 549–557

48 Mehta, C. and Patel, N. (1983) A network algorithm for performing Fisher's Exact Test in r×c contingency tables. *J. Am. Stat. Assoc.* 78, 427–434

49 Neale, B. and Sham, P. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* 75, 353–362

50 Cardon, L. and Bell, J. (2001) Association study designs for complex diseases. *Nat. Rev. Genet.* 2, 91–99

51 Roses, A. (2002) SNPs – where's the beef? *Pharmacogenomics J.* 2, 277–283

52 McCarthy, L. *et al.* (2001) Single-nucleotide polymorphism alleles in the insulin receptor gene are associated with typical migraine. *Genomics* 78, 135–149

53 Hewett, D. *et al.* (2002) Identification of a psoriasis susceptibility candidate gene by linkage disequilibrium mapping with a localized single nucleotide polymorphism map. *Genomics* 79, 305–314

54 Martin, E. *et al.* (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.* 67, 146–154

55 Rioux, J. *et al.* (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nat. Genet.* 29, 223–228

56 Carlson, C. *et al.* (2004) Mapping complex disease loci in whole-genome association studies. *Nature* 429, 446–452

57 North, B.V. *et al.* (2002) A note on the calculation of empirical P values from Monte Carlo procedures. *Am. J. Hum. Genet.* 71, 439–441

58 Davison, A.C. and Hinkley, D.V. (1997) The Basic Bootstraps. In *Bootstrap Methods and their Application* (Gill, R. *et al.*, eds), pp. 11–66, Cambridge University Press

59 Davison, A.C. and Hinkley, D.V. (1997) Tests. In *Bootstrap Methods and their Application* (Gill, R. *et al.*, eds), pp. 136–187, Cambridge University Press

60 Roses, A. (2004) Pharmacogenetics and drug development: the path to safer and more effective drugs. *Nat. Rev. Genet.* 5, 645–655

Related articles in other Elsevier journals

Target identification and validation through genetics
Allen, M.J. and Carey, A.H. (2004) *Drug Discovery Today: TARGETS* 3, 183–190

Translating pharmacogenetics and pharmacogenomics into drug development for clinical pediatrics and beyond
Leeder, J.S. (2004) *Drug Discov. Today,* 9, 567–573

Genome scans and candidate gene approaches in the study of common diseases and variable drug responses
Goldstein, D.B. *et al.* (2003) *Trends Genet.* 19, 615–622

SNP association studies in Alzheimer's disease highlight problems for complex disease analysis
Emahazion, T. *et al. Trends Genet.* 17, 407–413