Candidate genes and single nucleotide polymorphisms (SNPs) in the study of human disease

Stephen Chanock*
National Cancer Institute, Gaithersburg, MD 20877, USA

The genomic revolution has generated an extraordinary resource, the catalog of variation within the human genome, for investigating biological, evolutionary and medical questions. Together with new, more efficient platforms for high-throughput genotyping, it is possible to begin to dissect genetic contributions to complex trait diseases, specifically examining common variants, such as the single nucleotide polymorphism (SNP). At the same time, these tools will make it possible to identify determinants of disease with the expectation of eventually, tailoring therapies based upon specific profiles. However, a number of methodological, practical and ethical issues must be addressed before the analysis of genetic variation becomes a standard of clinical medicine. The currents of variation in human biology are reviewed here, with a specific emphasis on future challenges and directions.

Keywords: Variation, genome, genetic, mutation, disease susceptibility

1. Introduction

With the completion of the first composite map of the human genome, an extraordinary resource has become available to investigate the role of genetic variation in human diseases [1,2]. While the total number of genes is roughly 35,000, less than the initial estimates, it is nonetheless, a formidable task to organize and catalog the differences between any two genomes. The concept of a single human genome was useful for constructing a first generation map, but now, it is clear that there is no such entity. In fact, any two human genomes are estimated to differ by approximately 0.1% or less. Remarkably, it is within this tiny fraction of a genome, namely the collection of sequence variations, that we find an opportunity to decipher genetic determinants of disease susceptibility and outcome. Furthermore, the catalog of variations represents an unprecedented resource for investigating evolutionary and migratory events in human history. The combination of technical advances in genetic analyses coupled with the enormous resource of genetic information has set the stage for a new age in the study of human disease.

2. Variation in the human genome: The predominance of SNPs

The most common sequence variation in the human genome is the substitution of a single base, commonly referred to as a single nucleotide polymorphism (SNP). By definition, a SNP is a variation in sequence with a frequency of greater than 1% in at least one population. It has been estimated that the number of SNPs in an individual numbers in the millions, reflecting enormous sequence diversity across all human chromosomes [3, 4]. Initial estimates of the density of SNPs across the genome suggest that the average frequency of SNPs is between 1 in 1.3 and 1.9 kb overall [4–6]. It is notable that the density of SNPs varies between regions of chromosomes, as well as between chromosomes [7, 8]. These differences reflect a spectrum of selective pressures on genes as well as complex rates of mutation and recombination, which, also vary greatly across the genome. Less frequent variants (i.e., with less than 1% frequency) might also be informative in the mapping of complex-trait disorders as well as familial diseases. By definition, a mutation results in a significant phenotype, whereas a SNP possesses mild or no phenotypic changes. It is more than likely than some rarer variants

*Address for correspondence: Steven Chanock, M.D., Immunocompromised Host Section, Pediatric Oncology Branch and The Advanced Technology Center, National Cancer Institute, 8717 Grovemont Circle, Gaithersburg, MD 20877, USA. Tel.: +1 301 435 7559; Fax: +1 301 402 3134; E-mail:sc83a@nih.gov.

Disease Markers 17 (2001) 89–98
ISSN 0278-0240 / $8.00

act more like SNPs, conferring a mild phenotypic difference in rarer disorders. The challenge lies in identifying infrequent variants from public databases and applying them to directed studies of rarer diseases (i.e. childhood cancer). The study of SNPs, which are, by definition, stable genetic changes in the genome, permits us to look closely at the footprints of past generations. Thus, the study of SNPs within families or pathways of genes also offers a perspective on prior events in human evolution, ones that have shaped the diversity we appreciate today [9].

Several factors in combination have shaped the diversity we now observe as SNPs, fixed variations in gene sequence. One of the major contributing factors is selective pressure in response to challenges within discrete populations. Random mutation at each base in the 3.1 billion base pair genome is not equally tolerated; in the extreme, some changes are not compatible with survival (i.e., pre-terminal stop codons in critical genes). Since the rate of random mutational rate is roughly 2×10^{-8}, it is expected that each base has been mutated many times in the history of human evolution. So far, initial surveys indicate that transitional changes are more common than transversions, suggesting that deamination contributes to a higher likelihood of mutation, especially at sites of CpG dinucleotides [10]. Based on these assumptions and the data emanating from the SNP Consortium, it has been posited that there are as many 11 million SNPs per person [3].

Historically, the first single nucleotide polymorphisms were identified during 'gene-centric' studies, detected by restriction fragment length polymorphism (RFLP) and used to analyze single genes in population-based studies. The utility of the RFLP was replaced by the use of the simple tandem repeat (STR) as the preferred unit for genetic studies. STRs are highly polymorphic allelic repeats of 2, 3 and 4 nucleotide units strewn evenly through-out the genome, but with substantially lesser frequency than SNPs [11]. Rarer variations, namely large scale deletions, substitutions or duplications arise with lower frequency and are useful for study of informative genes [12]. In some cases, a SNP is in linkage with a more complex, stable variant. Interestingly, some of the strongest associations identified so far have been between complex variants and disease outcomes, and not simple SNPs, though many use the term SNP to indicate more than a single nucleotide variation.

Adequately spaced STRs have been be used to scan the whole genome for markers in linkage disequilibrium with a trait. Using PCR amplification technology, it is possible to amplify a collection of STRs distributed across the genome and map monogenic disorders to a unique locus using linkage analysis in family pedigrees. The success of this approach has been employed in the search for rare, familial mutations with high penetrance (in other words, when the phenotypic expression of the rare mutation is quite significant).

For more complex disorders, in which multiple genes each contribute a small effect, the utility of the whole genome scan approach has been more problematic for several reasons. The ability to discriminate the effect of single genes is difficult in complex disorders, which, by definition, are characterized by interactions between multiple genes. So far, inadequate coverage of the genome with mapped SNPs together with technical challenges have undermined the effectiveness of this approach. Previously, a whole genome scan identified one or more genetic markers within a specific region of a chromosome, but does not pinpoint the specific gene nor the significant change(s) in a gene's sequence, which ultimately alters the phenotype.

Recently, many investigators have turned back to SNPs, planning to capitalize on several important features. One, SNPs are more abundant than STRs. Second, they are more stable compared to short nucleotide repeats. The high degree of variation in STRs represents a formidable barrier to population-based studies, as is frequently employed in candidate genetic association studies. Lastly, a subset of SNPs (see below) effect the biological properties or expression of a gene product, thus providing a functional component that can neatly tie in with in vitro studies. In the future, the catolog of 'functional' SNPs will be useful in mapping complex trait diseases, while at the same time, providing insights into the mechanisms of disease or treatment outcomes.

3. A book of past lives: The catalog of SNPs

It is safe to say that we are still in the initial phase of discovery with respect to SNPs. There are a number of web-based tools designed to interrogate public databases in search of SNPs (see Table 1). The National Center for Biotechnology Information is curating a public website, db-SNP, for deposition of SNPs and tagging the SNPs to other webtools useful for investigating genetic information in silico [13]. Already, nearly 1.4 million possible SNPs have been deposited in the SNP Consortium Database for public use with more expected in the months to come (http://snp.cshl.org/) [4].

Table 1
Selected public resources for SNP analysis

Public databases of SNPs	
PubMed	Published Literature (search by gene)
(www. ncbi.nlm.nih.gov/entrez/query.fcgi)	– Resource for searching published data
dB SNP	NCBI Database of deposited SNPs
(www.ncbi.nlm.nih.gov/SNP/)	– Central repository of SNPs
The SNP Consortium	Database of predicted SNPs
(snp.cshl.org/)	– Deposited in db-SNP
Web-based tools for identifying SNPs	
Cancer Genome Anatomy Project- GAI	NCI-based SNP Discovery Project
(lpg.nci.nih.gov/GAI/)	– Gene lists, tools for SNP analysis including predicting non-synonymous SNPs
SNP pipeline	CGAP-GAI search of EST/Unigene
(lpgws.nci.nih.gov:82/perl/snp/snp_cgi.p)	– Search EST sequences for SNPs
Leelab SNP Database	UCLA search of EST/Unigene
(www.bioinformatics.ucla.edu/snp/)	– Search EST sequences for SNPs
Databases of SNPs with annotation	
HG Base	International Database
(Hgbase.cgr.ki.se/)	– Repository for SNP
Imunology -SNP database	Curated collection of immunologically significant SNPs
(www-dcs.nci.nih.gov/pedonc/ISNP/)	– SNP database of known genes with SNPs
University of Utah Genome Center GeneSNPs	Curated collection of SNPs derived from public databases
(http://www.genome.utah.edu/genesnps_old/)	

The importance of validating SNPs can not be overemphasized. Sophisticated web-tools have been developed to search public databases of expressed sequence tags (ESTs) in search of putative SNPs [14,15]. Recently, it has been suggested that roughly three-fourths of the SNPs in two large public databases can be validated at a high frequency ($> 20\%$ for the minor allele) [14–16]. Others have reported that using in silico tools yields relatively high specificity, but low sensitivity in SNP validation [17]. Still, candidate SNPs need to be validated in genomic assays that amplify the specific locus of interest from a unique chromosomal location, determine that the variant exists by one of several technologies (see below) and demonstrate its Mendelian inheritance. So far, this approach has been biased towards SNPs that arise in unique regions of the genome. On the other hand, using PCR technology, it is difficult to amplify regions replete with a high density of redundancies, where informative SNPs (or STRs) could be still positioned.

While most SNPs are neutral and do not alter the phenotype, there is a subset of SNPs, which are predicted to change the phenotype of the gene. These are the most interesting targets for investigating the contribution of SNPs to disease. Traditionally, SNPs in the coding region have been of particular interest because an amino acid substitution can alter the function of a protein. These are known as non-synonymous coding SNPs and are less common than synonymous SNPs in the coding region (i.e., those that do not result in the substitution of an amino acid). Early studies have confirmed that non-synonymous coding SNPs are rarer, and thus, support the hypothesis that greater selective pressure is required to effect a functionally significant amino acid substitution [5,6]. The same can be argued on behalf of variants critical for the regulation of a gene's expression. In the rush to catalog SNPs, many groups have concentrated on identifying non-synonymous SNPs and in some cases, developing sophisticated web-tools to browse public databases to display in silico translation predictions of validated SNPs (http://lpgws.nci.nih.gov:82/perl/snp2ref). On the other hand, it is harder to identify SNPs that alter the regulation of a gene, namely its expression. This class of SNPs are particularly important in complex pathways, such as an immunological cytokine network or the coagulation cascade, where small differences in expression can have pleotropic effects downstream, in either amplification or dampening of a response pathway [18]. Since the field of promoter analysis and gene regulation is highly specific to unique sequence motifs, it is unlikely that a comparable search program will facilitate identification of informative, promoter SNPs.

As mentioned above, there is great interest in the cataloging of "functionally" important SNPs, namely, those SNPs that might directly alter the phenotypic expression of the gene. The field is beginning to address the importance of correlating phenotype with SNPs, specifically within pathways or collections of genes that fit a biological paradigm. Based upon the num-

ber of genes in the human genome (roughly 35,000), a SNP density of 1 in 1.8 kb and the average size of the coding region, 5' and 3' UTRs (1.34 kb, 300 bp and 770 bp, respectively), it is possible to estimate that there are perhaps as few as 55,000 or as many as 250,000 functionally interesting SNPs. The challenge is to find these SNPs and apply them in well-designed molecular epidemiology studies. A public effort at the National Cancer Institute, the Genetic Annotation Initiative of Cancer Genome Anatomy Project is re-sequencing the entire coding region and 5'/3' untranslated regions of candidate genes of importance to cancer biology (http://lpg.nci.nih.gov/GAI/). This directed approach has been employed by other groups in an effort to validate common SNPs in genes of immediate importance to one or more common complex disorders (i.e., diabetes, neurodegenerative diseases or cancer) [5,6]. Taking the next step, several web-tools have been developed to curate SNPs within biological pathways; one such example is the first generation of the Immunology-SNP database (http://www-dcs.nci.nih.gov/pedonc/ISNP/), which catalogs polymorphisms, primarily SNPs in genes of immunological importance. Comparable efforts are underway to develop virtual links between in vitro biological observations, SNPs and genetic association studies. In parallel, an international web-based effort is underway to annotate genes, linking them with function in specific pathways and families; this is known as the Gene Ontology Project (http://www.geneontology.org) [19]. There is no doubt that this resource will be invaluable in the search to map SNPs, both as genetic markers and in informative cases, in which SNPs are defined as modifiers of disease outcome.

The SNP Consortium has deliberately sought to validate over 100,000 SNPs throughout the genome, without a bias towards coding or regulatory SNPs [4]. This collection will be invaluable for large scale linkage studies, and in selected circumstances, genetic association studies. A formidable number of SNPs, at least several hundred thousand are required to adequately provide a SNP density of one every 3 kb; this is based upon inferred linkage disequilibrium calculated for the whole genome [20–22]. As mentioned above, linkage disequilibrium may not be evenly present throughout the genome. Consequently, predictions for the number of SNPs required to conduct whole genome scans in population based studies have been decreased, but not by an order of magnitude. Another, more directed approach is to annotate SNPs in the roughly 35,000 known genes. This approach, known as an 'intelligent SNP' scan, favors identifying functionally significant SNPs. However, this bias, as attractive as it is to interrogate genes, does not provide sufficient coverage of intergenic regions for linkage analysis, which still could influence expression or less likely, function. Nonetheless, in the future, one could imagine a more refined form of an 'intelligent scan' – one that uses validated 'functional' SNPs to conduct scans of the genome in well-defined population-based studies. Or for that matter, one can envision studies that analyze SNPs derived from biologically critical pathways only.

4. Applying SNPs to the study of disease: The changing landscape

There is ample evidence to suggest that in common complex trait diseases, genetic factors contribute to the disease process but, it is most likely that multiple genes contribute to disease susceptibility. Although the effect of any single variant is probably small, combinations of SNPs, either as haplotypes or between distant genes, may coordinately contribute to disease risk. Still, the net effect of each SNP is small, which making it especially difficult to identify key contributors. There have been a few examples in which linkage studies have localized a variant, later shown to confer risk (e.g., APOE and Alzeihmers) [23,24]. Still, the utility of the whole genome scans is limited by the low-penetrance of many common variants. Genetic association studies have emerged as the primary method of studying the effect of SNPs on disease outcomes, whether it is susceptibility to a disease or differences in phenotypic expression of a known gene.

A genetic association study is designed to evaluate the contribution of one or more SNPs to well defined, clinical endpoints. Success is predicated on accurate determination of clinical outcomes in a well-characterized population based case control or cohort study. In either a cohort or case control study, statistical significance is determined by comparison of the distribution of genotypes in two groups, with one serving as a "control" and evaluates the effect of a variant on a clinical endpoint. Usually, these are conducted in population-based studies, consisting of unrelated subjects. Family-based analysis can be useful, too, especially in transmission disequilibrium studies.

Until recently, genetic association studies have been limited to studies involving one or a few candidate genes using classical epidemiological tools designed to evaluate the effect of a gene and its variant(s) in a

population-based case control or cohort study. Typically, an interesting variant was discovered in a gene and studied in a population-based study, if a plausible mechanism could be advanced. The emerging field of molecular epidemiology has been driven by the candidate gene approach, namely, designing studies with known variants in which prior biology or association studies justified study. The landscape has changed dramatically with the explosion of knowledge. The current problem lies in extracting useful and informative SNPs from the public databases, which include millions of putative SNPs. In this regard, the gap between understanding the biological significance of a variant and the identification of a variant is widening at an accelerated pace. It is the annotation and possible functional significance that highlights the importance of looking at candidate SNPs that will be of immediate interest to investigators conducting population-based studies.

Most experts would agree that until the cost and efficiency of high-throughput platforms for SNP analysis are markedly improved, the candidate gene(s) approach will be preferred. Already, new technologies have moved the field from the interrogation of single SNPs to as many as several dozen SNPs in a well-defined population study. The choice of SNPs for a new study should take into consideration several important factors. One is the frequency of the SNP in the population of study. There is great diversity in the frequency of common SNPs between different populations. Second, a plausible mechanism should underlie the choice of genes or families of genes. In other words, a sound biological basis is important in designing population-based genetic association studies.

The literature is filled with small pilot studies that are not confirmed by subsequent, better designed studies. Two major reasons account for this recurrent problem, small cohorts and poorly tracked collection of data points, often retrospective in execution. Pilot studies are useful to identify the likely candidate genes to be validated in larger, more focused studies. It can be argued that false positives in pilot studies are greatly preferred to false negatives, mainly because subsequent studies will determine the validity of a finding, as opposed to missing a potentially informative SNP. A series of stringent studies is required before the results of clinical association studies can be applied to clinical decision making [25].

It is also notable that common polymorphisms, especially SNPs are population specific and have to be viewed in the context of a particular population. Recently, some groups have argued that population stratification (i.e. heterogenous mixture of individuals) does not represent a major stumbling block to execution of well-designed case control studies [26]. Still, the importance of defining the population is evident in interpreting the significance of a SNP. For example, individuals who are heterozygous for the sickle cell mutation in the malaria belt of Western Africa are protected from complications of the infection. In other regions of the world, where the selective pressure of malaria no longer exists, it is viewed differently, as a life-threatening monogenic disorder. The context of the variant defines whether it is a 'balanced polymorphism', as in regions of Africa, or a highly penetrant, deleterious mutation.

Genetic association studies can be used to address two fundamentally different questions. The first seeks to identify genetic variants that influence susceptibility to a complex, multi-factorial disorder. The second type addresses differences in outcomes within a disease population, including differences in responses to medications. The latter is known as pharmacogenomics, a burgeoning field that promises to revolutionize medical treatment in the future [25].

5. SNPs and susceptibility to disease

Candidate gene selection based upon a proposed role for the variant in disease susceptibility has assumed a central role in conducting genetic association studies. More often than not, the genes are chosen a priori, based upon a hypothesis and studied in a suitable population based study. The net effect of any single SNP, especially for a common disorder is generally small, making it difficult to conduct linkage analysis, even with low confidence levels [27]. Still, a number of important studies have demonstrated the utility of examining candidate genes in a range of diseases, including cancer, and diabetes [28–30]. In some circumstances, the contribution of a functionally important SNP is population specific; this has been elegantly shown with a population-specific TNF SNP and susceptibility to malaria in a region of West Africa [31].

Common diseases, such as cancer or hypertension have the advantage of the availability of sufficiently large enough cohorts to examine the question in different populations. In rarer disorders, the collection of adequate numbers of subjects with firm clinical endpoints is a daunting challenge. After a sufficient number of studies have been published, it is possible to conduct case-series meta-analyses to assess the net effect of the

variant. For example, the importance of two separate genetic variants has been shown in bladder cancer [29, 32,33]. With respect to both genes, the odds ratio is greater than one, but less than two, indicating a modest, but real effect.

6. Understanding outcomes in a disease: SNPs as disease modifiers

It has long been appreciated that among individuals with a monogenic disorder, such as cystic fibrosis or chronic granulomatous disease, there can be heterogeneity in the disease course. Even within families, outcomes vary greatly between members with an identical primary mutation, suggesting that other genetic factors modify outcomes. Already, a number of seminal studies have identified modifying SNPs in different populations afflicted with a life-threatening monogenic disorder. If this trend is confirmed in further studies, our notion of a monogenic disorder will have to "modified" to include an appreciation of secondary genes, which influence outcomes. In the future, genetic counseling might include analysis of secondary modifying genes, particularly in an effort to institute preventive measures to intercede and avoid or ameliorate serious complications.

It is notable that under stress conditions, such as a primary immunodeficiency, common polymorphisms, mainly SNPs, can act as modifying genes for specific outcomes. In chronic granulomatous disease, a primary immunodeficiency of innate immunity (i.e., neutrophils and monocytes fail to effectively kill invading organisms), individuals who inherit a variant of a low affinity Fc gamma receptors (*FCGR2A* or *FCGR3B*) or the myeloperoxidase gene (*MPO*) are at increased risk for immunologically mediated gastrointestinal complications [34]. In a preliminary analysis, combinations of the three informative variants conferred a higher risk for this type of complication. This pilot study illustrates the possible significance of the phenotype of the SNP in a pathway not directly affected by the primary mutation is intensified and provides an opportunity to investigate the contribution of genes in vivo [35]. Certainly, it is possible that we will better appreciate the relative importance of genes within a pathway or biological process by examining its effect. Simply put, it is possible to observe the relative importance of specific genes by studying the effects of their variations in large population-based studies.

By no means, should this paradigm be restricted to monogenic disorders. The study of host genetic factors has been extremely informative in dissecting the molecular events associated with both susceptibility and progression of infection with HIV-1. Interestingly, the same variants that increase the likelihood of acquiring infection, also appear to accelerate the course of disease overall. Much of this work has focused on chemokines, namely, *CCR2*, *CCR5*, *SDF1* and *RANTES* which are critical co-factors for infection with HIV [36–39]. Interestingly, disease progression has also been associated with specific polymorphisms of the MHC complex. In particular, heterogeneity in class one MHC, namely HLA-A, B and C loci, was shown to provide a selective advantage, thus slowing progression of HIV infection [40]. Furthermore, genetic factors that alter the risk for developing one of several life-threatening complications of HIV-1 infection have been identified by genetic association studies. For example, a common variant of the low affinity Fc gamma receptor, *FCGR3A* increases the risk for developing Kaposi's Sarcoma (KS) in HIV infected men, also co-infected with the Kaposi's Sarcoma herpes virus, KSHV [41, 42]. KS afflicted roughly 20% of infected men in the USA prior to the development of highly active antiretroviral therapy.

7. Pharmacogenomics

The term, pharmacogenomics refers to field of study seeking an association between pharmacological phenotypes and common genetic variations, namely SNPs. The underlying principle is that genetic variations can account for differences in response to drugs. This is not to ignore other reasons for adverse or inadequate response to a drug. Certainly, the failure to respond can also be attributed to allergies, drug-drug interactions, and erroneous prescriptions (either inappropriate dosings or medications). Still, the SNP revolution offers an opportunity to generate genetic profiles that will be useful in the choice of medications. This pertains to not only choosing the best drug, based upon a genetic pattern, but also avoiding selected drugs in individuals in whom the risk for a serious adverse event is high. From a practical point of view, the applicability of this approach has been demonstrated in principle, with several key examples. Dissecting the role of different steps at the genetic level has been guided by deconstruction of the fate of medicinal drugs. These include the drug target, transport, and metabolism. So far, pilot studies

have not evaluated the drug as a foreign antigen, or a possible allergen. For instance, it has been shown that patients homozygous for the G17R variant in the B2 adrenergic receptor are at risk for exacerbation of asthmatic attacks, when using albuterol treatment [43, 44]. An example of a variant with a frequency of less than 1% but with important clinical consequences is the *TPMT* variant (roughly, 0.3%). Rare individuals who are homozygous for the *TPMT* variant develop severe toxicity when treated with azothioprine therapy for either leukemia or auto-immune diseases [45,46]. Thus, clinical decisions can be based upon knowledge of the genetic profile, and in the future, allowing physicians to tailor therapies on an individual basis.

It is conceivable that SNP studies will also define suitable targets for therapeutic intervention. Though early in its genesis, it is notable that studies in monogenic disorders could generate important insights that give rise to specific therapies. For example, in cystic fibrosis, several groups have confirmed that the presence of a common, non-synonymous SNP in a gene located on a separate chromosome from the *CFTR* gene, the mannose binding lectin, *MBL2*, is an adverse risk factor for pulmonary outcomes. Informative *MBL2* variants (clustered in exon 1) have an overall frequency of approximately 35% and alter both function and circulating levels of the protein, a C-type collectin critical for recognition of pathogens. In the setting of CF, a phenotypically insignificant variant in the normal host impacts pulmonary defenses against pathogenic bacteria [47]. Based on the available data, recombinant mannose binding lectin could be given to CF patients who co-inherit one of the common *MBL2* variants. Directing therapy, even in a rare, monogenic disorder holds tremendous potential for treatment in the future.

8. The labor of SNP detection

The technical platforms for detecting SNPs are rapidly changing. Currently, half a dozen different assay systems are commercially available to discriminate single base changes following amplification of a unique amplicon. The flanking region and the informative SNP are amplified by PCR technology, usually from genomic DNA. Unlike cDNA array studies, which capture the full complement of messenger RNA using a common oligo dT primer, a unique set of oligonuceotide primers is required for amplifying each individual SNP. In this regard, each assay has to be optimized, even before multiplexing, making large scale analysis more cumbersome in development and execution. A number of promising platforms (see Table 2) have been developed that can increase the throughput of SNP detection and, in some cases, analyze multiple SNPs in one reaction (but not at a scale comparable to the 50,000 messages analyzed in cDNA array studies). Price and effort have both been streamlined, but still are not sufficiently economical to support large scale genome wide studies, which require 100,000 or more unique SNPs. Consequently, to analyze this number of SNPs in a population studies, a million or more genotypes would be required. Therefore, it is not surprising that until now and for the foreseeable future, the candidate gene(s) approach will be favored.

In response to the technical challenges of large scale SNP analysis, many groups have turned to an approach called DNA pooling. Instead of analyzing individual SNPs in each sample, pooled DNA (e.g., equal aliquots of DNA from a large number of subjects) is analyzed for allelic frequency [48–50]. Naturally, labor and cost are substantially reduced, which is particularly advantageous in a pilot study, seeking to identify excellent candidate SNPs for confirmation in validation studies. At the same time, large scale screening can provide sufficient power to minimize the likelihood of identifying false positives. Pooling studies have the disadvantage of analyzing only the allelic frequency of SNPs, and in pools defined by the initial design. Thus, it is difficult to analyze more complex questions of outcomes or subgroups, unless chosen a priori in the design of the pooling studies. Still, this approach is extremely promising for the study of complex, multi-genic common disorders [49,51]. In this regard, pooling is an effective approach for screening candidate genes that could influence susceptibility to an exposure or a disease (e.g., in large scale cancer cohort studies).

The strategy of interrogating pools of DNA has been successfully reported in identifying SNPs contributing to complex diseases and the identification of rare, Mendelian mutations [21,52]. Since we are still in the early phase of developing this approach, there is intense interest in optimization of the number of subjects pooled in conjunction with the platform used to analyze SNPs. In addition, aliquoting equal amounts of genomic DNA is a major technical challenge, which is further magnified by amplification-based technologies for SNP detection. Therefore, the margin for error is small in the execution of DNA pooling studies, especially when differences in the distribution of allelic frequencies might be relatively small, a common finding in genetic association studies.

Table 2
Current technical platforms for SNP detection

Generic method	High-throughput platform	Type of analysis
Direct sequence analysis[a]	Generally not	Qualitative only
Single base extension	Promising if multiplexed	
Single strand sequence	Limited to single sites	
Hybridization methods[a]	Variable	Quantitative + Qualitative
Target amplification[b]	Moderately efficient	
Signal amplification[c]	Highly efficient	
Microarray[a]	Highly efficient	Quantitative + Qualitative
Restriction enzyme analysis	Not efficient	Qualitative only
Conformational analysis[a]	Moderately efficient	Qualitative only

These platforms are also applicable to use for detection of rare, highly penetrant mutations.
[a]Relies upon amplification technology (i.e., polymerase chain reaction, PCR) to generate amplicon for analysis.
[b]Example of platform is real-time polymerase chain reaction.
[c]Example of platform is the chip-based matrix associated laser desorption/ionization time-of-flight (MADLI-TOF).

9. The challenges in SNP analysis that lie ahead

The study of SNPs in human disease is a rich resource for dissecting the genetic contribution to complex-trait diseases and modifiers of monogenic disorders. The extraordinary spectrum of variation is also its weakest link, because each study has to be interpreted in the context of the population examined. The literature is filled with examples of informative SNPs that do not reproduce in different settings. This conundrum underscores the importance of combinations of SNPs, which for the purpose of any population-based study of unrelated individuals assumes no contribution from the background of SNPs. This problem raises difficult questions for identifying and validating gene-environment interactions as well as gene-gene interactions. Despite the fact that the effect of a SNP is measured over an extended period of time, it is difficult to dissect the temporal relationship between SNPs without a sound biological model. One of the major hurdles of the future is to develop suitable systems to analyze the complex interactions of SNPs and create a suitable (and reproducible model) that will be useful for clinical implementation of genetic risk factors. The field is moving towards examining collections of SNPs, which are derived from biological pathways or families of genes. An extension of this approach, known as 'neighboring' SNPs, expands the genes under study to include those that interact up and downstream from the core set. By saturating a pathway, one can evaluate changes in a pathway, one that might dampen or amplify a cascade (i.e. complement cascade).

We are early in the study of SNPs, which , for the purpose of initial studies, are viewed as individual units. However, SNPs form haplotypes and it is the investigation of haplotypes that will probably be most informative- both as genetic markers and as tools to correlate genetic variation with functional outcomes (i.e. clinical states). How many haplotypes exist for a given cluster of SNPs, either in one gene or in a group of closely positioned genes remains to be determined, but it is critical to pursue this approach [27,53]. Since haplotypes are defined by the blocks of genes which maintain variants already in place, such as SNPs, the cataolog of SNPs will be an invaluable tool in defining haplotypes, especially ones that include variants that functionally alter the gene or gene product. If indeed, large scale sequencing is possible in the near future, then the field will have to analyze (and re-analyze current studies when possible) haplotypes. The technical platforms will need to be more flexible and extend further distances between variants to capture the informative components of haplotypes.

Lastly, the most difficult task will be to consider the implementation of SNPs in clinical decision making, particularly as it relates to providing recommendations for interventional or preventional measures, based upon the concept of "risk". Together with ethicists, a dialogue must begin to address how and in what manner to use genetic information, especially when the consequences of the information have pleotrpoic implications for personal security, insurance and health.

10. Conclusion

We are at the beginning of an era when we can investigate the functional implications of single nucleotide

polymorphisms, SNPs, and other rarer variants. The potential usefulness in medicine is unprecedented. To define risk factors for disease and pharmacological outcomes based upon genetic profiles of SNPs could revolutionize medical care. It also comes at a dangerous cost of potential political and philosophical challenges, which must be addressed in parallel, or actually in advance, if we are to protect the rights and will of the individual. Still, these small differences cumulatively have a staggering effect, creating the individuality that we recognize in each person, while at the same time, reflect the changes that have taken place over generations, many in response to environmental and pathogenic challenges. The opportunity to annotate the differences between individuals has provided an extraordinarily rich resource for investigating complex genetic events, particularly as they relate to disease susceptibility and population genetics.

References

[1] E.S. Lander, L.M. Linton and B. Birren et al., Initial sequencing and analysis of the human genome, *Nature* **409** (2001), 860–921.

[2] J.C. Venter, M.D. Adams and E.W. Myers et al., The sequence of the human genome, *Science* **291** (2001), 1304–1351.

[3] L. Kruglyak and D.A. Nickerson, Variation is the spice of life, *Nat Genet* **27** (2001), 234–236.

[4] R. Sachidanandam, D. Weissman and S.C. Schmidt et al., A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms, *Nature* **409** (2001), 928–933.

[5] M. Cargill, D. Altshuler and J. Ireland et al., Characterization of single-nucleotide polymorphisms in coding regions of human genes, *Nat Genet* **22** (1999), 231–238.

[6] M.K. Halushka, J.B. Fan and K. Bentley et al., Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis, *Nat Genet* **22** (1999), 239–247.

[7] P. Taillon-Miller, I. Bauer-Sardina and N.L. Saccone et al., Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28, *Nat Genet* **25** (2000), 324–328.

[8] I.A. Eaves, T.R. Merriman and R.A. Barber et al., The genetically isolated populations of Finland and sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes, *Nat Genet* **25** (2000), 320–323.

[9] J.G. Hacia, J.B. Fan and O. Ryder et al., Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays, *Nat Genet* **22** (1999), 164–167.

[10] B.K. Duncan and J.H. Miller, Mutagenic deamination of cytosine residues in DNA, *Nature* **287** (1980), 560–561.

[11] R. Chakraborty, M. Kimmel, D.N. Stivers, L.J. Davison and R. Deka, Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci, *Proc Natl Acad Sci USA* **94** (1997), 1041–1046.

[12] E.S. Lander and N.J. Schork, Genetic dissection of complex traits, *Science* **265** (1994), 2037–2048.

[13] E.M. Smigielski, K. Sirotkin, M. Ward and S.T. Sherry, dbSNP: a database of single nucleotide polymorphisms, *Nucleic Acids Res* **28** (2000), 352–355.

[14] K.H. Buetow, M.N. Edmonson and A.B. Cassidy, Reliable identification of large numbers of candidate SNPs from public EST data, *Nat Genet* **21** (1999), 323–325.

[15] K. Irizarry, V. Kustanovich and C. Li et al., Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences, *Nat Genet* **26** (2000), 233–236.

[16] G. Marth, R. Yeh and M. Minton et al., Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet* **27** (2001), 371–372.

[17] D. Cox, C. Boillot and F. Canzian, Data mining: efficiency of using sequence databases for polymorphism discovery, *Human Mutation* **17** (2001), 141–150.

[18] C.B. Foster and S.J. Chanock, Mining variations in genes of innate and phagocytic immunity: current status and future prospects, *Curr Opin Hematol* **7** (2000), 9–15.

[19] M. Ashburner, C.A. Ball and J.A. Blake et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet* **25** (2000), 25–29.

[20] L. Kruglyak, Prospects for whole-genome linkage disequilibrium mapping of common disease genes, *Nat Genet* **22** (1999), 139–144.

[21] A. Collins, C. Lonjou and N.E. Morton, Genetic epidemiology of single-nucleotide polymorphisms, *Proc Natl Acad Sci USA* **96** (1999), 15173–15177.

[22] J. Ott, Predicting the range of linkage disequilibrium, *Proc Natl Acad Sci USA* **97** (2000), 2–3.

[23] M.A. Pericak-Vance, J.L. Bebout and P.C. Jr. Gaskell et al., Linkage studies in familial Alzheimer disease: evidence for chromosome 19 linkage, *Am J Hum Genet* **48** (1991), 1034–1050.

[24] W.J. Strittmatter, A.M. Saunders and D. Schmechel et al., Apolipoprotein E: high-avidity binding to beta-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease, *Proc Natl Acad Sci USA* **90** (1993), 1977–1981.

[25] F.S. Collins, Shattuck lecture – medical and societal consequences of the Human Genome Project, *N Engl J Med* **341** (1999), 28–37.

[26] S. Wacholder, N. Rothman and N. Caporaso, Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias, *J Natl Cancer Inst* **92** (2000), 1151–1158.

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

[28] S.J. London, T.A. Lehman and J.A. Taylor, Myeloperoxidase genetic polymorphism and lung cancer risk, *Cancer Res* **57** (1997), 5001–5003.

[29] P.M. Marcus, R.B. Hayes and P. Vineis et al., Cigarette smoking, N-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction, *Cancer Epidemiol Biomarkers Prev* **9** (2000), 461–467.

[30] D. Altshuler, J.N. Hirschhorn and M. Klannemark et al., The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes, *Nat Genet* **26** (2000), 76–80.

[31] J.C. Knight, I. Udalova and A.V. Hill et al., A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria, *Nat Genet* **22** (1999), 145–150.

[32] L.E. Johns and R.S. Houlston, Glutathione S-transferase mu1 (GSTM1) status and bladder cancer risk: a meta-analysis,

Mutagenesis **15** (2000), 399–404.
[33] P.M. Marcus, P. Vineis and N. Rothman, NAT2 slow acetylation and bladder cancer risk: a meta-analysis of 22 case-control studies conducted in the general population, *Pharmacogenetics* **10** (2000), 115–122.
[34] C.B. Foster, T. Lehrnbecher and F. Mol et al., Host defense molecule polymorphisms influence the risk for immune-mediated complications in chronic granulomatous disease, *J Clin Invest* **102** (1998), 2146–2155.
[35] S.J. Chanock and C.B. Foster, SNPing away at innate immunity, *J Clin Invest* **104** (1999), 369–370.
[36] M. Dean, M. Carrington and C. Winkler et al., Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study, *Science* **273** (1996), 1856–1862.
[37] M.W. Smith, M. Dean and M. Carrington et al., Contrasting genetic influence of CCR2 and CCR5 variants on HIV-1 infection and disease progression. Hemophilia Growth and Development Study (HGDS), Multicenter AIDS Cohort Study (MACS), Multicenter Hemophilia Cohort Study (MHCS), San Francisco City Cohort (SFCC), ALIVE Study, *Science* **277** (1997), 959–965.
[38] D.H. McDermott, M.J. Beecroft and C.A. Kleeberger et al., Chemokine RANTES promoter polymorphism affects risk of both HIV infection and disease progression in the Multicenter AIDS Cohort Study, *Aids* **14** (2000), 2671–2678.
[39] P.A. Zimmerman, A. Buckler-White and G. Alkhatib et al., Inherited resistance to HIV-1 conferred by an inactivating mutation in CC chemokine receptor 5: studies in populations with contrasting clinical phenotypes, defined racial background, and quantified risk, *Mol Med* **3** (1997), 23–36.
[40] M. Carrington, G.W. Nelson and M.P. Martin et al., HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage, *Science* **283** (1999), 1748–1752.
[41] C.B. Foster, T. Lehrnbecher and S. Samuels et al., An IL6 promoter polymorphism is associated with a lifetime risk of development of Kaposi sarcoma in men infected with human immunodeficiency virus, *Blood* **96** (2000), 2562–2567.
[42] T.L. Lehrnbecher, C.B. Foster and S. Zhu et al., Variant genotypes of FcgammaRIIIA influence the development of Kaposi's sarcoma in HIV-infected men, *Blood* **95** (2000), 2386–2390.
[43] E. Israel, J.M. Drazen and S.B. Liggett et al., The effect of polymorphisms of the beta(2)-adrenergic receptor on the response to regular use of albuterol in asthma, *Am J Respir Crit Care Med* **162** (2000), 75–80.
[44] E. Israel, J.M. Drazen and S.B. Liggett et al., Effect of polymorphism of the beta(2)-adrenergic receptor on response to regular use of albuterol in asthma, *Int Arch Allergy Immunol* **124** (2001), 183–186.
[45] L. Lennard, J.A. Van Loon and R.M. Weinshilboum, Pharmacogenetics of acute azathioprine toxicity: relationship to thiopurine methyltransferase genetic polymorphism, *Clin Pharmacol Ther* **46** (1989), 149–154.
[46] E.Y. Krynetski, H.L. Tai and C.R. Yates et al., Genetic polymorphism of thiopurine S-methyltransferase: clinical importance and molecular mechanisms, *Pharmacogenetics* **6** (1996), 279–290.
[47] P. Garred, T. Pressler and H.O. Madsen et al., Association of mannose-binding lectin gene heterogeneity with severity of lung disease and survival in cystic fibrosis, *J Clin Invest* **104** (1999), 431–437.
[48] N. Arnheim, C. Strange and H. Erlich, Use of pooled DNA samples to detect linkage disequilibrium of polymorphic restriction fragments and human disease: studies of the HLA class II loci, *Proc Natl Acad Sci USA* **82** (1985), 6970–6974.
[49] S.H. Shaw, M.M. Carrasquillo, C. Kashuk, E.G. Puffenberger and A. Chakravarti, Allele frequency distributions in pooled DNA samples: applications to mapping complex disease genes, *Genome Res* **8** (1998), 111–123.
[50] G. Breen, D. Harold, S. Ralston, D. Shaw and D. St Clair, Determining SNP allele frequencies in DNA pools, *Biotechniques* **28** (2000), 464–466, 468, 470.
[51] S. Germer, M.J. Holland and R. Higuchi, High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR, *Genome Res* **10** (2000), 258–266.
[52] L.F. Barcellos, W. Klitz and L.L. Field et al., Association mapping of disease loci, by use of a pooled DNA genomic screen, *Am J Hum Genet* **61** (1997), 734–747.
[53] P.H. Joosten, M. Toepoel, E.C. Mariman and E.J. Van Zoelen, Promoter haplotype combinations of the platelet-derived growth factor alpha-receptor gene predispose to human neural tube defects, *Nat Genet* **27** (2001), 215–217.