REVIEW

Online Resources for SNP Analysis

A Review and Route Map

Christopher Phillips

Abstract

The major online single nucleotide polymorphism (SNP) databases freely available as research tools for genetic analysis are explained, reviewed, and compared. An outline is given of the search strategies that can be used with the most extensive current SNP databases: National Centre for Biotechnology Information (NCBI) dbSNP and HapMap to help the user secure the most appropriate data for the research needs of clinical genetics and population genetics research. A range of online tools that can be useful in designing SNP genotyping assays are also detailed.

Index Entries: Single nucleotide polymorphism, SNP; genotyping; variation; polymorphism; online databases; linkage disequilibrium; haplotype; haplotype block; genome; HapMap; clinical genetics; population genetics.

1. Introduction

The validated set of human single nucleotide polymorphisms (SNPs) is one of the most valuable resources to have come from the human genome mapping project (HGP). This dataset was largely unforeseen in the initial phases of sequencing but as several genome equivalents were generated from different individuals during the project, the identification of SNPs became a by-product of growing importance. The discovery of new SNP sites by resequencing and the collation of detailed information about each locus has continued to develop both in scope and in depth as the simultaneous publication of draft sequence and first full SNP map 5 yr ago ***(1,2)***. Added to this, our knowledge of SNPs has expanded in parallel to an improved understanding of the structure and function of the genome and its gene content. The SNPs provide the most complete and densely spaced system of genome landmarks available. Above all, this enables researchers to improve the resolution of linkage maps, in particular in the precise mapping and study of genes contributing to complex disorders, where the contributory effect of individual genes is small or where genetics is compounded by multiple locus interaction and environment. As well as enabling greatly enhanced mapping, SNPs provide a new and fascinating layer of detail to our understanding of human variability and what this tells us about disease susceptibility, populations, and evolution. Because SNPs have low recurrent mutation rates compared to other polymorphic markers they provide the most stable and reliable indicators of the evolutionary history of populations. Last, but not least, SNPs are the very stuff of genetic variation as a substitution in or near a transcribed region can change an amino acid sequence, the control of transcription events, or the splicing pattern for the resulting RNA products. Such changes arise from

*Author to whom all correspondence and reprint requests should be addressed. The Spanish National Genotyping Centre CeGen, Santiago node, Genomic Medicine Group, University of Santiago de Compostela, Galicia, Spain. E-mail: c.phillips@mac.com.

 ISSN: 1073–6085/Online ISSN: 1559–0305/2006/35:1/065–098/$30.00

loci commonly termed coding SNPs, promoter–site SNPs, and splice–site SNPs, respectively. Therefore, it can be argued that SNPs form the principal part of the variability affecting each stage of gene expression and can justifiably be said to influence the transcriptome and proteome, as well as the genome itself. As such, SNPs are more than just ubiquitous marker points, they are a core genome feature that will ultimately help to explain one of the enduring mysteries of human genetics: why comparable numbers of genes in species at opposite ends of the evolutionary scale can create such profoundly different levels of complexity.

In the same spirit as that prevailing in the database management of HGP draft and final nucleotide sequence, SNP data arising from public genome analysis have been open source from the beginning; freely available in websites accessible by any researcher. Furthermore, the largest database, dbSNP, accepts submissions directly from the scientific community, allowing the rapid dissemination of newly discovered SNP polymorphisms. In fact all the SNP databases described in this review are based on open access and in this sense are a shared scientific resource. Added to this, during 2006 the bulk of privately held SNP data will become available to the whole research community. This means one of the original precepts of HGP—free access to the best possible information acts to accelerate scientific progress, equally applies to the overriding majority of human SNP data. Unhindered access to information is considered to be of such primary importance now that a system for data release has been formalized in the recommended framework for international community resource projects (http://www.welcome.ac.uk/doc_WTD003208.HTML). This framework secures, for all scientists, rapid dissemination, highest standards of database management, and the end use of data without restriction.

At the same time that genomic data started to become available, the fields of bioinformatics and database management were rapidly developing to keep pace with the scale of the information generated and the complexity of the searches required. These developments have meant that searching for specific SNPs among a total of 9 million or more human loci can now be achieved fairly easily, provided the researcher begins with clearly determined criteria for the markers required for a study. To this end, the I can firmly recommend, as starting points, two excellent textbooks: *Human Molecular Genetics* and *Human Evolutionary Genetics* ***(3,4)*** that each cover thoroughly the fields of genetic analysis and population genetics, respectively, these forming the two principal applications for SNP analysis. Both books provide, in nonscientific parlance, "one-stop shops" for understanding the extent and specific characteristics of SNPs needed for planning studies of inherited human disease (referred to as clinical genetics from this point on) or population genetics. I also consider that it is important for scientists in one area of speciality to be better acquainted with the other, so the use of *both* books is recommended. Careful project planning is an essential step, in addition to the scrutiny of appropriate reporting publications, before any database search starts in earnest. Therefore, the major steps of project design involving the access of online resources can be visits to PubMed (for the best current reports in the field), HapMap (for the review of SNP candidates in, or close to, genes or genome regions of interest), dbSNP (for the characterization of the chosen SNPs), and finally use of various web tools for optimizing the genotyping assay design. This is not intended to be a prescribed route—many of the sites outlined in this review have viable alternatives that can act, above all, to supplement the data obtained from the primary sources described here. In addition, particular areas of clinical genetics such as cancer studies have specialized databases that are not covered in this review but clearly serve their field with a more direct focus of information.

This review was originally written for a third potential application of SNP analysis, that of forensic identification, commonly termed DNA profiling. The SNPs offer the potential to greatly reduce amplified product size compared to existing profiling loci and therefore, could provide

much improved results with the highly degraded DNA commonly encountered in forensic analysis or disaster victim identification.

Forensic profiling requires a specific set of search criteria to find SNPs with the necessary properties for this field and readers interested in the forensic application of SNP analysis can find more specific guidelines in the original article along with several specialized reviews *(5–7)*. It is interesting to note that forensic SNP sets may find applications in genetic studies where measurement of stratification and admixture ratios could be achieved using a discrimination and geographic origin marker panel. These equally specialized areas of SNP use in genetic analysis are discussed in the overview of SNP biology.

Therefore, the purpose of this review is twofold: to compare SNP databases and to detail the methods that can be used to check the characteristics of SNP markers of most relevance to a proposed study. Emphasis has been placed on ways to narrow the selection of SNPs (or any other supporting data) to a manageable size—an increasingly important strategy given the huge scale of information available. In the final section several online resources such as Basic Local Alignment Search Tool (BLAST) and RepeatMasker, which provide invaluable tools for the design of robust and reliable SNP genotyping assays, are discussed as a supplementary resource to the purely genetic databases.

Finally, during the preparation of this article wholesale changes to the availability of a substantial proportion of private SNP data were taking place. Although the review was not intended to cover searching private SNP databases—in particular, the 4 million SNPs comprising the Celera database known as Celera Discovery System (CDS)—all the information previously held by Celera on these SNPs is in the process of transfer to dbSNP. This means the extent of freely accessible SNP data is certain to expand still more in the coming year. Since Celera SNPs were discovered using different approaches to the public SNP mapping initiative the expansion should provide a considerable number of completely new markers (potentially as many as 1 million), despite the majority of commercially held Celera loci being common to both public and private databases. Some idea of how much supplementary data may eventually be released for public access can be obtained from the time taken by National Centre for Biotechnology Information (NCBI) to scrutinize, validate, and reorganize the Celera SNPs. This has already occupied 6 months of 2005 and is likely to go beyond a full year of data processing on a large scale. This underlines the continuing growth in importance of SNPs to provide both sufficient coverage for the analysis of all parts of the genome and to help fully understand the complex mechanisms of gene expression.

2. The Biology of SNPs—A Brief Overview

The SNPs, as base modifications, represent changes to the ancestral DNA sequence comprising a genome. Because the cellular mechanisms for correcting a base mismatch are extremely effective, it is necessary to understand how these substitution events progress from a single sequence change that is promptly edited back to the correct base, to become allelic (i.e., the variant base is polymorphic with a minor allele frequency greater than 1%). Substitutions, once reaching this frequency level, become true single nucleotide polymorphisms, a self-explanatory term but note that the term SNP also encompasses single base insertions and deletions (commonly termed indels). Two processes create substitution-based SNPs: incorrect base incorporation during DNA replication and in situ chemical modification of a base. The first of these two processes, generating new SNPs from nucleotide misincorporation at DNA synthesis, is an extremely rare event given the fidelity of DNA polymerase enzymes and the elaborate proofreading mechanisms in place to check physical alignments in fresh sequence copies. The frequency of misincorporation has been estimated to be approx 10^{-9}–10^{-10} per nucleotide *(8)* and this event alone is not sufficiently common to account for the huge number of SNPs with detectable minor allele frequencies observed in all organisms studied so far. This is an important

point because a feature of SNPs often cited in favor of their use in both population genetics and association studies is the long-term stability of substitutions that have become fixed in a sequence, stemming directly from a very low recurrent mutation rate. Consequently, the alternative process of in situ modification of bases must account for the majority of SNPs generated. Although a range of factors external to the cell can affect a chemical transformation of a base, most notably high-energy radiation, the usual repair processes are, again, far too effective at correcting changes to account for observed SNP frequencies. The real clue to the most widespread SNP generation event comes from the fact that C-T and A-G SNPs far outnumber all other types of substitution. Furthermore, the rate of substitution observed within CG dinucleotides (usually termed CpG) is an order of magnitude higher than for all other dinucleotide motifs. The CpG dinucleotides have two critical characteristics. First, they are the target for methylation: a universal process of base modification affecting 75% of CpG dinucleotides that plays a central role in the control of gene expression. Second, they exhibit dyad symmetry, that is, the complementary base sequence is also CG, therefore methylation effects both strands. Once cytosine is methylated to become 5-methylcytosine it can undergo deamination to form a stable thymine base surviving on the uncorrected strand. In short, CpG can become TpG or CpA with equal likelihood and in half of these cases the original strand is corrected by the repair machinery rather than the deaminated strand, leading to a stabilized substitution event. Actually the mutability of CpG is such that these motifs only occur at 20% of the frequency predicted by normal base composition. When CpG motifs do occur at high frequency, in particular the transcription control regions at the 59 ends of genes, the tendency is for CpG to escape the destabilizing process of methylation, forming the characteristic CpG islands that can often act as telltale signatures for spotting putative genes. All this adequately accounts for the bulk of SNPs generated throughout the human genome and explains the overall high frequency of SNPs compared to other sequence changes such as indels (found at only ~10% of the frequency of substitution SNPs). It also predicts that the great majority of SNPs comprise C-T or A-G substitutions, which is indeed the case. Last, there is no reason why a substitution event cannot occur more than once at a given base position, however rare such an event is predicted to be and I have successfully located many nonbinary SNPs in online databases with the aim of developing sets of these loci for mixture detection in forensic analysis ***(9)***.

Although the mechanism behind the generation of SNP variation had been well characterized, it was not until the completion of the HGP that the distribution, density, and diversity of SNPs could be viewed on a scale large enough to assess how adequate, or otherwise, the marker coverage was going to be. If any extensive gaps in SNP distribution occurred in the genome, then SNPs could not be universally applied for the fine mapping of every gene candidate, or used to track a complete set of genes when large numbers of these acted with small additive effect in multigenic traits. The international SNP map working group described in 2001 ***(2)*** the first systematic study of the whole set of SNP variation as it stood at that time, following comprehensive sequence comparisons between the multiple donors used for the HGP sequence compilation. The analysis of the 1.42 million markers in this first map provided a detailed picture of the characteristics of human SNPs and allowed the appropriate planning of the future development of SNP maps, ultimately giving rise to the work of the SNP consortium Allele Frequency Project and the HapMap initiative. The most important finding with a bearing on the usefullness of SNPs as linkage markers was that the distribution of SNPs is constant throughout the autosomal chromosome set. In addition, the vast majority of the genome was seen to contain SNPs at suitably high density, with only 4% of sequence showing frequencies less than one SNP per 80 kb, much of this comprising incomplete SNP mapping at the time. This was an important finding, not just for the future prospects of SNP-based linkage analysis but as a contrast to the studies of gene density from HGP data that were

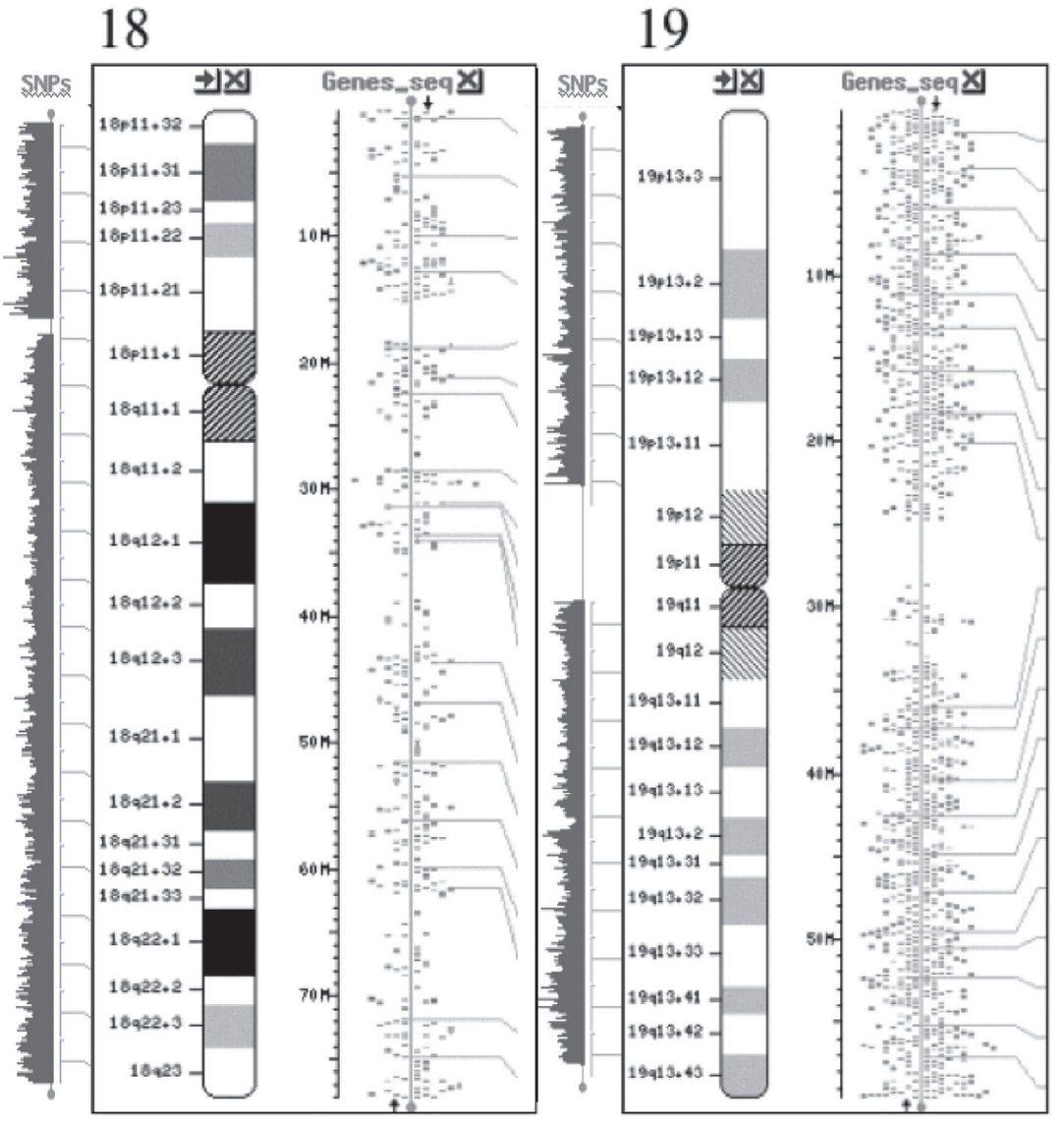

Fig. 1. SNP and Gene density plots for two autosomes of comparable size: chromosomes 18 and 19.

revealing considerable disparities between chromosomes. This is shown by the contrasting numbers of genes observed on chromosomes 18 and 19—a pair with almost identical lengths but at opposite ends of the gene density scale. **Figure 1** illustrates the gene distributions alongside the SNP density plots for both chromosomes. In line with distribution and density, patterns of SNP variability, measured as nucleotide diversity, were also remarkably consistent between the autosomes. All showed nucleotide diversities within 10% of the mean, except for the high and low value outliers of chromosomes 15 and 21, respectively, and the major histocompatibility complex (MHC) region of chromosome 6. The X and Y chromosomes differ from autosomes in both effective population size and mutation rate and the lower observed densities and levels of

heterozygosity for nonautosomal SNPs matched predictions based on the tendency of these chromosomes to show reduced diversity. Although SNPs are binary loci with a much lower average heterozygosity than any other polymorphisms, it was evident from the HGP map that their balanced distribution and levels of variability made the use of genome-wide high-density SNP mapping a viable prospect.

Although a vast number of well-characterized SNPs can provide a wealth of marker sets for association studies, the actual level of polymorphism shown by individual SNPs is limited because many closely neighboring loci are bound together in near-complete linkage as chromosome segments termed haplotype blocks (alternatively linkage disequilibrium or LD blocks). Haplotype blocks occur because the chromosome segments have shared ancestry, therefore that particular combinations of SNP alleles in close proximity comprising the haplotype are changed only slowly by recombination or accumulated mutation. Human evolutionary history has the unusual characteristics of involving very small population sizes and a relatively short total period of development of the species. As a result the human genome is particularly amenable to haplotype block analysis because the math dictates that the rate of erosion of haplotypes is slow ($\sim 10^{-8}$ per base, per generation) and the number of generations adter the disease variant mutation is relatively small ($\sim 10^{-4}$–10^{-5}). Using haplotype blocks as the basis for SNP analysis can have significant advantages in association studies, as genomic regions can be tested without the need to first pinpoint the location of the functional variants of the trait. However, haplotype block structure adds an additional layer of complexity because there are no guarantees that blocks are consistent in size or distribution, coincidental in position to the genes of interest or even very close ***(10)***. In fact, initial analysis using patterns of association between SNP pairs revealed a nonrandom distribution of linkage disequilibrium in the genome, although the blocks themselves consistently showed much reduced SNP variability with only a limited number of allele combinations per block. This latter characteristic had the potential to erode the power of SNPs as association markers, as the normal and variant gene might both share the same associated haplotype in a great many cases. For this reason, there is much interest in allele frequency differences between the major population groups, as it provides a chance to choose the best study group for a particular set of blocks. Haplotype blocks are clearly inconsistent with the established model of genetic distance measurement based on a predictable recombination rate and it is generally agreed that they represent areas of extremely low recombination bounded by smaller segments where recombination is much more frequent, so-called hotspots ***(11)***. For this reason an important phase of study after the SNP map publication was the analysis of haplotype block diversity and structure. Several studies examining chromosome-wide LD distribution found that blocks can span relatively large distances ***(12–14)***. Two studies obtained comparable block length values from different chromosomes: Daly et al. ***(14)*** reported a size range between 3 and 92 kb on 5q, whereas De La Vega et al. ***(15)*** found a wider range of block size for chromosomes 6, 21, and 22 (5 to 300 kb), but with an average size of just 26 kb and 18 kb in Europeans and Africans, respectively. The longest block discovered at this time was 800 kb ***(11)***, but blocks longer than 100 kb are normally expected to be rare (representing only 2% to 3% of the total found on chromosomes 6, 21, and 22) ***(15)***. Many of the principles applied to haplotype block mapping and underlining the HapMap approach still do not meet with universal support and the main arguments are more fully discussed by Wall and Pritchard ***(16)***. However, the outcome that would suit most researchers would be to be sure that a SNP at a given distance from a gene variant could adequately mark the variants underlying the phenotype and therefore act as a tag for comparing case subjects to control subjects. This is the principle of positional cloning, where progressively finer mapping can focus on the positions of contributing loci in the absence of clear information about gene action. Such an approach can potentially reduce the genotyping efforts needed to examine complex disorders to

the point where a very small number of tagging SNPs might enable the tracking of the multiple variant genes creating complex disorders. This can be further simplified by attempting analysis on populations with a history of reduced variability due to small founder numbers, bottlenecks, reproductive isolation, or endogamy, hence, the widespread interest in isolates such as island populations or small, culturally coherent groups like the Anabaptist sects of North America. At present, population genetics and association studies start to share a considerable amount of common ground. Furthermore, much can be learned from differences in the frequency of, and levels of susceptibility to, common diseases among the five major population groups. These are broadly based on the continental boundaries of Africa, Europe, Asia, the Americas, and Oceania (Pacific island groups). The HapMap initiative (of which more later) was set up to apply the principles outlined to delineate haplotype blocks in the genome. To develop the exploration of population differences, phase I of HapMap SNP analysis has genotyped all loci in 90 subjects each from Africa and North America (of European descent) plus 45 subjects each from China and Japan.

To conclude this section mention should be made of two important aspects of association studies using SNPs where a prior understanding of the characteristics of the study population is particularly important: admixture mapping and stratification effects. Admixture mapping (often termed MALD, for mapping by admixture linkage disequilibrium) is exploiting the differences in haplotype block boundaries between admixing populations, previously isolated, to gain higher levels of association ***(17)***. This is because recombination over the limited number of generations because admixture has less opportunity to disrupt the associations between SNPs and genes shown to different degrees by each contributing population. In particular, African Americans are an informative population for association studies with an estimated level of admixture with Europeans of about 20%, although the level has a broad range from 4%-30% depending on the US region ***(18)***. Stratification effects occur when the trait studied and the genetic variation as a whole are both clustered into strata within a population that is presumed to be homogeneous for both. As a result, the overall within-population variation approaches levels seen between groups of study and control subjects because the groups do not share identical ancestries. Stratification can potentially create spurious associations between the traits studied and a set of SNPs chosen to map them ***(19)***. An example of this effect is shown by the SNPs rs182549, which shows a strong allele frequency gradient (cline) from northern to southern Europe as does the trait studied—adult stature. The association between trait and variation in this case was entirely the result of an identical distribution of variation and was lost when individuals were rematched on the basis of geographic latitude within Europe ***(20)***. Interestingly, forensic discrimination SNPs can make effective measures of stratification, as such loci are required to be neutral, freely assorting, and highly polymorphic, characteristics not found in the association study SNPs. Once again these aspects highlight the importance of a thorough understanding of human population structure and history in the design and interpretation of clinical genetics studies.

3. NCBI: PubMed, Entrez, Boolean Principles, and Databases of Relevance to SNP Analysis (http://www.ncbi.nlm.nih.gov/)

Any search for information relating to genetics and medicine should begin at the homepage of the National Centre for Biotechnology Information (NCBI) website. NCBI is one of the institutes of the US National Institutes of Health (NIH) and has been the principal worldwide repository of genomic data for the past 18 years. The nucleotide sequence variation database housed at NCBI is known as dbSNP and comprises the largest collection of SNPs available with the most comprehensive set of supporting data for each locus. The following sections detail the structure and use of dbSNP specifically but the importance of NCBI outside of dbSNP is that genetics research involving SNPs should always be set in the context of supporting information. In particular, it is impor-

tant to gather, at the same time as SNP data, information about genes, phenotype, proteins (both chemical and structural variation), expression dynamics, and the contextual sequence surrounding the SNP. Given the extent of the biological databases at NCBI it is of no surprise that this institute has built the most extensive collections of gene (Gene), inherited disorder (OMIM), protein (Protein), gene expression (GEO), and nucleotide (GenBank) databases in the past 8 yr. Together with the Santa Cruz genome database (http://genome.ucsc.edu/), NCBI has managed the content of each of the draft versions of the human sequence since 1990 and now keeps the reference sequence, completed in April 2003. This comprises 2.9 billion bases (99% coverage of gene-containing DNA) with an error rate of 1 in 10,000 bases. To most biologists NCBI will already be familiar in the guise of Medline and PubMed: two bibliographical databases that collate all the principal citations from biomedical journals (~5000 journals in total). Medline was the main source of data for PubMed, but has largely been supplanted, the two are still distinguished by the fact that PubMed has a broader scope by including articles predating the Medline selection and by containing certain "out of scope" content (i.e., not biomedical). The NCBI bibliographic databases are by default predominantly text oriented. This means they work by matching text recognized in the query submission to text in the data records and to work efficiently the system needs to regulate vocabulary. The database of words used to index PubMed is MeSH (medical subheadings) and this can be searched itself using the search menu in the top left of each NCBI homepage. For association study research, checking MeSH is a useful step to help clarify terminology before a search begins and to review in brief possible related areas of study to the disease of interest. For example, the query Repetitive Strain Injury gives the three specific medical terms used to describe varieties of this condition with the year of introduction of each. For clarification at the top of the page are the alternative terms named suggestions that are also used in PubMed for text matching with the query term (e.g., repetition strain injury).

The importance of prior experience in the use of PubMed is that both PubMed and dbSNP are components of a unified NCBI database retrieval system termed Entrez. Therefore, familiarity with bibliographic search strategies that can enable a manageable list of published articles from the 12 million available at NCBI can help the user to develop the same principles for SNP searching. NCBI uses Entrez as a standardized query interface for all the major databases it manages. Therefore, using Entrez has two clear benefits for the user. First, the query system is the same for each Entrez database and second, the searches can be made global to return data that is then seen to be interlinked between many databases. Data returned from multiple sources are linked by cross-referenced hyperlinks termed linkouts (i.e., two-way connections between each database). To work efficiently and to delineate a search correctly Entrez relies on an understanding of Boolean terms and combinations of parameters, termed fields, that define the required characteristic from a piece of data. For instance, searching NCBI using just the search term *diabetes* gives a quarter of a million literature citations alone and equally daunting numbers from the other databases. When diabetes is used with the operator "AND" in combination with the term *chromosome 2*, the returned PubMed citations drop to a much more realistic 140 articles, summarizing several studies, among others, analyzing the interleukin 1 gene cluster on chromosome 2 implicated in diabetes susceptibility. Not surprisingly the number of genes listed in EntrezGene drops in similar fashion from 1114 to 3, again, acting to define a specific genome feature before any follow-up searches have begun. This simple predefinition of search terms can be particularly useful for the process of designing a new genetic study when it is important to check that the work to be undertaken is both manageable and has enough leads to instigate database research in earnest, or equally important, has not already been achieved elsewhere. Therefore in the initial stages the PubMed and OMIM text-based databases (termed literature databases in NCBI) are as important as the genetic content databases (termed molecular).

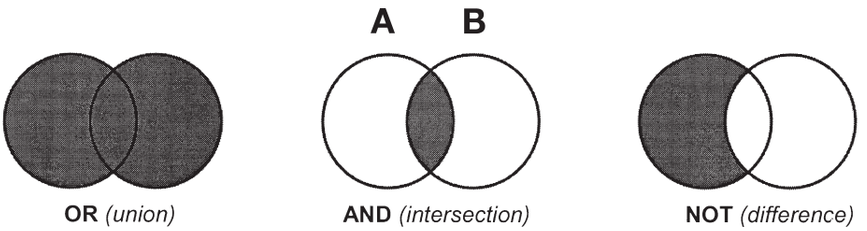

Fig. 2. Boolean operators. OR applies to all items in A **or** B, AND to items found in both A **and** B, and NOT to items in A **not** found in B.

Boolean terms govern the rules used in all database searching by applying the principles of logic that define the relationship between a set of inclusive or exclusive terms, ruled by the three operators (or operands): AND, OR, and NOT. The logic is summarized in venn diagram form in **Fig. 2** and each operator can be described as follows:

- **OR** (often termed union) is inclusive, therefore it returns all database entries that contain at least one of the provided search terms.
- **AND** (often termed intersection) is exclusive, therefore it only returns database entries that contain all the provided search terms.
- **NOT** (often termed difference) excludes from the returns all database entries with the provided search terms.

Nearly all Entrez database queries use the operator AND to narrow down a search using multiple combinations of an extensive array of fields, each set being tailored to the content of the database. This approach is sometimes known as a relational search, as it looks for relations or links between the search terms found in each data record. If search terms are to be confined to a specific field Entrez rules require that these are described using predefined tags termed field tags and set in square brackets. For example, entering "short [au]" returns publications in PubMed with Short as author and the list will not include studies of short tandem repeat loci (unless Short is an author!). Descriptions and details of the whole PubMed field tag list are outlined at (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=help pubmed.box.pubmedhelp.Box_1_Search_Field_D). Entrez searches default to searching all fields in the absence of field tags and to the operator AND, not OR, if there are spaces between fields, as is common practice elsewhere. For most of the molecular databases the text of a tagged field must be in the correct format, termed syntax to be automatically read by the NCBI search engine and this can initially be a source of frustration until Entrez formatting becomes second nature. To begin using Entrez it is better to use a menu of fields for a single database search from the limits tab that appears on each database homepage. These take two forms depending on the type of information held in the database: additional query description boxes or tick-boxes and query description boxes. An example of the first might be a PubMed search that can work with one of 29 limits such as author or text word. These still require an entry in the query box but some assistance is given for combining these with extra intersections by limiting certain fields at the same time using a small number of choices in seven additional areas such as text language or human/animal subject matter. The date range for publication or adoption to PubMed (Entrez Date) gives the most useful field limit in combination with text word as it ensures the returns concentrate on the most recent publications. Overall, the underlying theme is to allow simplified queries that still enable manageable numbers of returns from the largest databases. The second form of limits using tick-boxes is used for many of the molecular databases, notably SNP (termed EntrezSNP in this guise) where a clear level of categorization is possible with the data. Therefore, for example, ticking one of the 15 IUPAC substitution codes (e.g., Y to denote C/T substitutions [listed in

Table 1
IUPAC Codes Used With the [ALLELE] Tag Denoting SNP Base Substitutions

Code	Substitution	Code	Substitution
M	A or C	**V**	A or C or G
R	A or G	**H**	A or C or T
W	A or T	**D**	A or G or T
S	C or G	**B**	C or G or T
Y	C or T	**N**	A or C or G or T
K	G or T		(or indeterminate base)

A, **C**, **G**, and **T** can be used individually to select all SNPs exhibiting that base as an allele.

Table 1] and sometimes termed IUPack) starts the process of concentrating search terms down to a specific set of data to provide focus. The SNP field list is one of the most extensive and the use of terms is outlined in greater detail in the dbSNP section.

Three *modifiers* of operator function can be used in Entrez:

- **Ranging:** setting a range for a value in a field (e.g., SNP heterozygosity) using a colon (:) between the lower and upper limits for the value.
- **Parentheses:** combining related terms together as logical groups and forcing the order of operation for the search process or performing a common operation on a group of terms. In the first case this sets the order of searching to the bracketed terms first so that the next operation is performed on the results of the first operation. In the latter case brackets are commonly used to group together NOT items combined using OR. Entrez syntax uses curved parentheses for grouping and square parentheses for field tags. Brackets are helpful in ensuring descriptions lacking clear definition such as "learning disorders" are replaced by a broadly based set of more specific terms that in combination keep the focus but prevent false exclusions from returns. In place of the above query term, using (dyslexia OR attention deficit hyperactivity disorder) together ensures the combined returns with either term are available for the next operation in the query. The example used in PubMed help is apt because it illustrates that use of parentheses can mirror the logic of a sentence: so "find articles on the effects of heat and humidity on multiple sclerosis" takes the form: (heat OR humidity) AND multiple sclerosis.
- **Wild card:** using a star in place of missing text allows a partial entry to be used as a query term (e.g., using BRC* will find both BRCA1 and BRCA2). NCBI does not generally use adjacency searching in the molecular databases. This is based on the proximal operator NEAR, routinely used by web search engines like Google. A notable exception is the alternative text terms named suggestions that are handled by MeSH in PubMed. Using adjacency searches tends to lack focus for the majority of data in NCBI, as information is usually clearly and unequivocally categorized.

It is possible to use Boolean terms to combine individual Entrez searches performed at different times as an alternative to using parentheses, enabling more opportunity to monitor the number of returns with different search term combinations. This uses the clipboard and history tabs. The clipboard is a workspace for holding up to 500 items manually selected from search returns. *Note:* (before losing work) that contents are cleared after 8 hours of inactivity. History lists the database search activity as numbers prefixed by a hash (#). Because these records are, again, cleared after 8 hours of inactivity it is worth getting into the habit of transferring long multistep search records into the personal folders available as "My NCBI" (http://www.ncbi.nlm.nih.gov/entrez/login.fcgi?call=so.SignOn..Login&callpath=QueryExt.CubbyQuery..ShowAll&db=pubmed). Previous searches can be combined as hash fields and using Boolean operators (e.g., #1 AND #2 gives an intersection of the first two searches from the current active session). It is also possible to use hash fields together with normal fields, thus helping to build a stepwise record of the search process as it is modified to reduce return numbers in small stages. Finally, it is logical to fix the values of certain fields in Entrez to filter down the number

of returns. For instance, choosing human as the organism is wise as much SNP data are now held for the mouse genome. These options for EntrezSNP searching are discussed in more detail in **Subheading 7.**

Several key NCBI databases have an important place in clinical genetics research design or in support of dbSNP searches. These include GenBank, Gene, Online Mendelian Inheritance in Man (widely referred to as OMIM), and UniSTS. This is just a fraction of the total collection and represents the most useful databases for adding data already obtained from dbSNP, PubMed, Map Viewer, and MeSH. What follows is a brief outline of the structure and use of the four databases that can be accessed in combination with dbSNP searches.

3.1. GenBank

GenBank comprises the nucleotide sequence database of NCBI. This simple description belies the scale of the information held—a collection of sequences comprising 59 gigabases of data from more than 130,000 species (spanned by 17 different genetic codes), which is updated daily. The database organization involved is equally complex but the front end is simple enough if the user needs just a sequence segment, the coordinates, and the amino acid translation sequence. The example sequence file on the GenBank homepage illustrates a standard sequence report with the fields that can be used in searches. GenBank is part of Entrez, has its own specialized fields, and is a subset database under the "umbrella" group of EntrezNucleotide. This comprises sequence subsets for expressed sequence tag data (dbEST), genome survey sequence data (dbGSS), and Core Nucleotide—the subset of interest to most users containing genomic sequence data. This allows joint searches in Entrez and individual searches in BLAST, avoiding cross-referencing when it is not needed. To add more complexity and another name to keep in mind, there is also the RefSeq database. This can be thought of as the reference sequence set for the key study organisms (3244 in early 2006) with integration, meaning the included sequences can be optimally compared, as can the annotation, the process of characterizing a gene from the base sequence. RefSeq is not part of Entrez but the protein sequence portion can be searched in Entrez as EntrezProtein. For most users interested in SNP searches the main contact with nucleotide databases is in the process of designing genotyping assays. Therefore it is normally necessary to check flanking sequence for quality or presence of clustering SNPs and to check potential primer designs using BLAST. There is not a great need in most cases to go deeper than this.

3.2. Gene

Gene comprises the NCBI gene directory previously known as LocusLink. Maintaining this database is particularly challenging as the gene landscape is constantly changing in so many aspects. Definitions of function, interactions, association with a trait, activity, and many other characteristics are regularly revised and this must be collated with equal regularity. Gene works with unique identifiers assigned to three types of gene entities: genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information. These gene identifiers are tracked, and new information is added when available. The scope of Gene to encompass all organisms supersedes LocusLink, which was centered solely on human gene data. Searches generally begin with an identifier in the form of a letter/number combination standardized by HUGO (http://www.gene.ucl.ac.uk/nomenclature/). The term does not have case sensitivity but care is needed to avoid spaces (treated as an AND operator and usually failing to find the target). The list includes all the species with genes matching the query combinations but here case becomes an important point of distinction. Query "ABCC1" returns ABCC1 in humans and Abcc1 or ABCC1 in other mammal species—all essentially the same gene, but also abcC1 in *Dictyostelium* sp., which is a different gene. Each linkout in the list gives a single report page starting with a two-line header including the full name and a unique

geneID number that can be used in Entrez by itself. This is followed by sections: summary information; graphic summary of transcription structure; graphic summary of genomic context with a linkout to MapViewer; bibliography; general gene information; general protein information; RefSeq sequences; related sequences and additional links. This is comprehensive enough to provide most of the search directions needed to explore the context of the gene as well as the likely critical characteristics that can help the researcher assess its status as a candidate for study. Invariably, the most useful section is the bibliography—a comprehensive catalog of relevant studies of the gene that allows an easy check of current interest in the gene in the context of a particular disease. Reference to the gene symbol denotes the name written as upper case letter/number combinations, gene ID denotes the 5-digit number.

3.3. OMIM

Online Mendelian Inheritance in Man or OMIM is the NCBI phenotype database. In this role it catalogs traits, diseases, and disorders, although not always with reference to a gene if no association has currently been described. The list of entries in early 2006 reveals the paucity of understanding of the genetic basis for disease. Of a total of 16,612 entries only 384 list a gene with a known sequence and phenotype (unique OMIM number prefixed with $^{+}$). Another 2229 describe a phenotype with a suspected Mendelian inheritance pattern (no prefix), 1502 with a well-described phenotype lacking a described molecular basis ($^{\%}$), and 1862 with a known molecular basis ($^{\#}$). This leaves the remaining 10,635 entries listed as merely genes with known sequence (*), therefore the majority of OMIM data consists of genes lacking known phenotypes. Each entry has a unique number often used in the literature to denote a condition and usable as a search term throughout NCBI. OMIM compensates for the lack of concrete data by being very readable and informative about gene function. The noticeable difference in the character of the database is because OMIM is hosted by NCBI and developed independently at John Hopkins University. Using OMIM as the basis for a literature search is usually a fruitful approach, the text acts to review the quality of the associations suggested by studies in the area of interest under headings cloning, gene function and structure, mapping, molecular genetics, and animal models. The references at the end of the OMIM report are a selected list of the most relevant studies to provide the clearest direction for further investigation. As an example that highlights the difficulties of collating very extensive phenotype data with genetic data there is no mention in an exhaustive report for gene ABCC1 of the effect of coding SNP: rs17822931 on human earwax viscosity ***(21)***. The OMIM mapping section is particularly helpful as a starting point when planning an association study with SNPs or seeing whether further linkage analysis is needed as a preliminary stage of study.

3.4. UniSTS

UniSTS comprises the NCBI database of linkage markers termed Sequence Tagged Sites (STS) and leads on from the above point about linkage analysis. UniSTS content encompasses polymorphic loci other than SNPs available for linkage analysis (predominantly short tandem repeat sites) and can be searched using gene identifiers or chromosome position. The return page gives related information and helpfully recommends polymerase chain reaction (PCR) primers to simplify the development of linkage marker typing if this is required. The primer pair information is used to match alternative names for linkage markers, therefore, for example, "D2S2300" will retrieve the marker named in the database as "AFM261YB1." The easiest way to combine STS markers and SNPs to fully cover an area of interest with suitable linkage markers is to use the Between Markers search option in the dbSNP homepage (http://www.ncbi.nlm.nih.gov/SNP/index.html). This provides two query windows for inserting the STS ID's and a list of SNPs is returned that spans the distance in between.

4. The dbSNP Database (http://www.ncbi.nlm.nih.gov/SNP/index.html)

NCBI dbSNP is the principal database of SNP information generated from the HGP and the simultaneously published first SNP map. dbSNP has continued to collate all the data from various SNP validation initiatives that have followed since, including output from The SNP Consortium (principally the Allele Frequency Project), The Perlegen SNP genotyping initiative, and HapMap. The SNP data are regularly updated in synchrony with genome rebuilds, ensuring the highest quality of SNP locus mapping and scrutiny. In early 2006, the total dataset amounted to 40.6 million different SNP loci, with just under half of this number validated by genotyping a sample set to confirm polymorphism. The human content comprised 10,430,750 SNP clusters (i.e., rs-numbers) of which 4,236,590 were sited in genes. A total of 35 organisms have individual SNP databases, 12 of these from completed genomes. The Chimpanzee dataset is likely to be of growing importance and currently comprises 1.54 million SNPs clusters with just more than a third of these in genes. With this pace of data building it is a good idea to subscribe to the dbSNP-announce automatic e-mail update system to keep updated on developments (http://www.ncbi.nlm.nih.gov/mailman/listinfo/dbsnp-announce). As well as reporting the release of each new build, announcing newly added features, and outlining corrections or discovered problems with past or present builds, there is an archive for referencing possible problems with, or qualifications to, previously obtained search data (http://www.ncbi.nlm.nih.gov/mailman/pipermail/dbsnp-announce/). The rapid growth of human SNP data in dbSNP during the past 5 years is shown in **Fig. 3**.

Any reference to a SNP locus within NCBI (and elsewhere such as the scientific literature and alternative SNP databases) uses a unique identifier comprising a number prefixed with rs. All rs-numbers are listed in NCBI as linkouts, which will return a standard format summary page for the SNP termed the Cluster Report, which lists a full set of key parameters for the locus. This page can be thought of as the SNP locus homepage and from this standard point of reference different routes can be followed to database entries elsewhere in NCBI with content related to the SNP, such as Gene, OMIM, or GenBank. Of particular use is the link to Map Viewer (*see* **Subheading 7.**), which plots the SNPs chromosome position in relation to a variety of other genome features acting as landmarks for the locus. Interlinking in both directions is now standard practice, therefore clicking an rs-number in any other major SNP database outside of NCBI will connect to the dbSNP cluster report. This means that it is important to be familiar with the page layout and to know the limitations that exist for the user with the way data are presented. The cluster report layout has been revamped at the start of 2006 and after the summary header of four lines each for locus and allele information, the detail sections currently comprise submission, fasta, geneview, map, diversity, and validation.

4.1. Submission

Submissions for the SNP are listed, with the reports used to validate the locus marked by an icon. Each ss number linkout leads to a detailed breakdown of the genotyping performed and these in turn linkout to lists of genotypes with sample ID's and detailed population descriptions if these require checking. The sample details allow use of consensus controls as genotyping standards, for example, submission ss2316529 for SNP rs1490413 lists sample CEPH1331.01 as an A-G heterozygote.

4.2. Fasta

Fasta lists the flanking sequence around the substitution site. Minor problems can occur with recovery of sequence from this section requiring care. The amount of sequence can vary from tens to hundreds of bases, whereas the SNP can sometimes be found very close to the start or the end of the listed sequence (**Fig. 4**). In these cases recovery of sufficient sequence involves visits elsewhere, either to Entrez Nucleotide or the more user-friendly Santa Cruz genome assembly. In-

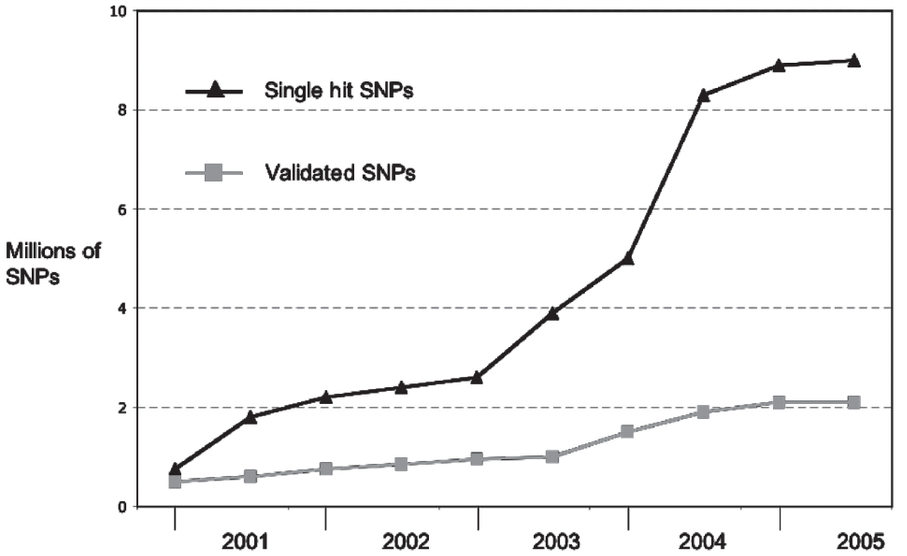

Fig. 3. Growth in SNP numbers since dbSNP began cataloging loci in 2000. Plot shows the cumulative number of unique SNP loci (*black line*) and the proportion of these SNPs validated by genotyping (*gray line*). Assimilation of loci into dbSNP is fast but proper assessment of the polymorphism much slower.

conveniently the NCBI linkout to Santa Cruz has recently been removed requiring a "manual trip" to the gateway page http://genome.uscs.edu/cgi-bin/hgGateway and entry of the rs-number. In the Santa Cruz map browser, click on the blue rs-number linkout under the various genome elements in the map view to get the summary page and then click on "View DNA for this feature" and choose the width of flank. It may be wise to choose 0 and 100 bases (i.e., upstream/downstream), for example, followed by 100 and 0, to keep track of the substitution site, as this is shown as a normal base and therefore its position can be difficult to locate. Note that Santa Cruz uses the term *simple* nucleotide polymorphisms. Back in fasta, sequence is arranged in 10-base blocks using different typesets: upper case/lower case and black/green. Upper case denotes normal, unique genomic sequence, whereas lower case is used for sequence identified by RepeatMasker (*detailed in* **Subheading 11.**) as low complexity or repetitive element sequence. Green is used to denote sequence used by the submitter lab during SNP identification. A common problem is a lack of consistency in the direction of the displayed sequence in fasta and Entrez Nucleotide Sequence Viewer. Therefore, users wishing to check sequence more carefully will need to be prepared to use sequence inversion macros in Excel or in stand-alone programs to "flip" between different strands in different locations. Helpfully Sequence Viewer allows a "view on minus strand" option. One final sequence tracking problem is that clustering SNPs are occasionally given the IUPAC code (e.g., CTAYGGA) within the sequence in fasta and this is easily overlooked, although it appears to be uncommon and largely ad-hoc in occurrence.

4.3. GeneView

For SNPs in genes the GeneView section outlines the gene context of the substitution with color codes for synonymous, not synonymous, and intronic SNPs (pale green, red, and yellow, respectively), includes the different amino acid residues and their position in the protein sequence plus linkouts to the nucleotide reference assembly for the coding regions of the gene.

4.4. Map

The Map section lists the NCBI and CDS positions where both exist, such as, reference and Celera plus linkouts to the contigs used in the assembly for each genome. This section also has linkouts to Map Viewer and to neighbor SNP details. The position and density of neighbor SNPs may be different between NCBI and CDS, therefore in assay design it is important to track all of these to be sure of clean primer-binding site sequence free from interfering substitutions nearby. This is especially true of primer extension chemistries that need ~20 bases of SNP free sequence immediately adjacent to the target substitution. If SNPs close together are in association it follows that a particular allele may carry a neighbor at identical frequency and always dropout from the assay producing an apparently monomorphic locus. It is important to note that this problem occurred in ~6% of cases during the HapMap phase I SNP genotyping ***(22)***. For a high density of neighbor SNPs it may be too problematic to find a clean sequence but the SNPs can be more easily tracked in Santa Cruz, which shows by far the clearest graphic arrangement of SNP grouping close to the study SNP. Unfortunately it requires a longhand process of clicking each linkout and obtaining the positions to construct a fully annotated sequence around the assayed substitution site.

4.5. Diversity (replaces Variation)

The Diversity section originally termed Variation has been the source of problems for several years and is now being completely revised to incorporate the detailed population data coming from HapMap. Until recently this section had been potentially misleading in the way it summarized allele validation information, because it used average allele frequencies and heterozygosity based on merged data from all submitting laboratories. For example, a submitters estimate of 0.2 minor allele frequency based on 100 individuals combined with another of 0.5 based on 20 individuals was summarized as 0.25 because dbSNP used total chromosome counts from all submissions to obtain the average values. Not only can large allele frequency differences arise from combining samples from different population groups but dbSNP previously made no distinction between random population samples and samples of individuals with particular conditions. This is illustrated by SNP rs2075745 where a C allele only occurs in subjects with type II diabetes. Despite this, a frequency estimate for C of 0.476 was given for many years in the cluster report, although all other populations tested to date exhibit an A/T substitution at this SNP. In this case the misleading frequency resulted from a comparatively large sample of 200 diabetic subjects tested by one submitting laboratory and this skewed the estimates. Since late 2005, dbSNP has begun the process of incorporating the detailed, population-based breakdown of frequency estimates as it now appears. This gives a vast improvement in clarity for a critical SNP characteristic and this is obviously being extended to all submitter estimates, not just HapMap, as the previous example of rs2075745 is now unambiguous in presentation. Although the averaged figures are retained, it is now straightforward to interpret these with reference to the population studies made by the submitting laboratories. Note, however, that alleles are listed in base-alphabetical order in dbSNP and do not use the HapMap convention of reference allele frequency first. This may seem to put undue emphasis on detail but a look at SNP rs176000, an example of a base inversion between two submitting laboratories compounded by three-allele variation, indicates there is still scope for confusion in this section.

4.6. Validation

Validation briefly summarizes the Mendelian status, PCR performance, and allele frequency distribution quality indicators of the SNP. In addition to SNPs, data are held for other polymorphisms loosely defined as simple. These include small-scale multibase insertions or deletions (alternatively termed deletion/insertion polymorphisms, indels, or DIPs), microsatellite repeat variation (also termed short tandem repeats or STRs), and retroposable element insertions. Because dbSNP

is an open database, there is a straightforward framework for receiving and checking SNP data sent in from submitting laboratories. The SNP locus information is either reported as a new discovery (rare for human SNPs now) or collated into an existing reference SNP set. Clearly the latter case, where different laboratories routinely report identical SNPs, requires careful scrutiny of the flanking sequence to check for previous submission to NCBI and unique location in the genome. Submission criteria are very effective at detecting nonunique SNPs, the checking process requires a minimum 25-bp context sequence each side of the substitution for the detection assay and uses a minimum total context sequence of 100 bp to position the SNP uniquely in the genome or otherwise. The proportion of nonunique SNPs is, however, small (about 5%) and they are more common in pericentromeric areas, therefore the use of loci from these chromosome regions generally needs more care to ensure that the SNP is unique and is flanked by a relatively small proportion of low complexity sequence (sequence containing portions of intra- or interchromosomal repeats, polybase, or short tandem repeats). When the context sequence can be uniquely positioned and the SNP is identified as previously observed, dbSNP will place the submission into the reference SNP group, hence the use of the familiar rs-number denoting the refSNP. The full group of submissions for the same locus is termed a cluster in NCBI, hence the single summary page for each SNP is termed the reference SNP cluster report. Submitted SNP and reference SNP details are distinguished by using multiple ID numbers prefixed with ss and a single rs-number, respectively, at each cluster report. The current ratio between submissions and clusters is approximately 2.5:1 (27.2 million to 10.4), therefore multiple institutions have independently confirmed the majority of NCBI SNPs. Clicking on an ss number (also termed the accession number) allows the scrutiny of the quality of the submission including statistical analysis of the individual genotypes to ensure they are in predicted ratio from the assumption of Hardy Weinberg equilibrium. As outlined previously, these detailed genotype listings can then be used to check the reliability of study assays by retyping the same controls.

5. The HapMap Project (http://www.hapmap.org/downloads/nature02168.pdf)

The international HapMap project was launched in late October 2002 with the stated aim of determining the haplotype structure of the human genome. This has broadened in scope slightly to encompass, in their own words, "all common human sequence variation, providing information needed as a guide to genetic studies of clinical phenotypes." What this means is that the mapping of haplotype blocks using set SNP positions has been extended to include any SNP landmark that could be equally useful. The full program of the HapMap project is ambitious in scale, in some senses approaching that of HGP, but it remains simple in concept—to begin by genotyping at least one SNP per 5 kb of sequence (just more than 1 million markers) in 269 individuals taken from 4 populations located in three continents and to conclude by consolidating SNP number and annotation so that a limited set of SNPs can be confidently assigned as markers that tag each haplotype block. Haplotype blocks create a large amount of redundancy in the use of SNPs to measure association, as the average haplotype can contain a considerable number of SNP markers that share exactly the same frequency and the same recent ancestry with nearby gene variants, therefore using more than one SNP per block to track the genes of interest by association does not necessarily add any more value to the study. This is, of course, a simplistic argument that assumes that blocks are easy to define and haplotype diversity is limited, but it emphasizes one of the tenets behind HapMap planning: to reduce the genotyping effort required for a clinical genetics study without loosing any quality in the association values obtained from using a small subset of the SNP variation available. Because the total genotype analysis needed far exceeds anything accomplished before, it was appropriate to start with manageable aims and build on these. Therefore, the phase I goals were to characterize and map

haplotype blocks; to collate haplotype diversity in each population; and to define every coding SNP (this last aim spans phase I and phase II). The four populations sampled were Yoruba in Ibadan, Nigeria (referenced in the HapMap data as YRI), Japanese in Tokyo (JPT), Han Chinese in Bejing (CHB in HapMap but HCB in dbSNP), and CEPH Utah residents with ancestry in Northern and Western Europe (CEU). The YRI and CEU samples comprised trios that ultimately allowed very precise analysis of haplotype phase, that is, whether the alleles of heterozygous SNPs reside on one chromosome or the other. Ten ENCODE (Encyclopaedia of DNA Elements) regions were analyzed with a 1—fold increase in SNP density to compare data quality between the HapMap genome coverage and a more complete SNP catalog. Sequencing of 48 subjects from each population for these regions has spanned the two phases and also acts as a test bed for low frequency SNPs. In contrast to the initial focus on 1 million SNPs in phase I, HapMap is now producing genotypes for the phase II goals of expanding the number of SNPs in dbSNP with adequate multiple population validation from 2.6 million to 9.2 million. Phase II also includes extended sample numbers in the four study populations, a broadening of study populations, and a focus on more detailed analysis of the ENCODE regions.

Three years after initiating the project, the group reported the phase I findings in October 2005 ***(22)***. The investigators highlight the fact that HapMap is intended to concentrate internationally coordinated resources on the characterization and understanding of the *variant* part of the genome sequence as a natural extension of the work of HGP in establishing the *invariant* sequence shared by all individuals. To summarize an extensive report:

1. SNP loci have proved to be highly correlated with their immediate neighbors.
2. Analyses to date show the generality of haplotype block structure and recombination hotspots in the human genome.
3. The redundancy of proximal SNP sets should yield efficiencies in association studies from the use of a catalog of tagging SNPs and coding SNPs.
4. The SNP data generated so far offers a means to study genomic variation without recourse to wholesale resequencing.
5. The findings have gained increased understanding and characterization of the human genome (notably the mapping of deletion mutations), natural selection events in the recent past and fine-scale recombination organization.
6. dbSNP has collected the vast majority of common SNP variability in the human genome, when SNPs have not been listed they show tight correlation to loci that are in dbSNP. SNP discovery using PCR-based sequencing is biased against low frequency SNPs (i.e., those with minor allele frequencies <0.05).

Two further points emerge from the report. First, it is increasingly clear that HapMap data founded, as it is, on the analysis of four populations, represents a valuable resource for population genetics analysis ***(23)***. Clear signals of natural selection have been found from the HapMap data in a number of genes that are not obvious candidates for adaptive responses in the immediate evolutionary past. Furthermore, a comparison of haplotype-based selection detection tests compared with classic methods that use individual loci indicates that the former approach can be more sensitive in detecting recent positive selection (notably in the analysis of G6PD and TNFSF5 genes) ***(24)***. Extension of HapMap genotyping to new population samples will bolster the dataset further and help to pinpoint which SNP loci are appropriate choices for more extensive sampling of human populations. Second, the report's conclusions include a demand for rigor in association studies through multiple replications and enlarged sample sizes. Furthermore, the investigators emphasize the need for an unbiased approach to the reporting and interpretation of association study results, regardless of outcome. Because the common diseases are almost all complex diseases it is worth taking heed of this advice as the investigators outline that such diseases require very careful control of environmental influences including lifestyle and behavior, of adequate clinical characterization of phenotype, and of sufficient replication of studies. All of these factors are

important to control correctly if the precision at the genetic level is to be fully exploited to understand complex disease.

Since phase I was completed, HapMap has become an essential complement to dbSNP for checking the haplotype positions of a chromosome region as defined by values measuring linkage disequilibrium between SNP pairs. In addition, access to high-quality allele frequency estimates for 1 million SNP markers represents a significant move forward for the validation status of a large proportion of human SNPs. This becomes particularly useful as a means to check ones own allele frequency estimates obtained from the subjects of a study, allowing greatly improved quality control of the user's own genotyping assays. This can be taken one step further by including CEPH trio-positive controls in an assay and referencing the genotypes obtained to the individually listed results in HapMap. The HapMap database contains structured access to all the genotypes generated in the form of SNP report pages, together with detailed maps of haplotype structure in the form of annotated LD plots using pairwise comparisons of SNPs in the chromosome interval using a stand-alone graphic browser ***(25)***. Finally, an equivalent approach to Entrez exists in HapMap, termed SNPmart, for filtering down the datasets to downloads of manageable size based on similar principles.

The relationship between the two principal SNP databases of HapMap and dbSNP is in the process of change and consolidation. They continue to be very closely interlinked but a number of differences should be emphasized. First, the process of dbSNP to provide haplotype map details independently of HapMap is progressing slowly and does not to match the quality and depth of the other components of dbSNP. For example, in Map Viewer a haplotype map option exists (termed dbSNP haplotype), but this is only available for chromosome 21 and takes the form of block positions and reference numbers listed as hyperlinks to reports from Perlegen. These proceed to detail the haplotype SNP allele composition, for example, SNPs rs2822549 and rs2822550 link to a block named B002180 with three haplotypes outlined in base color-coded plots. However, this gives the impression of work on hold because HapMap have begun data release. It seems that dbSNP is unlikely to provide an adequate basis for detailed haplotype mapping in its own right until sufficient data have been obtained. This process has started with the collection of Haplotype data from publications (http://www.ncbi.nlm.nih.gov/SNP/hap/dbSNP_haplotype_intro.html.). Second, dbSNP is still in the process of updating the Variation section of each SNP cluster report page to encompass the vast quantity of allele frequency data generated by HapMap and other contributors. HapMap provides a full list of dbSNP loci across the whole genome as link-outsto dbSNP cluster reports for each locus, with HapMap validated loci uppermost and all others below. The genome map also makes reference to the NCBI Entrez Gene report page for each gene but a third point of difference is that occasional differences in gene characterization are evident. For example, SNP rs11779952 is placed in gene SLC39A4, a solute carrier, by HapMap and in NFKBIL2, a nuclear factor kappa B inhibitor, by dbSNP. The disparity in the affiliation of SNP and gene in this case seems to have been caused by a difference in gene position between NCBI and Celera genome builds. Although it is difficult to know how this ambiguity has occurred, this example serves to illustrate the effect of different approaches that may be taken between HapMap and NCBI to the curation of data rather than the management of data. These two roles are quite distinct, curation in this context describing the interpretive activity required by large collections of data that need characterization to be properly cataloged (much like museum contents). Although the majority of NCBI database information can be described objectively, certain data must be characterized with the available knowledge and this can introduce disparities between databases run by different organizations. As this example suggests a principal source of curation ambiguities is gene description, more specifically termed gene annotation, that is, the characterization of a gene and

its function based on the interpretation of descriptive data. This is one step beyond management of information and requires interpretive judgments to be made from the data. For instance, is a new putative gene model recently placed in the database a real gene that has not yet been fully described or a pseudogene? Can the gene function be defined in terms of existing pathways? Is the ontology (defined as a set of terms that describe equivalence of function, role in a process, or sequence between a set of genes or proteins) a close match to a well-defined gene family or different enough to suggest a lack of relatedness? Ambiguities in data curation will continue to complicate matters when researchers routinely collect and compare information from more than one database. Last, any user familiar with dbSNP will have noticed the differences in public and Celera sequence position for SNP locations in the Integrated Maps section, when both databases hold the same locus. Because the genome sequences were assembled in different ways and problems like sequence inversions can occasionally arise, it makes sense for dbSNP to list both until there is a full consensus sequence assembly. This is not principally a problem of curation but of differences in the chromosome coordinates obtained by each sequencing project.

6. Finding SNPs for Clinical Genetics: Using HapMap and SNPbrowser®

Faced with the task of locating the genetic component of a disease, researchers will either begin whole genome linkage analysis and focus on the chromosome regions showing the clearest linkage signals or will have an idea of candidates from studies performed already. The strategy that best starts the SNP analysis proper is to concentrate the initial efforts of locus selection on coding and tagging SNPs as the information content from these loci can give the strongest pointers toward the genes underlying the disease of interest. Consulting the HapMap and SNPbrowser databases can form the principal part of selecting SNPs for this part of the study and any further fine mapping analysis of candidate regions.

Applied Biosystems (AB) SNPbrowser is a stand-alone database of 5 million loci comprising a compilation of public and private (i.e., CDS based) SNPs. The marker set is essentially a data dump direct to the user's PC for use offline, with an easy and intuitive front-end in the form of an annotated map. The advantage of this database is that it shows haplotype block information in a much easier to read format than HapMap and in a window independent of a web browser. The block maps are based on Celera's own pairwise analysis of 160,000 SNPs (termed the backbone validated SNPs) so it complements, as well as clarifies, HapMap haplotype block map annotations. SNPbrowser can be downloaded from the AB website (http://marketing.appliedbiosystems.com/mk/get/snpb_landing?isource=fr_E_RD_www_allsnps_com_snpbrowser) and once launched, can be configured to suit the user's needs in terms of haplotype block annotation displayed, SNP type, population studied, and extent of region shown. The first choice to make is the haplotype block map display. The options are to use HapMap or Celera and to display all study populations or just a single one. The Celera maps are constructed from the analysis of 45 individuals each from white (American European) and African American, plus a smaller number of validated SNPs (hence less reliable map definitions) in Chinese and Japanese. These clearly mirror the HapMap study populations, but offer an opportunity to compare HapMap Africans (YRI) with African Americans and explore the effect of admixture and the potential for MALD analysis, outlined under **Subheading 2**. The match to a HapMap style of presentation continues with the SNP information window, an example of which is given in **Fig. 5**. This needs to be unlocked: "View menu," "SNP details," "Show (ctrl + D)," then click the padlock icon upper right of the pop-up window. This additional window shows the Celera allele frequencies in identical pie charts making comparisons easy and works with a mouse-over allowing the map to remain uncluttered when SNP density is high. The NCBI button is a linkout if the SNP carries an rs-number, and it soon becomes

Fasta sequence (Legend)

>gnl|dbSNP|rs187992|allelePos=324|totalLen=705|taxid=9606|snpclass=1|alleles='C/T'|

```
CATTTATAAA TTAATCTCAG TGGTTTAATC ATAATCAACA GATGTTGGTA TGGCCCCTTC
CTTCTTATTC TCCCTGCCTT CCTTTAACAT ACCCATCTGG GGGTCAGTGG TTCCTATGCA
CCAGGCACCG CACATAGCAT TTTATCTGCA GAATTTCAAC TGACTCTCAC TGCAGCTCTA
TTTGTTTTTT GTTGTTTTAA ttttatattt tattttattt attgagacag ggtcttgctc
tgtcacccat gctggagtac agtgacacga tatcagctca ctgcaacctc cacttcctgg
gttcaagcaa ccctccctgc ctc
Y
gcctccggag tagctgggat tacaagtgct gccatcatgc ctggtgaatt tttgtatttt
tagtagagac ggggtttcac catgttggtc aggctggtct caaactcctg ggctcaggta
atccgcctgc cttggcctcc aacagagctg ggattagagg cgtgagtcac catgcctggc
cTTGCAGCTC TACCTGATGG GTACTTTATT AACACTTCTA TATCAGGGGG CTGTCTGACT
CCATGGCCTG CCCACACTCT TAAAAAACAC ACATTCTGTT CAGATTCAGG AAAGGTTCCA
GAGAGGGTGG GGGTGGGGGA CAGCTGAAAT GCTGGACTCT GAAGGTATGA CTCTTTAACA
ACAAACCAGT TTTGGGCTGG C
```

Fasta header containing SNP summary data

Upper case – normal sequence

Green sequence – confirmed by SNP detection assay

Lower case – repetitive sequence

Fig. 4. Typical fasta sequence report. The SNP detection assay only confirmed the substitution site because in this example, computational contig comparison techniques were used.

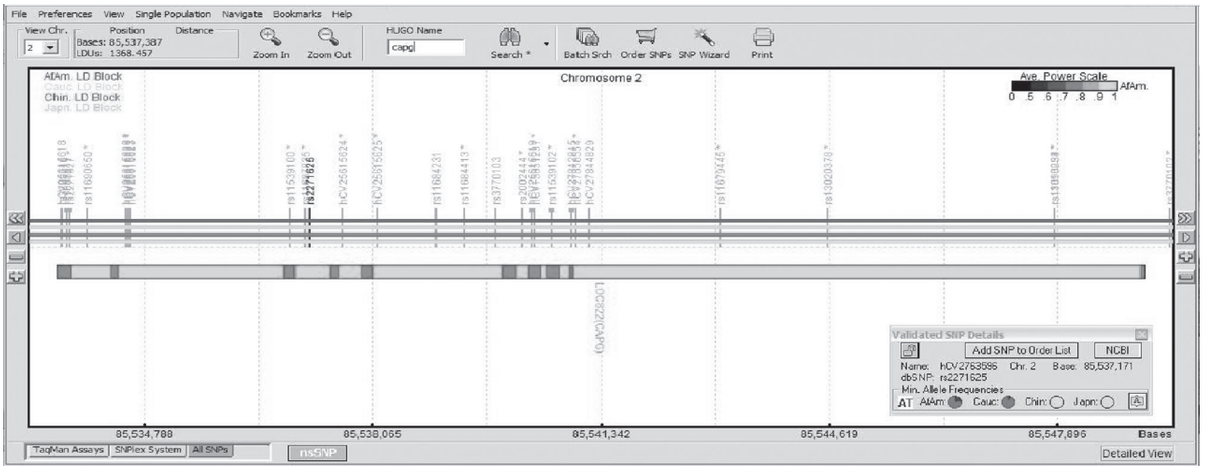

Fig. 5. SNPbrowser map view for the gene CAPG (*see* **Fig. 6**). Top four color bars show haplotype block distribution in African American, European, Chinese, and Japanese (top to bottom). The gene bar is color-coded black to light gray to denote the power based on the gene variant frequency, haplotype frequencies, and block positions (upper right scale). The pop-up box lower right gives SNP details linked to a mouse-over system for the dark and light bars positioning each SNP locus on the haplotype block bars above.

evident in any gene that a large proportion of the SNPs displayed are Celera only. The idea of SNPbrowser is to provide a shopping list for loci that can be genotyped with AB's proprietary Taqman® and SNPlex® technologies, but there is no commitment in downloading and using the browser other than e-mail registration. This same approach applies to the Taqman genotyping assay database described under **Subheading 9**. What is available to any user is a commercially orientated database with a significant amount of private SNPs, but each of these allows important infor-

mation to be obtained about the position and variability of the Celera markers ahead of their incorporation into dbSNP. The color-coded Haplotype block positions from the four study populations are displayed above the gene bar as a combination or individually. Certain caveats apply here: the block edges, although graphically sharp are tentative and any block definition used can only produce fuzzy boundaries at best. Subsets of SNPs, Taqman assay SNPs, SNPlex assay SNPs, All SNPs, and Coding SNPs only, can be selected with buttons on the lower left of the map. Finally of most interest, over and above the private SNP data obtained is the chance to review the power of the gene shown with a green–black scale (introns in purple–pink scale to match) and to see a measure of linkage disequilibrium in LDU (i.e., units), an arbitrary scale based on r^2 and *D'* calculations between sets of SNP pairs. Connecting lines link the SNPs true position on the normal kilobase scale to the LDU scale, therefore these merge in cases of SNPs sharing the same haplotype block. The two values of power and LDU, novel to the SNPbrowser map, are related because the power scale is a summary value based on the block distributions across the gene and is intended to provide a means of comparing different parts of a gene or different genes. The scale ranges from less than 0.5 (black) to five values between 0.5 and 1.0 (dark to pale green) and is intended to help predict the power of a block-based approach using different SNP combinations along with the effect of sample sizes and the minor disease/trait allele frequency. SNPbrowser amounts to a succinct and visually clean map browser that is an informative complement to HapMap browsing. The linkouts to dbSNP and to the AB Taqman genotyping summary pages allow quick follow-up of SNPs of interest and the SNP lists (shopping lists) can be saved or exported. Many researchers use SNPbrowser as the first snapshot of a gene and this does not necessarily mean avoiding the commercial pipeline attached, as Taqman has been a mainstay of SNP genotyping for medical genetics studies for many years. A set of poster presentations explaining the genetics in more detail can be downloaded from the AB website (http://docs.appliedbiosystems.com/pebiodocs/00112824.pdf, http://docs.appliedbiosystems.com/pebiodocs/00114486.pdf, http://docs.appliedbiosystems.com/pebiodocs/00112823.pdf).

HapMap browsing is initiated in a similar way to SNPbrowser, usually starting with a single gene name to locate a manageable segment of chromosome and view the genome landscape (**Fig. 6**). Begin by placing a landmark (SNP or gene) or chromosome region limits (the range in full bp numbers) in the genome browser page (Data linkout top right of the homepage, then the Generic Genome Browser linkout). Often HapMap will offer a choice of locations if the description is not a complete match to the data and the list gives full details to allow review of the user's own information. An excellent set of PowerPoint guides are listed in the tutorial page (http://www.hapmap.org/tutorials. html.en). The basics of genome browsing HapMap data are outlined by Lincoln Stein and explains the steps involves in finding SNPs in or near the gene (termed region of interest or ROI), viewing patterns of LD, selecting tag SNPs, and if required, downloading the SNP dataset generated by a search. The layout of the main HapMap view has already been detailed under **Subheading 5.** and the best strategy is for users to explore a region and become familiar with the acquisition of data with a level of depth that suits the particular purposes of the research stage. The other tutorial presentations from Michael Boehnke, Mark Daly, Augustine Kong, and Toshihiro Tanaka cover in much more depth than is possible in this review, the detailed aspects of association study design. One word of caution, using the scrolling arrows to browse a large gene at 5-kb scale or less can take a very long time and it will soon be noticed that even over long stretches of sequence the pie chart distributions start to look very familiar. At this stage it is worth cross-referencing to the SNPbrowser haplotype block definitions to determine whether the block is very long itself or whether the review of SNP data can fruitfully be focused on recommended tag SNPs for the region.

7. Undirected SNP Searches—EntrezSNP and Map Browsing

There are two ways to examine a group of SNPs in the absence of clear directions to a specific chromosome region: direct queries to EntrezSNP and genome map browsing. At present, there is little need to find SNPs without associated landmarks such as the position of a linkage signal or a candidate gene, but there can be particular reasons that a group of SNPs needs to be collected and studied with information other than position. An example of one such situation was the collection of SNPs for forensic analysis when it was important to find SNPs with appropriate levels of variability or substitution type ***(5,26)***. An EntrezSNP query is the best approach for obtaining a list of candidate SNPs with combinations of specific characteristics. The major drawback of EntrezSNP is that two criteria, linkage and flanking sequence quality, cannot be used as part of a query and the latter characteristic has a direct bearing on assay success in nearly all SNP genotyping methodologies. However, any SNP in dbSNP or HapMap suitable for study will have been validated by a submitting laboratory using one of the genotyping techniques. By default this tends to prevent SNPs that would be impossible to genotype from being returned from a query if validation status is used as a fixed field. The appropriate fixed fields would be "organism" **[ORGN] or [TAX_ID]** prefixed with **human** and "map weight" (the number of times a SNP maps to the genome) using the term **1[MPWT]** (ensures all SNPs are unique). The "validation" term **by frequency [VALIDATION]** is used to ensure SNPs have been validated by repeat genotyping rather than by contig comparison and sequencing of pooled donor DNAs. After this the list of search terms combined by operators can be taken from a tickbox list of limits (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Limits&DB=snp) or applied by the user who is familiar with the syntax and specific values required. The most important search field tags are listed in T**able 2**. As an example "all nonsynonymous, AT SNPs in segment 1 to 1.5 Mb of chromosome 1, true SNPs only" would comprise **coding nonsynon[FUNC] AND W [ALLELE] AND 1[CHR] AND 1000000: 1500000 [CHRPOS] AND snp[SNP_ CLASS] AND human [ORGN] AND by frequency [VALIDATION] AND 1[MPWT]**. EntrezSNP returns a list of SNPs that qualify with a graphic summary for each, shown in **Figure 7,** giving all the information necessary for a rapid scan of a large number of loci in one go. It is possible to sort the list in an alternative way to the default sort order of descending rs-number by selecting from the drop down menu (**sort**) to resort the list, for example, heterozygosity or map position. Using map position will list from q-arm telomere up to p-arm telomere but note that the first loci are unplaced SNPs. Ticking each SNP allows a list to be exported to a text-holding webpage, a file, or the clipboard (**Send to** drop down menu).

As an alternative to EntrezSNP, map browsing offers an intuitive way to review large numbers of SNPs in one session. Exploring a chromosome segment as a map gives the best way to scrutinize the position and characteristics of nearby genome features of importance: transcripts, genes, or clustering SNPs (neighbor SNPs in NCBI). Furthermore, the features around each SNP can be scrutinized easily through a series of linkouts embedded into the map view to the dbSNP cluster report page and Gene reports or other supporting databases. Both dbSNP and HapMap have a map-based system at the core of their SNP databases. HapMap Genome Browser (click Browse Project Data on left hand column of homepage) offers, above all, comprehensive SNP allele frequency data for all 1 million SNPs given in a succinct but clearly arranged pie chart graphic together with the position of any coincidental gene locus (these linkout to NCBI Gene) plus a chromosome scale. The reference allele is blue in each pie chart and the rs-number linkouts to the HapMap version of the cluster report with all the validation data for the four populations plus details of the assay used. At the top of the main map is the summary chromosome view aligned with SNP and gene density plots. In combination with a %GC plot these are a useful supplement to the map view given by the

Table 2
Important EntrezSNP Search Field Tags

Description	Tag	Search field used	Example
Observed alleles	[ALLELE]	IUPAC allele code (**Table 3**)	R[ALLELE] find SNPs with A/G substitutions
Chromosome	[CHR]	number / X, Y	21[CHR] OR 22[CHR] find SNPs on chromosomes 21 and 22
Base position	[BPOS]	ranged number (used with AND & [CHR])	18000:28000[BPOS] AND Y[CHR]—find SNPs in 10-kb section of Y-chromosome
Heterozygosity	[HET]	ranged number	30:50[HET] find SNPs with heterozygosity value in range 30%–50%
Function Class	[FUNC]	locus region, intron etc. (8 in total)	coding nonsynon[FUNC]
Build	[CBID]	number	125[CBID] search build 125
Gene location	[GENE]	gene symbol	CAPG[GENE] search for SNPs in actin capping protein, gelsolin-like gene
Genotyping method (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp - METHOD)	[METHOD] page below	description as listed SNPs found by chip hybridization	hybridize[METHOD] search for in
Map weight	[HIT], [MPWT]	number: 1 = once, 2 = twice, 3 = 3–9 times	NOT (2[HIT] OR 3[HIT]) exclude SNPs mapping twice or more in ge nome (NB better to avoid using NOT 2[HIT] to include CDS SNPs
Population (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp - POPULATION)	[POP]	description as listed page below	pacific[POP] search for SNPs in genotyped in Australasian & Oceanian samples

NCBI browser. The map elements, termed tracks, can be configured in a variety of ways with an emphasis, as would be expected, on SNP distribution and alignment with linkage disequilibrium measurements. Of most interest to users of HapMap comparing data to NCBI, is the process of annotating the default map view with LD and haplotype block information. Three LD measures can be plotted: *D'*, r^2, and *LOD* and haplotype block structures can be viewed as phased haplotypes (i.e., alleles are assigned to the most likely shared chromosome strand to denote the haplotype). These statistics require a more detailed outline than is possible in this review but the extensive help pages contain the descriptions and relevant publications. To obtain this additional map annotation download the plug-in (a java applet) and go to the reports and analysis drop down menu, right uppermost. Choose "Annotate LD Plot," click "configure" (with a variety of arrangements possible), then "configure" again, then "go". The same with "Annotate Phased Haplotype Display" from the same menu where configure just gives options to choose populations. The LD plots comprise red and gray scale pairwise block patterns. The plot gives marker-to-marker LD values, where the genotyped SNP are denoted as ticks and the marker pairwise information is plotted as boxes

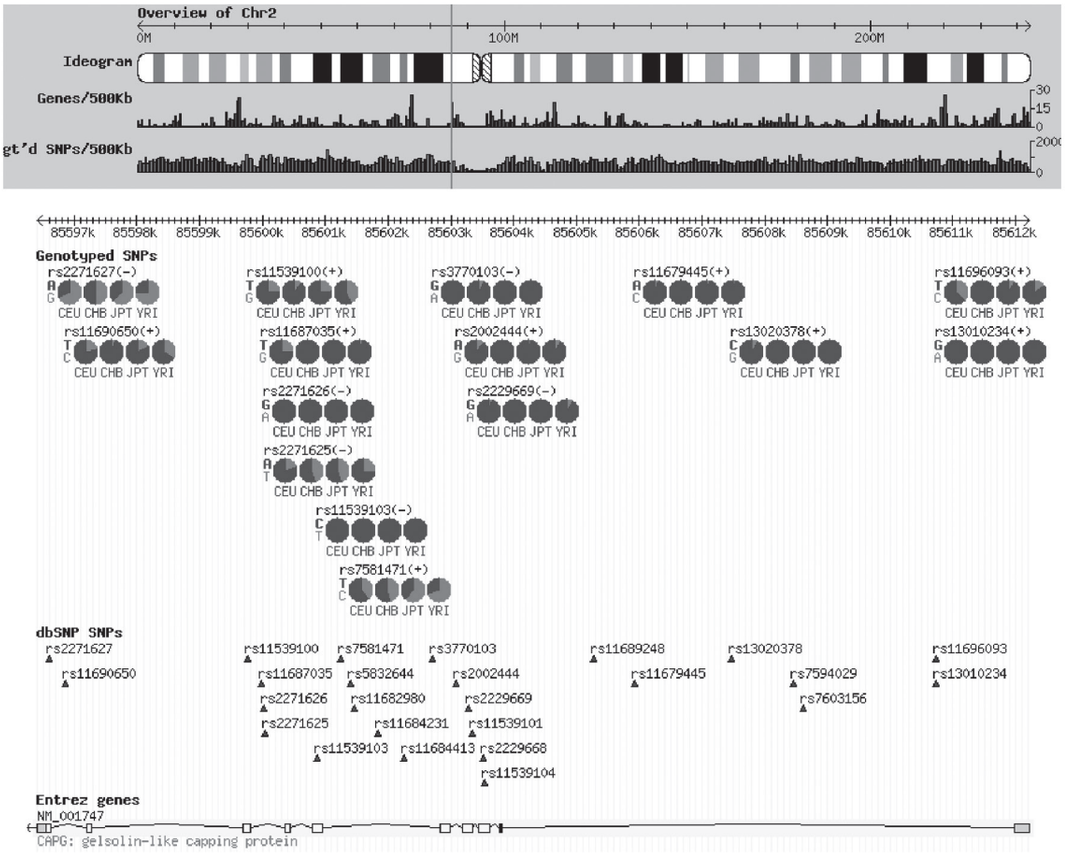

Fig. 6. Typical HapMap genome browser window for the gene CAPG (*see also* **Fig. 5**), with 15 SNPs genotyped by HapMap, shown in approximate position as pie-charts and below these, 25 SNPs shown as triangles (exact position) denoting loci not genotyped, but present in dbSNP. Each set of rs-numbers linkout to HapMap SNP reports and dbSNP cluster reports, respectively. The gene structure of CAPG is outlined as a line diagram at the base.

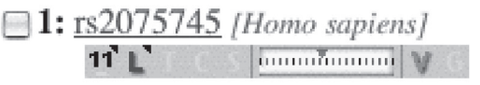

Fig. 7. Example locus return from EntrezSNP for rs2075745 showing annotation. The rs-number linkouts to the RefSNP cluster report, this human SNP occurs on chromosome 11, maps once (*underline*), is sited in a locus (*L*), exhibits 46% heterozygosity (*scale*), and has been validated by genotyping (*V*).

between these ticks. Phased haplotypes are given as blue and yellow blocks and unless the diversity is high it is the clearest way to view haplotype block boundaries in any current database.

Map Viewer is the NCBI map browsing tool allowing the simultaneous search and display of all the NCBI genomic information by chromosomal position. The interface provides a graphical overview of several databases in combination with user-controlled map arrangements. The map combinations and elements are arranged by clicking on the maps and options button mid-left and allows any of 48 different maps to be aligned in

Fig. 8. Symbols used to annotate the NCBI Map Viewer variation map.

the same segment view. The master map contains the linkouts to the matching database with SNPs placed in a map termed Variation in the sequence maps list. Densely packed SNPs are merged together and listed as "6 variations," etc, but only if the map scale is not fine enough to allow adequate spacing of the scale used and for these grouped SNPs, the linkouts are lost. Often it is better to start with a whole view before concentrating on one area and this is achieved by using the genome view (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606&query=). A landmark will be positioned on the relevant chromosome with a small red line, in most cases SNP queries will result in two marks for multiple positions due to the CDS coordinates being different. Multiple landmarks can be displayed together in a single view by using the OR operator in the search window. The individual whole chromosome view is obtained by clicking on the red number under the chromosome in the genome map. It soon becomes clear that dbSNP offers a much broader range of genome landmarks in the displayed region compared to HapMap and usefully all SNPs in the database have summary details summarized as icons placed against the map position and rs linkout shown in **Figure 8**. The smallest scale definable using the zoom scale box is 100,000th of total chromosome length (equivalent to 10 kb of chromosome 1). **Table 3** details the color-coding system used by Map Viewer to signify the annotation status of gene loci placed on the Genes map. NCBI labels all likely genes as "predicted gene models" until confirmed to have a role or be involved in a specific pathway.

8. Ensembl (http://www.ensembl.org/index.html)

Readers who routinely use Ensembl as their database of choice for genome browsing may be wondering why it has not been covered in this review up until now. Ensembl is one of the principal genome data repositories with extensive databases and search tools stemming from the integral role of The Sanger Centre and EMBL in HGP and the position of Ensembl in providing open-source software frameworks for data access and storage. Ensembl continues to provide the most up to date releases of annotated genome data and the widest range of species with genome analysis available. In the area of human SNP data and supporting information the content very closely mirrors that found in dbSNP and therefore there is a choice of approaches, NCBI or Ensembl, and each can be used to access, largely the same data, using comparable frameworks for searching and cross-referencing information for a study. There is little to choose between the two except a broader range of data in NCBI and closer integration to bibliographic data. For this reason

Table 3
Color-Coding System Used for Annotation of the Genes Map

Color	Gene model evidence used
Blue	Confirmed gene model based on alignment of mRNA or mRNA/ESTs
Green	EST evidence only
Brown	Predicted gene model (using Gnomon program) plus EST alignment evidence
Tan	Predicted gene model only
Orange	Conflicting evidence—discrepancy between mRNA sequence and model

EST, expressed sequence tag.

the researcher interested in just human SNP data can easily use the most familiar search environment and obtain data of comparable depth and scope from each database collection with visits to PubMed and OMIM when required. The features in Ensembl that can offer more detail and broader coverage are outlined.

Because Ensembl is both a collection of genome data and a software system that can be adapted for its organization it is important to highlight aspects of Ensembl content that complement the NCBI databases. First, Ensembl has had a pivotal role in the complex task of gene curation and has pioneered automated gene annotation techniques. Most genome content is now generated in such quantity that automated curation has become a necessity. In contrast, the primary importance of the human and mouse gene annotation process has meant that this has been completed manually using expert scrutiny, thereby enhancing gene recognition algorithms in the process. The VEGA (Vertebrate Genome Annotation) database homepage lists linkouts to human, mouse, dog, and zebrafish genome browsers ***(26)***. VEGA's mission statement is to provide high quality, frequently updated, manual annotation of vertebrate finished genome sequence ***(27)***. Without doubt this process will extend to chimpanzee, pufferfish, and agricultural species as the annotation process becomes more streamlined and benefits from established knowledge of gene character in vertebrates. Users interested in obtaining the best data relating to human gene architecture are encouraged to visit the linkouts from Vega human homepage (http://vega.sanger.ac.uk/Homo_sapiens/index.html) explaining involvement of VEGA in the CDDS and MHC Haplotype Projects and the Welcome HAVANA groups involvement in the analysis of ENCODE regions and the CORF projects (respectively, http://vega.sanger.ac.uk/info/data/ccds.html, http://vega.sanger.ac.uk/info/data/Homo_sapiens.html, http://vega.sanger.ac.uk/info/data/encode.html, and http://vega.sanger.ac.uk/info/data/corf.html).

Second, Ensembl has close ties, through EMBL-EBI, to the high-quality protein sequence database of swissprot/uniprot (http://www.ebi.ac.uk/swissprot/). This comprises manually annotated protein sequences with content that has been closely integrated with the gene annotation pipeline, therefore the two data sets, gene and protein, have considerable synergy within Ensembl. This makes swissprot/uniprot a better choice for detailed protein analysis than NCBI Protein, although, as with SNP and gene data, the content is shared and cross-referenced to a large degree. Third, Ensembl provides a gene-oriented search system in MartView (http://www.ensembl.org/Multi/martview), a system that, like HapMap, uses the biomart engine. Initiating a search of the human assembly with VEGA genes as the start point allows the interrogation of more than 22,100 genes with a full range of filters. This is potentially the best way to collate a set of candidate genes on the basis of function and role in processes, such as, a pathways approach to studying the relationship of gene families. OMIM may suggest candidates that can be extended to include all genes that share similar properties in terms of function. This can help to ensure that a search in

the initial phases is not too restricted in focus. Fourth, Ensembl allows access to an HGP sequencing trace repository at Trace Server (http://trace.ensembl.org/). Although listing single pass data and therefore it is both impossible and unrealistic to use this resource for SNP analysis, the breadth of data (1 million traces from 735 species) makes this a valuable resource for the analysis of sequence outside of the mainstream study species.

9 . Online Resources for Population Genetic Studies Using SNPs

Although the focus of clinical genetics database searching centers on the location of SNPs with reference to genes and haplotype structure, population genetic studies place more emphasis on the distribution of allele and haplotype frequencies. This can still involve the scrutiny of SNP position in the gene landscape as the effects of strong positive selection can leave a characteristic signature in the distribution of SNP variability around the selected gene. The discovery of haplotype variability reduction from selective sweeps, as this effect is described, has led to some exciting recent studies of genes hitherto not suspected to be the subject of selective pressure ***(28,29)***. The need for detailed and reliable frequency data in this field means HapMap and dbSNP hold centre stage as the two largest allele frequency datasets available. However, searching the data with the aim of exploring frequency distribution differences between populations is not easy in either database. Furthermore, frequency searches are limited in both by the 0.5 minor allele frequency ceiling. This is the limit to frequency-filtered searches that sets a frequency range for the minor allele between 0 and 0.5 for each class (in this case population). Although this is, of course, logical as the minor frequency cannot be more than 0.5, it prevents searches of SNPs where one population shows a minor allele frequency of, say, 0.2 for allele C and this is present at frequency 0.6 in another population, properly showing a minor frequency of 0.4 but for the other allele. This is not a major drawback as workarounds exist, but it prevents an easy search for loci showing the biggest frequency contrasts and these have proved to be among the most interesting SNPs. The problem of a maximum value of 0.5 for each allele applies to all the other databases and their frequency search systems.

A popular program for processing multiple databases with frequency filtered searches is Frequency Finder (https://mapgenetics.nimh.nih.gov/frequencyfinder/index.jsp) described as a frequency data acquisition tool for mining multiple public databases ***(30)***. This acts as a web portal or a stand-alone program, although both require a data upload in the form of a SNP list as a line delimited text file or alternatively as SNP identifiers or chromosome coordinates placed in a query box. Frequency Finder returns a table of rs-numbers, major and minor allele frequencies, and the data source. Although uploading a file of rs-numbers is a slightly clumsy way of initiating a search these days, the system is comprehensive in scope as it will locate and list frequency data that does not overlap between the database sources used (TSC, dbSNP, Celera, ALFRED, and HGVBase; discussed later). As dbSNP broadens the extent of its data collection from other sources, such as Celera and HapMap, this has become less important than it was previously. For example, the original report of the system in 2004 detected from a whole genome query (246,097 SNPs with data) that 5% of SNPs were unique to TSC, and 16% unique to AOD. The HapMap frequency data accessed was confined in early 2006 (v2.1) to European data only.

The proportion of SNPs described previously as unique to Celera are held in a publicly accessible subset of the CDS database known as Assays-on Demand (AOD) or Taqman SNP Genotyping Assays. When Celera and Applied Biosystems merged their interests to become Applera, a focus of much development was the combination of Taqman real-time PCR assay technology and the extensive SNP data in CDS, leading to a system described by the company as knowledge-based genotyping. With this system the frequency data for a SNP was as important as the availability of off-the-shelf Taqman assays to genotype the locus. The map browser to access this data SNPbrowser has been described, AOD is the equivalent of

Entrez, a search system front end that permits searches based on frequency (among other criteria). The importance of AOD compared to the vast array of SNPs that were available privately as CDS was the quality of validation. Celera used a polymorphism discovery resource (PDR) of six individuals to indicate whether a SNP detected by the private genome assembly was a valid SNP or not. This is an insufficient sample to provide reliable allele frequency estimates and means the CDS SNP details listed in AOD lack precision. In contrast, AOD data is based on pilot Taqman analysis of 45 individuals each from four populations as detailed under **Subheading 6.** for SNP browser. The AOD dataset has since been used for many important population genetics SNP studies that have extended the analysis to a wider range of populations by using the available Taqman assays on a broader base of well-defined populations. These studies provide valuable insights into population variability with the aim of improving the selection of subjects for MALD analysis and to gain a better understanding of human population dynamics. Similarly, I used AOD to begin the process of collecting loci for forensic SNP assays that can suggest a geographic origin for a sample of unknown donor ***(31)***.

The AOD search page (https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=ABGTKeywordSearch&catID=600769) allows access to HapMap, JSNP, and DME (drug metabolizing enzyme) data, in addition to AOD, using keyword searches with gene symbol, gene name, public accession number, biological process, or molecular function. With or without keywords, limits can then be set for intergenic, intragenic, or genic location plus SNP type and effect if genic, followed by the frequency filters for the four AOD or HapMap populations. The query returns a page listing the SNPs with linkouts to dbSNP and Gene plus two decimal place frequency estimates and chromosome location. If the user has an "hCV number" used to catalog Celera SNP data, then it is possible to gain information about the locus if it has been validated for AOD, and usefully to obtain the equivalent rs-number if one exists. One advantage that will become immediately obvious to the researcher is that this provides the easiest allele frequency search system for the 1 million HapMap SNPs despite the previously stated caveat that 0.5 is the frequency limit per allele. The simplicity of this approach for searching HapMap allele frequency data is that it gives a list that can be cross-checked with AOD estimates if applicable (both are listed even if one is searched) and ready access then to dbSNP and Gene. One additional point of reference if the SNP has been validated by HapMap and AOD: the African allele frequency estimates in AOD are based on an African-American sample, therefore by comparison with HapMap Africans it is possible to gain an insight into admixture levels in the AOD study population. Before HapMap these were the most reliable SNP allele frequency estimates available anywhere and were accurate enough to allow a detailed study of allele frequency-derived haplotype block definitions based solely on AOD SNPs showing tandem arrays of identical minor allele frequencies ***(32)***. The ability to search on frequency alone continues to make this database a powerful population genetics tool despite the recent incorporation of linkouts to AOD in the revamped dbSNP cluster reports.

Finally, as mentioned, recent interest in selection in the very recent past and its role in reducing haplotype diversity in the vicinity of the selected gene has led to a useful web tool, Haplotter (http://hg-wen.uchicago.edu/selection/haplotter.htm) to detect and study this effect. Based on an in-depth analysis of the HapMap phase I data release ***(29)*** it is likely to promote potentially interesting further study of genes that previously may not have been considered obvious candidates for positive selection. The tool scans HapMap data for signatures of low haplotype diversity and unusually long haplotype blocks that result from the rapid increase in frequency of the selected gene variation and bordering chromosome regions. The signatures are defined by the equivocal measure of *contrast* between the haplotypes and the surrounding genome landscape because the ancestral allele (positive contrasts as the allele increases in

frequency) can be the subject of selection as well as the variant allele (negative contrasts). Haplotter can work from gene identifiers or a single SNP landmark (more slow and varied in coverage). The program returns plots of *iHS*, the measurement of contrast to surroundings plus the selection signature or population diversity measures *H, D,* and F_{ST} followed by a table of adjacent genes that are colored light blue when they show significant evidence of selection effects. The major advantage of this tool is it allows an unbiased approach to finding regions with indications of recent selection pressure, so in use it is likely to reveal interesting and surprising candidates for more detailed study. In addition, it could focus studies on the phenotypes such loci exhibit with consequences for our understanding of the differences in susceptibility to disease between population groups. The disadvantage is that it appears to lack sensitivity to gene variation and associated SNPs that have reached, or are very close to, fixation (i.e., where a different allele is fixed, at a frequency close to 1, in different populations). Examples of genes, where coding SNPs are close to fixation, that fail to yield a detectable signal with Haplotter are FY (inferring resistance to malarial infection in African populations) and MATP (part of a depigmentation pathway in European populations). In contrast, LCT (creating hypolactasia in Europeans) showing balanced heterozygosity levels reveals one of the strongest selection signals of all, although this may also relate to how recently the selective sweeps have occurred at these loci.

10. Other SNP Databases and Resources

10.1. The SNP Consortium (http://snp.cshl.org/)

The SNP Consortium (TSC) is run by the Cold Spring Harbor Laboratory on behalf of a private/public partnership of 17 organizations. This database comprises 1.8 million loci, all of which are listed in dbSNP. Both databases are fully cross-referenced with linkouts, but TSC uses a different SNP locus identification system. TSC SNPs have been chosen for study specifically because of their proximity to genes, as a principal goal of the consortium was to construct the first high density SNP linkage map: the Allele Frequency Project. This project created the most significant feature of the TSC database for SNP research—detailed genotype frequency data for 55,000 loci from European (termed Caucasian), African, and Chinese population samples. This resource formed the core SNP validation data available to the public along with AOD data, before the initiation of the HapMap project. The findings of the Allele Frequency Project provide a detailed analysis of the nature of SNP variability in the genome and differences between the study populations ***(35)***. One word of warning about the interpretation of allele frequencies generated by pooled DNA techniques that form a proportion of the loci detailed in TSC. This technique is generally inaccurate and almost wholly so when the minor allele frequency is below 10%. An example is rs994174 with minor allele frequency estimates of 0, 1, and 0 for European, Asian, and African samples, respectively, using pooled DNA, whereas the same populations give estimates from repeat genotyping of 0.67, 0.58, and 0.24 (CEU, CHB, and YRI in HapMap).

10.2. HGVbase (formerly HGBase) (http://hgvbase.cgb.ki.se/)

The Human Genome Variability Database comprises nearly 9 million entries concentrated on human genome variants including SNPs, Indels, and STRs. HGVbase uses its own system of locus identification: a nine-digit number prefixed with SNP (if applicable to the locus). The database is in the process of adaptation to give much greater emphasis on phenotype/genotype collation so it will lose a large part of its focus on the cataloging of SNPs but gain increased importance as a means to link SNP variability to phenotype. In the website's own description: "sequence variations are presented with details of how they are physically and functionally related to the closest neighboring gene." This will make HGVbase an essential complement to NCBI OMIM and Gene for the analysis of SNP variation and its effect on the expression of traits. The citation list returned

from a query can provide a streamlined approach to starting a literature search of studies of gene variation resulting from coding SNPs. In addition, the strength of HGVbase has been in the emphasis on listing low frequency variants and new mutations that are on the periphery of mainstream SNP content elsewhere. Last, I thoroughly recommend the indispensable list of linkouts maintained here to no less than 46 other online SNP-related databases and resource sites (http://hgvbase.cgb.ki.se/cgi-bin/main.pl?page=databases_.htm). In large part the list covers much of this section of the review as a source list of specialized databases, each worth exploring but dependent on the particular aspect of SNP variability of interest to the user.

10.3. PolyPhen (http://tux.embl-heidelberg.de/ramensky/)

PolyPhen is a tool that provides Polymorphism Phenotyping—predicting the possible impact of an amino acid sequence change on the properties of a protein. It will not work with SNP data input directly but holds a nonsynonymous SNP database comprising 50,919 SNPs taken from dbSNP build 121 ***(34)***. The predictions on the effect of these SNPs make interesting reading: 9502 unknown, 27,991 benign, 7905 possibly damaging, and 5525 probably damaging, therefore 32.4% of SNPs with a known effect appeared to be detrimental to the protein. This data subset is the easiest way to use this tool and rs-numbers can be input as queries directly for comparison against the nonsynonomous SNP collection (http://genetics.bwh.harvard.edu/pph/data/index.html).

10.4. ALFRED (http://alfred.med.yale.edu/alfred/index.asp)

The ALlele FREquency Database is an extensive collection of frequency reports for polymorphic markers comprising 1501 loci, 475 populations, and 41,980 frequency tables. This forms a useful population analysis tool, in particular the map function, which provides an intuitive search interface based on geographic region. Unfortunately the SNP data held is currently patchy (only 841 rs-numbers), but this situation is certain to change. The rs-numbers that have been collated in ALFRED are matched to loci (either other polymorphic markers or genes) in "Summaries" then "Sites with dbSNP rs #."

11. Web-Based SNP Assay Design Tools

11.1. NCBI BLAST (http://www.ncbi.nlm.nih.gov/blast/)

BLAST is a tool for calculating sequence similarity that accesses the NCBI GenBank databases ***(35)***. Typically the SNP assay design process will query Nucleotide BLAST in two ways.

1. Finding a location for a submitted sequence, effectively the query being: does the submitted sequence exist in a GenBank database?
2. Checking for coincidental similarity in a sequence, normally a PCR primer, the query being: what is the degree of specificity of the submitted sequence?

There are three BLAST programs available for sequence comparison, the BLAST guide (http://www.ncbi.nlm.nih.gov/BLAST/producttable.shtml) can be consulted for the correct program choice. However, the alignment comparisons required for each of these queries are provided by MegaBLAST and standard BLAST (blastn), respectively. MegaBLAST is designed for long sequences and for a certain degree of mismatch, whereas blastn is designed to give a list of sequences in order of similarity. A third option, "Search for short and near exact matches" is recommended for sequence specificity checks with less than 20 bases. A BLAST query returns a three-part report: (1) a header with query sequence information plus summarizing graphic overview, (2) a set of single-line matching sequence descriptions, and (3) the matching alignments themselves. There are two statistics that annotate the returns from blastn: the bit score and E-value. The graphic overview shows the query sequence as a numbered red bar and below this the database hits as colored bars aligned to the query. The colors and proximity to the query represent the alignment scores from red (highest) through to black (lowest) and uppermost to lowest bars. The single-line descriptions give both a bit score indicating the goodness of fit of

each matched sequence and an expect value (E-value). The bit score is calculated from a formula that takes into account all matching nucleotides and gaps, the higher the score, the better the alignment (http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html). The E-value summarizes the statistical significance of the alignment, reflecting both the size of the database used to prepare the alignments and the score system used; the lower the E-value the more significant the hit. For example, a value of 0.05 equates to 5 in 100 or 1 in 20, signifying the probability of this match by chance alone. Overall, the routine use of BLAST to check primer sequence specificity should be a procedure familiar to all genetics researchers. Ensembl has a BLAST site (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/hgp) and sequence alignment tool, SSAHA (http://www.sanger.ac.uk/Software/analysis/SSAHA/) that provides an alternative to NCBI.

11.2. RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker)

RepeatMasker is a tool for screening submitted DNA sequences against a broadly based library of repetitive elements ***(36,37)***. A masked query sequence is returned that can be used for database searches plus a table annotating the regions of repetitive, low-complexity DNA. Users can expect about 50% of human sequence to be masked with this program. This system is used by NCBI for classifying the flanking sequence of dbSNP entries and shown in the fasta section of a cluster report. RepeatMasker tends to be quite aggressive in its annotation of sequence, therefore when designing primers for SNP genotyping assays, it can be safer to include masked sequence and then check the resulting designs for specificity in BLAST.

11.3. Primer³ (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi)

Primer³ is a well-established and popular primer design program and sequence analysis tool ***(38)***. The flanking sequence for a SNP is submitted and various PCR parameters such as amplicon size range and optimum T_m can be prescribed by the user before a list of suggested primer sequences are returned in order of optimum predicted performance in PCR. Despite this simple interface, an array of presets exists for the PCR conditions and the possibility to annotate the submitted sequence to direct the primer design process in useful ways. Primer³ provides particularly versatile secondary structure detection subroutines that can screen primer designs for such structures. This is an essential step as these can reduce the efficiency of PCR or even prevent obtaining SNP genotypes from an assay, particularly in large multiplex designs.

11.4. Santa Cruz In Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr)

In-silico PCR is a beautifully simple idea run by the Santa Cruz genome site for checking the specificity of the primer designs developed for the capture PCR in a SNP genotyping assay. When the forward and reverse primer sequences are inserted into the query boxes these are compared to the human sequence assembly (or 27 other species as options) and the sequence interval between the primer pair is returned to confirm that the correct segment is targeted. Each primer sequence is listed in fasta format as capitalized bases and the interval in lower case. The simplicity and ease of use of this web tool makes it a worthwhile alternative to waiting in the BLAST queue.

12. Concluding Remarks

This review is intended to provide some initial directions in which to point the mouse to ensure that a research project using SNP analysis is properly designed and framed. It is important to fully review as much of the relevant genetic data as possible and to consolidate the research aims on the basis of information gathered. Luckily, this has never been easier and the data never so extensive and detailed in content. However, it is appropriate to conclude with a last cautionary note. At any one time in the laboratory where I work, as many as four or five scientists of a team of 20 will be reviewing informational web pages from

PubMed, dbSNP, HapMap, or Ensembl. Computer-based research can often seem like the major part of the work, but users should resist the temptation to replace benchwork with a disproportionate amount of time following up the work of others or collating information about SNPs without investigating these loci for themselves. An often used dictum these days is "dry work should not be a substitute for wet work." Although usually describing the tendency to place undue emphasis on sequence analysis compared to investigations of cellular processes in live material, the phrase could equally well apply to excessive time spent searching online databases at the expense of generating original data for oneself in pursuit of the research aims. Only in the field of population genetics is there now the real possibility to perform primary research based solely on online data. The SNP genotype information generated by the HapMap phase I data release has opened up an interesting phase of SNP research as the exciting work of Voight et al. ***(29)*** has shown by finding unforeseen signatures of recent selection in human populations. This is an uncommon instance—online investigation is still just the *start* of the work in all other cases of genetic research.

References

1. International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921.
2. Sachidanandam, R., et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933.
3. Read, A. and Strachan, T. (2003) *Human Molecular Genetics* 3. Garland Science.
4. Jobling, M. A., Hurles, M. E., and Tyler-Smith, C. (2003) *Human Evolutionary Genetics*. Garland Science.
5. Phillips, C., Lareu, M., et al. (2004) Selecting SNPs for forensic applications, in *Progress in Forensic Genetics* 10(Doutremepuich, C. and Morling, N., eds.). Elsevier, Amsterdam.
6. Phillips, C. (2005) Using online databases for developing SNP markers of forensic interest. *Methods Mol. Biol.* **297,** 83–105.
7. Sobrino, B., Brion, M., and Carracedo, A. (2005) SNPs in forensic genetics: a review of SNP typing methodologies. *Forensic Sci. Int.* **154,** 181–194.
8. Nachman, M. W. and Crowell, S. L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* **156,** 297–304.
9. Phillips, C., Lareu, M., et al. (2004) Non binary single nucleotide polymorphism markers, in *Progress in Forensic Genetics* 10 (Doutremepuich, C. and Morling, N., eds.). Elsevier, Amsterdam.
10. Dawson, E,, et al. (2002) A first generation linkage disequilibrium map of chromosome 22. *Nature* **418,** 544–548.
11. Patil, N., et al. (2001) Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* **294,** 1669–1670.
12. Gabriel, S. B., et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296,** 2225–2229.
13. Phillips, M. S., et al. (2003) Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nat. Genet.* **33,** 382–387.
14. Daly, M., Rioux, J. D., Schaffer, D. F., Hudson, T. J., and Lander, E. S. (2001) High resolution haplotype structure in the human genome. *Nat. Genet.* **29,** 229–232.
15. De La Vega, F. M., et al. (2003) Selection of single nucleotide polymorphisms for a whole–genome linkage disequilibrium mapping set. CSH Genome Sequencing & Biology Meeting, Cold Spring Harbor, NY.
16. Wall, J. D. and Pritchard, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4,** 587–597.
17. Patterson, N., Hattangadi, N., Lane, B., et al. (2004) Methods for high–density admixture mapping of disease genes. *Am. J. Hum. Genet.* **74,** 979–1000.
18. Reed, T. E. (1969) Caucasian genes in American Negroes. *Science* **165,** 762–768.
19. Hinds, D. A., Stokowski, R. P., Patil, N., et al. (2004) Matching strategies for genetic association studies in structured populations. *Am. J. Hum. Genet.* **74,** 317–325.
20. Campbell, C. D., Ogburn, E. L., Lunetta, K. L., et al. (2005) Demonstrating stratification in a European American population. *Nat. Genet.* **37,** 868–872.
21. Yoshiura, K., et al. (2006) A SNP in the ABCC11 gene is the determinant of human earwax type. *Nat. Genet.* **38,** 324–330.
22. Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., Donnelly, P.; International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437,** 1299–1320.
23. McVean, G., Spencer, C. C., and Chaix, R. (2005) Perspectives on human genetic variation from the HapMap Project. *PLoS Genet*. **1,** e54.
24. Sabeti, P. C., et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837.

25. Thorisson, G. A., Smith, A. V., Krishnan, L., and Stein, L. D. (2005) The International HapMap Project Web site. *Genome Res*. **15,** 1592–1593.
26. Loveland, J. (2005) VEGA, the genome browser with a difference. *Brief Bioinform*. **6,** 189–193.
27. Ashurst, J. L., Chen, C. K., Gilbert, J. G., et al. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res*. **1,** D459–465.
28. Lao, O., Duijn, K., Kersbergen, P., Knijff, P., and Kayser, M. (2006) Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am. J. Hum. Genet.* **78,** 680–690.
29. Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006) A map of recent positive selection in the human genome. *PLoS Biol*. **4,** e72.
30. Nguyen, T. H., Liu, C., Gershon, E. S., and McMahon, F. J. (2004) Frequency Finder: a multi-source web application for collection of public allele frequencies of SNP markers. *Bioinformatics* **20,** 439–444.
31. Phillips, C., Lareu, M., et al. (2004) Population-specific single nucleotide polymorphism. *Progress in Forensic Genetics* 10 (Doutremepuich, C. and Morling, N., eds.). Elsevier, Amsterdam.
32. Costas, J., Salas, A., Phillips, C., and Carracedo, A. (2005) Human genome–wide screen of haplotype-like blocks of reduced diversity. *Gene* **11,** 219–225.
33. Miller, R. D., et al. (2005) High-density single-nucleotide polymorphism maps of the human genome. *Genomics* **86,** 117–126.
34. Ramensky, V., Bork, P., and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. **30,** 3894–3900.
35. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410.
36. Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1996) CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20,** 119–121.
37. Bedell, J. A., Korf, I., and Gish, W. (2000) MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16,** 1040–1041.
38. Rozen, S. and Skaletsky, H. J. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132,** 365–386.