SNP databases.

Article *in* Methods in molecular biology (Clifton, N.J.) · August 2009

DOI: 10.1007/978-1-60327-411-1_3 · Source: PubMed

1 author:

Christopher P Phillips
University of Santiago de Compostela
280 PUBLICATIONS **10,358** CITATIONS

Chapter 3

SNP Databases

Christopher Phillips

Abstract

Researchers interested in obtaining detailed information on SNPs now work in a golden age of online database availability: never has so much data and such a wealth of information been freely accessible for such a substantial proportion of the 18 million single nucleotide polymorphism (SNP) loci currently characterized in the human genome. This chapter describes the major SNP databases available for human genetics studies. Tools and strategies are outlined that can help researchers properly formulate a database query to be able to access the most appropriate information needed for their research aims, including medical or population genetics analysis – an approach that is getting increased attention given the expanding scale of online SNP data.

Key words: Single nucleotide polymorphism, database, search, query, National Center for Biotechnology Information, dbSNP Entrez, HapMap.

1. Introduction

In silico research as a part of the preparation for an experimental genetics study is now an essential preamble to the choice of genomic regions to analyze and markers to use, the design of genotyping approaches, and the listing of appropriate samples to characterize. This chapter provides a simple guide to the structure and use of the major online SNP databases, adapted to **Sections 2** and **3**, by linking each database to a particular research planning task: finding sets of single nucleotide polymorphisms (SNPs) that share common characteristics (NCBI Entrez); obtaining detailed information on a SNP locus and collating other genetically relevant data (dbSNP); exploring SNPs in coding regions (SNPper and PupaSuite); performing simple scrutiny of linkage

A.A. Komar (ed.), *Single Nucleotide Polymorphisms*, Methods in Molecular Biology 578,
DOI 10.1007/978-1-60327-411-1_3,

disequilibrium (LD) block structure and choosing SNP markers to tag chromosome regions (HapMap); and assessing population genetics parameters from online SNP data (Haplotter and SPSmart).

Some straightforward, common sense advice is given about Internet browsing (*see* **Notes 1** and **2**), processing of SNP data, once obtained, and direct use of generic search engines such as Google – to look across the Web space before focusing on known SNP databases. The latter approach can yield interesting results, but otherwise this chapter assumes the user will go directly to a particular SNP database gateway (*see* **Table 3.1**) to initiate a directed search of online data.

Table 3.1
The major online single nucleotide polymorphism (*SNP*) databases

Database	Host organization	Gateway URL for initiating SNP data searches
dbSNP	NCBI	http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp
HapMap	The HapMap Consortium	http://www.hapmap.org/cgi-perl/gbrowse/
Ensembl	EMBL-EBI/Sanger Center	http://www.ensembl.org/Homo_sapiens/index.html
Santa Cruz	University of California, Santa Cruz	http://genome.ucsc.edu/cgi-bin/hgGateway
Perlegen	Perlegen Sciences	http://genome.perlegen.com/browser/index_v2.html
Assays-on-Demand	Applera (Applied Biosystems)	https://products.appliedbiosystems.com/ab/en/US/adirect/ab?cmd=ABGTKeywordSearch&catID=600769
SeattleSNPs	US NHLBI (PGA)	http://gvs.gs.washington.edu/GVS/

NCBI National Center for Biotechnology Information, *NHLBI*, National Heart, Lung, and Blood Institute, *PGA* Program for Genomic Applications

2. Materials

2.1. The Major SNP Databases

Suggesting a definitive list of the major online SNP databases runs the risk of becoming out of date once done, but *dbSNP*, the SNP database of the National Center for Biotechnology Information (NCBI), and *HapMap* head the list in **Table 3.1**, which is otherwise not intended to indicate an order of size or usefulness. NCBI continues to be by far the most important and comprehensive set of genomic databases available, while the HapMap project has an

AQ1

ever-closer relationship with dbSNP in collating human SNP data. To summarize a complex and far-reaching project, HapMap was intended to concentrate global resources on the characterization of the *variant* part of the genome as a natural extension of the work of the original Human Genome Mapping Project in establishing the *invariant* sequence common to everyone and held in NCBI (1, 2). An important part of the initial work of HapMap was to check the efficiency of dbSNP, i.e., how well did the dbSNP catalogue represent the true extent of SNP variability in humans? This was achieved by resequencing ten ENCODE regions (detailed in **Table 3.2** of (1) and at http://www.hapmap.org/downloads/encode1.html.en) and extrapolating the SNP variability found to the genome as a whole. Two findings emerged from this comparison: firstly the false-negative rate of dbSNP (i.e., how often SNPs were present but not detected) although very low was significant for rare SNPs – loci with allele frequencies around 1% (0.01) or less; secondly the overriding majority of common variation SNPs had been captured by dbSNP or if absent had proxies in the same region in tight correlation and listed by dbSNP. It is

Table 3.2
HapMap study populations (*1–4*): phase I/phase II (*5–11*): added to phase III. Many published studies, including those of HapMap (1), merge CHB and JPT to a "population panel" abbreviated to ASN

	Abbreviation	Samples	Full description	Group
1	YRI	180	Yoruba in Ibadan, Nigeria	African
2	CEU	180	Utah residents with northern and western European ancestry	European
3	CHB	90	Han Chinese in Beijing, China	East Asian (ASN)
4	JPT	90	Japanese in Tokyo, Japan	East Asian (ASN)
5	ASW	90	African ancestry in southwest USA	African
6	CHD	100	Chinese in metropolitan Denver, Colorado, USA	East Asian
7	GIH	100	Gujarati Indians in Houston, Texas, USA	South Asian
8	LWK	100	Luhya in Webuye, Kenya	African
9	MEX	90	Mexican ancestry in Los Angeles, California, USA	Native American
10	MKK	180	Maasai in Kinyawa, Kenya	African
11	TSI	100	Tuscans in Italy	European

interesting that estimates of false-positive rates in dbSNP (i.e., incorrectly listing a nucleotide position as a SNP) were not detailed by HapMap, indicating that these were negligible and therefore dbSNP had developed very efficient systems for confirming that SNPs were real (*see* **Note 3**). In summary, dbSNP has proved to be both a comprehensive and a reliable catalogue of human SNP variability with an efficient system to cross-reference multiple submissions of the same SNPs from centers outside NCBI (*see* **Note 4**). Since 2003, HapMap has been the major contributor of SNP data to dbSNP. The other databases listed in **Table 3.1** both parallel and feed data into dbSNP, so they either provide an alternative system of browsing and searching the core human genome SNP data (Ensemble and Santa Cruz), or list the SNPs generated by their own independent genotyping initiatives with stand-alone browser systems dedicated to the data they have generated (Perlegen, Assays-on-Demand, and Seattle SNPs).

1. dbSNP. The strength of NCBI lies in the breadth of genomic databases held under the single umbrella. This means that queries to any of the NCBI databases can tap into the relationships that exist between the subject of interest and each of some twenty or more major databases within NCBI. So genetics research involving SNPs is easily set in the context of supporting information that details published studies of the SNP, context sequence of the SNP, gene structure and function (if this is where the SNP is sited), and how the SNP variation is expressed as a phenotype. These data are handled in NCBI by *PubMed*, *GenBank*, *Gene*, and Online Mendelian Inheritance in Man (*OMIM*) databases, respectively (*see* **Note 5**). In addition, NCBI benefits from a unified approach to constructing database queries, so once the user is familiar with the way to query one NCBI database, the same rules will apply to all other queries made. When accessing the most extensive NCBI data, comprising SNP, gene, protein, publications, phenotype, and sequence, one can execute data queries directly from a menu of choices in a global system termed "*Entrez*" (outlined in detail in **Section 3.1**). The SNP Entrez system *EntrezSNP* has a homepage menu listing the principal SNP criteria (http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp) that help to define a search. This is the main starting point of EntrezSNP and this SNP-focused menu differs from those of other Entrez databases such as EntrezGene and EntrezProtein. Therefore, dbSNP can be accessed in three ways: by using EntrezSNP; by following hyperlinks embedded in other NCBI databases, and by direct access to SNP summary pages, termed "*Cluster Reports*" – forming the core data page for each locus in dbSNP. A Cluster Report can be thought of as the SNP "homepage" listing a full set of the key parameters in a standardized format.

Reference to a SNP within NCBI, within all other SNP databases, and now, almost universally, in the scientific literature is made using a unique identifier: the *rs-number*, consisting of a number prefixed with "rs." As an open database, dbSNP receives submissions from genotyping centers and collates the data into a merged reference set (*see* **Note 4**). Since different centers routinely report identical SNPs to dbSNP, the submissions are clustered into reference SNPs (termed "*refSNPs*") based on genome-wide comparisons of the context sequence submitted. For these reasons the distinction between reference SNPs and submitted SNPs is made by rs and ss, respectively: prefixing a unique SNP identification number with rs, while creating a number for each submission prefixed with ss. All rs-numbers are displayed throughout NCBI as hyperlinks returning the Cluster Report.

2. SNP-related databases in NCBI: PubMed, GenBank, Gene, and OMIM.

   PubMed is the NCBI bibliographic database that provides the starting point for researchers to assess the current published state of the art in their chosen area of study. Data comprise ten million published articles from about 5,000 peer-review journals. PubMed is by default predominantly text-oriented so it works by matching text recognized in the query to the text in the data records, including key words in the article body text itself. Therefore, to work efficiently the system needs to carefully regulate vocabulary, which is done by a separate database of words used to index PubMed known as *MeSH* (i.e., medical subheadings) – searchable itself using the search menu at the top left of each NCBI homepage. It can be an important check to clarify the vocabulary relating to a trait or disease of interest before performing PubMed searches by subject. Searches using rs-numbers can be an efficient way to find studies related to a research aim, but note that the habit persists in many publications of identifying SNPs in genes by the amino acid substitution they create (e.g., shorthand such as MC1R V60L), so these will not be returned from queries (*see* **Note 6**).

   GenBank is the nucleotide sequence database of NCBI. This simple description belies the scale of the information held – a collection of sequences comprising 60 gigabases of data from more than 130,000 species updated daily. Despite the complexity of GenBank, most users interested in SNP analysis will simply require a specific context sequence segment of about 120–200 nucleotides to design a genotyping assay for the SNPs of interest. As explained later the

application of RepeatMasker and a neighbor SNP scan of ±100 bases of context sequence makes it more advantageous to collect it directly from the SNP Cluster Report.

Gene is the gene catalogue of NCBI and like dbSNP presents a single summary page format of relevant information for coding regions, including function summaries, transcription structure, genome maps, bibliography, protein data, sequences, and related links to supporting data. NCBI uses the near-standard gene identifiers that take the form of the letter/number combination standardized by the Human Genome Organization (HUGO) (http://www.gene.ucl.ac.uk/nomenclature/) or throughout NCBI by a GeneID number (*see* **Note 7**). Working with data that include a large proportion of text-based information can be difficult, so to view the context of a SNP or list of SNPs sited in genes it may be preferable to use a purely graphical display of SNP positions aligned with intronic, exonic, and 5′/3′-flanking region sequences such as that given by SNPper (*see* **Section 3.3.1**). Taking text-based data even further towards an article format, the OMIM database has summary pages written as articles describing a phenotype, trait, or disorder with a known or suspected genetic basis. As such, both Gene and OMIM are best consulted along with PubMed during the initial stages of a study design to gain an overview of the current understanding of a disease process. Luckily, OMIM is highly readable and can be described as an online textbook expanded and updated as knowledge of a trait or condition is consolidated. Searches of OMIM just provide the descriptive text and a list of articles without the benefit of the search items highlighted within the text; publications must then be read to gather the links to the area of interest. Similarly, rs-numbers are not regularly listed in the OMIM article body text.

3. HapMap. The original stated aim of the HapMap Project – to determine the haplotype structure of the human genome – has expanded to encompass the characterization of all common human sequence variation. The inclusion of copy number variation and the broadening of ENCODE resequencing efforts to capture rare variation will extend this even further, but HapMap remains dominated by common SNPs and their haplotypes: the correlated arrangement of loci in segments defined by highly variable recombination rates. HapMap data have been structured into study phases I–III with different ranges of SNPs, SNP characteristics, and study populations. It is not always easy to find how each phase was defined but, in short, phase I encompassed about one million SNPs in four populations to give one common SNP per 5,000 bases, phase II consolidated SNP coverage with a further 2.5 million

markers, and phase III has added another seven study populations. Current study population details are outlined in **Table 3.2** and at the time of writing phase III data have become publicly available.

The HapMap Web site provides a wide range of data, but of principal interest will be the genome browser, the SNP summary pages, and the HapMap data mart. HapMap has taken the view that the vast majority of users will start with a graphical overview of a chromosome segment and work outwards from there, so HapMap presents perhaps the best graphical genome browser for SNP variability currently available. Although the default map details (tracks) are relatively sparse, this provides clarity, while numerous other tracks can be added and kept as the user's default arrangement for future browsing. The chromosome coordinates and SNP positions stay as fixed tracks throughout. This representation usefully complements dbSNP since any SNP not characterized by HapMap has a hyperlink rs-number in position to gain the Cluster Report. SNPs characterized by HapMap are linked to their own summary pages, which are briefer in content, so again linking out to dbSNP can be the best approach here too. The HapMap graphical browser really becomes informative when used to study the haplotype structure around the sites of interest (*see* **Section 3.4**) – originally mainly coding regions, but increasingly including intergenic regions identified by genome-wide association studies. The methods of graphical representation of haplotype structure can be a challenge to the first-time visitor to HapMap and it is recommended that users familiarize themselves with approaches used by HapMap and in key papers to display LD and that they understand the characteristics of the principal SNP association metrics of r^2 and D' (3, 4).

4. Ensembl, Santa Cruz, Perlegen, and Assays-on-Demand.
Ensembl and *Santa Cruz* genome databases largely provide alternatives to NCBI to access most of the same SNP and genome data. Ensemble specializes in the analysis of genome features and sequence to best identify and annotate genes and has a large range of species under study. This provides the most informative approach for users interested in comparative genomic approaches: where commonality of nucleotide or protein sequence can be identified by comparing different species. Ensembl has had a pivotal role in the complex task of gene identification and characterization, pioneering automated gene annotation techniques. Hosted in Ensembl, the Vertebrate Genome Annotation (*VEGA*) database provides a range of genome browsers (5). The main aim of VEGA is in providing

the high-quality manual annotation of vertebrate genome sequence. Lastly, Ensembl provides close integration with the high-quality protein sequence database of *Swiss-Prot/UniProt* (http://www.ebi.ac.uk/swissprot/). This comprises manually annotated protein sequences with content that is fully linked with the Ensembl gene annotation pipeline. *Santa Cruz* has several features that can provide easier ways than NCBI to obtain information for SNP analysis – for example, the simple process of collecting extended context sequence for a SNP is more straightforward in Santa Cruz than from within dbSNP (*see* **Note 7**). Therefore, on occasions, working with two Web pages with different genome data browsers (essentially accessing the same underlying information) can be the optimum approach. The guide to Santa Cruz queries is at *Perlegen* and Applied Biosystems's *Assays-on-Demand* are private databases of SNP variability information that has been submitted to dbSNP and is publicly available, but can also be accessed from each company's Web site with dedicated filtered search pages. Filters parallel the query process of Entrez by offering a choice of criteria that reduce the data set returned to a small, more manageable group of items meeting the criteria. Both databases elected to study US European, US African-American, and US Chinese population panels that to a large extent mirror those of HapMap's CEU, YRI, and CHB, so data obtained can be combined from different sources to allow meaningful comparisons of population variability or less often directly between different samples but originating from the same population group (although comparing YRI Africans with African-Americans highlights the about 20–30% European admixture in the latter). The easiest way to compare SNP data from similar populations in different databases is to use SPSmart (*see* **Section 3.5.2**). Note that Perlegen uses an internal SNP identifier with the format "PS+8 digit no." (e.g., PS04631975) but accepts rs-number queries, while SPSmart provides a list of these numbers in its returned data.

Assays-on-Demand SNP data are in large part based on the Celera SNP database generated during the private genome annotation performed by Celera after the human sequence had been completed in parallel to the completion of the public Human Genome Mapping Project in 2000. Celera genome data were available on a subscription basis (as Celera Discovery System, or *CDS*) between 2002 and 2006, but now all Celera's SNP data have been incorporated into dbSNP and can be individually filtered in a search in Entrez with the inclusive term "AND Celera" or the exclusive term "NOT

Celera" (*see* **Section 3.3.1**). Accessing Celera SNP data is also possible through Assays-on-Demand; users in the latter case can utilize a stand-alone tool called "SNPbrowser" comprising five million SNPs from public and CDS sources. This allows access to some of the original CDS SNP and gene annotation but is of most use as an alternative to HapMap for the definition of haplotype structure in a particular chromosome segment (*see* **Section 2.2.2**). Particularly in the population genetics field, Assays-on-Demand allows a simple system to review a large data set of SNP allele frequency variability from three major population groups and so has provided a core search step for many studies seeking to isolate and develop ancestry informative marker SNPs (5).

5. SeattleSNPs. The SeattleSNPs initiative is funded as part of the US National Heart, Lung, and Blood Institute (NHLBI) Program for Genomic Applications (PGA) – the latter abbreviation is used by dbSNP to reference SeattleSNPs SNP submissions. The project has undertaken the resequencing of more than 300 genes identified as primarily important in the inflammatory response, but also including cardiovascular disease and the immunity (a full list of completed genes is at http://pga.gs.washington.edu/finished_genes.html). Although it is important to stress that the gene list mentioned above is not prescriptive – users are encouraged to nominate candidates for consideration. Therefore, SeattleSNPs provides a key opportunity to capture and characterize low-frequency SNPs from whole sequence data that would otherwise escape detection or be subject to acquisition bias (*see* **Note 9**). As sequencing technology has recently undergone one of the periodic quantum leaps in throughput, the chance to properly discover and catalogue new low-frequency SNPs by resequencing sufficiently large sample groups or individuals with a particular disorder will form the next major phase of SNP databasing. The extended ENCODE studies and the SeattleSNPs initiative stand at the vanguard of this work, with the 1,000 Genomes Project poised at the time of publication to take resequencing to the next level of resolution: that of full individual genomes. The evident drawback of SeattleSNPs comes from a focus on targeting a subset of genes or the pathways they occupy with the bias this might represent in attempting to understand the disease process. This is mainly due to the need to direct resources to the best areas for detailed SNP genotyping, and the fact that SeattleSNPs is actively engaged in association studies allows it to combine the knowledge this generates with new targets for resequencing in the genome. The SNP data from resequencing is fed

into a database known as the *Genome Variation Server* (GVS) and users are encouraged to access the tutorials that explain optimum use of SeattleSNPs and GVS at http://www.openhelix.com/downloads/seattlesnps/seattlesnps_home.shtml.

2.2. A Selection of Tools To Aid Analysis of SNP Data

The following tools are available to use as Web-based search systems or stand-alone programs that can help to make directed searches of the databases outlined previously.

1. NCBI tools: dbSNP-announce, MyNCBI, MapViewer, and Genome Workbench.

   Although not strictly online tools, *dbSNP-announce* (http://www.ncbi.nlm.nih.gov/About/news/announce_submit.html) and *MyNCBI* (http://www.ncbi.nlm.nih. gov/entrez/login.fcgi) are important subscription-based adjuncts to any use of dbSNP. Subscribing to dbSNP-announce provides automatic reports to the user's e-mail address of dbSNP updates. 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. MyNCBI, requires a single subscription step to provide a search workspace for the user that provides a clipboard permitting combined searches from stored results obtained at different times (*see* **Section 3.1.7**).

   MapViewer integrates the bulk of the NCBI databases into a customizable genome map of aligned components termed "*map elements.*" The SNP data map element, termed "*Variation*," can be included with any other genome feature in a custom map. A simple, clean icon set against each SNP marker positioned on the map showing a chromosome segment provides a clear summary description of the locus. Map browsing offers an intuitive way to review large numbers of SNPs in one session. Exploring a chromosome segment as a map is the best way to scrutinize the position and characteristics of nearby genome features of importance such as neighbor SNPs, genes, and their transcripts. Furthermore, the features around each SNP can be scrutinized easily through a series of hyperlinks embedded in many of the key map elements such as Genes.

   NCBI *Genome Workbench* (http://www.ncbi.nlm.nih.gov/projects/gbench/) is a stand-alone program that works locally, i.e., independently of individual online access to NCBI. Once installed, it can access and display genomic data from NCBI and combine this with the user's own data in a series of graphical representations. The program is available for download and installation in any operating system format,

and offers considerable flexibility in how the user chooses to align and compare genomic data. This extends to a range of alignment views, phylogenetic tree views, and tabular views of data. It can also align user's data to those of public databases, and retrieve BLAST results. A full guide is beyond the scope of this chapter, so users are encouraged to explore this tool and the five tutorials (http://www.ncbi.nlm.nih.gov/projects/gbench/tutorial.html) for themselves.

2. Checking SNP assay primer designs: BLAST and Santa Cruz In Silico PCR.

   BLAST is a tool for assessing/calculating sequence similarity between a query sequence and the target sequence(s) available in the NCBI GenBank nucleotide databases. Users interested in developing SNP assay designs will query Nucleotide BLAST in two ways: (1) finding the location of a submitted sequence that includes the SNP, as the query "does the submitted sequence exist in a GenBank database?", and (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?" These sequence comparisons can be made by choosing the standard BLAST (termed "*blastn*") and *Search for short and near exact matches* options, respectively. As a simple and quick alternative to BLAST, the Santa Cruz *In Silico PCR* tool (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start) offers a straightforward system that indicates the expected PCR product sequence from primer designs submitted by the user from comparisons to the current human reference nucleotide sequence. This tool is highly recommended since it provides a simple check before committing to primer purchases.

3. Exploring haplotype block structure maps: Haploview and SNPBrowser™.

   Haploview (http://www.broad.mit.edu/mpg/haploview/) is an essential adjunct to HapMap browsing comprising a Java applet tool that permits the analysis and visualization of haplotype block patterns in HapMap data, choosing tagSNPs (7, 8), and estimating haplotype frequencies (*see* **Section 3.4**).

   The Applied Biosystems *SNPBrowser™* tool (http://marketing.appliedbiosystems.com/mk/get/snpb_landing) provides a stand-alone database of five million Celera SNPs that is downloaded to the user's PC and can therefore be accessed offline or in parallel to online searches. The SNP data are presented as a chromosome segment map showing haplotype block distributions defined by Celera's own pairwise analysis of 160,000 SNPs (termed "*backbone validated*

SNPs"), so it provides an alternative to HapMap in the annotation of human haplotype blocks, although it can display both HapMap and Celera haplotype maps. Additionally, SNPBrowser™ is easily configured to tailor haplotype block annotation displayed, SNP type, population studied, and size of the region shown. SNPBrowser™ works along the same lines as Assays-on-Demand in providing a shopping list of SNPs based on user's criteria that can then be ordered as commercial single-plex (Applied Biosystems TaqMan™) or multiplex (Applied Biosystems SNPlex™) SNP genotyping assays.

4. Mapping SNPs and mutations in genes: SNPper.

SNPper provides a tool for the extraction and re-presentation of SNP data from public databases focused on coding regions, offering the clearest system for scrutinizing SNP positions in and around genes (9). Once the user has provided the gene identifier, SNPper will list exonic, intronic, and 5′/3′-regions, plus embedded SNP positions within these, either as a a plain nucleotide sequence or as triplet code groups with their amino acid codes. Although the same output can be achieved with GenBank and Santa Cruz nucleotide browsers, SNPper is a much quicker and simpler system for listing SNPs in a gene of interest with a clean and intuitive graphical summary of the gene. This particularly suits the cataloguing of mutation sites in coding regions since these are usually defined by the amino acid changes they produce and SNPper allows their identification in relation to the SNP landscape that surrounds them, providing a straightforward way to develop genotyping assays.

5. Exploring the effect of SNPs on gene action: PupaSuite, Polyphen, and ESEfinder.

PupaSuite ("Pupa" stands for putative phenotype alterations) encompasses two tools – PupaSNP and SNPeffect – that aid the identification of SNPs effecting the processing of genes (10, 11), namely, sites of intron/exon boundaries or exonic splicing enhancers (ESEs), predicted transcription factor binding sites, and amino acid sequence changes. PupaSuite works with the Ensembl gene annotation and SNP database and can process an uploaded SNP list, but the user can also provide individually identified SNP sites with the aim of exploring their effect on gene action. The utility of PupaSuite is the ability to explore the effect of SNPs on transcriptional activity and splicing as well as protein sequence – an increasingly important step when analyzing coding regions.

PolyPhen (http://genetics.bwh.harvard.edu/pph/data/index.html) is a tool that usefully predicts the possible impact of an amino acid sequence change on the properties of a

protein (12). Although it will not accept nucleotide input directly as it holds a nonsynonymous SNP database comprising about 50,000 SNPs from dbSNP, PolyPhen can check whether a SNP is nonsynonymous or not, using the site tool SNP2Prot. Effects on proteins are tentatively defined as unknown, benign, possibly damaging, and probably damaging. Users can input rs-numbers directly for comparison against the PolyPhen data, but they are advised to go directly to PupaSuite for novel coding SNPs discovered in their study. As with SNPper, this tool is particularly applicable to the characterization of mutations that are, by definition, SNPs at very low frequency.

Of the three tools that help define SNP effects, the most specialized is *ESEfinder* (http://exon.cshl.edu/ESE/), a tool dedicated to identifying precursor RNA splice site changes from SNPs sited at *exonic splicing enhancers* (ESEs) (13). As such, SNPs at the ESE positions of proteins that routinely undergo alternative splicing can profoundly affect the final protein structure. ESEfinder makes use of databases of different ESE sequence motifs to help identify putative SNPs influencing splice patterns.

6. Using SNP haplotypes to detect signatures of selection: Haplotter and SWEEP™.

   Compared with the tools available for studying gene and genome structure described above, population genetics tools are latecomers to SNP database analysis. Data of genome-wide patterns of polymorphic marker variation provide a powerful tool for studying the history of migration, bottlenecks/expansions, and adaptation in human populations. For those interested in analyzing such events, a major advantage in using SNP data is the distribution of SNPs at much higher densities compared with microsatellite or insertion–deletion variation and in the advanced characterization of SNP-based haplotypes. Therefore, SNPs are obvious candidate markers for the analysis of patterns of haplotype structure that can indicate signatures of past natural selection. Positive selection will amplify the frequency of a particular haplotype surrounding a favorable, novel gene variant because the haplotype accompanying the variant on the same chromosome strand also rises rapidly in frequency throughout the population. Before recombination disrupts this association, much higher SNP homozygosity is seen, as identical haplotypes are more likely to be found on each chromosome. Therefore, homozygosity is raised in the immediate vicinity of the selected gene variant and diminishes with distance, as recombination increasingly breaks up associations. This is the basis of the extended haplotype homozygosity (EHH) test that aims to

detect signatures of recent selection by analyzing irregularly long haplotype homozygosity patterns (14). Two tools are available for EHH analysis: *Haplotter* uses HapMap data and is accessed online, while *SWEEP* is a stand-alone program that can use data from any source that has been phased (i.e., allele combinations assigned to one of two strands).

Haplotter (http://hg-wen.uchicago.edu/selection/haplotter.htm) measures a value *iHS* (15) that expresses the contrast between haplotypes with changed frequencies and the surrounding genome landscape, so it can reveal frequency rises in ancestral alleles (positive contrasts as the allele increases in frequency) as well as in variant alleles (negative contrasts). Haplotter can work from gene identifiers or a single SNP landmark (slower and varied in coverage). The program returns plots of iHS, plus standard selection signature or population diversity measures *H*, *D*, and Fst, followed by a table of adjacent genes, colored light blue when showing significant evidence of selection. The major advantage of Haplotter is it allows an unbiased approach to finding regions with indications of recent selection, so in use it is likely to reveal interesting new candidates for more detailed study. This can enable studies to focus on the phenotypes such loci exhibit as a way to explore differences in susceptibility to disease between populations. An advantage of using HapMap data is that the study populations will be extended to allow examination of more widely distributed patterns of local selection.

The stand-alone program SWEEP™ (http://www.broad.mit.edu/mpg/sweep/) acts like Haplotter to measure the rate of decay of homozygosity with distance from putative regions subject to selection (14). Although it requires time and care to become familiar with use of the program, the graphical output, particularly diagrams termed "*bifurcation plots*," provides very good representations of results summarizing extended homozygosity versus genomic distance to the core haplotypes.

SPSmart (http://spsmart.cesga.es/) is a tool that performs the simple task of re-presenting SNP allele frequency data from multiple sources as pie charts identical to those of the HapMap browser. So SPSmart allows the user to review SNP variability across a wider range of populations than is feasible from single databases accessed one by one. This appears to offer little extra value if, for example, the study populations of HapMap phase II and Perlegen are considered, with only a comparison of YRI Africans and African-Americans of potential interest. However, SPSmart also processes data from the Stanford and Michigan University initiatives that have

AQ8

genotyped some 650,000 SNPs in the CEPH human genome diversity panel (HGDP) comprising over 1,000 samples from 51 global populations. Incorporation of HapMap phase III populations has also boosted the scope of global variability that can be accessed with SPSmart.

3. Methods

3.1. Finding Sets of SNPs That Share Particular Characteristics: NCBI Entrez and Boolean Rules of Database Searching

1. The NCBI Entrez system uses *Boolean* terms or *operators* to define searches. These include the principal operators: *AND*, *OR*, and *NOT*, summarized in **Fig. 3.1**. Operators are the key parameters that define the relationship between criteria that describe database entries. In Entrez these descriptive details or criteria are put in groups termed "*fields*" that are defined by *tags* (alternatively qualifiers). Field details can be written in lowercase letters (but following an appropriate format, or *syntax*) ahead of their tags, which are always given in capital letters with fixed syntax within square brackets, for example, to define search criteria "SNPs on chromosome 22" the field would be 22 denoted by the tag [CHR] written as 22[CHR], the chromosome field syntax being a number or X or Y. Users can either manually construct their own search with any combination of fields/tags and operators or simply choose tags from a menu on the EntrezSNP homepage (http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp) and provide fields to make a query using a default AND operator. For users unfamiliar with searches in NCBI, the latter option of choice from a menu can be easier to start with. The principal fields and their tags provided in EntrezSNP are given in **Table 3.3**. Fields separated by spaces alone also default to AND, e.g., query "HERC2[GENE] coding nonsynon[FUNC]" finds the nonsynonymous SNPs in HERC2 exons.

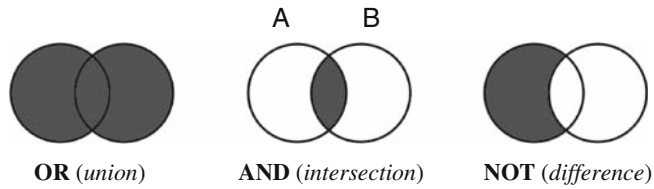

Fig. 3.1. Boolean operators. *OR* applies to all items in A or B, *AND* applies to items found in both A and B, and *NOT* applies to items in A not found in B.

Table 3.3
Key EntrezSNP fields and their tags

Description	Tag	Search field used	Example query
Observed alleles	[ALLELE]	IUPAC allele code (*see* **Table 3.4**)	R[ALLELE] find SNPs with A/G substitutions
Chromosome	[CHR]	Number/X, Y	21[CHR] OR 22[CHR] find SNPs on chromosomes 21 & 22
Base position	[BPOS]	Ranged number & AND & [CHR]	18000:28000[BPOS] AND Y[CHR]–find SNPs in 10 kb segment 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	DARC[GENE] search for SNPs in Duffy blood group, chemokine receptor
Genotyping method	[METHOD]	Description as listed at URL below	Hybridize[METHOD] search for SNPs found by chip hybridization
(http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=snp – METHOD)			
Map weight	[MPWT]	Number: 1=once, 2=twice, 3=3–9 times	NOT (2[HIT] OR 3[HIT]) exclude SNPs mapping twice or more in genome

SNP single nucleotide polymorphism.

2. As a worked example, a search could be written longhand: "find unique SNPs in dbSNP on human chromosome 22 that are AC substitutions and show heterozygosity of more than 45%." To perform a manual EntrezSNP search, place the following search description in the search box: 1[MPWT] AND human[ORGN] AND 22[CHR] AND M[ALLELE] AND 45:50[HET]. Note the field/tag items follow the same order as the longhand query, but this is not essential. The *IUPAC allele codes* applicable to the [ALLELE] field/tag are listed in **Table 3.4**. To perform the same search using the Entrez menu system, go to the limits menu – http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp&TabCmd=Limits – and choose tick boxes in (respectively) *Map Weight*, *Organism*, *Chromosome*, *Variation Allele*, *and Heterozygosity.*
3. Note that the heterozygosity tag in the above example search uses *ranging:* a range of values to define the field, described by a colon (:) in the middle of range limits. The menu-based

Table 3.4
IUPAC SNP substitution codes used in Entrez as the field with the [ALLELE] tag. In addition A, C, G, and T can also be used with [ALLELE] to select all SNPs showing that base as an allele

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		

N can also denote an indeterminate base.

system only allows heterozygosity ranges of 10%, so fine-tuned searches such as 49–50% heterozygosity require manual construction. Two other modifiers of operator function exist for manual searches: *parentheses* and the *wild-card asterisk* (*). Parentheses group search terms into logical sets to obtain items that further operations can search. To a large extent, the logic follows that used in a normal sentence, for example, in a PubMed search "find articles on the effects of heat and humidity on multiple sclerosis" is (heat OR humidity) AND multiple sclerosis, while "find articles on the effects of heat as well as the effects of humidity on multiple sclerosis" is heat OR (humidity AND multiple sclerosis). The wild-card asterisk in place of missing text allows a partial entry to be used as a query term, e.g., using BRC*[GENE] will find both BRCA1 and BRCA2 genes.

4. Each SNP in the EntrezSNP list that is returned from a query defaults to a summary graphic with components that describe the key parameters of the SNP, outlined in **Fig. 3.2a**. For the above example, query EntrezSNP lists 911 SNPs in order of chromosome position. If multiple chromosomes are listed, these are in order: Y, X, 22, 21, etc. A useful feature is the ability to change the default listing order amongst six options, including SNP ID (ascending rs-number) and heterozygosity. When assessing the role of particular SNPs in a disease process or by association with a candidate region, a particularly useful feature is the "Cited in PubMed" tab. Click the "Links/Pubmed (SNP Cited)" hyperlinks in this list to obtain each publication abstract.

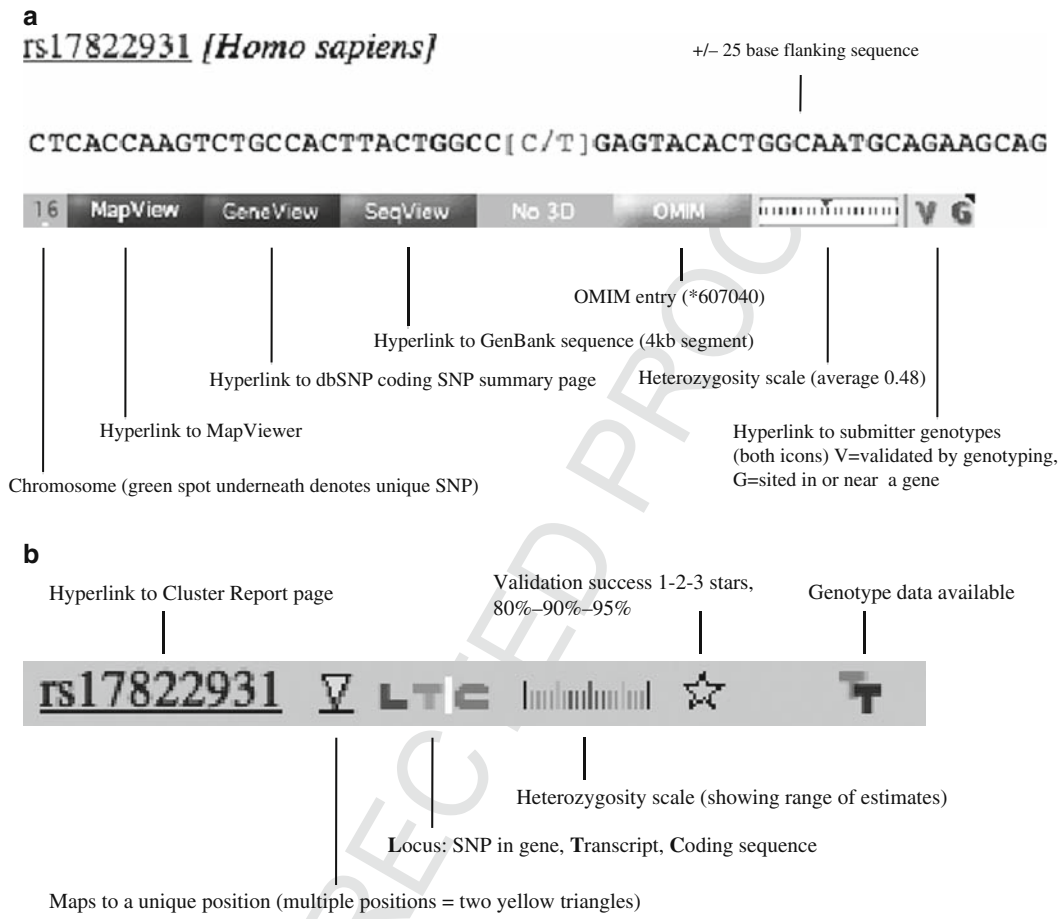

Fig. 3.2. Key to summary graphics for single nucleotide polymorphisms (SNPs). (**a**) Key to summary graphics for EntrezSNP search return of example SNP rs17822931. (**b**) Key to summary graphics for SNPs shown in the chromosome view in NCBI MapViewer of example SNP rs17822931.

5. It is important to note that not all possible search fields can be accessed from the EntrezSNP limits menu: some 14 from a total of 24 are available (the full list and details are at http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp). By far the most useful search field for medical genetics studies omitted from the limits menu is Gene. This allows the listing of SNPs found within and close to a gene (or multiple genes using the OR operator) described by the query. SNPs at the 5′-end and the 3′-end (that order) of the gene are also listed, but it is prudent to extend the search using chromosome/base position tags – [CHR] AND [BPOS] – to capture potential promoter SNPs further afield.
6. An Entrez search can be initiated from the NCBI Entrez homepage by selecting "all databases" from the Search drop-down menu or specifically from Entrez SNP ("SNP"

from the same menu). While searches from the EntrezSNP homepage return just a list of SNPs meeting the criteria, using the *all databases* option creates an NCBI-wide search with hyperlinked numbers of entries from 35 different databases that can be explored individually: making a good starting point in the early stages of a genetic study. The cross-database returns page groups six text-based databases, 26 genomic databases, and three catalogues (books, journals, and MeSH vocabulary) into three separate boxes. A cross-database search can be made using specific or general terms to obtain, respectively, a focused query of the broadest possible coverage within NCBI or a more open ended survey. For instance, "rs2293855" as a query returns a single PubMed reference to a possible role of this SNP in obesity, however with no reference to the gene MTMR9, where it resides, while "obesity" as a query lists no specific SNPs, but more than 111,000 publications and 681 genes, including MTMR9.

7. EntrezSNP gives the most efficient system for progressive searches as the lists generated can be stored in a clipboard and then sent to MyNCBI (avoiding an 8 h inactivity delete step for the clipboard), exported as a text file, combined with new searches, or re-searched itself. This uses the clipboard and history tabs at the top of the EntrezSNP page. The clipboard is a workspace for holding up to 500 items, while history lists the database search activity as numbers prefixed by a hash (#). Making Entrez searches at different times by exporting to MyNCBI allows the user to monitor the number of returns obtained with different search term combinations. Previous searches can be combined as hash fields with operators (e.g., #1 AND #2 gives items common to both searches). It is also possible to use hash fields together with normal fields, helping to build a stepwise record of the search process as it is modified in incremental stages.
8. To automatically reduce SNP numbers returned by a search, certain fields are best used as filters with fixed values including the organism and map weight. Therefore use of the human[ORGN] field/tag ensures only human SNPs are listed and 1[MPWT] ensures all SNPs are unique (i.e., single map weight). The SNP validation tag also provides a system to filter out SNPs detected by sequence comparisons alone, using by frequency[VALIDATION].

.

3.2. Obtaining Detailed Information on a SNP: dbSNP Cluster Reports

1. The Cluster Report page of dbSNP provides most, if not all, the information needed to assess the characteristics of a SNP and design a genotyping assay if required. Each page is broken down into seven sections with largely self-explanatory

headings: Submitter records; Fasta Sequence; GeneView; Integrated Maps; NCBI Resource Links; Population Diversity; and Validation Summary. This section of the chapter outlines the steps required to (1) obtain sufficient context sequence to design a genotyping assay, (2) scrutinize the map position of the SNP with accompanying genomic features, and (3) begin analysis of the population variation shown by the SNP.

2. The Fasta section is named after the fast-all sequence similarity program used by dbSNP to detect identical SNP submissions given in the Submitter section (*see* **Note 4**). Fasta lists the flanking sequence surrounding the SNP position – the quantity of nucleotides listed is variable in extent but always arranged as groups of ten bases, six groups per line, with the SNP positioned on a separate line as an IUPAC code base. The single Fasta header line contains summary locus details explained in the "Legend" hyperlink in the title above. Reliance on dbSNP Fasta sequence alone for primer designs can cause problems (*see* **Note 10**); however, one clear advantage of dbSNP Fasta is the inclusion of information from the initial submitter and from RepeatMasker analysis (*see* **Note 11**) which predicts the likely genomic uniqueness of the SNP context sequence. This can help avoid certain sequence segments that may occur in multiple genomic locations and therefore reduce the specificity of any genotyping assay designs.
3. An essential additional aid to the primer design process is the neighbor SNP detection tool found in the Integrated Maps section. Clicking on the "View" hyperlink under the Neighbor SNP heading of each assembly (for *ref_assembly*: i.e., the reference sequence is recommended) gives all SNPs within ± 100 bp of the reported SNP, including at 0 bp, the target substitution itself. This permits easy location and masking of variable sites that can interfere with the predictable binding of primers designed for the assay.
4. GeneView gives a graphical overview of the SNP if it is located in a gene, giving a position mark using nine color codes for one of 16 predicted functions (hyperlink guide: *color legend*). The cSNP radio button directs one to the coding SNP listing page for each gene (*see* **Note 6**).
5. The Integrated Maps section provides a "snapshot" genome view of the SNP position as a red mark-point by clicking the chromosome number hyperlink. This whole genome view can be used to map a series of SNPs by using the OR operator between each rs-number in the query string in the search box. Including "NOT Celera" eliminates the double mark-points and position listing (*see* **Note 7**). From this overall map any

chromosome can be viewed in detail in NCBI MapViewer, centered on the SNP position by clicking the number hyperlink of each chromosome showing a mark-point. Once in MapViewer, configure the view by clicking "maps and options" to show the "Variation" map as the master to ensure a summary icon set accompanies each SNP position (outlined in **Fig. 3.2b**). Any number of other tracks (map components) can be selected as part of the genome landscape around the SNP, although the single most useful of these is Gene.

6. The Population Diversity section summarizes SNP allele frequency distributions in different populations using gold (reference allele) and blue (alternative allele) bars. Clicking on "Genotype Detail" provides the Genotype and Allele Frequency Report (still in beta status at the time of publication). It is possible to obtain the individual SNP genotypes submitted from each contributing laboratory – data that can be particularly useful when using standard DNA samples, such as Coriell panels (http://ccr.coriell.org/), as genotyping controls in an assay. The easiest way to achieve this directly from HapMap, the major source of SNP population variability data in dbSNP, is to go to the individual SNP information page in HapMap (http://www.hapmap.org/cgi-perl/snp_details? name=rsnumber) and click the "retrieve genotypes" hyperlinks. Genotypes are always listed in the same sample-ID order and so can be downloaded directly to Excel and correctly ordered as rows per population per SNP using the text to columns option (*see* **Note 1**).

3.3. Exploring SNPs in Coding Regions: SNPper and PupaSuite

1. SNPper requires a subscription before the user can explore a gene of interest which can be found using "Gene Finder" by inputting standard identifiers (*see* **Note 8**). Once the gene has been obtained, click on the "Annotated" sequence hyperlink to obtain the nucleotide listing marked as follows: green, 5′/3′; black lowercase, exonic noncoding; black uppercase, coding; gray, intronic; blue underlined, SNPs. A useful approach for the reliable detection of mutations or scrutiny of coding SNPs is to click "View amino acid sequence" to obtain the coding nucleotides as triplet codes above their accompanying amino acids. To locate a novel mutation from the standard "*wild-type amino acid/codon/variant amino acid*" format as normally reported in the literature (e.g., V60L in MC1R) it is necessary to carefully count the relevant nucleotide and codon numbers from the leftmost reference numbers (pencil annotations of a printout are recommended).

2. PupaSuite can accept a list of genes using Ensembl or GeneID identifiers or can review a defined chromosome segment to search for SNPs and suggest an effect. PupaSuite is of particular interest if novel or uncharacterized SNPs are being studied as there is the opportunity to apply the same predictive tools to these loci. To explore the above three options, upload the relevant data to "Upload/paste file of genes," "Search a region," and "Have you got new SNPs?", respectively. There is an option to define gene flanking regions as numbers of nucleotides upstream of the translation start site to find SNPs that may affect transcription factor binding sites. Therefore, PupaSuite is a particularly useful tool for the identification of SNP sites associated with changes to intron/exon boundaries or transcription factor binding. Lastly, additional functional annotations are provided to help assess the impact of the uploaded SNPs, including gene ontology, homology data, and OMIM references.

3.4. Simple Reviewing of SNP Haplotype Block Structure: HapMap

1. Users new to SNP analysis may hesitate before undertaking the process of analyzing human haplotype block structure in regions of interest. The accurate mapping of haplotype blocks, interpretation of D' and r^2 values, selecting tag SNPs to track blocks (3, 4, 7, 8), and assessment of genome-wide patterns of association are all specialist tasks needing care and experience (16). However, all current genetic analysis approaches require an understanding of the likely patterns of association between a set of SNPs and correlating genes or regions of interest; therefore, using HaploView within the HapMap database browser can provide a simple overview to start this process. Once HaploView has been installed on the user's own PC as a Java applet, it is possible to work directly on data from HapMap or Perlegen, but it is easier to start by configuring and viewing LD maps in the HapMap genome browser.
2. Add a gene (or region) of interest to the "Landmark or Region" search box and tick the three "Analysis" tracks: *Phased Haplotype Display*, *LD Plot*, and *tag SNP Picker*. Clearer graphics can be obtained by initially selecting one population at a time by selecting each of "Annotate LD Plot/Phased Haplotype Display" and clicking "Configure...," then choosing a single population radio button.
3. The phased haplotype display presents the alternative haplotype blocks as blue and yellow segments matching the chromosome lengths occupied. The ease with which the user can interpret these depends on the number and length of the haplotypes in the region displayed. As an example, a very

AQ10

simple pattern is shown by ATM: a large but highly conserved gene (strong selective constraints apply to ATM, OMIM: 607585). The phased haplotype plot clearly shows that two haplotypes account for almost two equal halves of the CEU sample. No fewer than 27 of the 31 blocks define this division and the pattern is underlined by a series of identical equal-segment CEU pie charts for the genotyped SNPs in ATM. Note that at blocks 7 (left to right) and 13 a third and fourth common haplotype can be discerned and the third haplotype is characterized by different SNP alleles at blocks 17, 23, 25, 29, and 30. Finally, a singleton (literally a single CEU sample) and a minor-frequency haplotype can be seen in blocks 25 and 29, respectively. The pattern shown by ATM is, in fact, relatively common in the human genome and is termed "yin–yang haplotypes" (17).

4. The LD plots represent the extent of LD between SNPs in the region queried shown as inverted pyramid graphics. The default color scale, also in widespread use in the literature, shows maximum LD as dark red blocks and minimum LD as light gray blocks. Two example genes, CAPG and DTNBP1, illustrate how these plots can summarize both simple and complex predicted LD patterns: showing, respectively, a single, simple pyramid and multiple overlapping pyramids with heterogeneous LD values within each pyramid (checkerboards of red and gray blocks). While this partly reflects gene size and therefore SNP density (note the sevenfold difference between each gene), LD plots can provide a summary overview of recombination and SNP association patterns in the region.
5. The Tag SNP display, once configured, updates automatically between genes and many users may wish to rely just on this system to collect tag SNPs to combine with other core loci (nonsynonymous coding SNPs and translation/transcription-modifying SNPs identified by PupaSuite) to construct simple directed association studies. Although this process has largely been replaced by a standard two-stage approach of whole-genome scans then follow-up directed SNP genotyping, HapMap browsing can give a simple system for assessing the transportability, i.e., the applicability of a tag in multiple populations (8), power, and positioning of the tag SNPs that now form the core battery of markers in whole-genome analyses.

3.5. Assessing Population Genetics Parameters from Online SNP Data: Haplotter and SPSmart

1. Haplotter provides a useful way to begin exploring the population genetics parameters of iHS (outlined in **Section 2.2.5**), *H*, *D*, and Fst, in a genomic region. Queries are initiated by chromosome region, gene, or SNP and this will return four graphics which summarize the above-

mentioned parameters in the same order, with plots for each of the three HapMap panels (i.e., CHB and JPT populations are combined as panel ASN). The Fst graphic plots the three population comparisons to give a useful overview of genomic divergence – in particular the outliers plotted as single points can highlight those SNPs that show very strong interpopulation diversity. A table is given of iHS values around the region of interest with levels diagnostic of EHH highlighted in blue. An often-quoted example that users can investigate for themselves is the gene LCT (gene-ID 3938), this shows a very broad peak of elevated iHS in the CEU population extending well beyond the LCT chromosome region, underlined by high CEU-YRI and CEU-ASN Fst values and blue-labeled iHS levels in the accompanying table. Both the original study of selection patterns in LCT (18) and the Haplotter paper (1) ably explain these patterns.

2. In a fashion identical to Haplotter, SPSmart reviews a region, gene, or SNP list with the primary aim of summarizing the population variability found in multiple SNP databases as pie charts and key population metrics: observed *H* (heterozygosity), expected *H*, Fs, Fst, and *divergence* (In). Usefully SPSmart also pulls from dbSNP chromosome and position, validation status, reference and ancestral allele, and the minor allele frequency, providing alongside the population metrics a succinct one-line summary of each SNP. To explore the population variability of a set of SNPs, choose the SNP databases from HapMap phase II, HapMap phase III, Perlegen, and Stanford/Michigan CEPH-HGDP (4, 4+7, 3, and 51 populations, respectively) and provide the rs-numbers or locations. Clicking "metasearch" permits selection of a population or any combination from each of the five databases (but note the overlap between HapMap phase II and HapMap phase III) prior to uploading the SNPs of interest. For example, to review European frequency variability for the SNP rs12075, click each of the databases, tick the populations of interest, (e.g., CEU, TSI, European American, Italy-Sardinian, France Basque), add the rs12075 query to the "Search by SNPs" box, then (after choosing optional filters) click "search." Pie charts and population metrics (and their downloadable data) are returned as separate tabs, while missing data are clearly labeled. The evident Basque divergence for rs12075 demonstrates the simplicity but informative value of HapMap style pie charts as an aid to reviewing SNP variation across multiple population-based databases.

.

4. Notes

1. Several approaches to database searching using a PC can help the user considerably when manipulating the data obtained from a query. Using tabbed Web page holders in the Web browser of choice (Internet Explorer; Firefox; Safari) allows simple switching between pages. While much SNP data is numerical, all information can be uploaded to a simple individual database in Excel, which now also offers sophisticated text-handling capacity, for offline processing. Although it is rarely recommended by specialists, Excel can offer a simple stand-alone database system by adapting cells to use functions such as LOOKUP, COUNTIF, or those specifically geared to database searches prefixed with "D," such as DGET. Excel compensates for a lack of power by providing a simple and easily mastered set-up of small-scale personal databases suiting many SNP studies. The "Text to Columns" tool in the Excel "Data" menu is a straightforward way to directly process plain text files downloaded from the Web, while preserving the structure of data items separated by spaces, semicolons, or other delimiters. Simple text editors themselves are highly efficient systems for holding and searching data. For example, it is possible to find a single SNP amongst a list of 650,000 in real time using the "Find" function available in all text editors.
2. Google can be used directly to search for specific items such as rs-numbers or mitochondrial substitution sites – the latter being a particularly fruitful approach to finding medical or population studies reporting diagnostic mitochondrial haplotypes (19). For example, entering the search string "human mtDNA G6261A" into Google provides a list of papers reporting this mutation and a supposed role as a cancer risk factor. Care should be taken to ensure full use of the adjacency function of Google searches (known as the Boolean operator *NEAR*), which is not part of most genome database search engines. Therefore, to avoid very long lists of returns, it is advisable to include terms such as *human mtDNA* alongside the standard Cambridge Reference Sequence descriptions. As a compliment to PubMed queries, Google Scholar (http://scholar.google.com/advanced_scholar_search) should also be part of every researcher's online SNP query bookmarks.
3. HapMap experienced minor problems when collating project data generated in different genotyping centers for the same SNP sites, for example, SNP rs1355497 was amongst 37 SNPs reported as showing fixed-difference allele frequencies (1) but has since been shown to be an invariant, monomorphic SNP (also *see* **Note 9**).

4. Since a SNP is characterized by the context sequence each side of the nucleotide substitution site, it should be possible to uniquely define a SNP by referencing organism, chromosome, and base-pair position. However, a small but significant proportion of SNPs are nonunique, so the context sequence and its likelihood of repetition in multiple locations become critical in identifying whether a SNP is unique or not. The submission criteria of dbSNP are very effective at detecting nonunique SNPs, with a process that uses the FASTA program to check a minimum 100 bp flanking sequence to assess if the SNP can be positioned uniquely in the genome and can be matched with other submissions of the same SNP. The proportion of nonunique SNPs remains very small in dbSNP at about 5% and is much more common in certain regions, e.g., pericentromeric areas of each chromosome.
5. Very useful and readable guides to the routine use of the NCBI sites are detailed in a PDF handbook that can be downloaded chapter by chapter (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.part.1). Chapters particularly relevant to SNP research include Chapters 2 (PubMed), 5 (dbSNP), 7 (OMIM), 15 (Entrez), 16 (BLAST), and 20 (Map Viewer). Download them by clicking the PDF icon on each chapter summary page.
6. SNP sites are still routinely described by the amino acid substitution they create rather than an rs-number, particularly if they are mutations or rare enough to escape detection by dbSNP. The easiest way to obtain the rs-number (if it exists) is to record the gene identification number from NCBI Gene (e.g., query "MC1R AND human" gives GeneID 4157) then go to the coding SNP part of dbSNP using the following URL finishing with the ID number to obtain the listing of known coding substitutions and their affected amino acid residues: http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?locusId= number. For example, in gene MC1R, R151C is amongst the most commonly described coding SNPs, and was revealed to be rs1805007 using the procedure described above.
7. Human genes are consistently identified across different databases by a *gene symbol* (sometimes termed the "*gene name*") comprising a series of uppercase letters and numbers (http://www.genenames.org/), provided by HUGO. A *gene ID* consists of a number alone and refers to the number codes given to each gene by NCBI Gene. These can be used both within NCBI as the main point of reference for a gene (e.g., when reviewing coding SNPs) and in certain other databases. A useful gene ID converter tool is provided at http://idconverter.bioinfo.cnio.es/.

8. Use of the single map weight filter in EntrezSNP does not lead to the exclusion of all the SNPs in dbSNP with different Celera locations; however, the list of returns is headed by SNPs that carry the warning "*Mapped unambiguously on non-reference assembly only.*" Note, however, that when one uses the NCBI MapViewer (*see* **Section 3.2.5**) both reference sequence and Celera locations are marked on the genome-wide chromosome map. Therefore, it is advisable to include the term *NOT Celera* at the end of a multiple SNP list in MapViewer queries.
9. Resequencing is the only method of SNP characterization that avoids acquisition bias. This is the phenomenon where the characteristics of the SNP itself affect the chances of its detection by genotyping methods. Examples of SNP features that mean the loci are either not detected or incorrectly genotyped by large-scale projects such as HapMap include triallelic SNPs (the medically important CRP promoter rs3091244 being a notable example), SNPs with very low frequency minor alleles (also missed by resequencing if insufficient samples are typed), and SNPs with very dense arrays of closely neighboring SNPs such as those of the hypervariable major histocompatibility complex. Acquisition bias can also describe the process of selecting SNPs from databases using criteria which bias the SNP lists produced from a query.
10. Often the Fasta section lists less than 100 bp of context sequence each side of a SNP (e.g., rs1805009) – often owing to the fact that a submitting laboratory only provided short segments. The easiest way to obtain ± 100 bp of context sequence for assay primer design purposes is to use the Santa Cruz genome assembly. Add the SNP rs-number to the URL http://genome.ucsc.edu/cgi-bin/hgTracks?position=rs number (several dbSNP builds available), click the highlighted SNP in the map, click "view DNA for this feature," then opt for 100 bases upstream/downstream. The SNP base is the reference allele and is not marked so it is best to use 100+0 and 0 +100 in two separate sequence dumps. Another potential problem in the Cluster Report Fasta section is the occasional (and apparently ad hoc) listing of neighbor SNPs as IUPAC codes (*see* **Table 3.4**). For example, rs1805006 includes no fewer than six other SNPs in ± 100 bp of sequence, given as K, R, R, Y, R, R (in that order), that may cause problems once the sequence is inserted into primer design software. Processing the SNP context sequence directly from Santa Cruz avoids this problem.
11. Fasta section nucleotides are presented in two ways: in uppercase/lowercase letetrs and in black/green. Uppercase letters denote a normal, unique, genomic sequence, while lowercase letters are used for a sequence identified by RepeatMasker

(http://repeatmasker.genome.washington.edu/cgi.bin/RepeatMasker) as a low-complexity or repetitive element sequence. Green denotes a sequence identified by the submitter during the SNP assay process (a single green SNP base signifying identification by sequence comparison), while black denotes a flanking sequence used by NCBI from the nucleotide databases as part of the SNP submission checks.

Acknowledgements

The author would like to thank Maviky Lareu, Antonio Salas, and Angel Carracedo, University of Santiago de Compostela, for useful discussions in the preparation of this chapter. The work was in part supported by funding from Xunta de Galicia: PGIDTIT06PXIB228195PR and the Spanish MEC: BIO2006-06178.

References

1. The International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.
2. The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.
3. Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B. et al. (2002) The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229.
4. Wall, J. D. and Pritchard, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet* **4**, 587–597.
5. Ashurst, J. L., Chen, C. K., Gilbert, J. G., Jekosch K., Keenan S., Meidl P. et al. (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33**, D459–465.
6. Yang, N., Li, H., Criswell, L. A., Gregersen, P. K., Alarcon-Riquelme, M. E., Kittles, R. et al. (2005) Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers. *Hum. Genet.* **118**, 382–392.
7. de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.* **37**, 1217–1223.
8. de Bakker, P. I. W., Noel, N. P., Burtt, N. P., Graham, R. R., Guiducci, C., Yelensky, R., Drake, J.A. et al. (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* **38**, 1298–1303.
9. Riva, A. and Kohane, I. S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics* **18**, 1681–1685.
10. Conde, L., Vaquerizas, J. M., Santoyo, J., Shahrour, F., Ruiz-Llorente, S., Robledo, M. et al. (2004) PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.* **32**, W242–248.
11. Conde L., Vaquerizas J.M., Dopazo H., Arbiza L., Reumers J., Rousseau F. et al. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. *Nucleic Acids Res.* **34**, W621–625.
12. Ramensky, V., Bork, P., and Sunyaev, S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res* **30**, 3894–3900.
13. Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., and Krainer, A. R. (2003) ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res.* **31**, 3568–3571.
14. 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**, 446–458.
15. Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F. et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**, 832–837.

16. Haiman, C. A. and Stram, D. O. (2008) Utilizing HapMap and tagging SNPs. *Methods Mol. Med.* **141**, 37–54.
17. Zhang, J., Rowe, W.L., Clark, A.G. and Buetow, K.H. (2003) Genomewide distribution of high-frequency, completely mismatching SNP haplotype pairs observed to be common across human populations. *Am. J. Hum. Genet.* **73**, 1073–1081.
18. Bersaglieri, T., Sabeti, P. C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake J.A. et al. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120.
19. Bandelt, H. J., Salas, A. and Bravi, C. M. (2006) What is a 'novel' mtDNA mutation–and does 'novelty' really matter? *J. Hum. Genet.* **51**, 1073–1082.