Single-Nucleotide Polymorphism Bioinformatics
A Comprehensive Review of Resources

Andrew D. Johnson, PhD

Recent years have seen near exponential growth in knowledge regarding genetic and genomic variation as more genomes have been sequenced, and corresponding advances and economies of scale in sequencing and genotyping technologies have reduced their relative costs. In parallel with these developments, discoveries of genes contributing to monogenic and complex diseases have rapidly advanced, and bioinformatics databases and software relating to the collection and analysis of genetic data have increased in number, size, and scope. Single-nucleotide polymorphisms (SNPs), comprising the most abundant type of genetic variation, are now the principal raw material underlying most genetic studies and databases. Although other types of variation, including indels, microsatellites, copy number variants, and epigenetic markers remain important to consider and can impact disease, SNPs are largely the easiest to ascertain and the most useful and widely applied markers in genetic studies in the modern age.

Researchers and clinician-researchers are confronted with a dizzying array of software choices and increasingly large and complex datasets and databases relating to SNPs, sometimes working without assistance from a geneticist or a bioinformatician to help guide them. The principle aim of this review is to provide a comprehensive overview of available bioinformatics resources relating to human genetics research, with an emphasis on SNP-centered resources. The review also provides a resource for students seeking an introduction to SNP genetics resources and for wet laboratory molecular biologists conducting SNP-centered research who want to expand their knowledge on ways to apply SNP tools and databases. A number of important issues that affect users and developers of SNP bioinformatics resources are discussed throughout along with practical examples.

Although many of the resources described have relevance and origins in the study of nonhuman species, this review focuses on human clinical applications. The review discusses basic SNP bioinformatics issues, critical databases and their uses, basic strategies and queries using *APOE* examples, software and tools relating to association studies, the prediction and validation of functional SNPs, and miscellaneous SNP resources. The focus is primarily on academic resources that are widely available. Supplemental Table IV provides URL links for all resources described in the text sections in order of their appearance. Readers are encouraged to download the Data Supplement which contains nearly half of the full review article text including sections on practical examples relating to APOE and functional SNP prediction. Key abbreviations and definitions often encountered in this article and other SNP-related articles, databases, and informatics tools are given in the Table.

Basic SNP Bioinformatics Issues

There was a time when the existence of a reliable, comprehensive, centralized, and public resource on genetic variation was uncertain. That time passed with the progressive development of the National Center for Biotechnology Information (NCBI's) dbSNP into the definitive resource for this purpose, and its integration with other popular resources.[1] However, even with the establishment of reliable databases, there are a number of central issues to SNP bioinformatics that still exist. These issues can create problems for both users and designers of SNP tools and databases. A core issue is the need for updates to SNP databases or tools to keep up with current information and discoveries of new variants, which dbSNP addresses through periodic, sequentially numbered releases (called "builds"). Any user of SNP resources should be aware of when those resources were developed and last updated. In some cases, it will be important to tailor input information to software to a particular SNP database build.

A desirable feature in tracking SNP-related information would be to have a unique identifier associated with each SNP, which does not change over time and would be universally applied in all databases and publications. Identifiers known as reference SNP identifiers (SNPids), or rsIDs, exist in dbSNP and partially address the issue of unique identifiers, also including identifiers for indels and repeat polymorphisms. However, as dbSNP grew because of additional submissions of SNP discoveries and improved mapping of SNPs to a more complete reference sequence, it was realized that in some cases multiple rsIDs referred redundantly to the same SNP, resulting in the need to merge alias rsIDs. The result is that depending on when an article was

From the National Heart, Lung, Blood Institute's Framingham Heart Study, Framingham, Mass.
Handled by Calum A. MacRae, MB, ChB, PhD.
The online-only Data Supplement is available at http://circgenetics.ahajournals.org/cgi/content/full/CIRCGENETICS.109.872010.
Correspondence to Andrew D. Johnson, PhD, NHLBI Framingham Heart Study, 73 Mt Wayte Ave, Suite 2, Framingham, MA 01702. E-mail johnsonad2@nhlbi.nih.gov
(*Circ Cardiovasc Genet*. 2009;2:530-536.)

Table. SNP-Associated Abbreviations and Terminology Used Within Literature and Databases

aeSNP (allelic expression SNP): a SNP shown to affect gene expression through AE assay

cSNP (coding SNP): a SNP within the coding region of a gene

eSNP (expression SNP): a SNP that affects the expression of a gene, either putatively or functionally ascertained

nSNP, nsSNP (nonsynonymous SNP): a SNP within a protein codon that results in an amino acid change and may result in a frameshift. Abbreviations may also refer to nonsense SNPs (see below)

ncSNP (noncoding SNP): a SNP not within a coding region

rSNP (regulatory SNP): a SNP lying in a gene regulatory region or affecting regulatory function

sSNP (synonymous or "silent" SNP): a SNP within a protein codon that does not result in an amino acid change

srSNP (structural regulatory SNP): a SNP that affects the gene mRNA products through expression, splicing, or differential selection of gene isoforms

AIM (ancestry informative marker): a marker used in discerning ancestral groups

Ancestral allele: the allele that is believed to predate others in a lineage of sequences

Ascertainment bias: a phenomenon whereby sampling fewer chromosomes tends to underestimate the true proportion of rarer variants in a larger population

cis SNP, *cis*-acting SNP: a SNP that lies in *cis* to the region that it affects

Conserved SNP: a SNP that lies within a conserved sequence region across species or other meaningful groupings

Conservative SNP: a SNP in the coding region that does not result in an amino acid change or disruption

dbSNP: the largest database of SNP information (http://www.ncbi.nlm.nih.gov/SNP/)

Downstream SNP: a SNP that lies in a position that is 3′ of the referent object (eg, gene, intron)

Duplicon SNP: a SNP that lies within a segmental duplication or other region of high sequence homology repeated in a genome, making it difficult to determine if a SNP is a true positive, false positive or indicative of copy number variation

EST SNP: a SNP that was discovered from or lies within an expressed sequence tag

Frameshift SNP: a SNP resulting in a shift in the reading frame of a gene

Functional SNP: a SNP that has been shown in vitro and/or in vivo to exert a functional affect, or in some usages for which there is putative evidence indicating an effect on function

GWAS: genome-wide association study, currently largely SNP based

Imputed SNP: a SNP that has not been directly genotyped but is rather inferred within the specified study sample based on extended genotype information in a subset of the study sample or more often in a distinct study sample

IUPAC code: a code that allows ambiguous specification of SNP alleles with single characters, eg, R=A or G

LD (linkage disequilibrium): a measure of correlation between markers, including SNPs

LSDB (locus-specific database): a targeted data source of information on variation for a locus

MAF: (minor allele frequency) of a variant in a population sample

Monoallelic SNP: a SNP for which only one allele is observed in a database or population sample

MNP (multinucleotide polymorphism): eg, AC<>CT, ATG<>GTG<>del

Multimapped SNP: a SNP that physically maps to multiple regions of the genome

Nonconservative SNP: a SNP that results in an amino acid change or disruption

Nonsense SNP: a SNP that results in a premature termination codon possibly triggering nonsense-mediated decay

Promoter SNP: a SNP that lies within the promoter region of a gene

Pseudoautosomal SNP: a SNP that maps to a pseudoautosomal region or regions of the X and Y chromosomes that exhibit high sequence homology and demonstrates autosomal inheritance patterns

Putative SNP: a "candidate" SNP that is suspected or predicted to exert a functional/biological effect, or in other uses a predicted and unvalidated SNP based on initial results (eg, sequence clustering)

RNA editing: a process whereby DNA nucleic acids in cells are modified to other nucleic acids in mRNA, tRNA, and rRNA, which can be incorrectly interpreted to represent the presence of a DNA SNP

rsID (reference SNP ID no.): a SNP identifier assigned by dbSNP

Quad-allelic SNP: a SNP for which all 4 possible nucleotides are observed within a database or population

Somatic SNP: a SNP whose origin is in nongerm line cells, for example in cancer cells

SNP (single nucleotide polymorphism): in common definition, a single nucleotide variant observed at MAF $\geq$1% within a species population. In practice, SNPs may be variants with MAF $<$1% and may be a subpart of a complex variant (eg, an indel containing SNPs). SNP identifiers (rsIDs) may be assigned to indels, repeats, and even copy number variants

SNP masking: a process whereby SNP positions within sequences are "masked" in a variety of ways

Trans SNP, *trans*-acting SNP: a SNP that lies in trans to the region that it affects

Tri-allelic SNP: a SNP for which 3 nucleotide alleles are observed within a database or population

Typed SNP: a SNP that has been genotyped within the specified study sample or database

Upstream SNP: a SNP that lies in a position that is 5′ of the referent object (eg, gene, intron)

UTR SNP: a SNP that lies in a gene (5′ or 3′) untranslated region

Validated SNP: a SNP that has been validated by a specified approach(es) within a database or study

published or software was designed, a query using a particular rsID may be unsuccessful if aliases for that SNP are not taken into account. Tables that detail such historic merges are available from dbSNP. A Web-based tool, SNAP, takes aliasing into account and also has a feature that allows users to translate lists of rsIDs between current and historic dbSNP builds.[2]

Some SNP bioinformatics tools and databases are only queryable through gene identifiers. Gene identifiers also suffer from potential aliasing and versioning problems. Users can consult the HUGO gene nomenclature committees' online resource (http://genenames.org/) to translate their queries if necessary.[3] Finally, SNP databases and bioinformatics tools do grow obsolete and sometimes are no longer stably maintained at the original URL. This can be due to lack of utility, interest, additional funding support, or simply because the resource migrated to a different URL or was rereleased under a new name or version.[4] The next sections discuss specific SNP databases and strategies and considerations for their use and navigation.

SNP Databases

There are >800 databases of human genetic variation but only a few central databases that are most widely used. These data sources can be split into a few categories, including (1) common genetic variation; (2) rare genetic variation (discussed in the online-only Data Supplement); and (3) databases of variation with additional functional or curated information added or integrated.

Databases of Common Genetic Variation

The largest database of common genetic variation is the NCBI's dbSNP,[1] created after the Human Genome Project discovered a significant number of common variants. dbSNP has grown exponentially in its lifetime, at the time of this submission encompassing information on ≈18.4 million human variants and ≈34.9 million variants in >30 other species. With few exceptions, the other databases, bioinformatics tools, and experiments described in this review rely heavily on the underlying information from dbSNP. The database provides a central, freely available resource for tasks including but not limited to (1) mapping known variation to the human genome; (2) providing identifiers for known and novel variants; (3) ascertaining known variation within or around a gene and estimating the functional effects of variants; (4) designing assays to measure specific variants; (5) estimating prior support and validity that a variant, truly exists; and (6) estimating population allele frequencies of a variant in a variety of populations. The dbSNP variants are mapped to the genome and included in genome browsers (NCBI, University of Southern California Santa Cruz, and European Molecular Biology Laboratory), allowing users to integrate SNP information relatively easily with other features of genome annotation. dbSNP also features haplotype predictions, snpBLAST that allows users to query sequences against dbSNP, and targeted databases, including dbMHC, dbLRC and dbRBC. Information on individual SNPs can be retrieved, including gene-related annotation, information on sample assay types and validation, the SNP submitters, and allele frequencies in measured populations. Batch querying of many SNPs and download of all information in dbSNP is also available.

Another database of central importance to SNP bioinformatics is the International HapMap project. The HapMap project began as a collaborative effort to comprehensively survey allele frequency and linkage disequilibrium (LD) patterns among common human genetic variants across worldwide populations, and it now provides a critical platform of information for large-scale genetic association projects. The project has now progressed through 3 phases: phase I[5], phase II,[6] and phase III. The phase III data release of HapMap currently contains information for ≈1.6 million SNPs in 1115 samples from 11 worldwide populations, assayed on DNA derived from immortalized lymphoblastoid cell lines. This genotype information is available for download and can be viewed through the HapMap browser, other genome browsers and within dbSNP records. The HapMap information is valuable in a range of uses including but not limited to validating the presence and relative allele frequency of many SNPs in the genome, estimating and delineating haplotype blocks, estimating recombination rates and hotspots, providing the basis for genotype imputation, guiding selection of SNPs for genome-wide arrays, and providing reference samples for genotype assay design and in vitro experiments. An initiative underway, the 1000 Genomes Project, aims to sequence >1000 individual human genomes, including many HapMap samples. This project began releasing data in 2009 and will provide an even deeper resource on human genetic variation, not only capturing common variation but also discovering more rare variation than ascertained in earlier HapMap phases.

The HapMap and dbSNP provide a view of worldwide similarities and differences in allele frequency of human variation. There are a number of databases aimed at characterizing variation within or across human populations, including the Japanese SNP database,[7] the ThaiSNP database, the Taiwan-Han Chinese SNP database, SNP@ethnos,[8] the CEPH genotype database and ALFRED.[9] Many of these databases rely completely on or extend on SNP information from dbSNP and HapMap. ALFRED is notable because, although it contains information for only ≈18,000 variants, it has the most diverse sample encompassing >680 populations. Databases reporting on diverse samples have a variety of potential uses, including estimating expected population control frequencies for SNPs of interest, deriving power calculations for SNP studies, and estimating population ancestry measures.

Users of the common SNP databases that are mentioned above have multiple options to retrieve and organize SNP-centered information, often starting with simple downloads or tools available at the source websites. A number of "marts" allow relatively fast retrieval of SNPs that meet user-defined criteria (eg, population minor allele frequency thresholds), including (1) HapMart at the HapMap website; (2) BioMart developed by the OiCR and EBI; (3) SPSmart[10]; and (4) Genome Variation Server at National Heart, Lung and Blood Institute. Given a list of SNPs, a user can also conduct a "batch retrieval" from dbSNP to retrieve information avail-

able there. The UCSC Genome Browser also features easy viewing and downloading of SNP-centered information. For more complex SNP queries, BioMart or the Table Browser at the UCSC Genome Browser provides potential solutions. Although BioMart is currently limited to an older dbSNP version, it can provide filtering based on SNP validation status and SNP function (eg, all stop codon SNPs in the genome). The UCSC Table Browser allows users to construct open-ended queries based on UCSC annotations, for instance: retrieving all SNPs that are found in human micro-RNAs, all SNPs in conserved transcription factor–binding sites, or all SNPs found in Affymetrix U133 gene expression array probes. These queries are performed based on the relational data tables that underlie the UCSC Genome Browser annotation tracks. For individuals with interest in deeper analyses, most of the major databases of common variation (eg, dbSNP, HapMap, UCSC) include an option to download all data with minimal restrictions on use. An important consideration in any SNP informatics project is that each SNP data source contains potential ascertainment biases.

Additional Databases and Data Sources

There are a range of resources that provide useful information relating to specific variants, often integrating information from the literature, or multiple databases or datasets. The OMIM database is an excellent example, combining expert curated summaries of the literature with information on allelic variants and searchable by SNP identifier. The Human Genome Epidemiology Navigator provides flexible mechanisms to query for genetic associations in the literature based on phenotypes (Phenopedia) or genes (Genopedia).[11] Similarly, the Genetic Association Database at the National Institutes of Health also provides a resource to search >40 000 association studies through many mechanisms, providing information for some studies on populations studied, statistical associations with specific variants, and study conclusions.

Genome-wide association studies (GWAS) based on large-scale SNP genotyping have resulted in the generation and analysis of a previously unprecedented scale of data in the genetics literature, with >350 estimated GWAS published at this time and billions of genotypes analyzed. A number of recent efforts have made available access to an extended, albeit rather incomplete, proportion of GWAS results. The most extensive and centralized resource to date is NCBI's database of genotypes and phenotypes (dbGAP), although access to many results requires formal application. Separately, a list of top results from GWAS studies is currently maintained by the National Human Genome Research Institute. Another resource, HGVbaseG2P, an expansion of the previous HGVbase,[12] provides an informatics structure and "mart" for querying GWAS results with some restrictions. A number of available catalogs of GWAS scans for association with gene expression are publicly available and are highlighted in the Data Supplement.

We recently published a survey of the characteristics of top GWAS results and created an open access database of >56 000 SNP associations based on available results from 118 GWAS representing scans for >400 phenotypes.[13] The survey not only indicated that there may be significant insights to be made by open sharing of such genomic results, particularly by allowing them to be annotated in a standardized fashion to allow for additional analyses, but also showed that many investigators have chosen to share extensive results. At the same time, a recent effort revealed that it may be possible to identify the presence of individual participants in a cohort given availability of GWAS results, which has prompted caution and even retraction of the release of such results, particularly where the results included information on population allele frequencies.[14]

This raises important issues not always at the forefront in bioinformatics practice, namely ethical, social, and legal obligations to protect participants who have contributed data. SNPedia is an online tool that gives people information on risk based on their individual genotypes and an algorithm run over information from the literature. However, such initiatives are likely premature and possibly misleading, given our current level of understanding and the modest known risk contribution of most SNPs present on genotyping arrays.[15] Given the continued growth in large-scale genetic association studies, the release of 1000 Genomes Project data, and the expected imminent wave of cheaper personal sequencing, researchers will have to struggle with ethical questions of when and how to inform participants of genetic results, as well as ways to protect personal information while enabling the appropriate storage and access of large amounts of data for research purposes.

Software for the Conduct and Interpretation of Genetic Analysis Studies

The bulk of SNP-related software relates to genetic study design, collection and management of genetic information, and the statistical conduct, analysis, and interpretation of genetic studies. It is beyond the scope of this review to address the complement of software available in this area. An excellent online list is regularly updated, currently containing links and information for >480 programs (http://www.nslij-genetics.org/soft/). Tools to conduct statistical genetic analyses and to analyze LD patterns among markers and predict haplotypes are among the most frequently developed software areas. Here, I summarize popular and useful software and recent developments with a focus on SNP association software rather than linkage software. Many statistical geneticists and bioinformaticians also implement their own code for analyses, often relying on packages in the R programming language (eg, haplo.stats_R for haplotype association analysis). An extended version of this section highlighting additional software is found in the Data Supplement.

A first, and sometimes last, consideration in genetic analyses is a power calculation. Although such calculations are implemented in some genetic analysis software, an excellent standalone site exists for this purpose: http://pngu.mgh.harvard.edu/≈purcell/gpc/.[16] When samples and markers are determined, if prebuilt genotyping strategies are not applied, the next step is often assay design and validation. Careful use of software to assist in assay design and laboratory information management systems can help reduce errors and cut genotyping costs. Many programs exist to aid in assay design for various genotyping approaches, including some with specific

components for SNP design: PrimerBatch3, which includes multiple SNP assay types[17] and a popular general tool Primer3.[18] Good genotyping assay design principles should be applied when possible, including consideration of potential confounding effects from repetitive regions, SNPs that may hybridize to probe sequences, GC-rich regions, polynucleic acid stretches, and potential triallelic variants. For laboratories handling high volumes of genotyping results, a laboratory information management systems may be a desirable informatics capability.

For those undertaking GWAS analyses, an early concern is the careful application of genotyping calling algorithms. These algorithms have progressed over years with original algorithms largely displaced by algorithms that demonstrate improved accuracy. The major algorithms and software are largely platform specific (eg, Affymetrix versus Illumina) and in some cases array specific. Birdsuite[19] supports SNP, CNV, and CNP calling for the Affy 6.0 array. Current genotyping algorithms applicable to Illumina arrays include Illuminus[20] and GenoSNP.[21] Those conducting a DNA pooling approach to conserve samples and funds may apply pooling-specific calling algorithms, including GenePool.[22] For groups interested in integrating and managing genotype calls across Affymetrix and Illumina platforms, Integration of Genotypes from GeneChips (IGG) is specifically designed for this purpose.[23] Identifying overlapping SNPs and LD proxies across commercial arrays can also be done easily with SNAP.[2]

Once genotypes have been collected, cleaned, and called, depending on the scope of the project (eg, GWAS, candidate gene, or replication), a number of additional steps may be taken. Calculation of straightforward population genetic measures, such as Hardy-Weinberg equilibrium, may be informative. Such statistics are included in many programs or separate routines such as SNP-Hardy-Weinberg equilibrium[24] are also available. With genotypes fixed, another step can be to examine and potentially adjust for population structure and stratification, which can be a source of confounding in association analyses. Implementations are available for parametric approaches (STRUCTURE[25]) and nonparametric approaches (EIGENSTRAT[26]), which have gained favor in recent years. The PLINK whole genome association toolkit also includes a module for correction based on identical by state calculations for whole genome genotyping data.[27] Another approach often applied in whole genome level analysis is the use of genomic control calculations for adjustment.[28]

The use of inference based on measured SNP genotypes to estimate untyped SNPs, or allele dosages, has been an active area of development and application in recent years. Although also applicable to local and regional contexts, imputation is generally applied on a genome-wide scale. The methods for imputation generally take similar approaches, relying on LD relationships between SNPs in the HapMap, and are relatively computer intensive. Popular imputation programs include MACH,[29] IMPUTE,[30] PLINK,[27] BEAGLE,[31] BimBam,[32] and TUNA.[33] Application of these programs to most genome-wide genotyping datasets currently results in estimates for >2 million SNPs, increasing genomic coverage and allowing groups with distinct starting genotyping platforms to compare results or conduct meta-analysis. A review of imputation-driven meta-analysis gives a more detailed overview of important considerations.[34] Two recent empirical comparisons of imputation software have favored the use of MACH, IMPUTE, and BEAGLE.[35,36]

When a final genotyped or imputed set of SNPs is ready, the selection of appropriate tools for statistical association is a critical step. The selection of software and routines is influenced by many factors, including the nature of the phenotypes studied, the availability and selection of covariates, the extent of missing data, family structure and pedigree availability, cohort or case-control design, population stratification, the level of expertise of researchers involved, and the extent to which information can be harmonized if multiple populations or studies are combined. The implementation and sharing of association test routines in the R programming language is popular, with many available through Bioconductor (http://www.bioconductor.org/). Many specialized genetic analysis tools exist; I highlight only a few. The PLINK toolkit is arguably the most comprehensive and well-documented freely available system for conducting large-scale genetic analysis, including options for population-based tests under different models, family-based testing, haplotype tests, conditional tests, imputation, stratification, and annotation.[34] Additional comprehensive linkage and association software packages include Mendel,[37] MERLIN,[38] and Genomizer,[39] GHOST[40] (family-based) and GenAbel (genotype based) and ProbAbel (imputed based) for GWAS analysis. Family-based association tests are implemented as standalone software or as part of larger packages, including FBAT[41] and QTDT.[42] Two association software packages aimed at being relatively user-friendly with Windows GUI implementations are PowerMarker[43] and FamHap.[44] Combining evidence for association across multiple studies can provide evidence for replication of genetic effects. Considerations of power and design in the studies, nature and harmonization of the phenotype measurements and statistical tests, and matching of the genetic alleles modeled and direction of effect are all important meta-analysis considerations.[34] METAL (unpublished) is widely used to conduct genetic meta-analysis including on the genome-wide scale.

In particular, in the conduct of GWAS, where the scope of results handled is large, there is often more informatics to do after the primary analysis or meta-analysis is complete. One of the critical questions that arises when a significant association signal is detected in a GWAS or other study is what are the responsible, functional genes and variants? The peak marker associated is likely not the functional explanation and may even be located in or near a gene that does not have a role in the phenotype studied. A likely scenario is that the associated markers are in LD with one or more other markers, known or yet unknown, that are the functional explanation for the association signal. An immediate task is plotting results (eg, WGAViewer,[45] GWAS GUI,[46] AssociationViewer[47]), particularly regional LD and association plots that can be generated with SNAP through a Web interface[2] or the popular tool, Haploview.[48] Consideration of such plots can be helpful in evaluating the approximate genomic boundaries likely to contain functional variants. Identifying strongly associated

variants and those in LD informs further efforts like resequencing, molecular experiments on candidate genes in the region, and the prediction and validation of potential functional variants. The prediction of "functional SNPs" is an active and evolving area of SNP bioinformatics. Readers are encouraged to read the Data Supplement, which details bioinformatics tools and servers aimed at predicting functional protein and regulatory polymorphisms, respectively, along with important considerations for their use and interpretation. Functional prediction tools are described in detail in Supplemental Tables I through III. Practical bioinformatics examples are also discussed in relation to *APOE* variants in the Data Supplement along with additional areas of SNP bioinformatics, including tools relevant to sequence data, pathway mining, and literature searching.

Conclusion

Bioinformatics has been an integral part of genetics and genomics since relatively early studies on the effects of protein coding SNPs and the challenge of assembling and annotating early genome sequences. The growth in size and scope of SNP-related databases has been met with a growth in bioinformatics resources, and as a result, new opportunities for data analysis and integration have followed. The future impact of bioinformatics on SNP-related research is likely to continue to be great as decreased sequencing costs, technological advances, and large bio-bank projects not only lead to further insights and opportunities but also present difficult data management challenges.

Sources of Funding

Dr Johnson is supported by a National Institutes of Health Intramural Research Training Award position within the National Heart, Lung and Blood Institute.

Disclosures

None.

References

1. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29:308–311.
2. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*. 2008;24:2938–2939.
3. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO Gene Nomenclature Database, 2006 updates. *Nucleic Acids Res*. 2006;34:D319–D321.
4. Wren JD, Bateman A. Databases, data tombs and dust in the wind. *Bioinformatics*. 2008;24:2127–2128.
5. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437:1299–1320.
6. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449:851–861.
7. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y. JSNP: a database of common gene variations in the Japanese population. *Nucleic Acids Res*. 2002;30:158–162.
8. Park J, Hwang S, Lee YS, Kim SC, Lee D. SNP@Ethnos: a database of ethnically variant single-nucleotide polymorphisms. *Nucleic Acids Res*. 2007;35:D711–D715.
9. Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, Yeh CC, Miller PL, Kidd KK. ALFRED: the ALelle FREquency Database. *Update Nucleic Acids Res*. 2003;31:270–271.
10. Amigo J, Salas A, Phillips C, Carracedo A. SPSmart: adapting population based SNP genotype databases for fast and comprehensive web access. *BMC Bioinformatics*. 2008;9:428.
11. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. *Nat Genet*. 2008;40:124–125.
12. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ. HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Res*. 2004;32:D516–D519.
13. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. *BMC Med Genet*. 2009;10:6.
14. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*. 2008;4:e1000167.
15. Jannsens AC, Gwinn M, Bradley LA, Oostra BA, van Duijn CM, Khoury MJ. A critical appraisal of the scientific basis of commercial genomic profiles used to assess health risks and personalize health interventions. *Am J Hum Genet*. 2008;82:593–599.
16. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003;19:149–150.
17. You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*. 2008;9:253.
18. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol*. 2000;132:365–386.
19. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008;40:1253–1260.
20. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatowski DP, Clark TG. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics*. 2007;23:2741–2746.
21. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. *Bioinformatics*. 2008;24:2209–2214.
22. Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL, Webster JA, Beach T, Sando SB, Aasly JO, Heun R, Jessen F, Kolsch H, Tsolaki M, Daniilidou M, Reiman EM, Papassotiropoulos A, Hutton ML, Stephan DA, Craig DW. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. *Am J Hum Genet*. 2007;80:126–139.
23. Li MX, Jiang L, Ho SL, Song YQ, Sham PC. IGG: a tool to integrate GeneChips for genetic studies. *Bioinformatics*. 2007;23:3105–3107.
24. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*. 2005;76:887–893.
25. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*. 2003;164:1567–1587.
26. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904–909.
27. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81:559–575.
28. Devlin B, Roeder K. Genomic control for association studies. *Biometrics*. 1999;55:997–1004.
29. Li Y, Abecasis GR. Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet*. 2006;S79:2290.
30. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*. 2007;39:906–913.
31. Browning BL, Browning SR. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet*. 2009;84:210–223.
32. Guan Y, Stephens M. Practical issues in imputation-based association mapping. *PLoS Genet*. 2008;4:e1000279.

33. Wen X, Nicolae DL. Association studies for untyped markers with TUNA. *Bioinformatics*. 2008;24:435–437.
34. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet*. 2008;17:R122–R128.
35. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A. A comprehensive evaluation of SNP genotype imputation. *Hum Genet*. 2009;125:163–171.
36. Pei YF, Li J, Zhang L, Papasian CJ, Deng HW. Analyses and comparison of accuracy of different genotype imputation methods. *PLoS ONE*. 2008; 3:e3551.
37. Lange K, Sinsheimer JS, Sobel E. Association testing with Mendel. *Genet Epidemiol*. 2005;29:36–50.
38. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet*. 2002;30:97–101.
39. Franke A, Wollstein A, Teuber M, Wittig M, Lu T, Hoffmann K, Nurnberg P, Krawczak M, Schreiber S, Hampe J. GENOMIZER: an integrated analysis system for genome-wide association data. *Hum Mutat*. 2006;27:583–588.
40. Chen WM, Abecasis GR. Family-based association tests for genomewide association scans. *Am J Hum Genet*. 2007;81:913–926.
41. Horvath S, Xu X, Laird NM. The family based association test method: strategies for studying general genotype–phenotype associations. *Eur J Hum Genet*. 2001;9:301–306.
42. Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet*. 2000;66: 279–292.
43. Liu K, Muse SV. PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics*. 2005;21:2128–2129.
44. Herold C, Becker T. Genetic association analysis with FAMHAP: a major program update. *Bioinformatics*. 2009;25:134–136.
45. Ge D, Zhang K, Need AC, Martin O, Fellay J, Urban TJ, Telenti A, Goldstein DB. WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res*. 2008;18:640–643.
46. Chen W, Liang L, Abecasis GR. GWAS GUI: graphical browser for the results of whole-genome association studies with high-dimensional phenotypes. *Bioinformatics*. 2009;25:284–285.
47. Martin O, Valsesia A, Telenti A, Xenarios I, Stevenson BJ. AssociationViewer: a scalable and integrated software tool for visualization of large-scale variation data in genomic context. *Bioinformatics*. 2009;25: 662–663.
48. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*. 2005;21: 263–265.

Single-Nucleotide Polymorphism Bioinformatics: A Comprehensive Review of Resources

Andrew D. Johnson

Circ Cardiovasc Genet. 2009;2:530-536
doi: 10.1161/CIRCGENETICS.109.872010
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 1942-325X. Online ISSN: 1942-3268

SUPPLEMENTAL MATERIAL

Databases of rare genetic variation

While the databases of common variation discussed in the main text include some variation found at <1% minor allele frequency (MAF), they were largely designed to discover and make available SNPs with MAF ≥1%, and thus in current formats are currently under-representative of rarer variation and other classes of variation including indels, repeats, and structural variants. Since rare variation requires large sample sizes to reliably ascertain, we do not yet know the extent of rare variation in human populations and there is not yet an extensive or unbiased database of rare human variation. The 1,000 Genomes Project is designed to address these issues, and simulations indicate that many uncommon variants will be discovered[1]. However, based on decades of research we currently do have knowledge of many genetic variants that cause Mendelian inheritance of traits, sometimes with major clinical implications. These variants are generally relatively rare in populations, and often ascertained in relatively small sample sizes or pedigrees through direct sequencing approaches. Arising from a need to share information on mutations carefully curated by the experts in specific genetic disorders, locus-specific databases (LSDBs) were proposed and created beginning in the 1990s[2]. The Human Genome Variation Society (HGVS) maintains lists and URL links to hundreds of LSDBs at its website, with a similar list and a mechanism for centralized mutation submission available (www.centralmutations.org).

The standardization of information content in LSDBs is of particular importance given they can be utilized in making clinical decisions. Suggested recommendations for the design, implementation and content of LSDBs have been discussed[3]. Software systems have also been designed to facilitate the creation and standardization of new LSDBs including the Leiden Open Variation Database [4] and the Universal Mutation Database[5]. The standardization of mutation nomenclature is one of the major challenges in the genetics literature and in the design and

maintenance of variation databases, and can be vital to making diagnostic and clinical decisions[6,7]. A program called Mutalyzer does provide help for those trying to generate standardized nomenclature for variants[8], and dbSNP now contains a similar feature. The problem is central to rare variation but extends to common variants and is exemplified by scanning early genetics articles where the identity of a variant is often obscure, and may require considerable targeted efforts in order to define the variant, place it in current knowledge contexts or design assays to target it (for related strategies see *APOE* examples below).

A handful of databases are referred to as Central or General Mutation Databases (CMDBs,GMDBs) and have a major focus on rare variation but also include common variant and association literature findings. Among these is one of the oldest, most useful resources in genetics, the Online Mendelian Inheritance in Mammals (OMIM) database. OMIM provides extensive, curated and gene-centered information including alleles, dbSNP identifiers, descriptions of clinical reports of phenotypes, gene origins, biochemical features and animal models. As a result OMIM is often the first line resource of choice for researchers to gain an understanding of the literature surrounding a particular gene. For those interested in potential functional variation culled from the literature the Human Gene Mutation Database (HGMD) provides a database and information on more than 80,000 variants that alter protein sequence, splicing or other mechanisms, many of which are rare variants. HGMD does not include mitochondrial or somatic variants. Mitochondrial variants are included in OMIM, specific LSDBs, and a number of databases that focus on mitochondrial variants (e.g., MitoMap). Somatic variants are generally covered in specific LSDBs, although a few large databases exist (e.g., COSMIC, GAC, TP53). Researchers have commented on limitations, inconsistencies and mistakes among sources of information on rare and pathogenic variation[9-11]. These cases highlight the fundamental difficulty of establishing a definitive source on genetic variation, which requires frequent updating, and depending on decisions about content may lack some information that will be found at another source. From a user standpoint, this means that a

working knowledge of multiple data sources will likely be important to many projects[11]. It also belies again the important caveat that each data source, whether focused on common or rare variation, has its own potential ascertainment biases that may influence analyses and interpretations, particularly surveys based on complete databases.

Recently a number of genetic testing databases have been established which provide information on certified testing labs, along with the diseases, genes and variants they test for, and contact information for the labs. In the United States, GeneTests provides information on CLIA-approved labs and the variants they test. Similar databases have been established elsewhere in the world including EuroGentest. In a recent survey of the variants in the GeneTests database, we found that <13% of tested variants were SNPs (Johnson et al., unpublished results). We searched for these CLIA-tested SNPs employing a variety of bioinformatics resources including dbSNP, OMIM, and ExPASy protein sequence annotations along with the UCSC Blat tool to localize the precise coordinates where variants should be annotated. We used the SNAP tool[12] to account for potential alias identifiers. Overall we found that SNP identifiers do not exist for >70% of these clinically tested SNPs (Johnson et al., unpublished results). This highlights the significant challenges currently faced in the description, storage and distribution of information on rare variation where the majority of variants have not been yet been assigned unique identifiers that map to specific sequences and locations in the genome. Although there is hope for future improvement, for instance from the Gen2Phen project which aims to unite LSDBs and other genetic, genomic and phenomic data sources, given the current state of resources on rare variation, researchers should be prepared to consult with and move between multiple data sources.

Basic SNP bioinformatics tasks and strategies

This section introduces and provides examples for a number of common SNP-related bioinformatics tasks using the example of variants in the *APOE* gene. Alleles in the *APOE* gene

are among the strongest and consistently replicated common genetic associations known, with a highly studied *APOE4* gene variant (comprising two SNPs), being associated with Alzheimer's disease, as well as dyslipidemia and coronary artery disease. Locating a SNPid for variants of interest is often a primary strategy to enable the gathering of more information for further tasks (e.g., in order to obtain surrounding DNA/RNA/protein sequences to design targeted assays for experiments, or to conduct *in silico* functional predictions as discussed in other sections below). The commonly studied *APOE* alleles correspond to combinations of SNPs in amino acids 112 and 158 of the mature protein. Specifically, these are the E2 (Cys112, Cys158), E3 (Cys112, Arg158), and E4 alleles (Arg112, Arg158). One of the first line resources for locating SNPids is a search of dbSNP. In the case of *APOE* a search of dbSNP (Build 130 version) for SNPs in the coding region does not reveal any variants listed at positions 112 or 158, but instead shows substitutions encoding Cys or Arg changes at positions 50, 130, 163, and 176. This case provides an excellent example of typical problems that can be encountered in searching for genetic variants in databases – they may often be named in the literature and even databases by something other than a SNPid or amino acid position, and may be joined with other SNPs in named (e.g., *APOE* "E2 Christchurch") or numbered (e.g., *APOE**4) alleles or haplotypes. Furthermore, even when a position is given for a genetic variant it may refer to a position within a particular sequence, transcript or protein isoform, of which there can be one to many per gene. Relying on genome coordinates, if available, to identify SNPs in data or genome browsers has the potential pitfall that as the human genome reference sequence has been filled in and changed (e.g., most recently from Build 35 to Build 36) the relative coordinates of SNPs have also changed. If old coordinates are known for a SNP(s), the UCSC LiftOver tool can be used to map to the coordinates to the current genome framework. Thus, a search of multiple data sources may be required to locate information on target variants.

In this case, dbSNP reports variants relative to their position in the *premature* APOE protein isoform. An invaluable source of information on protein isoforms, their cleavage and

primary genetic variants is the ExPASy/Swiss-prot/UniProt database, which is searchable by gene/protein ID. A search of the ExPASy database for human APOE indicates the first 18 N-terminal amino acids of APOE are a signal peptide that is cleaved to form the mature protein, thus designating the 112 and 158 amino acid positions referred to by the E2/E3/E4 nomenclature. Combining this information one can infer that the variants in dbSNP at positions 130 (rs429358) and 176 (rs7412) refer to the variants of interest for E2, E3 and E4 alleles.

Another strategy for locating or verifying the identity and localization of genetic variants is to use primary sequence information for the variants, or amplicons or oligonucleotide primers targeting them, and run a search for matches against the annotated genome. Referring to an article the following sequences are found surrounding the *APOE4* alleles, for positions 112 and 158 respectively: "ATGGAGGACGTGTGCGGCCGCCTGGTG" and "GACCTGCAGAAGCGCCTGGCAGTGTAC"[13]. While many tools could be used to match these sequences against the human genome, including a SNPblast tool from dbSNP, one of the most rapid and effective strategies is to employ UCSC Genome's BLAT tool[14] to visualize perfect or near-perfect sequence matches against the genome. For sequences of length ~20 nucleotides or more BLAT has a high probability of finding matches against the genome, and has the added advantage of being able to match across exon boundaries for RNA and protein sequences. A perfect or near-perfect match is expected in most cases, but the presence of mismatches between query sequences and the reference genome may occur due to polymorphisms. Multiple perfect, near-perfect or partial matches to the genome may reveal false positive SNPs, or "duplicons", SNPs within repetitive regions, or SNPs that legitimately map to multiple regions (e.g., the pseudoautosomal region), any of which may be a concern for researchers. In such cases, the quality of SNP alignments to the genome can be checked via a dbSNP weighting score that can be accessed by clicking on individual SNPs within the UCSC Genome Browser. In this case copying, pasting and querying the sequences above in BLAT reveals single perfect matches to the *APOE* gene, and clicking "browser", with SNP annotation tracks selected,

reveals the SNPids rs429358 and rs7412 align perfectly to the query sequences as expected. Clicking on the individual SNPs reveals more information including perfect alignment weighting scores for these SNPs along with relevant links.

Retrieving DNA, RNA or protein sequence containing variants is often a next step to executing further bioinformatics tasks like primer design for assays, or predicting potential functional consequences of polymorphisms like disruption of splicing, transcription factor binding, protein structure or post-translational modification. A variety of approaches exist to retrieve sequence around variants. In the BLAT queries executed above for *APOE* variants, or by entering the rsIDs (rs429358, rs7412) into UCSC Genome Browser, we arrive at a convenient graphical view of the sequence around the variants of interest. Clicking “zoom out” buttons or entering expanded genome coordinates allows us to view a larger piece of surrounding genome sequence. By clicking “DNA” at the top of the UCSC Genome Browser window one can obtain the reference genome DNA sequence corresponding to the region shown in the graphical window. The output can be tailored to user needs and saved in different formats for further use. Extended case/color options allow options for highlighting additional information including known SNPs within the retrieved sequence. An alternative approach is to employ the rsIDs to query dbSNP. Individual SNP records in dbSNP have FASTA formatted sequence flanking the SNP in each direction. The flanking sequence given varies in size depending on the length of the sequences in the cluster of dbSNP submissions that supports a particular SNP. The flanking sequence can be copied and pasted for further use, or multiple SNP flanking sequences can be retrieved through a batch submission. Protein sequences are available through NCBI, ExPASy and other sources, and can also be derived from DNA/mRNA sequences.

It is important to note that a DNA sequence encompassing a region with a SNP(s) retrieved from a database may contain major and/or minor alleles. The reference human genome sequence itself includes minor alleles at some positions. Thus, if your goal is to create a pure

representation of the major alleles for a gene or sequence with minor alleles at only one or a few specified positions, then careful attention must be paid to the starting sequence, and comparison against a known “wild type” reference sequence from the literature or Genbank is suggested first, followed by careful text editing to change nucleic acids to minor alleles as desired. A nomenclature code, the IUPAC code[15], is often employed in polymorphic nucleic acid sequences because it allows the ambiguous specification of major and minor alleles in single characters. This code is employed in dbSNP. For instance, although the *APOE* SNPs rs429358 and rs7412 are T>C and C>T changes, respectively, relative to the forward (or plus) strand of the genome, both are represented in the IUPAC code as ‘Y’ which indicates that both ‘C’ and ‘T’ have been observed. Sequences in the IUPAC code format may be required for additional software uses, whereas others may require users to translate IUPAC codes into their representative sequences. Another consideration that needs to be taken into account is the genome strand to which a sequence corresponds. Entries in dbSNP and other databases may not represent sequences and variants relative to the forward strand, and genotyping assays can be designed to target either strand. This can create confusion particularly with A<>T and C<>G transversion SNPs where the strand may be difficult to infer since reverse complemented SNP bases are similar. In such cases comparison of flanking or probe sequences to the reference genome sequence often helps to resolve the SNP alleles relative to their strand. A number of sequence manipulation tools can help format sequences for further uses (e.g., to take reverse complements, translate DNA sequences to protein sequences) and are available as simple web applications from EMBL-EBI under the heading of SRS tools.

Beyond the identification of SNPs and retrieval of accompanying sequences, a number of other common bioinformatics task involve gathering more information on SNPs of interest. Most of these tasks are satisfied through querying the databases discussed elsewhere in this text, and mainly dbSNP. It is important to note that information such as SNP allele frequencies, validation status and functional categorization may vary between builds of dbSNP and thus also

among third party sources. One basic question often posed is: what functional category is a SNP in? This is mainly determined via dbSNP records. SNPs in dbSNP are aligned to contigs and one or more RefSeq mRNA/proteins to determine the functional category of the alleles. This results in categorizations that include “locus-region” (near but not in a coding region), “coding” (in a coding region but the further class is unknown), “utr” (untranslated region), “intron”, “splice site” (in the first 2 or last 2 bases on an intron), “synonymous”, and “nonsynonymous” (“missense”, “nonsense” or “frameshift”). No functional category indicates that a SNP is intergenic relative to the applied contig and RefSeq annotations. Both *APOE* SNPs mentioned here are classified as missense SNPs. Since a gene can have multiple isoforms, and genes themselves can overlap, SNPs can truly fall within multiple functional categorizations. Functional categorizations for lists of SNPs can be retrieved through dbSNP batch retrieval. A rapid alternative is to upload rsIDs to the SNAP tool[12] and select Output to include Gene Annotations from GeneCruiser[16] and select the proxy distance limit as ‘0’. At this time SNAP is limited only to HapMap SNPs. SNAP has additional SNP annotation features including the ability to search for alias rsIDs which can be useful in querying other tools and databases, as well as options to filter and return information on SNP membership on commercial genotyping arrays. Another query mechanism is to use NHLBI’s Genome Variation Server to conduct batch queries to retrieve SNP information.

In deciding whether to pursue a targeted experiment or analysis of a SNP, or undertaking a bioinformatics analysis of multiple SNPs, an important consideration can be the level of confidence in the true existence of the SNPs of interest. SNP databases do contain records for SNPs that are truly artifactual due to sequencing and submitter errors, errors in computational identification, or generally weak evidence; or which are extremely rare or monomorphic in most or all populations currently included in the database. Five categories of validation exist in dbSNP, thus SNPs can have 0-5 types of validation: 1) multiple, independent submissions to refSNPcluster support the SNP, 2) minor alleles have been observed on 2 or more

chromosomes by frequency or genotype data, 3) by submitter confirmation, 4) all alleles have been observed on 2 or more chromosomes each, and 5) by HapMap genotyping. It is important to note SNPs genotyped in HapMap may be monomorphic in all samples typed. In dbSNP Build 129 the *APOE* SNPs rs429358 and rs7412 contain 4 and 3 types of validation, respectively, indicating well validated SNPs. Most importantly the RefSNP cluster records show many independent submissions supporting these variants.

Another common task is to retrieve all known SNPs that meet some criteria within specific genes of interest (e.g., all SNPs with allele frequency information, or all coding SNPs, in *APOE*). A simple approach is to search dbSNP by the gene name – *APOE* -- and click the tab for "Human SNPs". After selecting any SNP in the gene to view its RefSNP cluster report, one can then select under the "GeneView" heading: "in gene region", "cSNP", "has frequency" or "double hit". This generates a sorted list of SNPs in the gene, or gene region, that meet the selected criteria. Using the current version of dbSNP is often the safest strategy for this task since it is up to date and maintained. A number of other tools developed to facilitate candidate gene-SNP studies allow users to retrieve and display SNPs in genes with various filtering criteria including NHLBI's Genome Variation server and SNPLogic[17]. SNPLogic is a flexible tool that allows users to build, save, filter, score and share SNP lists based on a wide range of criteria. Older resources may also meet some users' needs include SNP Function Portal[18], SNPper[19], TAMAL[20], MutDB[21], PicSNP[22] and GeneSNPs from the NIEHS.

Additional software for the conduct and interpretation of genetic analysis studies

The main text summarized many of the most commonly used types of software and popular programs often employed by researchers. Here I provide discussion of some additional software not already mentioned in the main text. Some additional tools for researchers embarking on a genetic study who wish to do power calculations for sample size estimation are

PAWE-3D[23] and CaTS[24]. Targeted genotyping projects may involve assay design by a researcher and there are many programs available to assist with this process including PrimerBatch3 which includes multiple SNP assay types[25], Primer Z[26], SNPbox[27], SNPcutter[28] and SNPicker[29] (as part of SeqVista) for design of restriction enzyme digest-based assays, and a popular general tool Primer3 [30]. For labs handling high volumes of genotyping results a lab information management system (LIMS) may be a desirable informatics capability, but commercial solutions can be expensive. In an attempt to fill this need, a few freely available LIMS have been developed to manage genotype and phenotype data including IGS[31] and T.I.M.S. for TaqMan assays[32].

For those undertaking GWAS analyses an early concern is the careful application of genotyping calling algorithms. The major algorithms and software are largely platform specific (e.g., Affymetrix versus Illumina) and in some cases array-specific. Algorithms for Affymetrix arrays include the chronological progression of RLMM[33], BRLMM (Affy 500K) and Birdseed (Affy 6.0 and unsupported calling for Affy 500K, Affy 5.0) both available from Affymetrix, and CRLMM[34]. The authors of CRLMM find their algorithm equal to or better than the commercially supplied BRLMM and Birdseed algorithms. Birdsuite[35] supports SNP, CNV and CNP calling for the Affy 6.0 array. Additional genotyping software for Affymetrix platforms includes CHIAMO[36] (Affy 500K), AccuTyping[37], and SNiPER-10 [38] (Affy 10K) and SNiPER-HD[39] (Affy500K). Those conducting a DNA pooling approach should apply pooling-specific calling algorithms including GenePool[40] and MPDA[41]. Genotyping algorithms applicable to Illumina arrays include Illuminus[42] and GenoSNP[43].

After genotypes are collected, cleaned and called researchers may want to examine and potentially adjust for population structure and stratification, which can be a source of confounding in association analyses. Implementations of parametric approaches are available in STRUCTURE[44], L-Pop[45], POPGEN and LAMP[46]. Non-parametric approaches have gained favor in recent years, particularly via the application of principal components analysis, and a number

of software options are available including EIGENSTRAT[47], AWclust[48], POPGEN, and PCASNPS[49].

When a final genotyped or imputed set of SNPs is ready the selection of appropriate tools for statistical association is a critical step. A number of popular software packages are reviewed in the main text. Additional academic software packages exist (e.g., SNPTEST[36]) and can be located on a comprehensive website frequently updated (see http://www.nslij-genetics.org/soft/). Commercial packages are also available for genetic analysis and include HelixTree and SASgenetics. Thresholds for test significance in genetic analysis are often determined *a priori* or implemented within statistical software. An R program "p_ACT.R" is available for users who want perform sequential test correction among correlated SNPs[50].

Particularly in the conduct of GWAS there is often more informatics to do after association p-values are generated, including plotting results (e.g., SNP-VISTA[51], GeneWindow[52], GenMAPP[53], and VizStruct[54]). Additional tools may be used to attempt pathway (GSEA-SNP[55], PrioritizerWGA[56], DAVID[57]), ontology and literature (GRAIL[58]) mining of results, or to search for epistasis (SNPHarvester[59], PIA[60]). The use of GRAIL was recently shown to predict an increased proportion of replicated true positive SNPs among GWAS results[58]. Analyses of over-represented pathways among significant GWAS results have also indicated expected and novel disease-specific pathways[61]. Eventually investigators may want to attempt to identify a functional explanation for highly associated polymorphisms, and potentially plan functional molecular experiments. Identifying strongly associated variants and those in LD informs further efforts like re-sequencing, molecular experiments on candidate genes in the region, and the prediction and validation of potential functional variants. The prediction of "functional SNPs" is an active and evolving area of SNP bioinformatics which is discussed in more detail in the next two sections.

SNP function prediction: protein altering variants

Since the earliest understanding of protein sequences it has been recognized that one of the most likely mechanisms of functional variation is a change in protein sequence[62]. Thus, this has been one of the most active areas of development in SNP bioinformatics. Software addressing this question not only takes into account predicted effects on protein structure, but also considers the specific protein cellular environments, evolutionary conservation of protein sequence and structure, the physiochemical properties of the changed amino acids, critical interacting residues and potential post-translational modifications[62]. Software relating to protein altering variants is summarized in Supplemental Table 1. Some of these tools are static and limited to a pre-calculated set of proteins or SNPs at the time of last development. Others allow, or require, users to specify protein sequence, structure, multiple sequence alignments (MSA) and/or SNP information. ExPASy is a reliable source for wild type protein sequences, accessions for PDB (protein data bank) structures and for locating SNP positions in protein sequence. The careful generation of MSAs can be a more complex task requiring users to search for similar sequences across databases and generate and trim alignments. The availability of homologous sequences and/or protein data bank structures can both limit and affect the potential for nSNP bioinformatics analysis. Discussion of the nuances and recommendations on these software can be found in recent reviews[63,64], however in general most software employs one or more of the approaches described above, but few attempt to be comprehensive in their approach. Thus, it is often the experience of users that predictions are not always in step across different software[65], and it is important to remember that predictions may not be robust, and may be parameter driven (e.g., dependent on the information available for a particular protein). Among the available tools, SNPs3D offers one of the most innovative and detailed interfaces to explore results[66].

The best strategy when focusing on a limited number of genes and variants may be to combine information from multiple prediction tools[67], and carefully integrate results with what is known about the protein in question. If the protein in question has been well studied and/or a

crystal structure solved, then researchers may be able to determine through studying the literature, or databases like ExPASy, whether the SNP is positioned to potentially effect an important region of the protein (e.g., a soluble receptor surface with an interaction domain, a residue that participates in a salt bridge, or a site that is phosphorylated in a signaling step). More general physicochemical properties regarding particular types of amino acid changes can also be considered to provide a subjective feeling of the likelihood a variant produces a major change. A useful site has been established with information on this topic (http://www.russell.embl-heidelberg.de/aas/).

Polymorphisms resulting in premature termination codons (PTCs) may be found in the SNP2NMD database[68]. Post-translation modifications which overlap with known SNPs can be viewed in dbPTM[69], PhosphoPOINT[70], or in one a few multi-function prediction tools including F-SNP[71] and SNPeffect/Pupasuite[72]. Bioinformatics tools aimed at predicting functionality across many categories are a popular area of recent development, and usually are constructed to simply integrate information from multiple independent tools and databases. These tools can save on user time and the need for expertise in having to collect sequences and IDs and run analyses through independent software. Some of the tools attempt to provide SNP functionality scores for prioritization, however the scoring systems are generally simplistic and have not been systematically tested for their ability to truly predict functionality. A number of such multi-function tools and their features are described in Supplemental Table 2.

SNP function prediction: regulatory variants

Efforts at predicting and annotating functional regulatory variation have been more diffuse than those at predicting protein level functional variation. Variation can affect gene regulation via a large number of potential mechanisms, and in some cases functional variation is located in *trans*, or in *cis* and quite distant, from the regulated gene. Effects may also be very context-specific (e.g., disease state, tissue, and epigenetic influences) making them more

difficult to isolate and study than protein-coding variants. The ENCODE project has surveyed in a deep manner approximately 1% of the genome leading to new understanding of the complexity of regulation at the transcript level, indicating that the genome is pervasively transcribed[73]. The expansion of such detailed information to more of the genome, including information on highly conserved sequences, and integration with data on genetic variation will improve predictions of potential functional variants for further validation. This expansion of the ENCODE project is underway and new data releases can be tracked at the UCSC Data Coordination Center. The term structural regulatory SNP (srSNP) was introduced to describe functional SNPs which affect the amount or character of RNA produced or translated, including SNPs which affect transcription, splicing, degradation and translational efficiency[74]. Functional surveys of genes in human tissues with allele-specific approaches indicate that srSNPs are common and likely to affect the regulation of many genes[74,75]. A number of groups have reported genome-wide efforts at characterizing SNP associations with gene expression (eSNPs) in different tissues[76-80]. Combining SNPs and next generation expression arrays will lead to enhanced understanding at the genome-wide level of the genetic regulation of splicing and exon level expression[79]. There is increasing awareness that regulatory variants seem to contribute to many of the association signals underlying common disease GWAS[79,81], and thus srSNPs represent an important functional class.

Principally bioinformatics approaches aimed at predicting functional regulatory variation have focused on attempts to identify srSNPs that affect transcription or splicing (e.g., splicing enhancer or repressor domains), largely via characterizing the creation or disruption of consensus binding sequences (tools listed in Supplemental Table 3). There are numerous examples of experimentally validated srSNPs that affect transcription or splicing, but a limitation of bioinformatics predictions in these areas is that they generally predict a high rate of false positives. Thus, proper experimental validation of functional mechanisms is always an important

next step. Use of expanded functional data from later stages of the ENCODE project may prove valuable in this area.

Abnormal splice variants can be very damaging, thus, a number of splice site prediction tools have been developed that are geared toward evaluating *de novo* rare variants based on input sequences. Another area of functional regulatory genetics involves effects on RNA structure. Mutations disrupting tRNA structure or iron response elements (IREs) can lead to disease[82]. A number of examples of common srSNPs affecting mRNA structures with phenotypic consequences have also been reported[83-85], and the importance of mRNA structure in the regulation of transcription, degradation and translational efficiency is becoming more apparent[86], however, high-throughput mechanisms for studying RNA structure *in situ* currently limit attempts at validation of such predictions. An additional area of recent interest is genetic effectors of miRNA regulation in disease[87]. Although few proven examples of variants in miRNAs or their targets are known to affect disease[87], a number of bioinformatics tools have been established to facilitate such studies (see Supplemental Table 3). The expansion of tissue collections, gene expression databases, related bioinformatics tools and databases like ENCODE, and techniques such as allele specific expression approaches[74,75] for assaying regulatory mechanisms is likely to lead to further identification and validation of srSNPs which impinge on human disease.

Additional areas of SNP bioinformatics

A small, but significant number of SNP bioinformatics tools do not fit well into other categories in this review. These tools enable solutions to specific SNP-related problems, and also provide potential ways to gain more from SNP-related data. Many software programs related to SNP discovery and annotation from Sanger sequencing results and from EST databases remain relevant including: PolyPhred[88], PolyBayes[89], SNPdetector[90], ssahaSNP[91], DNannotator[92], InSNP[93], novoSNP[94], SNP-PHAGE[95], SeqDoC[96] and PolyFreq[97]. An area of

recent growth and activity involves software for rapid SNP discovery and annotation from next generation sequencing reads including commercial software like Zoom, and academics efforts like PyroBayes[98], SeqMap[99] and Rolexa[100]. Compression formats and whole genome sequence representations have been suggested with likely implications for SNP representation in the coming wealth of genome sequences[101]. Those with a need to create SNP masked sequences of various types may find SNPmasker useful[102].

Resources like OMIM and Pubmed are often useful in further pursuing biological understanding of SNP associations. Furthermore, a number of SNP-literature search oriented tools exist that may help researchers further mine and explore the potential biological underpinnings of significant SNP associations including GAD and HuGE Navigator[103], as well as GRAIL[58], G2D[104], Prioritizer WGA[56], Prospectr[105], OSIRIS[106], and MarkerInfoFinder[107]. Those interested in intellectual property relating to SNPs can query the patent database with nucleotide sequences related to the SNPs by selecting the patent database (pat) with NCBI BLAST.

References

1. Ionita-Laza I, Lange C, Laird M. Estimating the number of unseen variants in the human genome. Proc Natl Acad Sci. 2009; 106:5008-5013.

2. Horaitis O, Talbot CC, Jr., Phommarinh M, Phillips KM, Cotton RG. A database of locus-specific databases. Nat Genet. 2007; 39:425.

3. Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehvaslaiho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT. Recommendations for locus-specific databases and their curation. Hum Mutat. 2008; 29:2-5.

4. Fokkema IF, den Dunnen JT, Taschner PE. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat. 2005; 26:63-68.

5. Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. UMD (Universal Mutation Database): 2005 update. Hum Mutat. 2005; 26:184-191.

6. den Dunnen JT, Paalman MH. Standardizing mutation nomenclature: why bother? Hum Mutat. 2003; 22:181-182.

7. Ogino S, Gulley ML, den Dunnen JT, Wilson RB. Standard mutation nomenclature in molecular diagnostics: practical and educational challenges. J Mol Diagn. 2007; 9:1-6.

8. Wildeman M, van OE, den Dunnen JT, Taschner PE. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat. 2008; 29:6-13.

9. George RA, Smith TD, Callaghan S, Hardman L, Pierides C, Horaitis O, Wouters MA, Cotton RGH. General mutation databases: analysis and review. J Med Genet. 2008; 45:65-70.

10. Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet. 2008; 45:124-126.

11. Bandelt HJ, Salas A, Taylor RW, Yao YG. Exaggerated status of "novel" and "pathogenic" mtDNA sequence variants due to inadequate database searches. Hum Mutat. 2008; 30:191-196.

12. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008; 24:2938-2939.

13. Wallis SC, Rogne S, Gill L, Markham A, Edge M, Woods D, Williamson R, Humphries S. The isolation of cDNA clones for human apolipoprotein E and the detection of apoE RNA in hepatic and extra-hepatic tissues. EMBO J. 1983; 2:2369-2373.

14. Kent WJ. BLAT--the BLAST-like alignment tool. Genome Res. 2002; 12:656-664.

15. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res. 1985; 13:3021-3030.

16. Liefeld T, Reich M, Gould J, Zhang P, Tamayo P, Mesirov JP. GeneCruiser: a web service for the annotation of microarray data. Bioinformatics. 2005; 21:3681-3682.

17. Pico AR, Smirnov IV, Chang JS, Yeh RF, Wiemels JL, Wiencke JK, Tihan T, Conklin BR, Wrensch M. SNPLogic: an interactive single nucleotide polymorphism selection, annotation, and prioritization system. Nucleic Acids Res. 2009; 37:D803-D809.

18. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics. 2006; 22:e523-e529.

19. Riva A, Kohane IS. A SNP-centric database for the investigation of the human genome. BMC Bioinformatics. 2004; 5:33.

20. Hemminger BM, Saelim B, Sullivan PF. TAMAL: an integrated approach to choosing SNPs for genetic studies of human complex traits. Bioinformatics. 2006; 22:626-627.

21. Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD. MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res. 2008; 36:D815-D819.

22. Chang H, Fujita T. PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. Biochem Biophys Res Commun. 2001; 287:288-291.

23. Gordon D, Haynes C, Blumenfeld J, Finch SJ. PAWE-3D: visualizing power for association with error in case-control genetic studies of complex traits. Bioinformatics. 2005; 21:3935-3937.

24. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006; 38:209-213.

25. You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD. BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics. 2008; 9:253.

26. Tsai MF, Lin YJ, Cheng YC, Lee KH, Huang CC, Chen YT, Yao A. PrimerZ: streamlined primer design for promoters, exons and human SNPs. Nucleic Acids Res. 2007; 35:W63-W65.

27. Weckx S, De RP, Van BC, Del-Favero J. SNPbox: a modular software package for large-scale primer design. Bioinformatics. 2005; 21:385-387.

28. Zhang R, Zhu Z, Zhu H, Nguyen T, Yao F, Xia K, Liang D, Liu C. SNP Cutter: a comprehensive tool for SNP PCR-RFLP assay design. Nucleic Acids Res. 2005; 33:W489-W492.

29. Niu T, Hu Z. SNPicker: a graphical tool for primer picking in designing mutagenic endonuclease restriction assays. Bioinformatics. 2004; 20:3263-3265.

30. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol. 2000; 132:365-386.

31. Fiddy S, Cattermole D, Xie D, Duan XY, Mott R. An integrated system for genetic analysis. BMC Bioinformatics. 2006; 7:210.

32. Monnier S, Cox DG, Albion T, Canzian F. T.I.M.S: TaqMan Information Management System, tools to organize data flow in a genotyping laboratory. BMC Bioinformatics. 2005; 6:246.

33. Rabbee N, Speed TP. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics. 2006; 22:7-12.

34. Lin S, Carvalho B, Cutler DJ, Arking DE, Chakravarti A, Irizarry RA. Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays. Genome Biol. 2008; 9:R63.

35. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008; 40:1253-1260.

36. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007; 447:661-678.

37. Hu G, Wang HY, Greenawalt DM, Azaro MA, Luo M, Tereshchenko IV, Cui X, Yang Q, Gao R, Shen L, Li H. AccuTyping: new algorithms for automated analysis of data from high-throughput genotyping with oligonucleotide microarrays. Nucleic Acids Res. 2006; 34:e116.

38. Huentelman MJ, Craig DW, Shieh AD, Corneveaux JJ, Hu-Lince D, Pearson JV, Stephan DA. SNiPer: improved SNP genotype calling for Affymetrix 10K GeneChip microarray data. BMC Genomics. 2005; 6:149.

39. Hua J, Craig DW, Brun M, Webster J, Zisman V, Tembe W, Joshipura K, Huentelman MJ, Dougherty ER, Stephan DA. SNiPer-HD: improved genotype calling accuracy by an expectation-maximization algorithm for high-density SNP arrays. Bioinformatics. 2007; 23:57-63.

40. Pearson JV, Huentelman MJ, Halperin RF, Tembe WD, Melquist S, Homer N, Brun M, Szelinger S, Coon KD, Zismann VL, Webster JA, Beach T, Sando SB, Aasly JO, Heun R, Jessen F, Kolsch H, Tsolaki M, Daniilidou M, Reiman EM, Papassotiropoulos A, Hutton ML, Stephan DA, Craig DW. Identification of the genetic basis for complex disorders by use of pooling-based genomewide single-nucleotide-polymorphism association studies. Am J Hum Genet. 2007; 80:126-139.

41. Yang HC, Huang MC, Li LH, Lin CH, Yu AL, Diccianni MB, Wu JY, Chen YT, Fann CS. MPDA: microarray pooled DNA analyzer. BMC Bioinformatics. 2008; 9:196.

42. Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatowski DP, Clark TG. A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics. 2007; 23:2741-2746.

43. Giannoulatou E, Yau C, Colella S, Ragoussis J, Holmes CC. GenoSNP: a variational Bayes within-sample SNP genotyping algorithm that does not require a reference population. Bioinformatics. 2008; 24:2209-2214.

44. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567-1587.

45. Purcell S, Sham P. Properties of structured association approaches to detecting population stratification. Hum Hered. 2004; 58:93-107.

46. Sankararaman S, Sridhar S, Kimmel G, Halperin E. Estimating local ancestry in admixed populations. Am J Hum Genet. 2008; 82:290-303.

47. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904-909.

48. Gao X, Starmer JD. AWclust: point-and-click software for non-parametric population structure analysis. BMC Bioinformatics. 2008; 9:77.

49. Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, Drineas P. PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet. 2007; 3:1672-1686.

50. Conneely KN, Boehnke M. So Many Correlated Tests, So Little Time! Rapid Adjustment of P Values for Multiple Correlated Tests. Am J Hum Genet. 2007; 81.

51. Shah N, Teplitsky MV, Minovitsky S, Pennacchio LA, Hugenholtz P, Hamann B, Dubchak IL. SNP-VISTA: an interactive SNP visualization tool. BMC Bioinformatics. 2005; 6:292.

52. Staats B, Qi L, Beerman M, Sicotte H, Burdett LA, Packer B, Chanock SJ, Yeager M. Genewindow: an interactive tool for visualization of genomic variation. Nat Genet. 2005; 37:109-110.

53. Salomonis N, Hanspers K, Zambon AC, Vranizan K, Lawlor SC, Dahlquist KD, Doniger SW, Stuart J, Conklin BR, Pico AR. GenMAPP 2: new features and resources for pathway analysis. BMC Bioinformatics. 2007; 8:217.

54. Bhasi K, Zhang L, Brazeau D, Zhang A, Ramanathan M. Information-theoretic identification of predictive SNPs and supervised visualization of genome-wide association studies. Nucleic Acids Res. 2006; 34:e101.

55. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics. 2008; 24:2784-2785.

56. Franke L, van BH, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet. 2006; 78:1011-1025.

57. Huang da W, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, Guo Y, Stephens R, Baseler MW, Lane HC, Lempicki RA. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007; 35:W169-W175.

58. Raychaudhari S, Plenge RM, Rossin EJ, Ng ACY, International Schizophrenia Consortium, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ. Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletion. PLOS Genetics. 2009; 5:e1000534.

59. Yang C, He Z, Wan X, Yang Q, Xue H, Yu W. SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. Bioinformatics. 2009; 25:504-511.

60. Mechanic LE, Luke BT, Goodman JE, Chanock SJ, Harris CC. Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. BMC Bioinformatics. 2008; 9:146.

61. Johnson AD, O'Donnell CJ. An open access database of genome-wide association results. BMC Medical Genetics. 2009; 10:6.

62. Betts MJ, Russell RB. Amino-acid properties and consequences of substitutions. In: Barnes MR, ed. Bioinformatics for Geneticists. 2nd ed. West Sussex, UK: John Wiley & Sons Ltd; 2007:311-342.

63. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet. 2006; 7:61-80.

64. Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat. 2008; 29:1327-36.

65. Burke DF, Worth CL, Priego EM, Cheng T, Smink LJ, Todd JA, Blundell TL. Genome bioinformatic analysis of nonsynonymous SNPs. BMC Bioinformatics. 2007; 8:301.

66. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7:166.

67. Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat. 2007; 28:683-693.

68. Han A, Kim WY, Park SM. SNP2NMD: a database of human single nucleotide polymorphisms causing nonsense-mediated mRNA decay. Bioinformatics. 2007; 23:397-399.

69. Lee TY, Huang HD, Hung JH, Huang HY, Yang YS, Wang TH. dbPTM: an information repository of protein post-translational modification. Nucleic Acids Res. 2006; 34:D622-D627.

70. Yang CY, Chang CH, Yu YL, Lin TC, Lee SA, Yen CC, Yang JM, Lai JM, Hong YR, Tseng TL, Chao KM, Huang CY. PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. Bioinformatics. 2008; 24:i14-i20.

71. Lee PH, Shatkay H. F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res. 2008; 36:D820-D824.

72. Reumers J, Conde L, Medina I, Maurer-Stroh S, Van Durme J, Dopazo J, Rousseau F, Schymkowitz J. Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. Nucleic Acids Res. 2008; 36:D825-D829.

73. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the genome by the ENCODE pilot project. Nature. 2007; 447:799-816.

74. Johnson AD, Zhang Y, Papp AC, Pinsonneault JK, Lim JE, Saffen D, Dai Z, Wang D, Sadee W. Polymorphisms affecting gene transcription and mRNA processing in pharmacogenetic candidate genes: detection through allelic expression imbalance in human target tissues. Pharmacogenet Genomics. 2008; 18:781-791.

75. Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KC, Koka V, Dias J, Gurd S, Martin NW, Mallmin H, Nilsson O, Harmsen E, Dewar K, Kwan T, Pastinen T. Targeted screening of cis-regulatory variation in human haplotypes. Genome Res. 2009; 19:118-127.

76. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WO. A genome-wide association study of global gene expression. Nat Genet. 2007; 39:1202-1207.

77. Myers AJ, Gibbs JR, Webster JA, Rohrer K, Zhao A, Marlowe L, Kaleem M, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, Holmans P, Heward CB, Reiman EM, Stephan D, Hardy J. A survey of genetic human cortical gene expression. Nat Genet. 2007; 39:1494-1499.

78. Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D, Montgomery S, Tavare S, Deloukas P, Dermitzakis ET. Population genomics of human gene expression. Nat Genet. 2007; 39:1217-1224.

79. Heinzen EL, Ge D, Cronin KD, Maia JM, Shianna KV, Gabriel WN, Welsh-Bohmer KA, Hulette CM, Denny TN, Goldstein DB. Tissue-specific genetic control of splicing: implications for the study of complex traits. PLoS Biol. 2008; 6:e1.

80. Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarkis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R. Mapping the genetic architecture of gene expression in human liver. PLoS Biol. 2008; 6:e107.

81. Chen R, Morgan AA, Dudley J, Deshpande T, Li L, Kodama K, Chiang AP, Butte AJ. FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. Genome Biol. 2008; 9:R170.

82. Wittenhagen LM, Kelley SO. Impact of disease-related mitochondrial mutations on tRNA structure and function. Trends Biochem Sci. 2003; 28:605-611.

83. Wang D, Johnson AD, Papp AC, Kroetz DL, Sadee W. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. Pharmacogenet Genomics. 2005; 15:693-704.

84. Zhang Y, Wang D, Johnson AD, Papp AC, Sadee W. Allelic expression imbalance of human mu opioid receptor (OPRM1) caused by variant A118G. J Biol Chem. 2005; 280:32618-32624.

85. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science. 2006; 314:1930-1933.

86. Shabalina SA, Ogurtsov AY, Spiridonov NA. A periodic pattern of mRNA secondary structure created by the genetic code. Nucleic Acids Res. 2006; 34:2428-2437.

87. Sethupathy P, Collins FS. MicroRNA target site polymorphisms and human disease. Trends Genet. 2008; 24:489-497.

88. Bhangale TR, Stephens M, Nickerson DA. Automating resequencing-based detection of insertion-deletion polymorphisms. Nat Genet. 2006; 38:1457-1462.

89. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitziel NO, Hillier L, Kwok PY, Gish WR. A general approach to single-nucleotide polymorphism discovery. Nat Genet. 1999; 23:452-456.

90. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, Rowe W, Liu PP, Gibbs RA, Buetow KH. SNPdetector: a software tool for sensitive and accurate SNP detection. PLoS Comput Biol. 2005; 1:e53.

91. Ning Z, Cox AJ, Mullikin JC. SSAHA: a fast search method for large DNA databases. Genome Res. 2001; 11:1725-1729.

92. Liu C, Bonner TI, Nguyen T, Lyons JL, Christian SL, Gershon ES. DNannotator: Annotation software tool kit for regional genomic sequences. Nucleic Acids Res. 2003; 31:3729-3735.

93. Manaster C, Zheng W, Teuber M, Wachter S, Doring F, Schreiber S, Hampe J. InSNP: a tool for automated detection and visualization of SNPs and InDels. Hum Mutat. 2005; 26:11-19.

94. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, De Johnghe P, Van Broeckhoven C, De Rijk P. novoSNP, a novel computational tool for sequence variation discovery. Genome Res. 2005; 15:436-442.

95. Matukumalli LK, Grefenstette JJ, Hyten DL, Choi IY, Cregan PB, Van Tassell CP. SNP-PHAGE--High throughput SNP discovery pipeline. BMC Bioinformatics. 2006; 7:468.

96. Crowe ML. SeqDoC: rapid SNP and mutation detection by direct comparison of DNA sequence chromatograms. BMC Bioinformatics. 2005; 6:133.

97. Wang J, Huang X. A method for finding single-nucleotide polymorphisms with allele frequencies in sequences of deep coverage. BMC Bioinformatics. 2005; 6:220.

98. Quinlan AR, Steward DA, Stromberg MP, Marth GT. Pyrobayes: an improved base caller for SNP discovery in pyrosequences. Nat Methods. 2008; 5:179-181.

99. Jiang H, Wong WH. SeqMap: mapping massive amount of oligonucleotides to the genome. Bioinformatics. 2008; 24:2395-2396.

100. Rougemont J, Amzallag A, Iseli C, Farinelli L, Xenarios I, Naef F. Probabilistic base calling of Solexa sequencing data. BMC Bioinformatics. 2008; 9:431.

101. Christley S, Lu Y, Li C, Xie X. Human genomes as email attachments. Bioinformatics. 2009; 25:274-275.

102. Andreson R, Puurand T, Remm M. SNPmasker: automatic masking of SNPs and repeats across eukaryotic genomes. Nucleic Acids Res. 2006; 34:W651-W655.

103. Yu W, Gwinn M, Clyne M, Yesupriya A, Khoury MJ. A navigator for human genome epidemiology. Nat Genet. 2008; 40:124-125.

104. Perez-Iratxeta C, Bork P, Andrade-Navarro MA. Update of the G2D tool for prioritization of gene candidates to inherited diseases. Nucleic Acids Res. 2007; 35:W212-W216.

105. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS. Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics. 2005; 6:55.

106. Furlong LI, Dach H, Hofmann-Apitius M, Sanz F. OSIRISv1.2: a named entity recognition system for sequence variants of genes in biomedical literature. BMC Bioinformatics. 2008; 9:84.

107. Xuan W, Wang P, Watson SJ, Meng F. Medline search engine for finding genetic markers with biological significance. Bioinformatics. 2007; 23:2477-2484.

108. Tavtigian SV, Byrnes GB, Goldgar DE, Thomas A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum Mutat. 2008; 29:1342-1354.

109. Masso M, Vaisman II. Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics. 2008; 24:2002-2009.

110. Cheng TM, Lu YE, Vendruscolo M, Lio' P, Blundell TL. Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comput Biol. 2008; 4:e1000135.

111. Kaminker JS, Zhang Y, Watanabe C, Zhang Z. CanPredict: a computational tool for predicting cancer-associated missense mutations. Nucleic Acids Res. 2007; 35:W595-W598.

112. Kono H, Yuasa T, Nishiue S, Yura K. coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res. 2008; 36:D409-D413.

113. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res. 2006; 34:W239-W242.

114. Guex N, Diemand A, Peitsch MC. Protein modeling for all. Trends Biochem Sci. 1999; 364-367.

115. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005; 33:W382-W388.

116. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. 2005; 33:W306-W310.

117. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005; 21:2814-2820.

118. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005; 15:978-986.

119. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins. 2006; 62:1125-1132.

120. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res. 2005; 33:W480-W482.

121. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics. 2007; 8:450.

122. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics. 2006; 22:2729-2734.

123. Mi H, Guo N, Kejariwal A, Thomas PD. PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res. 2007; 35:D247-D252.

124. Jegga AG, Gowrisankar S, Chen J, Aronow BJ. PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. Nucleic Acids Res. 2007; 35:D700-D706.

125. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002; 30:3894-3900.

126. Kwasigroch JM, Gilis D, Dehouck Y, Rooman M. PoPMuSiC, rationally designing point mutations in protein structures. Bioinformatics. 2002; 18:1701-1702.

127. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de lC, X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005; 21:3176-3178.

128. Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L. Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics. 2007; 23:1444-1450.

129. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31:3812-3814.

130. Han A, Kang HJ, Cho Y, Lee S, Kim YJ, Gong S. SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences. Nucleic Acids Res. 2006; 34:W642-W644.

131. Uzun A, Leslin CM, Abyzov A, Ilyin V. Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. Nucleic Acids Res. 2007; 35:W384-W392.

132. Liu CK, Chen YH, Tang CY, Chang SC, Lin YJ, Tsai MF, Chen YT, Yao A. Functional analysis of novel SNPs and mutations in human and mouse genomes. BMC Bioinformatics. 2008; 9 Suppl 12:S10.

133. Yuan HY, Chiou JJ, Tseng WH, Liu CH, Liu CK, Lin YJ, Wang HH, Yao A, Chen YT, Hsu CN. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res. 2006; 34:W635-W641.

134. Dantzer J, Moad C, Heiland R, Mooney S. MutDB services: interactive structural analysis of mutation data. Nucleic Acids Res. 2005; 33:W311-W314.

135. Freimuth RR, Stormo GD, McLeod HL. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. Hum Mutat. 2005; 25:110-117.

136. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, Rousseau F, Schymkowitz J, Dopazo J. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res. 2006; 34:W621-W625.

137. Li S, Ma L, Li H, Vang S, Hu Y, Bolund L, Wang J. Snap: an integrated SNP annotation platform. Nucleic Acids Res. 2007; 35:D707-D710.

138. Wang P, Dai M, Xuan W, McEachin RC, Jackson AU, Scott LJ, Athey B, Watson SJ, Meng F. SNP Function Portal: a web database for exploring the function implication of SNP alleles. Bioinformatics. 2006; 22:e523-e529.

139. Xu H, Gregory SG, Hauser ER, Stenger JE, Pericak-Vance MA, Vance JM, Zuchner S, Hauser MA. SNPselector: a web tool for selecting SNPs for genetic association studies. Bioinformatics. 2005; 21:4181-4186.

140. Nalla VK, Rogan PK. Automated splicing mutation analysis by information theory. Hum Mutat. 2005; 25:334-342.

141. Georges M, Coppieters W, Charlier C. Polymorphic miRNA-mediated gene regulation: contribution to phenotypic variation and disease. Curr Opin Genet Dev. 2007; 17:166-176.

142. Bao L, Zhou M, Wu L, Lu L, Goldowitz D, Williams RW, Cui Y. PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res. 2007; 35:D51-D54.

143. Zhao T, Chang LW, McLeod HL, Stormo GD. PromoLign: a database for upstream region analysis and SNPs. Hum Mutat. 2004; 23:534-539.

144. Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics. 2008; 9 Suppl 1:S2.

Supplemental Table 1.

nSNP tool	Description	SNP input(s)	URL
Align GV-GD[108]	Grantham differences, MSA, classifiers	MSA	http://agvgd.iarc.fr/index.php
AUTO-MUTE[109]	Multiple classifiers, stability changes	PDB	http://proteins.gmu.edu/automute/
Bongo[110]	Residue interaction analysis, classification	PDB	http://www.bongo.cl.cam.ac.uk/Bongo/
CanPredict[111]	Cancer mutation classification	seq	http://www.cgl.ucsf.edu/Research/genentech/canpredict/
Cn3D	Visualization of MSA, structures, SNPs	129 (via dbSNP)	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
coliSNP[112]	Visualization of SNPs, solvent accessibility	129 (PDB)	http://yayoi.kansai.jaea.go.jp/colisnp/
CUPSAT[113]	Protein stability changes	PDB	http://cupsat.tu-bs.de/
Swiss-PdbViewer[114]	Visualization of structures, mutation effects	PDB	http://spdbv.vital-it.ch/
Foldx[115]	Protein stability changes	PDB	http://foldx.crg.es/foldx.jsp
I-mutant[116]	Protein stability changes	seq, PDB	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi
LS-SNP[117]	Sequence, structure, MSA, SVM classifier	126	http://modbase.compbio.ucsf.edu/LS-SNP/
MAPP[118]	MSA, physicochemical, classifier	MSA	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html
Mupro[119]	Multiple attribute SVM classification	seq, PDB	http://www.ics.uci.edu/~baldig/mutation.html
nsSNPAnalyzer[120]	Generates MSA, classifier, accessibility, SIFT, environment, 2^o structure	seq	http://snpanalyzer.utmem.edu/
Parepro[121]	Multiple attribute SVM classification	MSA	http://www.mobioinfor.cn/parepro/index.htm
PhD-SNP[122]	Multiple attribute SVM classification	seq, PDB	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
PANTHER[123]	MSA classification	seq	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PolyDoms[124]	SIFT, PolyPhen, LS-SNP, OMIM, pathways, Reactome interactions, visualization (PDB)	125	http://polydoms.cchmc.org/polydoms/
PolyPhen[125]	Generates MSA, multiple attribute classifier	126 (or seq)	http://genetics.bwh.harvard.edu/pph/
PoPMuSiC[126]	Protein stability changes	PDB	http://babylone.ulb.ac.be/popmusic/
PMUT[127]	Multiple attribute neural network classifier	MSA	http://mmb2.pcb.ub.es:8080/PMut/
SAPRED[128]	Multiple attribute SVM classification	seq, PDB	http://sapred.cbi.pku.edu.cn/
SIFT[129]	MSA, physicochemical classification	129 (or seq)	http://sift.jcvi.org/
SNP@domain[130]	SNPs in Pfam/SCOP protein domains	123	http://variome.kobic.re.kr/SnpNavigator/
SNPs3D[66]	Multiple attribute SVM classification	128	http://www.snps3d.org/
StructureSNP[131]	Visualization of MSA, structures, SNPs	125	http://glinka.bio.neu.edu/StSNP/

Supplemental Table 2.

Multi-function tool	Description	dbSNPbuild	URL
F-SNP (v. 1.0)[71]	nSNP (PolyPhen, SIFT, SNPeffect, LS-SNP, SNPs3D), splicing, TFBS, microRNA, conservation, post-translational modifications, (559,322 SNPs analyzed)	126	http://compbio.cs.queensu.ca/F-SNP/
FANS[132]	similar to VisualSNP but allows *de novo* SNP analysis, multi-species	n/a	http://fans.ngc.sinica.edu.tw/fans/
FastSNP[133]	nSNP (PolyPhen), TFBS, splicing	128	http://fastsnp.ibms.sinica.edu.tw/
mutDB[134]	nSNP (SIFT), Swissprot and literature annotation of nSNPs	126	http://mutdb.org/cgi-bin/mutdb.pl
PolyMAPr[135]	nSNP (PolyPhen), TFBS, splicing, *de novo*	n/a	requires download (author contact)
PupaSuite (v. 2.0.0)[136]	nSNP categorization, TFBS, microRNA, Promoter flexibility, conservation, splicing, multi-species, *de novo* SNP analysis (4,965,073 SNPs analyzed)	126	http://bioinfo.cipf.es/pupasuite/www/
Snap[137]	visualization of SNPs with overlapping annotations (protein domains, splicing, post-translation modification sites)	128	http://snap.humgen.au.dk/
SNPeffect (v. 3.0)[72]	nSNP affects on stability, folding, dynamics, aggregation, interaction sites, post-translation modifications, turn-over, cellular localization (133,505 SNPs analyzed)	125	http://snpeffect.vib.be/
SNP Function Portal[138]	nSNP (LS-SNP), TFBS, conservation, splice site, CpG island, methylation, microRNA, gene ontology, literature, Tajima's D, repeat mask/info	126	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx
SNPseek	TFBS, conservation, splicing	126	http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi
SNPselector[139]	conservation, regulatory potential, repeat info	126	http://snpselector.duhs.duke.edu/
VisualSNP (v. 2.4)	only SNPs in exons or splice sites are analyzed, nSNP (SIFT), splicing	126	http://genepipe.ngc.sinica.edu.tw/visualsnp/

Supplemental Table 3.

Regulatory function tool	Description	dbSNPbuild	URL
Automated Splice Site Analyses[140]	Multiple analyses of mutation effects on splicing	121	https://splice.uwo.ca/
Human Splicing Finder	Multiple analyses of mutation effects on splicing	n/a	http://www.umd.be/HSF/
mRNA bySNP browser[76]	Database of gene expression associated SNPs	limited to 410K Affymetrix SNPs	http://www.sph.umich.edu/csg/liang/asthma/
Patrocles[141]	SNPs in miRNAs, 3'UTR miRNA targets, genes in involved in miRNA processing, multi-species	128 (in beta release)	http://www.patrocles.org/
PolymiRTS[142]	SNPs in 3'UTR of predicted miRNA targets (using TargetScan), multi-species	126	http://compbio.utmem.edu/miRSNP/home.php
PromoLign[143]	Multi-species alignment of promoters, TFBS, SNPs (6,471 genes, 80,436 SNPs)	118	http://polly.wustl.edu/promolign/main.html
SNPExpress[79]	Database of gene expression associated SNPs (93 cortex samples, 80 PBMC samples)	limited to Illumina 550K SNPs	http://people.genome.duke.edu/~dg48/ SNPExpress/ (available for download)
SNP@promoter[144]	SNPs in predicted TFBS (Transfac 7.0)	126	http://variome.kobic.re.kr/SNPatPromoter/

See also most of the multi-function tools in Supplemental Table 2 which include regulatory prediction

Supplemental Table 4. SNP bioinformatics related tools and databases in order as mentioned.	
Overview of SNP bioinformatics.	
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
SNAP	http://www.broad.mit.edu/mpg/snap/
HUGO/HGNC	http://genenames.org/
SNP databases.	
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
HapMap	http://www.hapmap.org/
1000 Genomes Project	http://www.1000genomes.org
JSNP	http://snp.ims.u-tokyo.ac.jp/
ThaiSNP database	http://www4a.biotec.or.th/thaisnp/db
Taiwan-Han Chinese SNP	http://genepipe.ngc.sinica.edu.tw/thcsd/
SNP@ethnos	http://variome.kobic.re.kr/SNPatETHNIC/
CEPH genotype database	http://www.cephb.fr/en/cephdb/
ALFRED	http://alfred.med.yale.edu/
BioMart	http://www.biomart.org/
SPSmart	http://spsmart.cesga.es/
GVS	http://gvs-p.gs.washington.edu/GVS/
UCSC Table Browser	http://genome.ucsc.edu/cgi-bin/hgTables
HGVS	http://www.hgvs.org/
The Waystation	http://www.centralmutations.org/
LOVD	http://www.lovd.nl/2.0/
UMD	http://www.umd.be/
Mutalzyer	http://www.humgen.nl/mutalyzer/1.0.1/
OMIM	http://www.ncbi.nlm.nih.gov/sites/entrez?db=omim
HGMD	http://www.hgmd.cf.ac.uk/ac/index.php
MitoMap	http://www.mitomap.org/
COSMIC	http://www.sanger.ac.uk/genetics/CGP/cosmic/
GAC	http://www.niehs.nih.gov/research/resources/databases/gac/index.cfm
TP53 database	http://www-p53.iarc.fr/
GeneTests	http://www.genetests.org/
EuroGenTest	http://www.eurogentest.org/
Swissprot/ExPASy	http://ca.expasy.org/
UCSC BLAT tool	http://genome.ucsc.edu/cgi-bin/hgBlat
SNAP	http://www.broad.mit.edu/mpg/snap/
Gen2Phen	http://www.gen2phen.org/
HuGE Navigator	http://www.hugenavigator.net/
GAD	http://geneticassociationdb.nih.gov/
dbGAP	http://www.ncbi.nlm.nih.gov/sites/entrez?Db=gap

NHGRI Catalog of GWAS	http://www.genome.gov/gwastudies/
HGVbaseG2P	http://www.hgvbaseg2p.org/index
Open Access GWAS db	http://www.biomedcentral.com/1471-2350/10/6
SNPedia	http://www.snpedia.com/
Basic SNP bioinformatics tasks and strategies.	
dbSNP	http://www.ncbi.nlm.nih.gov/projects/SNP/
UCSC LiftOver tool	http://genome.ucsc.edu/cgi-bin/hgLiftOver
Swissprot/ExPASy	http://ca.expasy.org/
snpBLAST tool	http://www.ncbi.nlm.nih.gov/SNP/snp_blastByOrg.cgi
UCSC BLAT tool	http://genome.ucsc.edu/cgi-bin/hgBlat
UCSC genome browser	http://genome.ucsc.edu/
Genbank	http://www.ncbi.nlm.nih.gov/Genbank/
EMBL-EBI SRS tools	http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+applSelect
SNAP	http://www.broad.mit.edu/mpg/snap/
GeneCruiser	http://genecruiser.broad.mit.edu/genecruiser3/
GVS	http://gvs-p.gs.washington.edu/GVS/
SNPLogic	http://www.snplogic.org/
SNP Function Portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx
SNPper	http://snpper.chip.org/
TAMAL	http://neoref.ils.unc.edu/tamal/
MutDB	http://mutdb.org/
PicSNP	http://plaza.umin.ac.jp/~hchang/picsnp/
GeneSNPs	http://www.genome.utah.edu/genesnps/
Software for the conduct and interpretation of genetic analysis studies.	
Web collection of software	http://www.nslij-genetics.org/soft/
haplo.stats_R	http://mayoresearch.mayo.edu/mayo/research/schaid_lab/software.cfm
Genetic Power Calculator	http://pngu.mgh.harvard.edu/~purcell/gpc/
PAWE-3D	http://linkage.rockefeller.edu/pawe3d/
CaTS	http://www.sph.umich.edu/csg/abecasis/CaTS/
PIAGE	http://www.dkfz.de/en/klepidemiologie/software/software.html
BatchPrimer3	http://probes.pw.usda.gov/batchprimer3/
Primer Z	http://genepipe.ngc.sinica.edu.tw/primerz/
SNPbox	http://www.snpbox.org/
SNP Cutter	http://bioinfo.bsd.uchicago.edu/SNP_cutter.htm
SNPicker (part of SeqVISTA)	http://zlab.bu.edu/SeqVISTA/
Primer3	http://primer3.sourceforge.net/
T.I.M.S.	http://www.bioinformatics.org/macroshack/prog_list.html
RLMM	via Bioconductor or the authors

BRLMM	commercial
Birdseed	commercial
CRLMM	via Bioconductor or the authors
BirdSuite	http://www.broad.mit.edu/mpg/birdsuite/
CHIAMO	http://www.stats.ox.ac.uk/%7Emarchini/software/gwas/chiamo.html
AccuTyping	http://www2.umdnj.edu/lilabweb/Publications/AccuTyping.html
SNiPER-10	http://www.tgen.org/neurogenomics/data
SNiPER-HD	http://www.tgen.org/neurogenomics/data
GENEPOOL	http://genepool.tgen.org/
MPDA	http://www.stat.sinica.edu.tw/hsinchou/genetics/pooledDNA/mpda.htm
Illuminus	http://www.well.ox.ac.uk/~tgc/illuminus_documentation.htm
GenoSNP	http://www.stats.ox.ac.uk/~giannoul/GenoSNP/
IGG	http://bioinfo.hku.hk:13080/iggweb/home.htm
SNAP	http://www.broad.mit.edu/mpg/snap/
SNP-HWE	http://www.sph.umich.edu/csg/abecasis/Exact/
STRUCTURE	http://pritch.bsd.uchicago.edu/structure.html
L-Pop	http://pngu.mgh.harvard.edu/~purcell/lpop/
POPGEN	http://www.stats.ox.ac.uk/%7Emarchini/software.html#popgen
LAMP	http://lamp.icsi.berkeley.edu/lamp/
EIGENSTRAT	http://genepath.med.harvard.edu/~reich/Software.htm
AWClust	http://awclust.sourceforge.net/
PCASNPS	http://www.cs.rpi.edu/~drinep/PCASNPS/
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/
MACH	http://www.sph.umich.edu/csg/abecasis/MaCH/
IMPUTE	http://www.stats.ox.ac.uk/~marchini/software/gwas/impute.html
BEAGLE	http://www.stat.auckland.ac.nz/~browning/beagle/beagle.html
BimBam	http://stephenslab.uchicago.edu/software.html
TUNA	http://www.stat.uchicago.edu/~wen/tuna/
Bioconductor	http://www.bioconductor.org/
Mendel	http://www.genetics.ucla.edu/software/mendel
MERLIN	http://www.sph.umich.edu/csg/abecasis/Merlin/
SNPTEST	http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html
Genomizer	http://www.ikmb.uni-kiel.de/genomizer/
GHOST	http://www.sph.umich.edu/csg/chen/ghost/
GenAbel	http://mga.bionet.nsc.ru/nlru/GenABEL/
ProbAbel	http://mga.bionet.nsc.ru/~yurii/ABEL/
FBAT	http://www.biostat.harvard.edu/~fbat/fbat.htm
QTDT	http://www.sph.umich.edu/csg/abecasis/QTDT/
PowerMarker	http://statgen.ncsu.edu/powermarker/
FamHap	http://famhap.meb.uni-bonn.de/

HelixTree	commercial
SASgenetics	commercial
p_ACT	http://csg.sph.umich.edu/boehnke/p_act.php
METAL	http://www.sph.umich.edu/csg/abecasis/Metal/
WGAViewer	http://people.genome.duke.edu/~dg48/WGAViewer/
SNP-VISTA	http://genome.lbl.gov/vista/snpvista/
GeneWindow	http://genewindow.nci.nih.gov/
GenMAPP	http://www.genmapp.org/
VizStruct	http://www.cse.buffalo.edu/DBGROUP/bioinformatics/supplementary/vizstruct/vizstruct.html
GWAS GUI	http://www.sph.umich.edu/csg/weich/browser/
AssociationViewer	http://associationviewer.vital-it.ch/
SNAP	http://www.broad.mit.edu/mpg/snap/
Haploview	http://www.broad.mit.edu/mpg/haploview/
GSEA-SNP	http://www.nr.no/pages/samba/area_emr_smbi_gseasnp
PrioritizerWGA	http://bio-informatics.gr/content/view/24/39/
DAVID	http://david.abcc.ncifcrf.gov/
GRAIL	http://www.broad.mit.edu/mpg/grail/
SNPHarvester	http://bioinformatics.ust.hk/SNPHarvester.html
PIA	http://www3.cancer.gov/intra/lhc/PIA2-distribution.zip
SNP function prediction: protein altering variants	
Align GV-GD	http://agvgd.iarc.fr/index.php
AUTO-MUTE	http://proteins.gmu.edu/automute/
Bongo	http://www.bongo.cl.cam.ac.uk/Bongo/
CanPredict	http://www.cgl.ucsf.edu/Research/genentech/canpredict/
Cn3D	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
coliSNP	http://yayoi.kansai.jaea.go.jp/colisnp/
CUPSAT	http://cupsat.tu-bs.de/
Swiss-PdbViewer	http://spdbv.vital-it.ch/
Foldx	http://foldx.crg.es/foldx.jsp
I-mutant	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant2.0/I-Mutant2.0.cgi
LS-SNP	http://modbase.compbio.ucsf.edu/LS-SNP/
MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html
Mupro	http://www.ics.uci.edu/~baldig/mutation.html
nsSNPAnalyzer	http://snpanalyzer.utmem.edu/
Parepro	http://www.mobioinfor.cn/parepro/index.htm
PhD-SNP	http://gpcr2.biocomp.unibo.it/cgi/predictors/PhD-SNP/PhD-SNP.cgi
PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp
PolyDoms	http://polydoms.cchmc.org/polydoms/
PolyPhen	http://genetics.bwh.harvard.edu/pph/

PoPMuSiC	http://babylone.ulb.ac.be/popmusic/
PMUT	http://mmb2.pcb.ub.es:8080/PMut/
SAPRED	http://sapred.cbi.pku.edu.cn/
SIFT	http://sift.jcvi.org/
SNP@domain	http://variome.kobic.re.kr/SnpNavigator/
SNPs3D	http://www.snps3d.org/
StructureSNP	http://glinka.bio.neu.edu/StSNP/
Swissprot/ExPASy	http://ca.expasy.org/
PDB	http://www.wwpdb.org/
Amino acid change	http://www.russell.embl-heidelberg.de/aas/
SNP2NMD	http://variome.kobic.re.kr/SNP2NMD/
dbPTM	http://dbptm.mbc.nctu.edu.tw/
PhosphoPOINT	http://kinase.bioinformatics.tw/
F-SNP	http://compbio.cs.queensu.ca/F-SNP/
SNPeffect	http://snpeffect.vib.be/
Pupasuite	http://bioinfo.cipf.es/pupasuite/www/
FANS	http://fans.ngc.sinica.edu.tw/fans/input.do
FastSNP	http://fastsnp.ibms.sinica.edu.tw/pages/input_CandidateGeneSearch.jsp
mutDB	http://mutdb.org/
PolyMAPr	via the author
Snap	http://snap.humgen.au.dk/
SNP function portal	http://brainarray.mbni.med.umich.edu/Brainarray/Database/SearchSNP/snpfunc.aspx
SNPseek	http://snp.wustl.edu/cgi-bin/SNPseek/index.cgi
SNPselector	http://snpselector.duhs.duke.edu/
VisualSNP	http://genepipe.ngc.sinica.edu.tw/visualsnp/
SNP function prediction: regulatory variants	
ENCODE project data	http://genome.ucsc.edu/ENCODE/
Automated Splice Site	https://splice.uwo.ca/
Human Splicing Finder	http://www.umd.be/HSF/
mRNA bySNP browser	http://www.sph.umich.edu/csg/liang/ asthma/
Patrocles	http://www.patrocles.org/
PolymiRTS	http://compbio.utmem.edu/miRSNP/home.php
PromoLign	http://polly.wustl.edu/promolign/main.html
SNPExpress	http://people.genome.duke.edu/~dg48/ SNPExpress/
SNP@promoter	http://variome.kobic.re.kr/SNPatPromoter/
Other areas of SNP bioinformatics	
PolyPhred	http://droog.gs.washington.edu/polyphred/
PolyBayes	http://genome.wustl.edu/tools/software/polybayes.cgi

SNPDetector	http://lpg.nci.nih.gov/
ssahaSNP	http://www.sanger.ac.uk/Software/analysis/ssahaSNP/
DNannotator	http://bioinfo.bsd.uchicago.edu/DNannotator.htm
InSNP	http://www.mucosa.de/insnp/
novoSNP	http://www.molgen.ua.ac.be/bioinfo/novosnp/
SNP-PHAGE	http://bfgl.anri.barc.usda.gov/ML/snp-phage/
SeqDoC	http://research.imb.uq.edu.au/seqdoc/
PolyFreq	as a supplement
ZOOM	commercial
PyroBayes	http://bioinformatics.bc.edu/marthlab/
SeqMap	http://biogibbs.stanford.edu/~jiangh/SeqMap/
Rolexa	http://bbcf.epfl.ch/view/BBCF/BBCFResources
SNPmasker	http://bioinfo.ebc.ee/snpmasker/
GAD	http://geneticassociationdb.nih.gov/
HuGE Navigator	http://www.hugenavigator.net/
GRAIL	http://www.broad.mit.edu/mpg/grail/
G2D	http://www.ogic.ca/projects/g2d_2/
Prioritizer WGA	http://bio-informatics.gr/content/view/24/39/
Prospectr	http://www.genetics.med.ed.ac.uk/prospectr/
OSIRIS	http://ibi.imim.es/OSIRISv1.2.html
MarkerInfo Finder	http://brainarray.mbni.med.umich.edu/Brainarray/DataMining/MarkerInfoFinder/
NCBI BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi