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.**)

***Circ Cardiovasc Genet* is available at http://circgenetics.ahajournals.org** **DOI: 10.1161/CIRCGENETICS.109.872010**

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 ≥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 available 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.