Scientific software: seeing the SNPs between us

Steven David Buckingham

The results of large genome-wide association studies (GWASs) are being deposited in public databases with increasing frequency. But the software to analyze and interpret GWAS datasets can be difficult to use. Could a new generation of user-friendly programs fill the gap?

Genome-wide association studies (GWASs) are an exciting new approach in the hunt for genes affecting our health and wellbeing. GWASs look for associations between underlying genetic features—such as single-nucleotide polymorphisms (SNPs) and copy-number variations (CNVs)—and phenotypes such as health, illness and even behavior. The last two years have witnessed an explosion in the number of published GWAS studies, revealing new genes involved in diseases such as late-onset Alzheimer's[1], type-2 diabetes[2,3], schizophrenia[4] and cancer[5,6].

But if these datasets are to be of use to all biologists, not just GWAS experts, then new, user-friendly software is urgently needed. Happily, software designers are on the case, coming up with new ways of making GWAS data easy to explore and share among researchers, and designing analysis packages that deal with the increasing computational demands posed by these datasets.

Julie Williams of the Cardiff University School of Medicine in Cardiff, UK sees a clear need for new software. Last year, Williams was awarded £1.3 million (about $2.6 million) from the Wellcome Trust to lead one of the biggest GWAS studies of Alzheimer's disease to date. "One of the major challenges for GWAS software designers is keeping up with the changing demands of the research field," notes Williams. "These demands can be statistical, such as incorporating SNP imputation, or bioinformatic, as with the integration of large-scale gene-expression data."

Williams' view is shared among GWAS researchers. Hakon Hakonarson of the Children's Hospital of Philadelphia in Pennsylvania, USA, who recently headed up a GWAS that identified a new type-1 diabetes gene, thinks there is a need to improve software and computational power on all fronts to gain more speed when exploring these larger datasets. "Some of these analyses can take days to run," he notes.

Visualization tools worth a thousand words

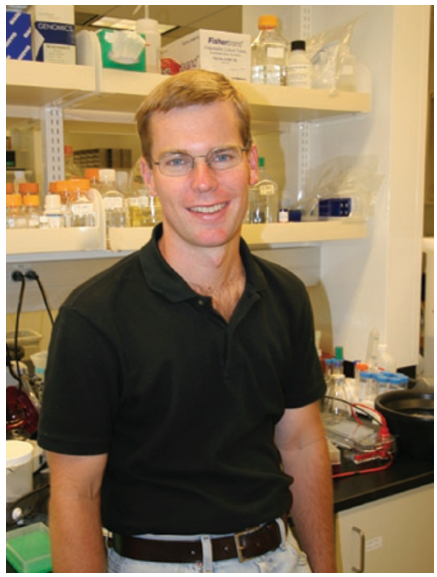
There is little doubt that the currently available GWAS software tools present a barrier to the novice user. The associations that GWASs look for can be quite subtle, with several genes potentially acting together to exert an effect. This requires advanced statistical methods to tease associations out of datasets, without misleading researchers with false positives.

GWAS experts have traditionally used their own in-house scripts written in command-line environments, such as the R statistical package, a computational and graphics environment developed at Bell Laboratories in Murray Hill, New Jersey, USA. Although statisticians and computer programmers are familiar with these powerful toolkits, to most biologists they are like a foreign language. But researchers like Fredrik Pettersson from the Wellcome Trust in Cambridge, UK, who developed a GWAS program called Goldsurfer2, are creating a new generation of visually based, interactive GWAS programs.

"With complex datasets, it is useful to use our superior cognitive abilities by interpreting pictures instead of text," says Pettersson. He notes that these genomic datasets, with their references to physical positions and the ability to link to annotations, are well suited to interactive environments. "All sorts of filters, colors, shapes and sizes can be used to represent slices of the data."

The size of GWAS datasets means the visualization software must also be capable of doing two things at once: provide the user a broad overview of the data in a way that reveals trends as well as have the ability to focus on areas of interest once identified." Stephen Miller, European Director of Business Development for Progeny Genetics based in South Bend, Indiana, USA, says, "Visualization of targeted genes is really important, especially in family-based studies. For example, once you get to the gene level you might want to see patterns of haplotypes and how that reflects the disease."

Certainly with gene chips capable of scanning a million SNPs and a typical GWAS study incorporating thousands to tens of thousands of individuals, standard tables and plots are of no use with datasets of this size. Goldsurfer2 overcomes this problem by arranging the data in a hierarchical



Trey Ideker is working to apply his Cytoscape pathway analysis program to GWAS data.

structure and linking it to plots and tables that update themselves as users focus in on nodes of interest. The raw data are presented in a tree structure, making it easy to create subsets based, for instance, on population stratification. The user can import several sets of samples, merge them into a common file and use principal component analysis (a statistical method for reducing multidimensional datasets with minimal information loss) to see whether stratification is skewing the results. Presenting data as tables side by side with interactive graphical plots is an approach also taken by Partek Inc. of St. Louis in their Genomics Suite, which offers the possibility of exploring data with biplots, heat maps and frequency plots.

For biologists, such interactive visual environments can make standard GWAS analysis techniques quite intuitive. For example, Pettersson's Goldsurfer2 takes the traditional two-dimensional linkage disequilibrium plot into three dimensions, where the contours and coloring of the plot can be used to represent different measurements of linkage disequilibrium, and zooming in on the plot shows the calculated values for cases and controls. A similar approach was adopted for the "Genomics Module" of the JMP7 software package from SAS Institute Inc. of Cary, North Carolina, USA, which plots principal component analyses in three dimensions, thereby allowing the user to reduce a multidimensional dataset to three dimensions to see how the data group.

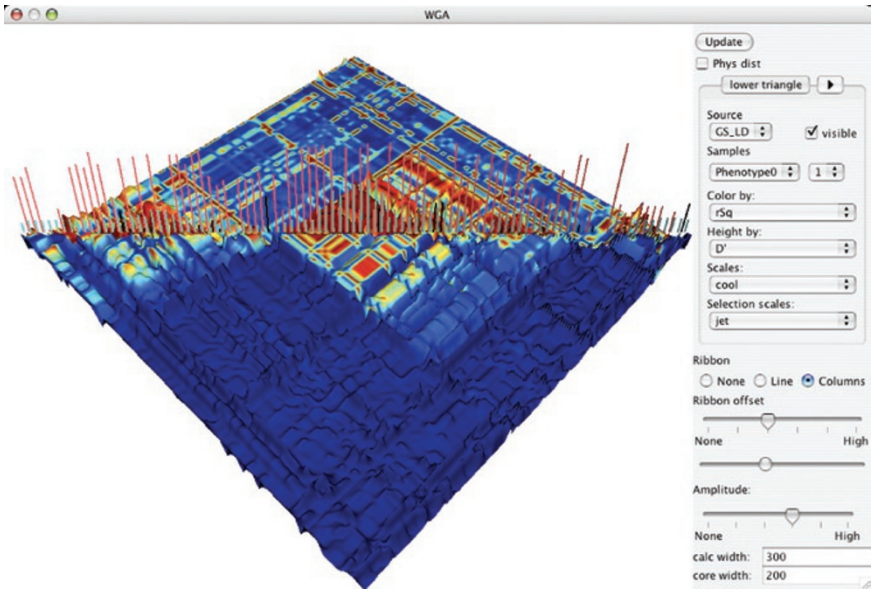
Progeny Genetics

Progeny Genetics' Stephen Miller says visualization of target genes is important when analyzing GWAS datasets.

Still, designing visualization approaches for GWAS data has lagged far behind other areas of GWAS software development. "There are two reasons for this reluctance to tackle the visualization challenge," says Kristin Tolle, Senior Research Program Manager for Biomedical Computing for External Research in Microsoft Research in Redmond, Washington, USA. "First, it is not clear exactly how to represent multidimensional data in a way that will lead to discoveries. And in the second place there are immense computational barriers in visualizing such large datasets." In April 2008, Microsoft awarded over $850,000 to six teams of researchers to jumpstart the development of visualization tools.

Visualizing the right path

The complexities of the visualization challenge are not lost on Trey Ideker, associate professor of Bioengineering at the University of California, San Diego's Jacobs



Fredrik Pettersson

The Goldsurfer2 software package allows three-dimensional visualization of linkage disequilibrium.

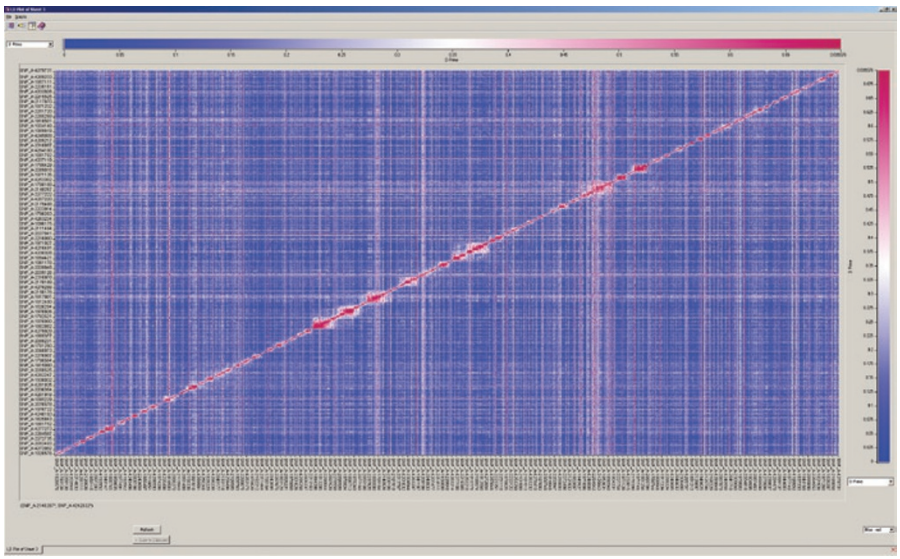
School of Engineering. But Ideker, who developed a software program for the visualization of protein interaction networks called Cytoscape, believes that in the future GWAS datasets will be visualized in terms of pathways. "We think that GWAS needs to transcend SNP-based thinking and move on to pathways," he argues. His team was one of the six groups funded by Microsoft, and he envisions a day when a user will take a result from a GWAS study and use a program such as Cytoscape to query protein-interaction databases and identify pathways and subnetworks involved with the phenotype.

"The gold nuggets of genome-wide significant SNPs only explain a small fraction of heritability," says Christophe Lambert, chief executive officer of Golden Helix in Bozeman, Montana, USA, a data-analysis software company he founded 10 years ago in anticipation of the present need for user-friendly software. "We need to look at the less significant SNPs—the gold dust mixed in the dirt."

Some existing commercial packages specializing in pathway analysis can already be used for the analysis of GWAS datasets. Programs such as Metacore from Genego of St. Joseph, Michigan, USA and Ingenuity Pathway analysis from Ingenuity of Redwood City, California, USA allow any large-scale dataset, including GWAS datasets, to be imported and analyzed with regard to protein interactions and other pathways, such as metabolic and signaling networks. Following a slightly different approach, the JMP Genomics Module uses clustering tools to identify pathways highly represented in datasets. Imaginative and computer-aware users could even use the current version of Ideker's Cytoscape to superimpose GWAS findings onto protein-protein interaction networks, although Ideker says future developments will incorporate this functionality into the main Cytoscape program.

But not every GWAS software developer thinks visualization is the answer to all our needs. "At the end of the day, the real value of analysis is in reducing the massive datasets to a small set of hypotheses that the investigator can interpret," says Hakonarson. "While a user-friendly interface is appealing to most users, sometimes command line–driven programs are better, such as for rapid and parallel computation of large data sets."

Lambert also notes that although the human mind is hard-wired to see patterns, you still need good statistics to sort out whether those patterns arose by chance alone. "Visualization is an important trend, but won't get you anywhere if the underlying analysis isn't right," says Lambert. Indeed, with datasets comprising over a million SNPs, a standard *P*-value cut-off of 0.05 that does not account for multiple testing would show ~50,000 SNPs as significant just by sheer chance. But programs like Golden Helix's SNP and Variation Suite include a genome-wide permutation test that efficiently deals with the false discovery problem by examining the proportion of the time that the genome-wide minimum *P*-values



Helix Tree, from Golden Helix, uses color to visualize linkage disequilibrium.

generated from tests on randomly shuffled phenotypes are at least as significant as the *P*-values generated from the original phenotype. This provides a model-free estimate of the *P*-values, more closely approximating the true significance of potential findings.

"The top hits in the first round of analysis from any genome screen are mostly artifacts that result from genotyping problems or very low minor allele frequencies," points out Bill Cookson of the National Heart and Lung Institute in London, who used a GWAS to discover genes involved in asthma. "The key tools at this stage are a strong sense of cynicism and the desire to root out false positives by examination of genotyping traces and raw allele counts." And he adds that only after this process does modern software help navigate through the biological complexities that underlie association signals.

Cloud genomics

Visualizing increasingly large datasets is not the only problem being addressed with the new GWAS software. Programmers are also facing up to the intense computational demands placed by the sheer size of these datasets. When interrogating a million SNPs per sample, a 10,000-sample study will be simply too big to load into the memory of a normal desktop computer. Goldsurfer2 solves this problem by using a swap file, and can comfortably load datasets containing 5,000 markers and 5,000 samples. But things can get even trickier when it comes to the analysis. Access times can become intolerable, and advanced statistical procedures that involve pair-wise comparisons can become prohibitively slow. However, by applying a battery of the latest computing techniques, some software packages can cope with these problems surprisingly well.

"Golden Helix has successfully performed both whole-genome SNP and CNV association studies on over 11,000 samples with over 500,000 markers apiece," claims Lambert. "Starting from terabytes of raw data, final processed matrices of 40 GB or more can be efficiently analyzed with desktop computers. This includes principal component analysis for the correction of batch effects and of population stratification."

But the challenges are only going to get tougher. Up to now, most GWASs have explored simple 'yes' or 'no' phenotypes: does the patient have diabetes or not? Is the patient overweight? But there is increasing interest in applying GWAS to study more graded phenotypes, such as blood pressure or insulin levels. These expression quantitative trait loci will inevitably increase the size of datasets by an order of magnitude.

Another reason GWAS datasets are going to get bigger is the increasing interest in CNVs. "[CNVs] may account for more variation than SNPs," argues Lambert. CNVs are regions of repeated sequence in the genome that differ from one individual to the next. The latest SNP chips, including the SNP 6.0 chip from Affymetrix in Santa Clara, California, USA and the Infinium HD Bead Chips from Illumina of San

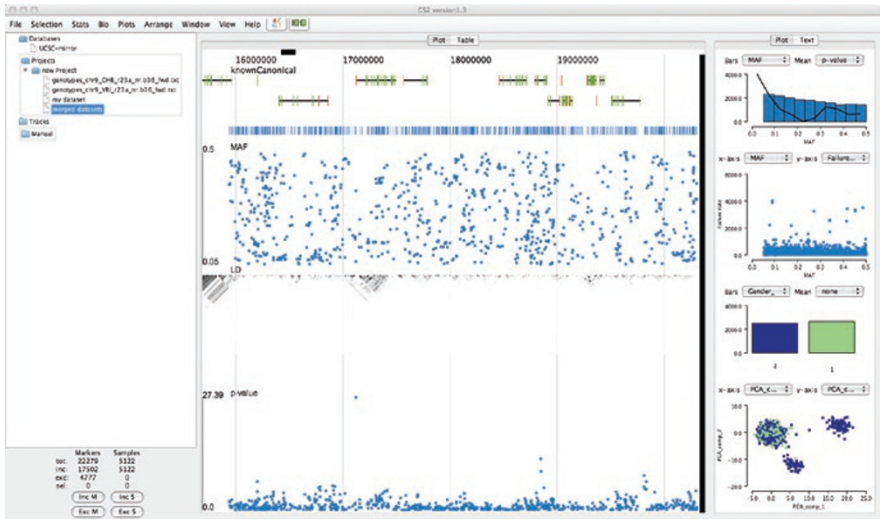
Diego, have already begun to incorporate CNVs—with more on the way. "Whole-genome sequencing will increase data sizes by several orders of magnitude," according to Lambert, who talks of a forthcoming "fire hose of information."

Software companies are not that far behind in rising to this CNV challenge. Affymetrix genotyping console software includes algorithms and visualization tools for analyzing CNVs, and Golden Helix also includes a CNV module. The Affymetrix Genotyping Console implements Canary, a new CNV-calling algorithm that was developed in collaboration with the Broad Institute in Cambridge, Massachusetts, USA. The Partek Genomics Suite is also suitable for exploring CNV data, allowing copy-number and loss-of-heterozygosity data to be explored in the same region and visualized with the original mapping data.

In the end, the increasing dataset size from GWAS is likely to outstrip the capacity of the standard laboratory computer, forcing some developers to look to the internet for solutions. Microsoft is planning a web-based, end-to-end virtual GWAS environment for researchers. Offered in conjunction with the British Library, the Research Information Centre—which Tolle likens to a "Facebook for geneticists"—can simplify the process of information search, facilitate discovery, effectively manage research objects, and enable versioning and archiving. "The time is right for this," says Tolle, "all the pieces are there; it is just a matter of putting them all together."

The collaboration environment resides within a hosted Microsoft Office SharePoint Server 2007 platform, which is accessible from a web browser, and allows researchers to collaborate and share data all the way from the beginning of a GWAS study to the final publication. Users will be able to select and run workflows online, which will run Microsoft's own in-house analysis routines behind the scenes. This service is currently in beta testing.

User-friendly, graphical programs for analyzing GWASs are becoming a reality, and imaginative visualization approaches are making it ever easier for users to explore GWAS datasets. But it remains to be seen whether the exponential increase in computational power demanded by ever-growing datasets will ultimately drive analysis off the desktop and onto the web.



Fredrik Pettersson

Goldsurfer2 presents users with linked tables and plots to facilitate interaction.

1. Grupe, A. *et al. Hum. Mol. Genet.* **16**, 865–873 (2008).
2. Sladek, R. *et al. Nature* **445**, 881–885 (2007).
3. Hakonarson, H. *et al. Nature* **448**, 591–594 (2007).
4. Stefansson, H. *et al. Nature* advance online publication 30 July 2008 (doi:10.1038/nature07229).
5. Thorgeirsson, T.E. *et al. Nature* **452**, 638–642 (2008).
6. Hung, R.J. *et al. Nature* **452**, 633–637 (2008).

Steven David Buckingham is a researcher in the Functional Genetics Unit at the University of Oxford, UK (steven.buckingham@dpag.ox.ac.uk).

SUPPLIERS GUIDE: COMPANIES OFFERING CHEMICAL GENOMICS REAGENTS AND INSTRUMENTATION

Company	Product
Affymetrix	http://www.affymetrix.com/
Agilent Technologies	http://www.chem.agilent.com/
Ariadne Genomics	http://www.ariadnegenomics.com/
Array Genetics	http://www.arraygenetics.com/
Attagene	http://www.attagene.com/
Bio-Rad	http://www.bio-rad.com/
deCODE genetics	http://www.decode.com/
Enzo Life Sciences	http://www.enzo.com/
Expression Analysis	http://www.expressionanalysis.com/
Genego	http://www.genego.com/
Geneservice	http://www.geneservice.co.uk/
Genizon Biosciences	http://www.genizon.com/
GenoLogics	http://www.genologics.com/
Genolyze	http://www.genolyze.com/
Geopspiza	http://www.geospiza.com/
Golden Helix	http://www.goldenhelix.com/
Illumina	http://www.illumina.com/
Infoquant	http://www.infoquant.com/
Ingenuity	http://www.ingenuity.com/
Jivan Biologics	http://www.jivanbio.com/
JMP software	http://www.jmp.com/
Marligen Biosciences	http://www.marligen.com/
Miltenyi Biotec	http://www.miltenyibiotec.com/
MiraiBio	http://www.miraibio.com/
Molecular Devices	http://www.moleculardevices.com/
NimbleGen	http://www.nimblegen.com/
Ocimum Biosolutions	http://www.ocimumbio.com/
Oxford Gene Technology	http://www.ogt.co.uk/
Oxford University	http://www.stats.ox.ac.uk/
Partek	http://www.partek.com/
PerkinElmer	http://www.perkinelmer.com/
Perlegen Sciences	http://www.perlegen.com/
Phalanx Biotech Group	http://www.phalanxbiotech.com/
Premier Biosoft	http://www.premierbiosoft.com/
Progeny	http://www.progenygenetics.com/
Quantiom	http://www.quantiom.de/
Rosetta Biosoftware	http://www.rosettabio.com/
Sequenom	http://www.sequenom.com/
Signature Genomic Laboratories	http://www.signaturegenomics.com/
Softgenetics	http://www.softgenetics.com/