A FRAMEWORK TO LINK WHOLE GENOME SNP ASSOCIATION STUDIES TO SYSTEMS GENETICS

S.J. Goodswen [1, 2], H.N. Kadarmideen [1], C. Gondro [2] and J.H.J. van der Werf [2]

[1] CSIRO Livestock Industries, JM Rendel laboratory, PO Box 5545, Rockhampton Mail Centre, Rockhampton, QLD 4702
[2] School of Environmental and Rural Sciences, University of New England, Armidale, NSW 2351

SUMMARY

One of the main outcomes from whole genome association studies (WGAS) are statistically significant single nucleotide polymorphisms (SNPs). This outcome would be enhanced if either biological meaning or potential functional roles for these SNPs are revealed. Such knowledge would enhance the process of marker- or gene-assisted selection (MAS/GAS), or at least would provide more understanding of the achieved genetic change. We develop and present an R package called *FunctSNP*, interfaced to a relational database, to add potential functional information to SNPs that have shown to be statistically significant in a WGAS. The approach and developed software has been designed in such away that the database can contain any model species.

INTRODUCTION

Whole genome association studies with genotyped SNPs are a set of methods to identify which SNPs are associated with variation in a particular complex trait of interest. There are essentially two reasons why WGAS has gained in popularity. Firstly, the huge increase in the availability of SNPs. For example, the dbSNP database housed by the National Center for Biotechnology Information (NCBI) contains millions of SNPs for many model species. Secondly, the falling cost and increased capacity of high-throughput SNP genotyping chips has made WGAS feasible and affordable.

The significant SNPs derived from WGAS can in effect be classified into 4 types: (1) a SNP that contributes to variation in the complex trait; (2) a SNP with no known biological effect but in linkage disequilibrium (LD) to an untagged SNP (not genotyped) that contributes to variation in the complex trait; (3) a SNP with an association only - no known biological effect or linkage to a causal SNP; (4) a SNP *not* associated with the complex trait (a false positive). These significant SNPs are currently used in two conventional, albeit alternative, approaches depending on the requirements of the researcher: (1) a molecular biologist usually searches for candidate genes in the region of the significant SNP and then does functional experiments, which could include targeted sequencing; (2) an animal breeder or geneticist uses significant SNPs as DNA markers for associated trait selection (MAS/GAS).

Although WGAS are the best methods we currently have, literature states that the most notable limitation of WGAS is the potential for false-positive results (Pearson and Manolio 2008). Since the type of significant SNP following WGAS is not distinguishable to the researcher, there is a potential to perform functional experiments based on a false-positive SNP and/or use a SNP as a marker that is neither the causal variant nor in LD with the causal variant.

At present, the only approach to truly validate a significant SNP is to repeat the WGAS on a different population. *FunctSNP*, however, can assist the researcher in distinguishing the type of significant SNP by providing information as to whether a SNP is: (1) a potential causal variant; (2) close to a SNP that is a potential causal variant; (3) close to a gene that has the potential to affect the complex trait [see Program functionality section for more details].

DESIGN AND IMPLEMENTATION

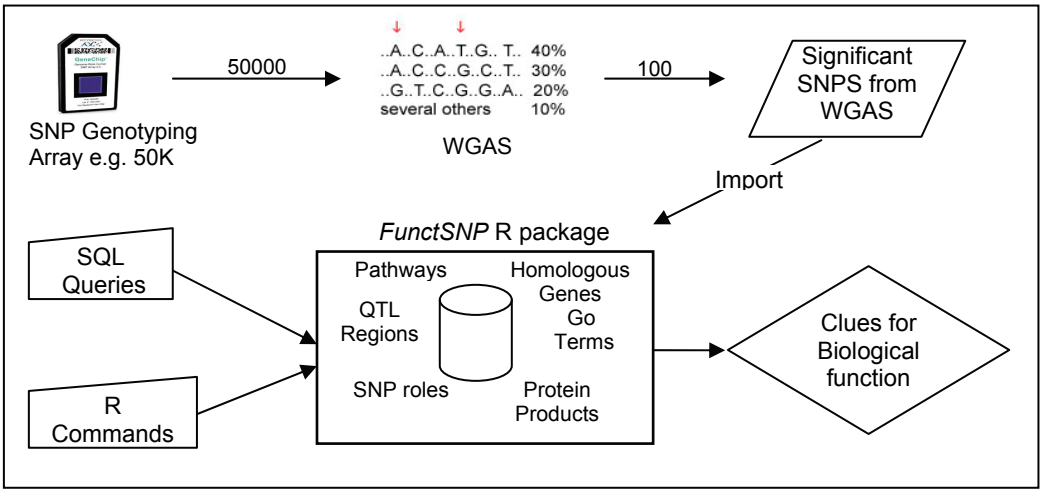

Figure 1. Schematic of *FunctSNP* R package post-WGAS

The *FunctSNP* R package (Figure 1) interfaces to a localised relational database that contains the most pertinent SNP related information for a chosen model species from all the sites listed in Table 1. Most of these sites are intuitive and informative. A researcher, through the use of Internet addresses, hyperlinks, and cross-linked data can navigate from site to site with ease; however, they cannot integrate related data in one query. So there are essentially two main reasons for using *FunctSNP*: The power of a statistical computing program, interfaced with a *relational* database. R is a free software environment for statistical computing and graphics (http://www.r-project.org/). A relational database is structured data in the form of two-dimensional tables. Each table is linked by a defined relationship and consequently all data is integrated. The power comes from a language designed to query the database called Structured Query Language (SQL).

Program creation. Our first database design consideration was which Relational Database Management System (RDMS) to use. We opted for the open source SQLite which stores data in a file that is platform independent and is the preferred RDMS for R (SQLite available from http://www.sqlite.org/). Next, we were faced with the challenge that there is currently no standardized format for transporting data from one database to the next, although extensible markup language (XML) and BioMart are gaining prominence. The data formats from each site were: NCBI - Microsoft SQL Server, GO – MySQL, KEGG – XML, UniProt – XML QTLdb – ASCII text file, OMIA – MySQL, and Homologene - ASCII text file. Generic C++ programs and Perl scripts were written to convert the data formats to a format compatible with SQLite.

There are essentially 4 manual steps to create a local database: (1) download the data from the online resource; (2) decompress file(s) if required; (3) convert data into a format that can be imported to a relational database; (4) import the data into the database. These 4 steps need to be repeated for each online resource. We have developed software to provide a framework in which these 4 steps are automatically repeated to create a local database with data extracted from any number of resources. To ensure the local database is always up-to-date the automatic creation can

be scheduled to run on a weekly or even daily basis. The principle idea is that once the schedule is set up no further human intervention is required.

At the time of writing this paper, *FunctSNP* is not publically available. It is anticipated that *FunctSNP* will be available for download from a CSIRO HTTP server. To the best of our knowledge there are only a few SNP software tools designed specifically for the livestock industry. Therefore it is expected that the first version of *FunctSNP* will be for *Bos taurus.*

Table 1. Internet sites containing biological data

Acronym	Name	Link	Resource
dbSNP	Single Nucleotide Polymorphism database	http://www.ncbi.nlm.nih.gov/	SNPs
GO	Gene Ontology	http://www.geneontology.org/	Genes and gene product attributes
KEGG	Kyoto Encyclopaedia of Genes and Genomes	http://www.genome.jp/kegg/	Biological pathways
UniProt	Universal Protein Resource	http://www.uniprot.org/	Protein sequences and functional information
QTLdb	Animal Quantitative Trait Locus database	http://www.animalgenome.org/QTLdb/	Quantitative Trait Loci data (QTL)
OMIA	Online Mendelian Inheritance in Animals	http://omia.angis.org.au/	Genes, inherited disorders and traits
HomoloGene		http://www.ncbi.nlm.nih.gov/homologene	Homolog detection

Program functionality. The primary function of *FunctSNP* is to provide information about the potential biological functions of SNPs or genes close to SNPs. Information such as SNP chromosomal location, exon/intron status, synonymous/non-synonymous effect, SNPS in Quantitative Trait Loci (QTL) regions, and biological pathways, GO terms, and protein products for related genes. There are two ways to access the information: (1) Direct access to the database using SQL queries through the SQLite program, and (2) using a set of R functions (currently 20) to answer preset questions. For example, "Give me all gene IDs associated with significant SNPs" and the output is a 2 dimensional array for further analysis with R commands.

The database is both SNP ID (using NCBI's dbSNP rs# cluster ID) and gene ID (using NCBI's gene ID) centric. For example, given a SNP ID as an input, the output is the gene ID on which the SNP resides. The gene ID provides the link to *every* gene attribute in the database such as name, function, protein products, and biological pathways. Alternatively, given a gene ID as an input, the output is a list of SNPs located on the gene. The SNP ID is then the link to every SNP attribute such as exon/intron status, and chromosomal location.

Not all significant SNPs from WGAS are located on genes. Therefore *FunctSNP* provides the functionality to find the nearest genes to such SNPs. The search distance is base pairs (bp) and is user defined. For example, enter 100 bp and the program returns the gene ID for any SNP residing less than 100 bp (e.g. a promoter region) from the gene's transcription start site. In a similar manner to searching for nearest genes, *FunctSNP* provides information as to whether a SNP is close to another SNP that is a potential causal variant. For example, search for all non-synonymous SNPs within a user defined distance from a significant SNP.

There is much more publically available data on some species than others. So for some species there may be no insightful data on the genes identified as linked with significant SNPs. *FunctSNP* provides the ability to find homologous genes across all species or a specified species. For example, for identified *Bos taurus* genes, we can obtain gene IDs for homologous genes from

Homo sapiens. These gene IDs can then be uploaded into other analysis programs such as DAVID (http://david.abcc.ncifcrf.gov/summary.jsp).

To aid interpretation of the output from *FunctSNP*, we intend to provide a scoring system for WGAS significant SNPs e.g. SNPs which reside on an exon region with a non-synonymous effect will be given the highest rank. A normalised score (taking into account such factors as sample size, SNP chip capacity) will be adopted to make it comparable with WGAS of the same population. There will also be the option to search for the highest scoring SNPs within a user defined distance from a significant non-causal SNP.

As part of the ongoing development, the Internet will be monitored for new or overlooked sites which contain informative publically available biological data that may be incorporated into *FunctSNP*. Also, future development will be focused on how best to interpret and present the results. For example, taking the genes associated with the SNPs and applying a gene set enrichment analysis (GSEA).

Program testing. To test *FunctSNP*, we used real 10k SNP genotype data from a yet to be published WGAS. From this 10K SNP data, 165 SNPs had been identified as significantly affecting a particular trait of economic interest. The NCBI reference numbers (rs#) for the 165 significant SNPs were imported into our database and in less than five minutes a series of reports using *FunctSNP* R functions were generated. From the 165 SNPs, 49 were located in a gene region. One of these 49 SNPs was located on an exon and was synonymous. For the 49 genes, there were 81 GO terms, 20 KEGG pathways, 355 homologous genes from 17 different species, and 32 homologous genes from the same species. A total of 12 genes downstream and 10 genes upstream were found less than 10,000 bp from the non-coding significant SNPs.

CONCLUSIONS

To take full advantage of WGAS, we need to make that essential link between the outcomes from WGAS and the information that exists about the function of genes and pathways. We conclude that *FunctSNP* is a post-WGAS tool that provides an opportunity to screen and select for SNPs that have a higher likelihood to be related to variation in a particular complex trait of interest.

ACKNOWLEDGEMENTS

Office of the Chief Executive (OCE) of CSIRO for postgraduate scholarship.

REFERENCES

Pearson, T. A. and Manolio, T. A. (2008) *JAMA* **299**:1335.