InforSense White Paper

InforSense GenSense

High throughput GWAS analysis

June 2007

InforSense Embedding Intelligence Throughout the Enterprise

Disclaimer and Confidentiality Notice

This document is provided on an as-is basis, without any warranty, expressed, implied or otherwise, as to satisfactory quality or fitness for a particular purpose. In particular, any technical details of the software shown in this InforSense™ document may change without notice. The information provided herein is confidential.

Table of contents

A CHALLENGE FOR SCIENTIFIC RESEARCHERS

Genome Wide Association Studies, or GWAS for short, are gradually becoming more commonplace. The main goals of these studies are to find genetic variants, or SNPs, that are correlated between Case or Control individuals. There have been recent publications announcing successful results in this field, including some of the first results from consortia who are currently running some of the largest GWA studies.

There continues to be rapid advances in the genotyping technology, which means that the ability to generate ever larger amounts of data has increased dramatically over the last year. Experiments are now being planned in the range of measuring a million SNPs across tens of thousands of samples.

To deal with these potentially scientifically groundbreaking and large data sets there is a growing need for better analytical software solutions to capture the value generated from these studies. Such software needs to be highly flexible to enable *ad hoc* integration of many different types of data and algorithms and to compare outputs from different approaches. Additionally software needs to facilitate the integration of custom built tools and analysis methods and support them in an environment for data management and analysis.

Once interesting SNPs have been identified, scientists need to be able to bring content and context in from other areas, like genomics and pathways, together with the GWA case/control studies and combine that with phenotypic information from clinical trials.

Current challenges for dealing with the GWA data include:

- How to handle and process large data sets.
- How to flexibly assess the quality of the data that is generated.
- How to cope with an area whose analysis methods are continually evolving.
- How to bring in context from other areas.
- How to build flexible analyses and comparing different statistical methods.
- How to integrate scripts and 3rd party tools in a maintainable environment.

In this paper we'll briefly describe how InforSense has met some of these challenges which scientists have to solve when faced with analysing the results from a GWA study.

GWAS AND INFORSENSE GENSENSE

InforSense GenSense is an extension to the InforSense Platform and enables the analysis of Genome Wide Association Studies and the deployment of those analyses to a customisable portal environment.

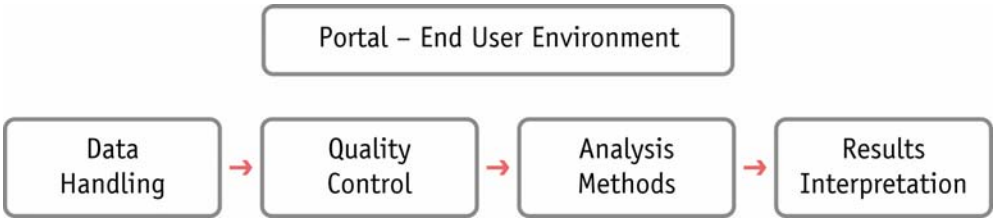
It is a solution that has been designed to deal with the high volume analysis of genotyping data coming from platforms like Affymetrix and Illumina and can unlock the value of the data that is being generated in, sometimes very expensive, genotyping experiments.

Core Concepts

The system is setup for data handling to manage the large volume of genotyping data that can be produced in a GWA study. GenSense provides high throughput analysis as it can stream large data sets of information through your analytical process. This helps in allowing it to scale to whole genome wide analyses.

Once the data is in the system quality control workflows are used to assess the data quality using a range of different methods and reports. When high quality data is assured, the multiple analytical and statistical methods can be applied to test for association between cases, controls and other phenotypes.

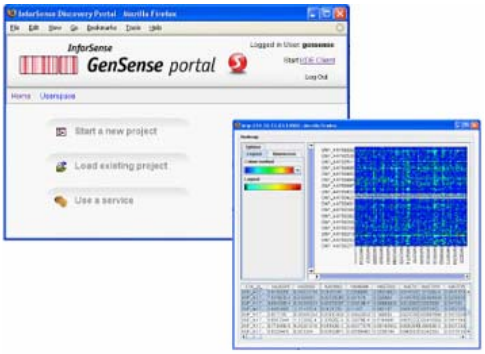
Finally, carrying out interpretation of the significant SNPs is achieved by enabling links to other products, systems and 3rd party tools that help place the results in biological context.



The system has been built using visual workflows rather than at the programming level. It is therefore easy to adjust the system and to standardise on the right analysis methods for your lab. With these workflow definitions for each analysis step it is straightforward to reuse and share your analyses with colleagues or collaborators.

Portal

These core concepts are delivered in an easy to use portal environment that makes up the end user interface. In the portal all the data is organised into a simple project structure where a user or group can be associated with a particular analysis that is underway. For each user, we lead them through the system in an interactive or 'wizard based' manner. The collection of underlying workflow services are automatically linked together and the analysis is done according to the path chosen by the user.



In general the user will define a project, associate or load data into that project and then assess the quality of the samples and SNPs before performing whichever statistical analyses they would like. This generates a set of interesting SNPs which may be sent on to an annotation service, such as SNP to gene mapping.

The portal provides interactive help at each step and should the portal not conform with the process in your lab then it is easy to customise and re-design the portal using the workflow building platform.

Data Structure

The two major vendors, Affymetrix and Illumina, who have the technology to genotype large numbers of SNPs, have been continually increasing the coverage of their arrays. In recent months both have announced new products giving the ability to genotype up to 1 million SNPs, and have included the ability to look at Copy Number Variation (CNV). The inset shows some of the recent products from both vendors.

Due to the ability of labs to generate large amounts of genotyping data, a key part of designing the system was around performance, throughput and scalability. It was with these in mind that the underlying data model in the system was built.

The proprietary data representation in GenSense is designed to be scalable and fast so that all the GWAS analyses can be run in memory.

The internal data structure is made up as follows:

- SNP Metadata
  - SNP name, alleles, counts of homo and heterozygosity, unknown and a customisable score
- Sample Metadata
  - Sample name, pedigree information, total genotype counts
- Genotype Data
  - Highly condensed binary representation

Affymetrix

- GeneChip Human Mapping 500K Array
  - 500,000 SNPs, comprising of 2 chips
- Genome-Wide Human SNP Array 5.0
  - 500,000 SNPs & 420,000 CNV probes
- Genome-Wide Human SNP Array 6.0
  - 906,000 SNPs & 946,000 CNV probes

Illumina

- humanhap550+ genotyping beadchip
  - > 550,000 SNPs, 4300 CNV SNPs, 7800 nsSNPs & 1800 tagSNPs
- human650Y genotyping beadchip
  - > 655,000 SNPs, 4300 CNV SNPs, 8000 nsSNPs & 1800 tagSNPs
- human1M beadchip
  - >1 million SNPs in total including 25000 nsSNPs

Data Scalability

Once the data structure was implemented we tested the system performance and scalability of the internal data representation. We used sample data from the Illumina HumanHap650Y Genotyping BeadChip. The test set contained 655,352 SNPs and we benchmarked from 159 to 2544 samples.

Based on the testing, the core statistical algorithms scale in a linear fashion with respect to the amount of samples and even studies with large sample sizes could be quickly calculated on a small desktop machine.

Quality Control

This is a key area, particularly for core genotyping facilities, who are interested in being able to develop and re-use standard procedures for

- **Duplicate checking**: categorized as to whether they are expected or unexpected duplicates, where information on the relationship between samples and individual is provided
- **Hardy-Weinberg tests**: test all genotypes at each locus for deviation from Hardy-Weinberg proportions
- **Genotype completion**: For locus and sample find the amount of non-missing and missing data in a genotype dataset
- **Mendelian inheritance**: Per sample and locus deviations from Mendelian inheritance patterns observed between parents and offspring

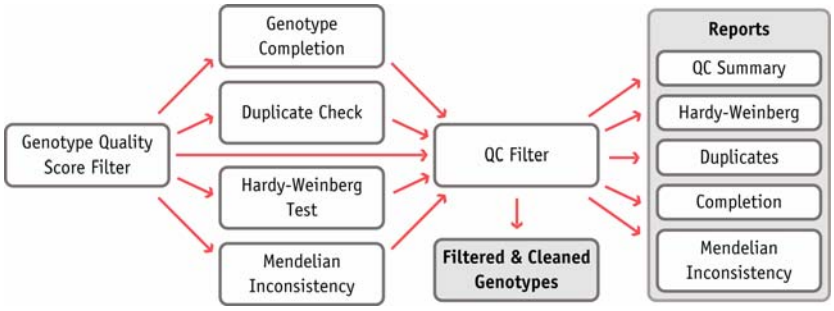


Figure 1: This is an example of a high level flow of how sets of QC services can be built to generate an automated workflow for checking data quality. Poorly scoring genotypes can be filtered on import; they are then assessed (in a parallel fashion) for completion, sample duplication and correspondence to Hardy-Weinberg proportions. Based on the results filters can be applied and a series of reports chosen to reflect the interest of the researcher.

GWA Statistical Analysis

In the case study at the end of this paper, you will see particular mention to two of the analysis nodes. One is the 'SNP analyzer'; the other is the 'Case Control SNP Tester'.

Using the *SNP Analyzer* with the input of the set of GWA genotype data results in a table of statistics that includes the following for each SNP:

- Major/Minor Alleles {A,C,G,T}
- MAF, heterozygosity, measures of variability
- Genotype Counts (AA, Aa, aa)
- Expected Counts
- HWE Chi^2 and Exact p-values

The *Case Control SNP Tester* identifies which SNPs, in a GWA, are associated according to the designation of Cases and Controls. As input the SNP tester node takes the set or subset of samples which are cases together with the set of control samples. Based on this data it performs a number of score type tests and other statistics that are relevant in such analysis. These include the following:

- Trend Tests – Additive / Dominant / Recessive
- Genotypic and Allelic Tests
- Dominant-Recessive Test
- Odds Ratio (Aa vs. AA, aa vs. AA)
- Relative Risk (Aa vs. AA, aa vs. AA)

As well as these statistical tests the solution also provides several different methods for correlation such as Pearson's R, Spearman Rank coefficient and Kendall's Tau methods. Linkage disequilibrium, or LD, can be calculated by different correlation methods, co-segregation or association measures, such as D, D prime, uncertainty coefficients, Cramer's V and others.

The system has interfaces to HaploView (from the Broad Institute), PLINK (from the Massachusetts General Hospital) together with generic methods for integrating Perl, other command line tools, R and SAS scripts.

Interpretation

There are many different ways of bringing extra context and biological information into the interpretation of the statistical results. There are several out-of-the-box workflow services supplied within the solution including:

- Bio WebServices annotation, via the NCBI eUtils
- SNP Gene finder
- SNP visualisation in UCSC Genome Browser
- SNP retrieval from UCSC Genome Browser
- SNP to Gene and Ontology annotation (see figure 5)

Together with other products running on the InforSense Platform, such as ClinicalSense, BioSense, TextSense and the Pathway Connector, it is possible to bring in a wealth of extra information for helping understand the mechanisms of disease.

CASE STUDY: PARKINSON'S DISEASE

Parkinson's disease, or PD, is caused by a reduction in nerve cells from the *substantia nigra* part of the brain, which is responsible for dopamine production. The disease causes a reduction in dopamine which leads to motor and coordination difficulties for the patient. It affects around 1 in 500 of UK population with symptoms typically appearing from the age of 50 onwards. It is not known why the disease occurs, but there are already some genes linked to the disease, such as the PARKIN genes with the familial form of the disease. There is a large amount of research underway to further understand the role genetics plays in the development of Parkinson's.

This case study uses data from the SNP Database at the NINDS Human Genetics Resource Center DNA and Cell Line Repository. This is publicly available at http://ccr.coriell.org/ninds/. The original genotyping was performed in the laboratory of Drs. Singleton and Hardy (NIA, LNG), Bethesda, MD USA.

They used the Illumina Infinium I and Infinium II assays to carry out the genotyping, which gave a combined set of 408,803 unique SNPs. There were 271 controls [1] and 270 individuals who had idiopathic Parkinson's Disease [2]. The data was provided in two sets, one for the control and one for the case, each broken down by chromosome. For import we merged the data into a generic sample major format containing both alleles.

The example workflow in figure 2 shows how the case and control data sets for Chromosome 10 were loaded into the system, using the 'Genotype Set File Importer' node. The format of the data was loaded using a the sample major format.

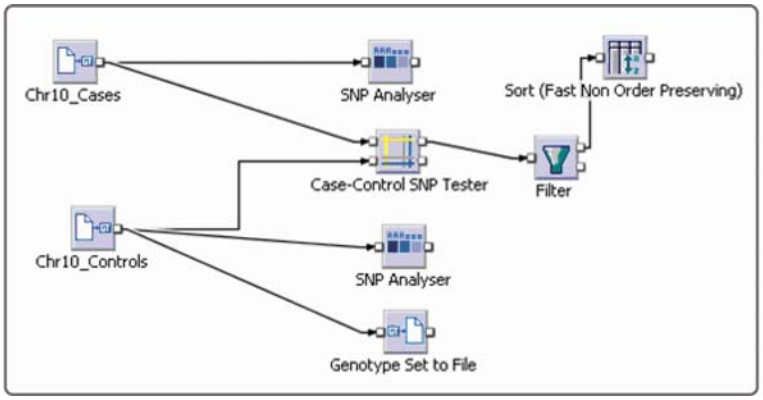


Figure 2: This figure shows a workflow taking the cases and controls for chromosome 10 and running them through a SNP Analyzer and Case-Control SNP Tester. The results are filtered and sorted.

In this example we calculate the MAF, Heterozygosity and Exact test for cases and controls together with other information using the SNP Analyzer node. The main test used in this case study was the Trend test calculated by the 'Case-Control SNP Tester'. We filter results based on a Trend Test, with a –log10 p-value of greater than 20. This resulted in a list of 42 highly significant SNPs.

Part of the output of the workflow is shown in figure 3. The SNPs are ranked according to the results of a Trend test.

SNP_ID	Trend_Test	Trend_A	Trend_D	Trend_R	Genotypic_Test	Allelic_Test	Dominant_Recessive_Test
rs1100828	33.6007	0.6886	0.4824	33.6007	0.4824	2.5259	1.5754
rs7898352	29.7235	12.491	9.7462	29.7235	9.7462	0.2691	12.1749
rs7901829	29.4951	0.4019	0.2827	29.4951	0.2827	3.5741	1.2053
rs1099795	28.8576	57.5755	21.2499	28.8576	21.2499	0.5763	0.1614
rs4746675	28.545	0.2906	2.3272	28.545	2.3272	4.2488	0.9572
rs1538564	28.0057	61.7527	12.907	28.0057	12.907	0.4519	0.324
rs303436	27.8155	8.6034	0.5258	27.8155	0.5258	2.0518	0.2336
rs2039709	27.7365	4.4082	0.8444	27.7365	0.8444	3.8269	1.5788
rs1181943	27.0786	26.928	11.998	27.0786	11.998	2.1113	0.5317
rs4917434	26.7665	0.1074	0.6922	26.7665	0.6922	2.4361	0.2191

Figure 3: This is the output table of the InforSense Platform workflow, showing the SNP ID together with the results of the Trend, Genotypic, Allelic and Dominant Recessive tests. All values are p-values with units of –log10

The complete data set can be filtered interactively using a GWA browsing tool. This allows the SNP test results to be shown in the context of genomic location together with other information, such as recombination rates and location of known genes. The Trend test data is shown in figure 4.

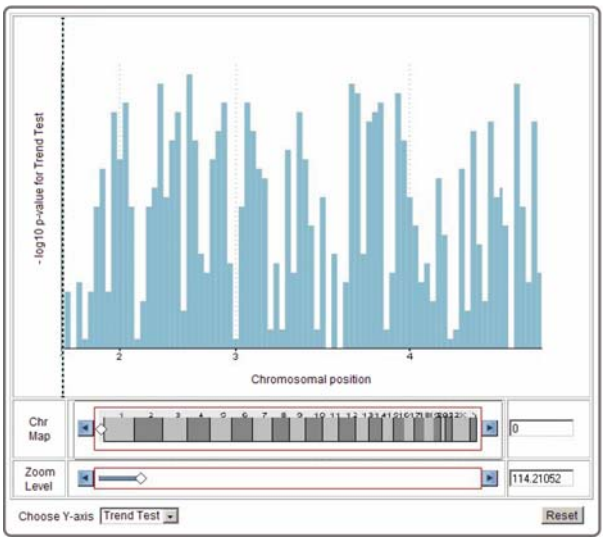


Figure 4: A plot looking at the p-values for Trend Test of all the SNPs across chromosomal position.

The top 10 hits were run through a 'SNP to Gene' annotation workflow, shown in figure 5 below.

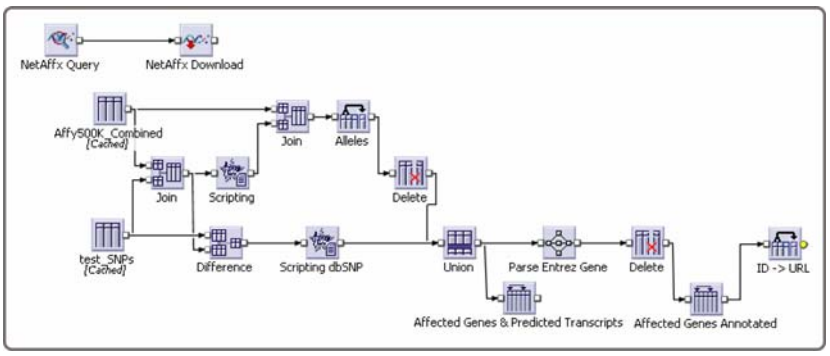


Figure 5: This workflow result uses the Affymetrix Probe Set ID or dbSNP ID to pull out information from either NetAffx annotations or dbSNP. Once there is information about the Allele, Chromosome and Topology etc, we use the gene symbol to further query EntrezGene and to build links to databases like UniGene, HPRD, OMIM and also the Gene Ontology.

Results of the first part of the 'SNP to Gene' workflow are shown in figure 6. Out of the top 10 hits we ran through the annotation workflow, we found five that were annotated as being associated to specific genes.

Hit Number	dbSNP ID	Chromosome	Alleles	Accession	SNP Topology	Gene Symbol	Gene ID
3	rs7901829	10	G/T	NM_018027	intron	FRMD4A	55691
4	rs10997954	10	C/T	NM_006258	intron	PRKG1	5592
5	rs4746675	10	A/C	NM_013266	intron	CTNNA3	29119
7	rs303436	10	A/C	NM_005204	intron	MAP3K8	1326
10	rs4917434	10	A/C	NM_014978	intron	SORCS3	22986

Figure 6: Five of the top 10 hits with gene symbol information and annotations after running the 'SNP to Gene' workflow which maps the RefSNP (rs) ID's to genes.

After a very preliminary investigation using workflows connecting to resources at the NCBI, using their WebServices, the following was noted for each of the SNP associated gene symbols:

- **FRMD4A** – FERM domain containing protein 4 (FRMD4A) is a cytoskeletal associated protein, involved in Cell growth and/or maintenance.
- **PRKG1** – PRKG1 protein kinase is involved in cell communication and signal transduction.
- **CTNNA3** – Alpha-T-catenin is a binding partner of beta-catenin (CTNNB1) which in turn interacts with PSEN1 that has many mutations which elevate A-beta-42 and cause early-onset familial Alzheimer's disease.
- **MAP3K8** – Mitogen-Activated protein kinase kinase kinase 8 is involved in Cell communication and Signal transduction. It is rarely associated with lung tumorigenesis.
- **SORCS** – these genes are known to be strongly expressed in the central nervous system.

SUMMARY AND CONCLUSIONS

After only a preliminary investigation of the results of the Parkinsonism study we identified 5 genes that are statistically associated with locus represented by SNPs whose signals were strongly associated with PD. Of these genes, two (CTNNA3 and SORCS) have been previously associated with disorders of the nervous system. Further and more detailed analysis is currently being carried out.

This paper has highlighted some of the unique capabilities of GenSense in how it can help you meet the data handling challenges that Genome Wide association studies can create. How the quality of that data can be assured in an easy, and automated way. Together, with the fact, that as the field develops, the analysis and data management tools also need to adjust. Here, using *ad hoc* custom workflows, it is easy to build and extend the system and also make use of the wealth of external information and existing tools that have been developed in the field of statistical genetics and beyond.

Genome Wide Association studies and the analysis thereof, will no doubt continue to produce valuable scientific results for several years to come.

REFERENCES

1. Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, Wavrant de Vrieze F, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A. Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals. Hum Mol Genet. 2007 Jan 1;16(1): 1-14

2. Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, Gibbs JR, Langefeld C, Stiegert ML, Schymick J, Okun MS, Mandel RJ, Fernandez HH, Foote KD, Rodriguez RL, Peckham E, De Vrieze FW, Gwinn-Hardy K, Hardy JA, Singleton A. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2006 Nov;5(11): 911-6

For more information
or a demonstration,
please call us today
to talk to one of
our consultants.

www.inforsense.com
information@inforsense.com

Europe:
InforSense Limited
Colet Court,
100 Hammersmith Road,
London, W6 7JP
United Kingdom
Phone: +44 (0)20 8237 8440
Fax: +44 (0)20 8237 8441

North America:
InforSense LLC
155 Second Street,
Cambridge, MA 02141
USA

Phone: +1 617 547 2500
Fax: +1 617 547 2772