



## **LarKC**

*The Large Knowledge Collider*

*a platform for large scale integrated reasoning and Web-search*

**FP7 – 215535**

---

# **D7b.3.1a Version 1 iteration report**

---

**Coordinator: Angus Roberts**

**With contributions from: Mattias Johansson, Paul Brennan,  
James McKay, Jon Wakefield, Yaoyong Li, Mark Greenwood,  
Thomas Heitz, Ian Roberts, Hamish Cunningham**

**Quality Assessor: Bosse Andersson**

**Quality Controller: Angus Roberts**

Document Identifier:	LarKC/2008/D7b.3.1a/V1.0
Class Deliverable:	LarKC EU-IST-2008-215535
Version:	version 1.0
Date:	October 19, 2009
State:	final
Distribution:	public



## EXECUTIVE SUMMARY

The Large Knowledge Collider (LarKC) project is building a platform for scaleable reasoning over terabytes of scientific data, using massive distributed incomplete reasoning. One of the use cases is carcinogenesis research. This has two scenarios, as described in LarKC Deliverable D7b1.1a *Requirements summary*. First, improved literature search is required to assist with the production of carcinogenesis reference works (known as Monographs). Second, literature knowledge mining is required to assist with predicting gene-disease associations in Genome Wide Association Studies (GWAS).

In the first 18 months of LarKC, we have built prototype software that uses LarKC to assist with the GWAS scenario. This is Version 1 of the use case software, and is described in LarKC Deliverable D7b.3.1b *Version 1 prototype*.

This document gives a report of the use of the prototype software. The report is provided as actual and draft research papers. First, one of several abstracts presented to international genetics audiences is provided. These shows that the software has use and credibility in the use case user community. Second, a draft journal paper is provided. This includes a quantitative evaluation of the prototype performance.

These papers replace the iteration evaluation report described in D7b.1.1b *Iteration evaluation methodology and report template*.



## DOCUMENT INFORMATION

<b>IST Project Number</b>	FP7 – 215535	<b>Acronym</b>	LarKC
<b>Full Title</b>	The Large Knowledge Collider: a platform for large scale integrated reasoning and Web-search		
<b>Project URL</b>	<a href="http://www.larkc.eu/">http://www.larkc.eu/</a>		
<b>Document URL</b>			
<b>EU Project Officer</b>	Stefano Bertolo		

<b>Deliverable</b>	<b>Number</b>	7b.3.1a	<b>Title</b>	Version 1 iteration report
<b>Work Package</b>	<b>Number</b>	7b	<b>Title</b>	Carcinogenesis reference production

<b>Date of Delivery</b>	<b>Contractual</b>	M18	<b>Actual</b>	19-Oct-09
<b>Status</b>	version 1.0		final	<input checked="" type="checkbox"/>
<b>Nature</b>	prototype <input type="checkbox"/> report <input checked="" type="checkbox"/> dissemination <input type="checkbox"/>			
<b>Dissemination Level</b>	public <input checked="" type="checkbox"/> consortium <input type="checkbox"/>			

<b>Authors (Partner)</b>	Mattias Johansson, Paul Brennan, James McKay (all International Agency for Research on Cancer); Jon Wakefield (University of Washington); Yaoyong Li, Mark Greenwood, Thomas Heitz, Ian Roberts, Hamish Cunningham, Angus Roberts (all University of Sheffield)			
<b>Resp. Author</b>	Angus Roberts		<b>E-mail</b>	a.roberts@dcs.shef.ac.uk
	<b>Partner</b>	University of Sheffield	<b>Phone</b>	+44 (114) 222 1917
















<b>Abstract (for dissemination)</b>	<p>Given advances in human genome sequencing, genetic testing, and the availability of samples from large population studies, it is now possible to carry out new types of study on the association between genes and diseases - Genome-Wide Associations Studies. In these, samples are tested from thousands of subjects with the disease in question, and thousands of disease-free controls. Each sample is tested with many hundreds of thousands of gene markers. If a marker is found more frequently in disease samples as opposed to control samples, then perhaps genes close to that marker are associated with the disease. Analysis of raw experimental data uses common statistical models to find the relevance of each marker, and rank them in order of relevance to the disease. Genes close to the top few markers are then studied in more depth. This last bit is expensive, and improving rankings could improve both the efficiency and the economics of the technique. Analysis could be improved if we incorporated knowledge we already have about genes - prior knowledge. Such knowledge is available e.g. in the vast numbers of research databases and research publications that now exist in the Life Sciences. The research papers attached to this document reports on, and evaluates, a prototype version of software to assist with this. The prototype software was produced in the context of the Large Knowledge Collider (LarKC) project.</p>
<b>Keywords</b>	Evaluation, Genome Wide Association Study, GWAS, carcinogenesis, Bayesian statistics, Bayesian False Discovery Probability, Single Nucleotide Polymorphism, SNP, GeneRIF



<b>Version Log</b>			
<b>Issue Date</b>	<b>Rev No.</b>	<b>Author</b>	<b>Change</b>
15/09/2008	1	Angus Roberts	Created document, front matter
15/07/2009	2	Angus Roberts	Summaries and outline
24/09/2009	3	Angus Roberts	Incorporate initial material
12/10/2009	4	Angus Roberts	Introduction and include abstract
16/10/2009	5	Angus Roberts	Version for review
19/10/2009	6	Angus Roberts	Review corrections



## PROJECT CONSORTIUM INFORMATION

Participant's name	Partner	Contact
Semantic Technology Institute Innsbruck, Universitaet Innsbruck	 	Prof. Dr. Dieter Fensel Semantic Technology Institute (STI), Universitaet Innsbruck, Innsbruck, Austria Email: dieter.fensel@sti-innsbruck.at
AstraZeneca AB		Bosse Andersson AstraZeneca Lund, Sweden Email: bo.h.andersson@astrazeneca.com
CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA		Emanuele Della Valle CEFRIEL - SOCIETA CONSORTILE A RESPONSABILITA LIMITATA Milano, Italy Email: emanuele.dellavalle@cefriel.it
CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O.		Michael Witbrock CYCORP, RAZISKOVANJE IN EKSPERIMENTALNI RAZVOJ D.O.O., Ljubljana, Slovenia Email: witbrock@cyc.com
Höchstleistungsrechenzentrum, Universitaet Stuttgart		Georgina Gallizo Höchstleistungsrechenzentrum, Universitaet Stuttgart Stuttgart, Germany Email : gallizo@hlrs.de
MAX-PLANCK GESELLSCHAFT ZUR FOERDERUNG DER WISSENSCHAFTEN E.V.		Dr. Lael Schooler, Max-Planck-Institut für Bildungsforschung Berlin, Germany Email: schooler@mpib-berlin.mpg.de
Ontotext AD		Atanas Kiryakov, Ontotext Lab, Sofia, Bulgaria Email: naso@ontotext.com
SALTLUX INC.		Kono Kim SALTLUX INC Seoul, Korea Email: kono@saltlux.com
SIEMENS AKTIENGESELLSCHAFT		Dr. Volker Tresp SIEMENS AKTIENGESELLSCHAFT Muenchen, Germany Email: volker.tresp@siemens.com
THE UNIVERSITY OF SHEFFIELD		Prof. Dr. Hamish Cunningham, THE UNIVERSITY OF SHEFFIELD Sheffield, UK Email: h.cunningham@dcs.shef.ac.uk
VRIJE UNIVERSITEIT AMSTERDAM		Prof. Dr. Frank van Harmelen, VRIJE UNIVERSITEIT AMSTERDAM Amsterdam, Netherlands Email: Frank.van.Harmelen@cs.vu.nl
THE INTERNATIONAL WIC INSTITUTE, BEIJING UNIVERSITY OF TECHNOLOGY		Prof. Dr. Ning Zhong, THE INTERNATIONAL WIC INSTITUTE Mabeshi, Japan Email: zhong@maebashi-it.ac.jp
INTERNATIONAL AGENCY FOR RESEARCH ON CANCER		Dr. Paul Brennan, INTERNATIONAL AGENCY FOR RESEARCH ON CANCER Lyon, France Email: brennan@iarc.fr
INFORMATION RETRIEVAL FACILITY		Dr. John Tait, INFORMATION RETRIEVAL FACILITY Vienna, Austria Email: john.tait@ir-facility.org



## TABLE OF CONTENTS

ABBREVIATIONS	7
1 INTRODUCTION	8
1.1 Summary of the use case . . . . .	8
1.2 Carcinogenesis use case iterations . . . . .	8
1.3 Substituted research papers . . . . .	9
REFERENCES	10
A ACCEPTED ABSTRACTS AND PRESENTED TALKS	11
B DRAFT PAPER FOR SUBMISSION	14



## LIST OF ABBREVIATIONS

<b>B FDP</b>	Bayesian False Discovery Probability
<b>GWAS</b>	Genome Wide Association Study
<b>IARC</b>	International Agency for Research on Cancer
<b>LarKC</b>	The Large knowledge Collider project
<b>SNP</b>	Single Nucleotide Polymorphism
<b>WHO</b>	World Health Organisation



## 1. Introduction

This deliverable reports on the development of software to support the LarKC WP7b Carcinogenesis use case. The software is developed iteratively over the life of the LarKC project. This report covers the first iteration.

The first iteration of the software itself has been delivered as *LarKC deliverable D7b.3.1b, Version 1 prototype* [1], and is described in the report accompanying that deliverable. The prototype software provides support for data analysis in Genome wide association studies (GWAS). When describing the software, the attached research paper makes certain assumptions, given audiences in genetics and bioinformatics journals. Simplified descriptions can be found in other deliverables. The use case is described in detail in *D7b.1.1a Requirements summary and data repository* [7], and a short, illustrated, description is given in *LarKC deliverable D7b.3.1b, Version 1 prototype* [1]. The following introduction to the use case is taken from the latter.

### 1.1 Summary of the use case

GWAS use bioprobes (SNPs - gene markers) to look for higher levels of association between genes in a diseased subjects as opposed to controls. The large numbers of markers mean that huge numbers of samples are needed to achieve sufficient statistical power.

Analysis of raw experimental data uses common statistical models to find the relevance of each marker, and rank them in order of relevance to the disease. Genes close to the top few markers are then studied in more depth. This last bit is expensive, and improving rankings could improve both the efficiency and the economics of the technique. Analysis could be improved if we incorporated knowledge we might already have about genes - prior knowledge. Such knowledge is available e.g. in the vast numbers of research databases and research publications that now exist in the Life Sciences, sometimes known as the data-ome and bibli-ome.

LarKC WP7b aims to apply LarKC technology to this problem, scaling knowledge discovery across the large amounts of biomedical knowledge now encoded in the data- and bibli-ome, and applying it to the millions of data points in a typical GWAS. We have prototyped a technique with the WHO's cancer research unit, IARC, to combine prior knowledge about a gene with experimental data, thus improving statistical power. The prototype uses early versions of LarKC plugins, and uses the LarKC data layer.

### 1.2 Carcinogenesis use case iterations

The general approach for reporting each use case prototype iteration is described in *D7b.1.1b Iteration evaluation methodology and report template* [6], which describes a system evaluation based on usability. By presenting research papers, we have departed from the usability evaluation plan. This departure is justified because:

- The usability evaluation gives a qualitative, user centric view. In this iteration, the software produces quantitative data. This can therefore be evaluated quantitatively - as is done in the attached papers.



- The usability evaluation assumes a large amount of interaction between the user and an interface. In this iteration, there is very little end user software. The interface is minimal. The software delivers data sets to the end-user, which they then manipulate in their existing tools.
- Usability is shown by the papers presenting results from this iteration to end user conferences and workshops, given in Appendix A.

### 1.3 Substituted research papers

The substituted papers are included in the appendices to this report.

- Appendix A includes a list of papers presented to end-user conferences and workshops, and gives an example of one.
- Appendix B gives a draft journal paper.



## REFERENCES

- [1] A. Roberts, M. Greenwood, D. Damjanovic, H. Cunningham, T. Heitz, I. Roberts, Y. Li, M. Johansson, and J. McKay. D7b.3.1b version 1 prototype. Technical report, LarKC project deliverable, 2009.
- [2] M. Johansson, Y. Li, J. Wakefield, M. Greenwood, T. Heitz, I. Roberts, H. Cunningham, P. Brennan, A. Roberts, and J. McKay. Using prior information attained from the literature to improve ranking in genome-wide association studies. In *The American Society of Human Genetics (ASHG) 59th Annual Meeting*, Honolulu, Hawaii, USA, October 2009.
- [3] M. Johansson, Y. Li, J. Wakefield, M. Greenwood, T. Heitz, I. Roberts, H. Cunningham, P. Brennan, A. Roberts, and J. McKay. Using prior information attained from the literature to improve ranking in genome-wide association studies. In *International Genetic Epidemiology Society (IGES) 18th Annual Conference*, Kahuku, Hawaii, USA, October 2009.
- [4] Mattias Johansson. Genome-wide association studies using prior information to boost power. 3rd Annual Collaboration Meeting on Cancer Epidemiology of Northern Sweden, ACME-NS, January 2009.
- [5] Mattias Johansson. Genome-wide association studies using prior information to boost power. Open door workshop on Human Genome Sequence, Wellcome Trust, January 2009.
- [6] A. Roberts, H. Cunningham, and A. Funk. D7b.1.1b iteration evaluation methodology and report template. Technical report, LarKC project deliverable, 2008.
- [7] A. Roberts, K. Straif, J. McKay, M. Stetter, and H. Cunningham. D7b.1.1a requirements summary and data repository. Technical report, LarKC project deliverable, 2008.



## A. Accepted Abstracts and presented talks

This appendix gives details of abstracts for talks and posters presented at end-user conferences. These papers describe the underlying technique, of:

- Mining the literature to assign a prior probability to SNPs;
- Using a Bayesian model (BFDP) to rank SNPs.

The appendix gives, in the following pages, the abstract of a paper presented at The American Society of Human Genetics (ASHG) 59th Annual Meeting [2].

Similar material has also been presented at:

- International Genetic Epidemiology Society (IGES) 18th Annual Conference [3]
- 3rd Annual Collaboration Meeting on Cancer Epidemiology of Northern Sweden, ACME-NS [4]
- Open door workshop on on Human Genome Sequence, Wellcome Trust [5]

# Using prior information attained from the literature to improve ranking in genome-wide association studies

Mattias Johansson<sup>1</sup>, Yaoyong Li<sup>2</sup>, Jon Wakefield<sup>3</sup>, Mark Greenwood<sup>2</sup>, Thomas Heitz<sup>2</sup>, Ian Roberts<sup>2</sup>, Hamish Cunningham<sup>2</sup>, Paul Brennan<sup>1</sup>, Angus Roberts<sup>2</sup>, James McKay<sup>1</sup>

1) International Agency for Research on Cancer (IARC), Lyon, France

2) Department of Computer Science, University of Sheffield, Sheffield, UK

3) Departments of Statistics and Biostatistics, University of Washington, Seattle, USA

Advances in high-throughput genotyping have made it technically possible to analyze hundreds of thousands of single nucleotide polymorphisms (SNPs) across the whole genome. Using this technology it is now feasible to conduct genome-wide association studies (GWAS) aiming to investigate the majority of common genetic variation and relate it to some phenotypic differences, often to risk of some disease. Whilst the price of GWAS assays are decreasing rapidly, conducting a GWAS is still a very expensive exercise, typically requiring genotyping several thousands of subjects at several hundreds of euros per sample in order to gain sufficient statistical power to distinguish the true association signals from the background noise.

Recognizing that a large proportion of GWAS findings reside near potential candidate genes for many of the investigated phenotypes, we here explore means to incorporate prior information attained from the literature to improve ranking in GWAS. We use this information to assign a crude prior probability of association for each SNP. The prior probabilities are thereafter integrated with the association result from the GWAS and the SNPs are re-ranked according to Bayesian false-discovery probability (BFDP). We show that this methodology improves the ranking for many known susceptibility loci with examples from studies on lung cancer and cancer of the upper aero digestive tract (UADT), see [Table 1](#). We have implemented this methodology in a web application where a user can specify a list of keywords and receive priors for all SNPs of interest. These priors can thereafter be used to rank the SNPs according to the BFDP.

**Table 1. Comparison of p-value-based and BFDP-based ranking for SNPs previously robustly implicated in lung cancer**

SNP	Proportion of data sampled	Median rank (power to rank SNP within the top 100)	
		P-value	BFDP
rs1051730	50%	959 (17%)	793 (18%)
Endpoint: Lung cancer	75%	10 (80%)	8 (81%)
15q25.1	100%	2	2
rs2736100	50%	17989 (3%)	1350 (16%)
Endpoint: Lung cancer	75%	2359 (4%)	222 (31%)
5p15.33	100%	77	8
rs3117582	50%	20033 (3%)	1038.5 (13%)
Endpoint: Lung cancer	75%	2717 (6%)	184.5 (35%)
6p22.33	100%	124	10
rs401681	50%	25446 (2%)	1866 (10%)
Endpoint: Lung cancer	75%	2775 (8%)	249 (32%)
5p15.33	100%	74	6
rs4324798	50%	7495 (3%)	6178 (3%)
Endpoint: Lung cancer	75%	844.5 (25%)	545 (28%)
6p22.1	100%	4	4
rs8034191	50%	502 (24%)	425 (28%)
Endpoint: Lung cancer	75%	4 (87%)	3.5 (89%)
15q25.1	100%	1	1



## B. Draft paper for submission

This appendix gives the text of a draft paper, to be submitted to BMC Bioinformatics. Note that the format is as required for submission to this journal.

The draft is awaiting further experimental data, duplicating the results with a second data set, before completion.

# Using prior knowledge to boost the statistical power of genome-wide association studies

Email:

\*Corresponding author

## Abstract

---

**Background:** Genome-wide association studies use Single Nucleotide Polymorphism markers to look for higher levels of association between genes in a diseased subjects as opposed to controls. The large numbers of markers mean that large numbers of samples are needed to achieve sufficient statistical power. Analysis of raw experimental data uses common statistical models to find the relevance of each marker, and to rank them in order of relevance to the disease. Genes close to the top few markers are then studied in more depth. Further study of high-ranking genes is expensive, and improving rankings could improve both the efficiency and the economics of the technique. Analysis could be improved if we use knowledge we might already have about genes - prior knowledge. Such knowledge is available in the many research databases and research publications that exist in the Life Sciences. The knowledge can be incorporated with experimental data using a Bayesian False Discovery Probability model. In order to harness this knowledge in a flexible, extensible and scaleable manner, we need an experimental platform that can mine the natural language literature, that can link that literature to multiple other data sources, and that can select and reason over the linked whole. Such a data layer and platform is provided by the LarKC project.

**Results:** We show that using prior knowledge mined from the research literature improves the ranking of known markers in two genome-wide association study datasets. We discuss how the use of prior knowledge can be extended using the LarKC platform.

**Conclusions:** Using prior knowledge in the analysis of genome-wide association study data brings improvements to statistical power. The technique has the potential to find new gene-disease associations, with cost and efficiency savings over existing techniques.

---

## Background

In the recent past, studies of gene disease association either concentrated on looking at those genes in particular families susceptible to the disease, or at those genes for which we had some strong hypothesis based on prior knowledge. This is problematic, as the search for a gene is based on the availability of specific family groups, and on the scientist's own preconceptions and biases.

Given advances in human genome sequencing, genetic testing, and the availability of samples from large population studies, it is now possible to carry out new types of study in the association between genes and diseases - Genome Wide Association Studies (GWAS). In these, samples are tested from thousands of subjects with the disease in question, and thousands of disease-free controls. Each sample is tested for many hundreds of thousands of Single Nucleotide Polymorphisms (SNPs), which are used as probes or markers. The relative frequency of a marker in disease and controls, then Odds Ratio (OR) is measured. If a marker is found more frequently in disease samples as opposed to control samples, then perhaps genes close to that marker are associated with the disease. The emphasis is on genome-wide evidence, and less on the biased selection of an individual researcher.

Of course, analysis of raw experimental Odds Ratios is not quite so simple as a straightforward frequency measure. Common statistical models are used to find the significance of each marker, and to rank them in order of relevance to the disease. Figure 3 shows the results of such a study. The horizontal axis gives position on the human genome. The vertical axis gives log of significance. Each dot represents a single SNP. Those above the threshold line are the ones considered significant enough to warrant further investigation. These markers turned out to be clustered near two genes that are now shown to be associated with lung cancer. Genes close to the top few markers are studied in more depth. This last bit is expensive, and improving rankings could improve both the efficiency and the economics of the technique. We have developed a technique to incorporate knowledge we might already have about genes - prior knowledge - into the analysis. For example, if we are studying lung cancer, and if we already know that a marker is close to a gene expressed in lung tissue, then we could boost the ranking of that marker. Such prior knowledge can be found in the vast numbers of research databases and research publications that now exist in the Life Sciences.

In order to implement this technique, we have two requirements. First, for each SNP, we need to be able to find the prior knowledge that is relevant to the disease in question. Second, we need a model with which to integrate the prior knowledge with experimental data from the GWAS.

We have met these requirements by building a prototype for the LarKC platform [6]. We have met the first requirement through the use of Linked Life Data (LLD) - a data schema and repository implemented on the LarKC project data layer [7]. In our prototype, we use the presence of keywords in curated literature descriptions, linked via LLD, as an indication of prior knowledge. We have implemented the second requirement using a Bayesian model to combine prior probability with new evidence from experimental GWAS data. We use Bayesian False Discovery Probability (BFDP) [1,2], which is explained in the next section. The BFDP model has been implemented as plugins for the LarKC platform.

The next section describes our methods, and explaining the way in which prior knowledge is found and detailing the Bayesian model used. Following this, we give results for experiments using a simple prototype, showing that the technique has greater statistical power than simple significance tests, and that it is capable of finding significant genes with less data. The Discussion describes a web service and

interface for calculating prior probabilities, looks at how the technique can be further refined, and how it may be extended using the LarkC platform. Finally, we conclude by stating the benefits of the method.

## Methods

This section describes the methods used to boost the statistical power of GWAS, using prior knowledge. GWAS looks for the genes relevant to one particular disease. In practice, the genotyping technique used detects differences at SNP positions in the genomes of the case and the control. The GWAS is therefore testing the relevances of SNPs to a disease. This in turn gives the genes covering the same position as the SNP, or in the region of the SNP. When using prior knowledge, we therefore set priors for each SNP. Prior knowledge is combined with experimental data using a Bayesian model. This section describes the Bayesian model first, followed by a description of how prior knowledge is obtained. Next, we discuss how the prior knowledge is used to assign a probability for the Bayesian model. Finally, we examine the experimental GWAS data used.

### Bayesian False Discovery Probability

In order to integrate prior knowledge from research databases and literature, with experimental data from a GWAS, we adopt a Bayesian approach. Bayesian inference computes a new (or posterior) probability for a hypothesis by updating the prior probability (or the prior) using new evidence or observation. Our priors are computed from previous information, which may exist in the literature, databases and other sources. These priors are updated using the new evidence provided by experimental data from a GWAS. In the experiments described in this paper, information from biomedical literature is used to determine the priors. Bayesian False Discovery Probability (BFDP) is defined in [2]. We computed a BFDP for each SNP in a GWAS data set. The smaller a BFDP is, the more likely the corresponding SNP is relevant to the disease. Summarising from [2], we assume that the prior is  $\pi_0$ . The prior odds is then defined as

$$PO = \frac{\pi_0}{1 - \pi_0} \quad (1)$$

Then BFDP is computed as

$$BFDP = \frac{ABF \times PO}{ABF \times PO + 1} \quad (2)$$

where ABF is the Approximate Bayes Factor. Bayes factor (BF) is a ratio of the conditional probabilities of the observed data  $\gamma = (\gamma_1, \dots, \gamma_n)$  respectively on the null hypothesis  $\mathcal{H}_0$  and the alternative hypothesis  $\mathcal{H}_1$ , i.e.

$$BF = \frac{p(\gamma|\mathcal{H}_0)}{p(\gamma|\mathcal{H}_1)} \quad (3)$$

ABF is an approximation of BF by using logistic regression on the observed data and some normalisation and independence assumptions (for details see [2]). Therefore the ABF is computed solely from the observed data and BFDP is a combination of ABF and the prior.

### Finding prior knowledge

We assign priors to SNPs, for the above BFDP calculation, based on the occurrence of prior knowledge in research databases and literature. This prior knowledge is retrieved for each SNP, from Linked Life Data (LLD) [8] as implemented in the LarKC data layer. This is shown in Figure 4, and described below.

Several genes may be in the region of a gene marker. We retrieve the IDs of all genes within a certain distance of the marker (100 000 base pairs in all reported experiments). Each gene has structured knowledge associated with its ID, in several different knowledge sources incorporated in LLD. Some of these knowledge sources also contain references to research papers of relevance to the gene. We collect a set of key terms of relevance to the disease, chosen by a domain expert. E.g. for lung cancer, these might be words such as “lung”, “cancer”, “tobacco”. We search for these key terms in the associated knowledge and research papers, and assign a prior based on the presence or absence of these key terms. Key term search and prior assignment is implemented as plugins for the LarKC platform. The precise details of prior assignment are given in the next section.

In the prototype, we use a single knowledge source, Entrez Gene <sup>1</sup>. Given a SNPs position, this is used to determine the genes within 100 000 base pairs of a SNP. Each gene in Entrez Gene may have associated with it short pieces of curated text, GeneRIFs. Each of these GeneRIFs describes some functional aspect of the gene, as derived from a single research paper, to which the GeneRIF is linked. In the prototype, prior knowledge is derived by searching these GeneRIFs for key terms. Key terms were split into three groups:

- **Group A:** terms referring to the organ that the disease resides in (e.g. lung for lung cancer).
- **Group B:** terms referring to etiological factors, such as smoking for lung cancer. This group is also used for general terms such as “genome-wide association study”.
- **Group C:** mechanistic terms such as “DNA damage” and “genetic disease”.

### Setting the priors

For each SNP we want to evaluate  $Pr(H_0|y, z)$  where  $H_0$  is the null of no association,  $y$  is the GWAS data (ie the confidence interval), and  $z$  is the search information. The posterior odds of  $H_0$  is given by:

$$PosteriorOdds(y, z) = BF(y)BF(z)PriorOdds \quad (4)$$

where  $BF(y)$  and  $BF(z)$  are the Bayes factors based on the GWAS and search information which (crucially) we have assumed are independent. The decision rule is then to reject the null hypothesis if the Posterior Odds are less than  $\frac{L_2}{L_1}$  where  $L_2$  is the loss of a Type II error (false positive) and  $L_1$  is the loss of a Type I error (false negative).

We can also write:

$$PosteriorOdds(y, z) = BF(y) \frac{Pr(H_0|z)}{P(H_1|z)} = BF(y)PosteriorOdds(z) \quad (5)$$

We can evaluate  $PosteriorOdds(z)$  in the following way.

---

<sup>1</sup><http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>

$$PosteriorOdds(z) = \frac{P(H_0|z)}{P(H_1|z)} \quad (6)$$

For each SNP, we assign a prior class according to the existence of prior knowledge. In our prototype case, this is the existence of key terms from our three groups **A**, **B**, **C** in the GeneRIF text. For each SNP, this gives  $z$  as one of one of the following eight classes, based on the presence of key terms from the three groups:

- $C_1$  = none of the three key term groups.
- $C_2$  = group A only
- $C_3$  = group B only
- $C_4$  = group C only
- $C_5$  = group A and group B only
- $C_6$  = group A and group C only
- $C_7$  = group B and group C only
- $C_8$  = all three groups.

So now we need to evaluate

$$PosteriorOdds(C_j) = \frac{P(H_0|C_j)}{P(H_1|C_j)} = \frac{P(C_j|H_0)P(H_0)}{P(C_j|H_1)P(H_1)} \quad (7)$$

where  $j = 1 \dots 8$

For the denominator we could think about how many SNPs we believe to be truly associated, and then how many fall in each category. For example, if a prior guess at the number of non-null SNPs in an experiment is 200, and we have 500 000 SNPs,

$$P(H_1) = \frac{200}{500000} \quad (8)$$

Further, if we guess the numbers in each category to be:  $\{100, 20, 0, 0, 20, 20, 0, 10\}$  then we have

$$P(C_1|H_1) = \frac{100}{200} \dots P(C_8|H_1) = \frac{10}{200} \quad (9)$$

For the numerator  $P(C_j|H_0)$  could we approximate by  $P(C_j)$  which could be based on the number of overall searches you find for all SNPs, ie if 95% out of 500 000 SNPs have  $C_1$  only then  $P(C_1|H_0)$  is approximately 0.95. Or, we can calculate more exactly using the identity:

$$Pr(C_j) = Pr(C_j|H_0)Pr(H_0) + Pr(C_j|H_1)Pr(H_1) \quad (10)$$

which can be rearranged to give:

$$Pr(C_j|H_0) = \frac{Pr(C_j) - Pr(C_j|H_1)Pr(H_1)}{Pr(H_0)} \quad (11)$$

To simplify, we merged prior classes to give five classes, denoted  $C_{m1...5}$  below:

- $C_{m1} = C_8$
- $C_{m2} = C_5|C_6$
- $C_{m3} = C_7$
- $C_{m4} = C_2|C_3|C_4$
- $C_{m5} = C_1$

These five levels are used throughout this paper unless stated otherwise.

### Experimental data

The genotyping data used was produced by the International Agency for Research on Cancer (IARC) Central Europe Lung Cancer Study, as reported in [3,5]. The data consists of 310023 SNPs over 2971 cases and 3746 controls. The association between the SNP and disease risk was estimated with Odds Ratio (OR) and the 95% confidence interval (CI), using a multivariate unconditional logistic regression and assuming a co-dominant genetic model (the effect of the variant by log-additive model with 1 degree of freedom). We used PLINK25 for this calculation. Study matching variables of age, sex, and country of recruitment were included in the regression as covariates. See [3,5] for more details about the statistical analysis of the experimental data.

The only difference between our experiments and the results presented in [5] is that we used prior knowledge from the literature combined with genotyping data, while [3,5] only used genotyping data. When finding prior knowledge, we excluded GeneRIFs from literature published after the reference article (03 April 2008). This ensures that evidence found from the original GWAS experiment, with which we are comparing our method, does not become prior knowledge for our method. As GeneRIFs have no associated date, we used MEDLINE to find the publication date of the GeneRIFs' associated articles. We adopted a conservative strategy for those MEDLINE entries for which publication dates were not complete. For example, if a publication date had only month mm and year yy, we set the date as 31/mm/yy.

We used 11 key terms which are related to lung cancer, partitioned into three groups as explained above:

- **Group A:** *lung*
- **Group B:** *smoking, tobacco, genome wide association study, carcinogen, adenocarcinoma, non-small cell carcinoma*
- **Group C:** *DNA repair, DNA damage, genetic disease, genetic trait*

## Results

The section presents the experimental results of applying BFDP to the lung cancer study data from [3]. We will first show the SNP rankings produced by BFDP where all the SNPs were assigned the same prior, and where no key term groups or search were used. This is shown in Table 0.1. We can see that if we do not use any keyword at all and set the prior to be the same for each SNP, the BFDP ranks are similar to the P-values based ones. BFDP therefore gives little improvement over P-value in this case.

Next, we show the effect of assigning different priors in the BFDP model, using the keyword groups from the previous section. These results are shown in Table 0.2, for six SNPs which are known to have a highly probable association with lung cancer. Results for the full data set are shown in the 100% row for each SNP. SNP rankings based on P-Value and BFDP are shown. The BFDP with priors ranks all of these SNPs, which we know have an association with lung cancer, at either the same rank or higher than the typically used P-Value. BFDP ranks all six within the top 10 ranks, whereas P-Value ranked three of the six SNPs at ranks 74, 77 and 124. BFDP is more able to find the association between these three SNPs and lung cancer.

### Varying the size of the genotyping dataset

We ran experiments varying the size of the genotyping dataset. We produced two smaller data sets by sampling a three-quarters and half of the cases and controls, respectively. The two smaller data sets are denoted as *75%* and *50%* in Table 0.2. In each case, we examined the power of both BFDP and P-Value to produce the same ranking as the 100% data set. It is clear that BFDP has a better power to rank SNPs than P-value, even with half as much data.

### Different keyword lists

We also ran some experiments to check the sensitivity of results to the keyword lists. The original keyword list contains 11 terms, as discussed in the previous section. We carried out experiments with different keyword lists, as follows:

- **Experiment A** used six terms in three groups
  - **Group A** *lung*
  - **Group B** *carcinogen, non-small cell carcinoma*
  - **Group C** *DNA repair, genetic disease*
- Experiment B used three terms in three groups:
  - **Group A** *lung*
  - **Group B** *smoking*
  - **Group C** *genetic disease*
- Experiment C used just one keyword, in one group: *lung*.

The ranks of the six SNPs in the three experiments are presented in Table 0.3.

We can see that using a smaller keyword list changes the ranking. However, even just using “lung” as keyword in the Experiment C, the ranks were still better than those from P-value.

## Discussion

The results in the previous Section show that BFDP using prior knowledge can better rank known gene-disease associations, and that it has greater statistical power than P-Value. In this Section, we look at how the selection of prior knowledge might be further improved. We then describe how implementing the technique on the LarKC platform gives us the flexibility to examine other sources of prior knowledge. Finally, we present a prototype web interface to a GWAS BFDP service, based on LarKC software.

### Further improving the selection of prior knowledge

We have considered several extensions to the technique, to improve both performance and theoretical soundness. We present these below.

It is argued that one advantage of GWAS over other techniques, is that there is no selection bias. The investigator has not selected genes for examination using any preconceptions. By allowing the investigator to select a set of keywords, however, it could be argued that we have re-introduced selection bias. It would be straightforward to amend the BFDP with prior knowledge technique to remove this bias. We can envisage an investigator inputting only the genetic endpoint that they are investigating - such as *lung cancer*. We would use this as a seed to find other relevant key terms for prior knowledge mining, using for example a term frequency measure such as TFIDF. Further to this, we could assign weights to key terms based on their frequency, and use these to create a continuous prior, as opposed to the current set of discrete levels.

We also make no distinction when calculating the prior as to where the keywords appear, and the frequency of keywords. There is clearly much scope for experimentation.

We currently use a fixed window size of 100 000 base pairs, selecting prior knowledge for all genes within this distance of each SNP. The strength of association of a SNP to a gene is not, however, linear. In some cases, distant SNPs may be more closely related to a gene than in other cases. This strength can be measured with *Linkage Disequilibrium* (LD). We could use LD to vary window size.

It is possible that the technique falsely raises the prior of SNPs that fall distant from, but between, two genes by including them within the window for both. This would give a higher prior than a SNP that occurs inside one or other of the genes. A simple solution would be that the prior of a SNP is simply the maximum of the priors calculated for each gene in its window.

### Using other sources of prior knowledge

The reported experiments use a very simple model of prior knowledge: keywords found in manually curated text. One advantage of using the LarKC data layer, is that multiple other sources can be included, with unified schemas. The data layer also has scope for storing *semantic annotations* – linking of terms in

text to concepts in other knowledge sources. We can envisage much richer prior knowledge sources that involve such conceptual markup and selection over multiple life science data sources and ontologies. For example, we might want to assign a higher prior, when searching for “kidney”, if we were to find that a gene is expressed in some anatomical structure that has a part-whole relation to kidney. This would clearly involve handling multiple relationships in multiple data sources. Similarly, what if we need to assign a higher prior if a gene is involved in a pathway known to be important in the metabolism of a component of a carcinogen?

We illustrate these types of prior knowledge search in Figure 5. The intention of using the LarKC data layer and platform, is to give us the flexibility, adaptability and scalability to move beyond our current key term model, to examine these richer searches.

### **A web interface for BFDP calculation**

We have implemented the current method using LLD and LarKC plugins. We have provided a simple web interface with which GWAS investigators can access the service. This is shown in Figures 1 and 2. Figure 1 shows the page through which an investigator can set up an analysis. The investigator provides a list of SNPs, or the name of a known set of SNPs (current sets include a set for all SNPs, and sets for some common micro-array chips). The investigator then sets keywords, and sets the experiment running. Figure 2 shows details of a running experiment.

### **Conclusions**

We have demonstrated a method for ranking SNPs in GWAS, incorporating prior knowledge about genes close to the SNP, with experimental data. When compared to P-Value, the method improves the ranking of SNPs from a test set already known to be associated with lung cancer.

The method finds prior knowledge by searching for key words in text associated with the SNP, via nearby genes. Prior knowledge is integrated with experimental GWAS data using a Bayesian model, BFDP, and a reasoned method for assigning priors.

We have implemented the method as a service using LLD and LarKC software, and provided a web interface. We are exploring ways in which further use can be made of the LarKC platform

Our method could bring significant cost savings to GWAS, by improving statistical power, this reducing the size of data required to identify new associations.

### **Authors contributions**

Paul Brennan and Hamish Cunningham conceived of the idea of using prior knowledge from the literature to give statistical power to GWAS analyses. Angus Roberts, Mattias Johansson and James McKay elaborated this to the current algorithm. Mattias Johansson provided data and analysis. Jon Wakefield provided the calculation of theoretical priors. Yaoyong Li wrote an initial prototype, contributed to experimental design, and wrote first drafts of the Methods and Results sections; Angus Roberts wrote first drafts of the other sections. Mark A. Greenwood and Angus Roberts wrote the current implementation of

the algorithm. Mark A. Greenwood and Thomas Heitz wrote the current interface. Ian Roberts provided data integration. All authors contributed to the writing of the paper.

## Acknowledgements

This work was funded by a EU Large-Scale Integrating Project, Number FP7 - 215535, LarKC - The Large Knowledge Collider.

## References

1. Wacholder S, Chanock S, Garcia-Closas M, El-ghormli L, Rothman N: **Assessing the probability that a positive report is false: an approach for molecular epidemiology studies.** *J Natl Cancer Inst* 2004, **96**(6):434–442.
2. Wakefield J: **A Bayesian Measure of the Probability of False Discovery in Genetic Epidemiology Studies.** *The American Journal of Human Genetics* 2007, **81**(2):208–227.
3. McKay JD, Hung RJ, Gaborieau V, Chabrier A, Byrnes G, et al: **Lung cancer susceptibility locus at 5p15.33.** *Nature Genetics* 2008, **40**(12):1404–1406.
4. Cunningham H: **Information Extraction, Automatic.** *Encyclopedia of Language and Linguistics, 2nd Edition* 2005, :665–677.
5. Hung RJ, McKay JD, Gaborieau V, Boffetta P, et al: **A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25.** *Nature* 2008, **452**:633–637.
6. Fensel D, van Harmelen F, Andersson B, Brennan P, et al: **Towards LarKC: a Platform for Web-scale Reasoning.** *Proceedings of the IEEE International Conference on Semantic Computing*, August 2008, Santa Clara, CA, USA.
7. Momtchev, V et al: **Prototype v1, LarKC project deliverable D7a.3.1**, 2009.
8. Momtchev V, Psychev D, Primov T, Georgiev G: **Expanding the Pathway and Interaction Knowledge in Linked Life Data.** Submitted to *International Semantic Web Challeng* 2009.

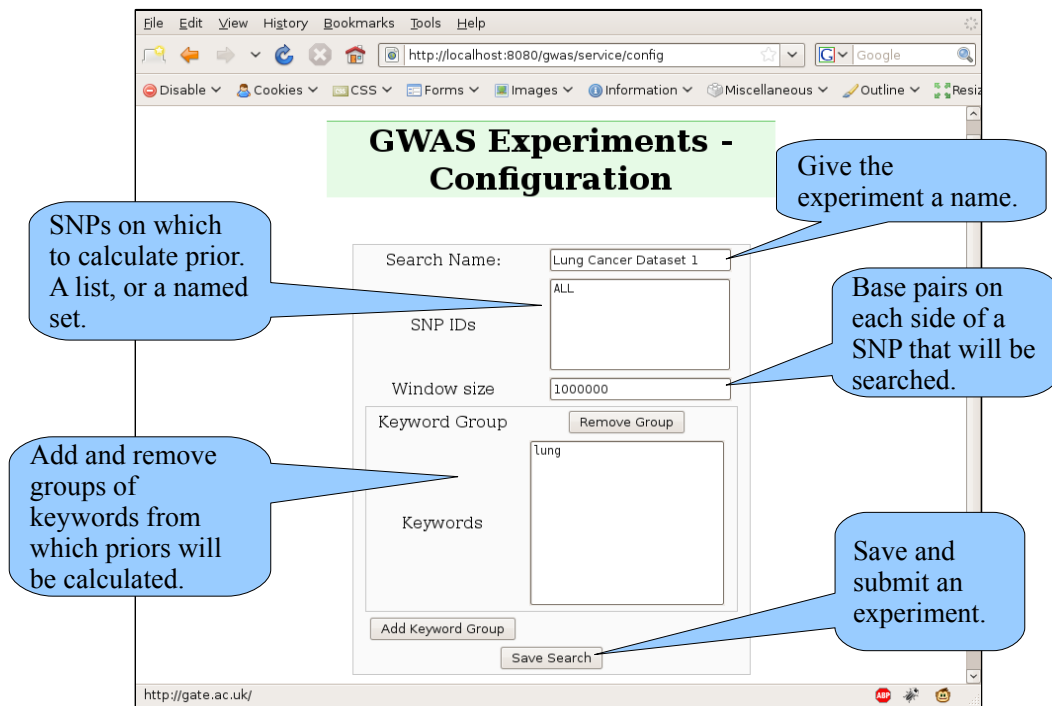


Figure 1: Configure an experiment.

**GWAS Experiments - Details**

**Lung Cancer Dataset 1** 1.7% complete

Number of SNPs: 283338  
Window Size: 1000000

**Keywords**

Group 1	Group 2	Group 3
lung	smoking carcinogen non-small cell carcinoma	DNA repair genetic disease

Prior	SNP count
0.3	211
0.1	4178
0.6	64
0.4	301
0.8	63

refresh | return to list

© 2009 The University of Sheffield - GATE Group.

Done

**Description.**

**Keyword groups.**

**Priors, and number of SNPs assigned these priors.**

**On completion, results are downloaded and combined with wet lab results to give BFDP.**

Figure 2: Examine experiment details.

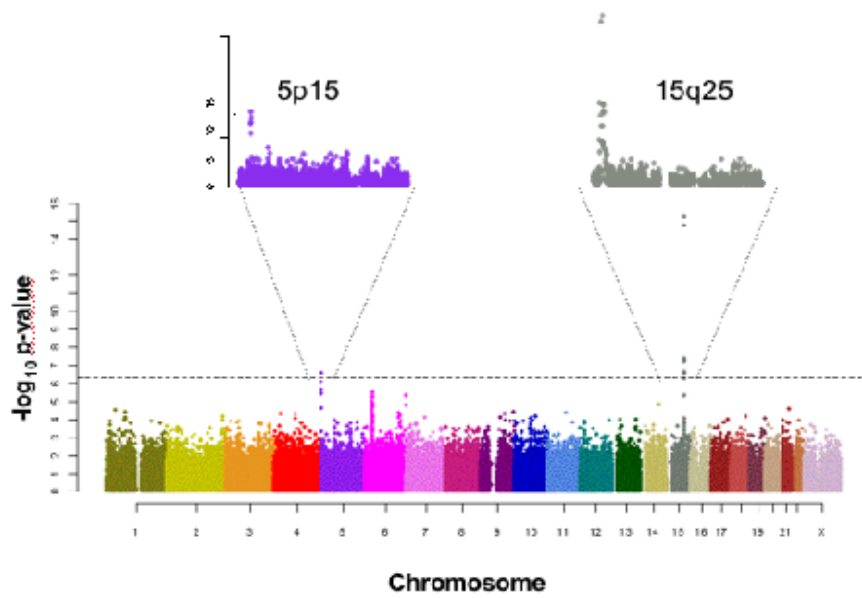


Figure 3: SNP significance plot of a genome-wide association study in lung cancer

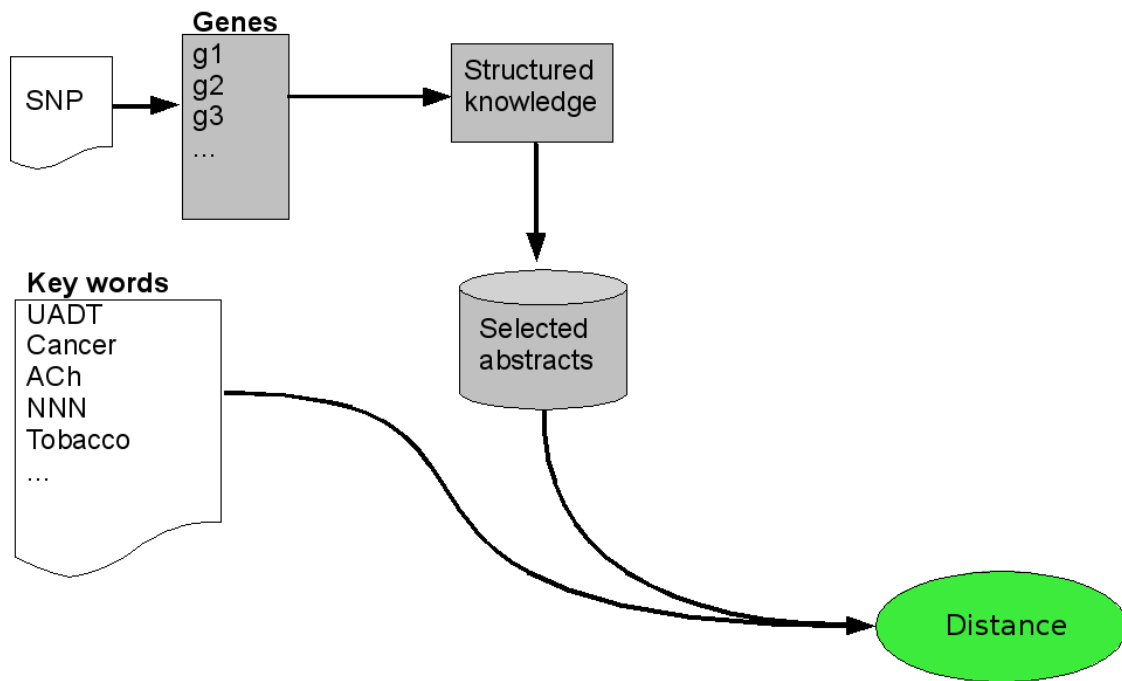


Figure 4: We can search for keywords in these abstracts and calculate the prior of a SNP using the presence of these keywords

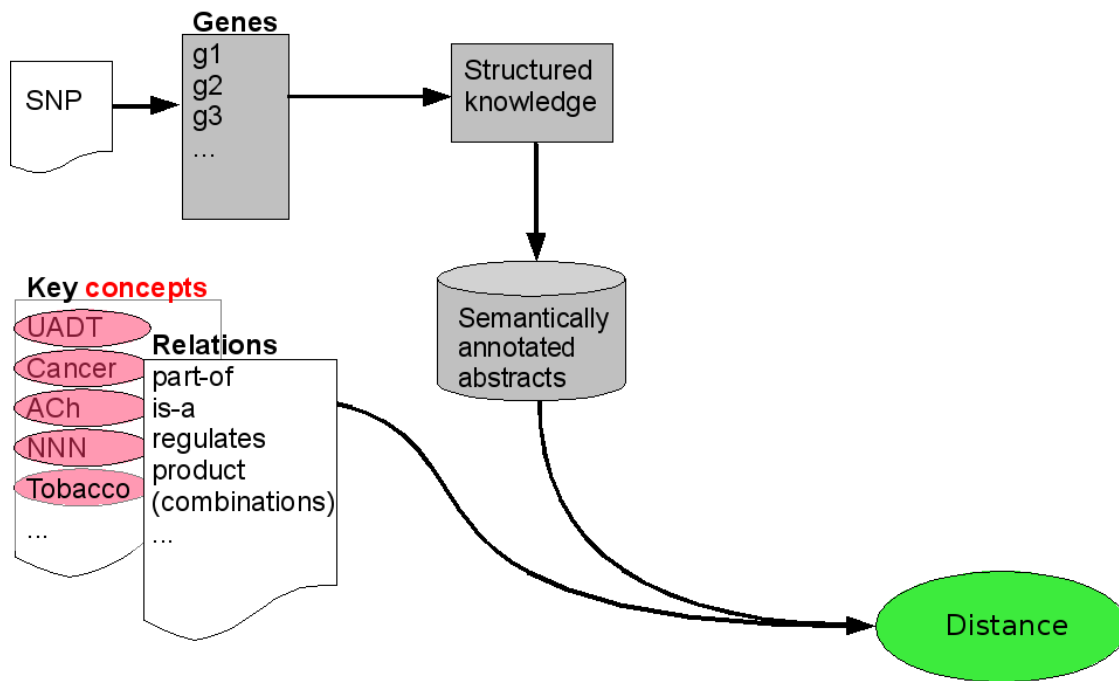


Figure 5: In the future: We could also include semantic and relational information from the associated knowledge sources

### 0.1 Table - fixed priors

The BFDP and the ranks for the six SNPs in the experiment where all the SNPs had the same prior 0.9.

SNP	Prior	BFDP	Rank
rs8034191	0.1	6.155E-6	1
rs1051730	0.1	5.541E-5	2
rs4324798	0.1	9.996E-4	3
rs3117582	0.1	0.051	67
rs2736100	0.1	0.065	75
rs401681	0.1	0.316	491

## 0.2 Table - comparison of P-Value and BFDP ranking

Comparison of p-value and BFDP based ranking for SNPs which have previously been robustly implicated in lung cancer.

SNP ID	Gene	Proportion of data samples	P-value		BFDP	
			Rank	Power	Rank	Power
rs1051730	15q25.1	100%	2	-	2	-
		75%	10	80%	8	81%
		50%	959	17%	793	18%
rs2736100	5p15.33	100%	77	-	8	-
		75%	2359	4%	222	31%
		50%	17989	3%	1350	16%
rs3117582	6p22.33	100%	124	-	10	-
		75%	2717	6%	184	35%
		50%	20033	3%	1038	13%
rs401681	5p15.33	100%	74	-	6	-
		75%	2775	8%	249	32%
		50%	25446	2%	1866	10%
rs4324798	6p22.1	100%	4	-	4	-
		75%	844	25%	545	28%
		50%	7495	3%	6178	3%
rs8034191	15q25.1	100%	1	-	1	-
		75%	4	87%	3	89%
		50%	502	24%	435	28%

### 0.3 Table - varying keywords

The BFDP based ranks of the six SNPs in the three experiments which use different keyword lists.

SNP	Experiment A	Experiment B	Experiment C
rs8034191	1	1	1
rs1051730	2	2	2
rs4324798	3	3	3
rs3117582	4	5	20
rs2736100	6	28	26
rs401681	27	51	158