BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 9 2007, pages 1132–1140
doi:10.1093/bioinformatics/btm001

Data and text mining

A computational system to select candidate genes for complex human traits

Kyle J. Gaulton[1,2,3,*], Karen L. Mohlke[3] and Todd J. Vision[4]

[1]Curriculum in Genetics and Molecular Biologly, [2]Bioinformatics and Computational Biology Training Program, Departments of [3]Genetics and [4]Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27516

Received on October 30, 2006; revised on January 2, 2007; accepted on January 8, 2007

Advance Access publication January 19, 2007

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Identification of the genetic variation underlying complex traits is challenging. The wealth of information publicly available about the biology of complex traits and the function of individual genes permits the development of informatics-assisted methods for the selection of candidate genes for these traits.

Results: We have developed a computational system named CAESAR that ranks all annotated human genes as candidates for a complex trait by using ontologies to semantically map natural language descriptions of the trait with a variety of gene-centric information sources. In a test of its effectiveness, CAESAR successfully selected 7 out of 18 (39%) complex human trait susceptibility genes within the top 2% of ranked candidates genome-wide, a subset that represents roughly 1% of genes in the human genome and provides sufficient enrichment for an association study of several hundred human genes. This approach can be applied to any well-documented mono- or multi-factorial trait in any organism for which an annotated gene set exists.

Availability: CAESAR scripts and test data can be downloaded from http://visionlab.bio.unc.edu/caesar/

Contact: kgaulton@email.unc.edu

1 INTRODUCTION

Unlike Mendelian traits, in which a mutation in one gene is causative, or oligogenic traits, where several genes are sufficient but not necessary, complex traits are caused by variation in multiple genetic and environmental factors, none of which are sufficient to cause the trait (Peltonen and McKusick, 2001). The contribution of any given gene to a complex trait is usually modest. In addition, complex traits often encompass a variety of phenotypes and biological mechanisms, making it difficult to determine which genes to study (Newton-Cheh and Hirschhorn, 2005).

As a result, traditional methods of genetic discovery, such as linkage analysis and positional cloning, while widely successful in identifying the genes for Mendelian traits, have had more limited success in identifying genes for complex traits. Candidate gene studies have had encouraging success, yet this approach requires an effective method for deciding a priori which genes have the greatest chance of influencing susceptibility to the trait (Dean, 2003). Recent advances in genotyping technology have provided researchers with the ability to test association in hundreds of genes relatively quickly, and even the entire genome through a genome-wide association study. Genome-wide association studies are promising, yet not always economically feasible or statistically desirable (Thomas, 2006). Therefore, one of the greatest challenges in disease association study design remains the intelligent selection of candidate genes.

To this end, we have developed a computational methodology, named CAESAR (CAndidatE Search And Rank), that uses text and data mining to rank genes according to potential involvement in a complex trait. CAESAR exploits the knowledge of complex traits in literature by using ontologies to semantically map the trait information to gene and protein-centric information from several different public data sources, including tissue-specific gene expression, conserved protein domains, protein–protein interactions, metabolic pathways and the mutant phenotypes of homologous genes. CAESAR uses four possible methods of integration to combine the results of data searches into a prioritized candidate gene list. In effect, CAESAR mimics the steps a researcher would undertake in selecting candidate genes, albeit faster, potentially more thoroughly, and in a more quantitative manner.

CAESAR represents a novel selection strategy in that it combines text and data mining to associate genetic information with extracted trait knowledge in order to prioritize candidate genes. In contrast to a number of existing approaches (Adie *et al.*, 2006; Turner *et al.*, 2003; van Driel *et al.*, 2003), gene selection is not limited to one or more genomic regions, as all genes annotated in one of our databases are potential candidates. CAESAR is ultimately designed for traits in which the relevant biological processes may not be well understood and potentially hundreds of reasonable candidate genes exist.

The potential benefits to a researcher in adopting a computational approach to gene selection such as CAESAR include the ability to quickly and systematically process several hundred thousand biological annotations, many of

*To whom correspondence should be addressed.

which require highly specialized domain expertise to interpret. This benefit will continue to grow in importance as the volume and technical detail of annotation data increases. Relevant gene annotations can easily escape human consideration due to biases that investigators bring to the task of prioritization and that are difficult to overcome even by conscious effort. This is particularly valuable for complex traits, which may be affected by a wider array of biological processes, some of which may not have been directly implicated by previous studies. CAESAR also reports the evidence supporting the prioritization rank of each gene, allowing an investigator to trace the line of reasoning and to exercise his or her own judgment as to its validity. Thus, it can be seen as a very sophisticated aid to manual prioritization.

Though designed to help with the design of an association study involving a few hundred genes, CAESAR can also be used to prioritize a smaller number of candidates within a region of linkage, or to prioritize among polymorphisms annotated with ranked genes that show significant association in a genome-wide study.

We have tested CAESAR on 18 susceptibility genes for 11 common complex traits in humans including type 1 and type 2 diabetes mellitus, schizophrenia, Parkinson's disease, cardiovascular disease, age-related macular degeneration, rheumatoid arthritis and celiac disease. Test genes were ranked higher than 95.7% of all ranked genes on average, and higher than 99.7% in the best case.

2 METHODS

CAESAR is comprised of three main steps. First, previously implicated genes mentioned in the input text are identified and ontology terms are ranked based on their similarity to an input text. Second, genes are ranked for each data source independently based on the relevance of the ontology terms with which they are annotated. Third, the individual gene lists are integrated to provide a single ranked list of candidate genes that combines evidence from all data sources. We refer to these three steps as text mining, data mining and data integration, respectively. The approach of CAESAR is presented as a schematic diagram in Figure 1a.

2.1 Text mining

CAESAR requires a user-defined body of text (referred to as a corpus) as input. This text is ideally an authoritative and comprehensive source of biological knowledge about the trait of interest. If an online Mendelian inheritance in man (OMIM) (Hamosh *et al.*, 2005) identifier is supplied, CAESAR will use the OMIM record as input. Alternately, the user can provide any other body of text, for instance one or more review articles.

Since the corpus is written in natural language, the information must be converted to machine-readable form. This is done in two ways. First, human gene symbols are identified within the corpus. If an OMIM record is used as input, gene identifiers can be extracted directly from the OMIM database. Otherwise, gene symbols are extracted by matching to a reference list. Genes are weighted based on frequency of occurrence in the corpus, f_g, where the weight c_g of extracted gene g is calculated as f_g divided by the sum of all f_g across n total extracted genes. The reference list of standard names, symbols, database identifiers and corresponding mouse homologs for each gene is compiled from Entrez Gene (Maglott *et al.*, 2005) and Ensembl (Birney *et al.*, 2006). The extracted genes are assumed to be relevant

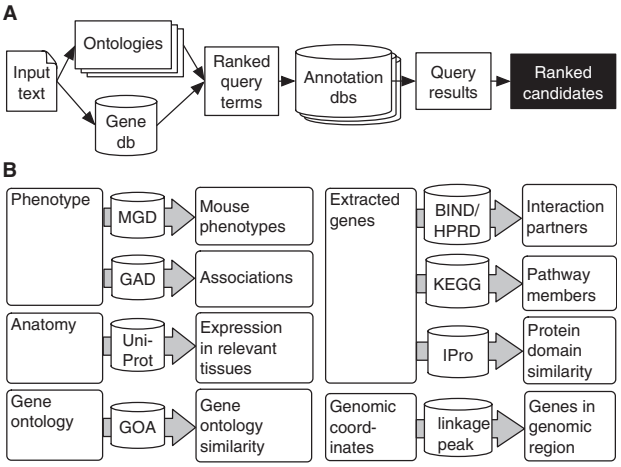

Fig. 1. CAESAR overview. (**a**) Text mining is used to extract gene symbols and ontology terms from the input. In the data-mining step, genes within each gene-centric data source are ranked based on the relevance to the trait-centric terms. In the data-integration step, the results from each source are combined into a single ranked list of candidates. Db = database. (**b**) Eight types of functional information (GO molecular function and biological process listed together) are queried using extracted genes and anatomy, phenotype and gene ontology terms. Genomic regions of interest represent optional user input. See text for abbreviations.

Table 1. Data sources and ontologies used in CAESAR

Source[a]	Version[b]	URL	Records	Content
Ontology				
MP	01/23/06	www.informatics.jax.org/	3850	Phenotype
eVOC	2.7	www.evocontology.org/	394	Anatomy
GO bp	01/23/06	www.geneontology.org/	9687	Function
GO mf	01/23/06	www.geneontology.org/	7055	Function
Database				
OMIM	01/23/06	www.ncbi.nih.gov/	16564	Disease
Gene	10/01/05	www.ncbi.nih.gov/	32859	Gene
Ensembl	37.35j	www.ensembl.org/	20134	Gene
SwissProt	48.8	www.ebi.ac.uk/uniprot/	13434	Expression
TrEMBL	31.8	www.ebi.ac.uk/uniprot/	57551	Expression
InterPro	12.0	www.ebi.ac.uk/interpro/	12542	Domain
BIND	10/01/05	www.bind.ca/	35661	Interaction
HPRD	10/01/05	www.hprd.org/	33710	Interaction
KEGG	41.0	www.genome.jp/kegg/	209	Pathway
MGD	3.41	www.informatics.jax.org/	7705	Phenotype
GAD	01/23/06	hpcio.cit.nih.gov/gad.html	8176	Association
GOA	01/23/06	www.ebi.ac.uk/goa/	27768	Function

[a]See text for abbreviations.
[b]Download date reported where version information is not available.

to the biology of the trait, but do not necessarily contribute to the genetic variation of the trait.

Second, the corpus is used to quantify the relevance of terms within several different biomedical ontologies. Four ontologies are used as part of CAESAR, the gene ontology biological process (GO bp) and molecular function (GO mf) (Harris *et al.*, 2004), the mammalian phenotype ontology (MP) (Smith *et al.*, 2005) and the eVOC anatomical ontology (Kelso *et al.*, 2003) (Table 1). Relevance is

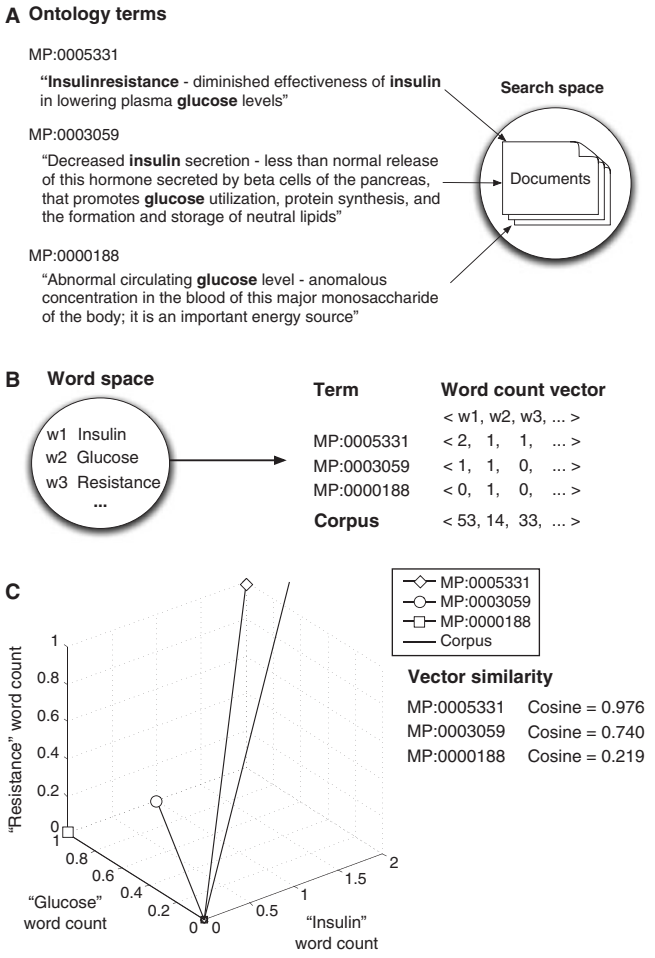

Fig. 2. Vector-space similarity search. (**a**) Each ontology term and its description comprise a document, as in this example of three terms from the mammalian phenotype ontology. (**b**) The word space consists of all unique words. For illustration, here the word space is ('insulin', 'resistance', 'glucose'). Each document, including the corpus, describes a vector in word space, where the elements of the vector are weighted counts within the document of each word in the word space. (**c**) The similarity of each of the documents to the corpus is measured as the cosine of angle formed by the document and corpus vectors. High-ranking ontology terms have document vectors that are similar in both direction and magnitude to the corpus vector. In this example, MP:0005331 is the highest-ranking document.

quantified using a similarity search under a vector-space model (Salton *et al.*, 1975), as follows (Fig. 2). For each ontology, the individual terms are split into separate documents containing the term name and term description if available. These documents together comprise a document database, or search space, against which the corpus is queried (Fig. 2a). The corpus and each document are converted to vectors $v_i =< w_{i1}, w_{i2}, ..., w_{in} >$ with dimensionality equal to the size of the word space n, which is the total number of unique words in the document database. Commonly used stop words such as 'and' and 'the' are removed from the word space. Each element of the vector for document i is calculated as $w_{ij} = e_{ij}$, where e_{ij} is the number of occurrences of word j in the document.

The similarity of the corpus to each document is calculated as the cosine of the angle between the vectors, which is equal to the dot product of the vectors divided by the product of the magnitudes of the vectors. A larger cosine indicates vectors with greater similarity. Using this measure, ontology terms are weighted based on their similarity to the corpus (Fig. 2c), where the weight c_t of term t is directly equal to the cosine.

2.2 Data mining

Eight sources of gene-centric information are used to map ranked ontology terms to the genes annotated with them (Fig. 1b). The resulting output is eight lists of gene scores, one for each functional category.

Mammalian phenotype ontology terms are used to query the mouse genome database (MGD) (Blake *et al.*, 2003) for genes producing a given phenotype when mutated and to query the genetic association database (GAD) (Becker *et al.*, 2004) for genes showing positive evidence of association with a phenotype in a human population. The eVOC anatomical ontology terms are used to query the UniProt database (Bairoch *et al.*, 2005) for genes expressed in a given tissue. Gene ontology terms are used to query the gene ontology annotation database (GOA) (Camon *et al.*, 2003) for genes annotated with a given gene ontology biological process or molecular function term. Finally, the extracted genes are used to query the biomolecular interaction network database (BIND) (Alfarano *et al.*, 2005) and the human protein reference database (HPRD) (Peri *et al.*, 2004) for genes encoding proteins that interact with the protein products of the extracted genes, query the Kyoto encyclopedia of genes and genomes (KEGG) pathway database (Kanehisa *et al.*, 2004) for other genes involved in the same human cellular pathways and query the InterPro protein domain database (IPro) (Apweiler *et al.*, 2000) for genes sharing conserved protein domains with the extracted genes.

The user may also optionally input one or several genomic sequence regions to include genes in chromosomal regions implicated through genetic linkage as an additional list of genes (Fig. 1b).

The score r_{ij} of gene i for source j is then calculated as either the maximum, sum or mean of the weights of the k matching ontology terms or extracted genes $c_1 \ldots c_k$. The three alternatives weigh the combined evidence for relevance in different ways, as described below for data integration from multiple sources.

2.3 Data integration

The gene scores from the eight sources are integrated to produce one combined score for each gene. Integration is accomplished using one of four methods. Each method represents a different approach that an investigator might choose when manually prioritizing candidate genes on the basis of evidence from several data sources.

The first three methods involve taking the maximum, sum or mean of the z-transformed r_{ij} scores for each gene. The maximum favors genes with strong evidence from one data source, the sum favors genes with evidence in many data sources and the mean favors genes with strong evidence only, penalizing genes with any weak evidence. The maximum mean and sum are referred to as int1, int2 and int3, respectively. Transformed scores are calculated as $z_{ij} = (r_{ij} - \bar{x}_j)/s_j$, where $\bar{x}_j$ is the mean and s_j the SD of the scores from source j. The combined score $\phi_{\cdot,i}$ is then obtained by calculating the maximum

$$\phi_{\mathrm{int1},i} = \max z_{ij}$$

average

$$\phi_{\mathrm{int2},i} = \sum_{j=0}^{n} z_{ij}/n$$

Table 2. Tests using susceptibility genes for complex human traits

Complex trait	OMIM	Review(s)[a]	Gene[b]	Reviews				OMIM			
				Rank	Total	Percent	Enrich	Rank	Total	Percent	Enrich
Age-related macular degeneration	603075	15094132; 15350892	*CFH*	7263	13771	47.3	2	10450	12608	17.1	1
			LOC387715	–	13771	–	–	–	12608	–	–
ARMD (second run)	603075	N/A[c]	*C2*	–	–	–	–	766	12875	94.1	17
			CFB	–	–	–	–	44	12875	99.7	293
Alzheimer's disease	104300	15225164	*LOC439999*	–	13550	–	–	–	13709	–	–
Asthma	600807	12810182; 14551038	*NPSR1*	1117	13881	92.0	12	2835	13120	78.4	5
Autism	209850	11733747; 12142938	*EN2*	98	13610	99.3	139	98	13213	99.2	135
Celiac disease	212750	12907013; 12699968; 14592529	*MYO9B*	234	13039	98.2	56	168	12703	98.7	76
Myocardial infarction	608446	15861005; 16041318	*LTA4H*	122	14043	99.1	115	–[d]	–	–	–
Parkinson's disease	168600	16026116; 16278972	*SEMA5A*	4548	13477	66.2	3	879	13329	93.4	15
Rheumatoid arthritis	180300	15478157; 12915205	*PTPN22*	333	13279	97.5	40	2156	13038	83.5	6
			FCRL3	3743	13279	71.8	3	2230	13038	82.9	6
Schizophrenia	181500	15340352; 16033310	*ENTH*	10013	14603	31.4	1	8065	13572	40.6	2
Type 1 diabetes mellitus	222100	12270944; 11921414	*SUMO4*	12123	14272	15.1	1	7675	13130	41.5	2
		11237226; 11899083	*PTPN22*	165	14272	98.8	86	833	13130	93.7	16
			IL2RA	130	14272	99.1	110	528	13130	96.0	25
			CTLA4	78	14272	99.5	183	324	13130	97.5	40
Type 2 diabetes mellitus	125853	15662000; 15662001; 15662002; 15662003	*TCF7L2*	2911	13922	79.1	5	4013	13586	70.5	3
Totals				725[e]	13826[e]	94.7[e]	54[f]	879[e]	13130[e]	93.4[e]	43[f]

[a]PubMed ID(s) of review articles used in corpus.
[b]For references see Methods section. HUGO approved gene symbols used to identify genes.
[c]No suitable review corpus available (see Methods section).
[d]The OMIM record is insufficiently detailed and was not used.
[e]Median result.
[f]Mean result.

or sum

$$\phi_{\text{int3},i} = \sum_{j=0}^{n} z_{ij}$$

of the transformed scores for gene *i*.

The fourth method, referred to as int4, differs from the other three by considering both the score of a gene within a data source as well as the number of genes returned for that data source. First, a transformed score s_{ij} is obtained.

$$s_{ij} = \frac{r_{ij}}{\sum_{i=0}^{n} r_{ij}}$$

The transformed gene scores are then summed together to provide a final score for each gene.

$$\phi_{\text{int4},i} = \sum_{j=1}^{J} s_{ij} \frac{g_j}{G}$$

where g_j is the number of genes returned for source *j* and

$$G = \sum_{j=1}^{J} g_j$$

2.4 Implementation

The CAESAR algorithms were written using Perl version 5.8.1 and Java version 1.4.2. The vector space similarity searches were performed using a modified version of the Perl module Search::VectorSpace by Maciej Ceglowski (http://www.perl.com/pub/a/2003/02/19/engine.html). Databases and ontology schemas were downloaded and parsed into XML under a custom XML schema. Intermediate text and data-mining results were also stored as XML under the same schema.

2.5 Selection of the tests for complex traits

To assess the ability of CAESAR to choose valid candidates, 18 test genes were selected from recently published reports providing strong evidence of statistical association with known complex human disorders. The test genes included *CTLA4* (Ueda *et al.*, 2003), *PTPN22* (Bottini *et al.*, 2004), *PTPN22* (Begovich *et al.*, 2004), *SUMO4* (Guo *et al.*, 2004), *FCRL3* (Kochi *et al.*, 2005), *ENTH* (Pimm *et al.*, 2005), *EN2* (Gharani *et al.*, 2004), *TCF7L2* (Grant *et al.*, 2006), *CFH* (Klein *et al.*, 2005), *LOC387715* (Rivera *et al.*, 2005), *LTA4H* (Helgadottir *et al.*, 2006), *C2* (Gold *et al.*, 2006), *CFB* (Gold *et al.*, 2006), *NPSR1* (Laitinen *et al.*, 2004), *MYO9B* (Monsuur *et al.*, 2005), *IL2RA* (Vella *et al.*, 2005), *SEMA5A* (Maraganore *et al.*, 2005) and *LOC439999* (Grupe *et al.*, 2006).

Each disorder required a custom corpus, either an OMIM record or one or more review articles describing the biology of the disorder (Table 2). Review articles were selected by searching PubMed (Wheeler *et al.*, 2006) for articles published before the year of discovery of each gene association. Where multiple suitable review articles were available, the texts were concatenated to produce the corpus. We removed any direct reference to the testing gene in the input text. In addition, entries in the GAD containing the test genes were removed. Thus, the input data closely mimicked the state of knowledge prior

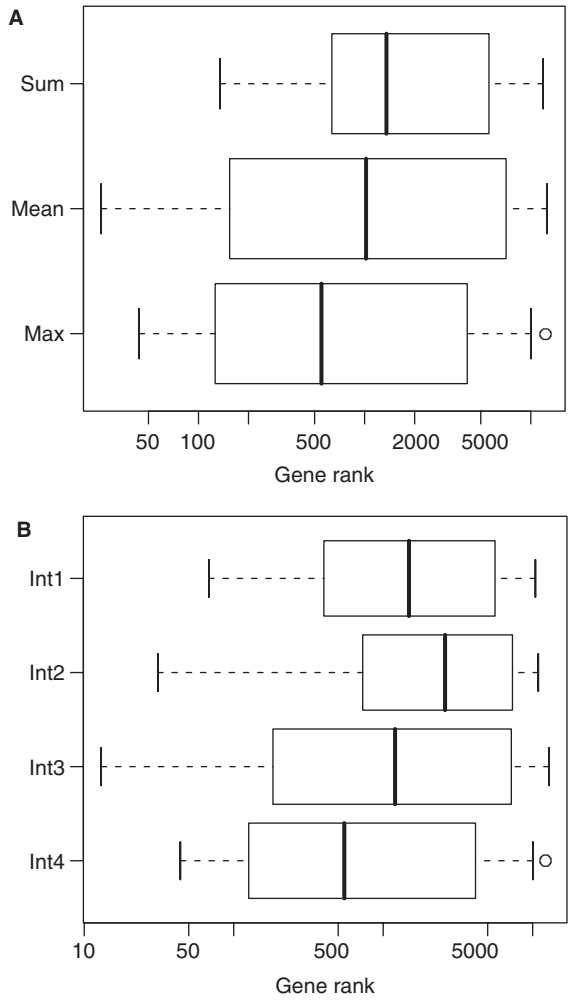

Fig. 3. Box and whisker plot distributions of the ranks of 18 test genes in Table 2 using different CAESAR parameters. Ranks are plotted on a log scale. Plots are constructed so that the bounds of the box are the upper and lower quartile medians, the line inside the box is the median, the whiskers extend to the last value no more than 1.5 times the length of the box, and all remaining values are outliers. (**a**) Distribution of ranks using the max, mean and average data-mining methods (int4 method for integration). (**b**) Distribution of ranks using the four different integration methods (max data-mining method).

to the discovery of the positive association between the disease and the test gene.

In the case of age-related macular degeneration (ARMD), positive associations for the two test genes, *CFB* and *C2*, were reported after the discovery of *CFH* as a susceptibility gene for the disease. Due to the absence of a suitable review article incorporating the discovery of *CFH*, results for these two test genes employ the ARMD OMIM corpus only.

A common way of summarizing the performance of previous candidate gene selection algorithms is to calculate 'fold enrichment', which is the total number of ranked genes divided by the rank of the test gene. Fold enrichment must be interpreted with caution, because it is not calculated relative to random expectation. Nonetheless, we report this statistic in order to facilitate comparison with other methods.

3 RESULTS

3.1 Testing of recently discovered complex trait genes

We tested the performance of the algorithm on a set of test genes previously reported to be associated with 11 complex human diseases (Table 2). For each disease, we selected one or more genes for which recent population genetic studies have reported a significant association with the disease phenotype. Nearly 15 000 genes had sufficient information from one or more data sources to be ranked. Table 2 summarizes results of the 18 test genes by separately considering tests using review articles and OMIM records as input, although not all genes were tested using both input types. In order to report the success of CAESAR using all 18 genes, we combined review article tests for 16 genes with OMIM record tests for 2 genes, *CFB* and *C2*, which were not tested using review articles (see Methods section). The following results using all 18 test genes are thus not summarized in Table 2.

First, we evaluated the choice of data-mining method for determining the score r_{ij} of each gene i for each data source j (see Methods section). The distributions of the ranks are shown in Figure 3a. Each data-mining method used the int4 integration method (data for other integration methods not shown). The maximum method had a smaller median rank (549.5) than both the sum (1353) and mean (1020) methods.

Second, we evaluated the four different methods for the integration of data from different sources (Fig. 3b). Int4 yielded the smallest median rank (549.5) compared to the results for int1 (max), int2 (mean) and int3 (sum), which were 1488, 2594 and 1201, respectively. Furthermore, int4 had smaller upper and lower quartile ranks than int1, int2 and int3. We thus report the results for the maximum data-mining and int4 integration method in what follows.

Overall, 16 of 18 test genes were ranked with a median rank of 549.5 and 67-fold average enrichment. Seven of the 18 test genes (39%) were ranked higher than 98% of all ranked genes for the trait in question, while five (28%) ranked in the 99th percentile. The highest rank seen in our tests was 44 for *CFB*, a susceptibility gene for age-related macular degeneration, which corresponds to a 293-fold enrichment. Two of the genes, *LOC387715* and *LOC439999*, were unranked due to a lack of information on these genes in any of the data sources.

We compared the observed distribution of the ranks for the 18 test genes to that expected by chance, which is a minimal test for the effectiveness of the method. The expected mean percentile for a random gene would be 50. The observed mean percentile is 80.5 and, under a binomial expectation, the 95% confidence interval is 66–95. Thus, the observed distribution of ranks for the test genes is significantly displaced relative to random expectation.

3.2 Comparison of input texts

We next examined the effect of the choice of corpus on the ranks for the test genes. Using review article corpus tests only, 14 of 16 test genes were ranked, with a median rank of 725 and

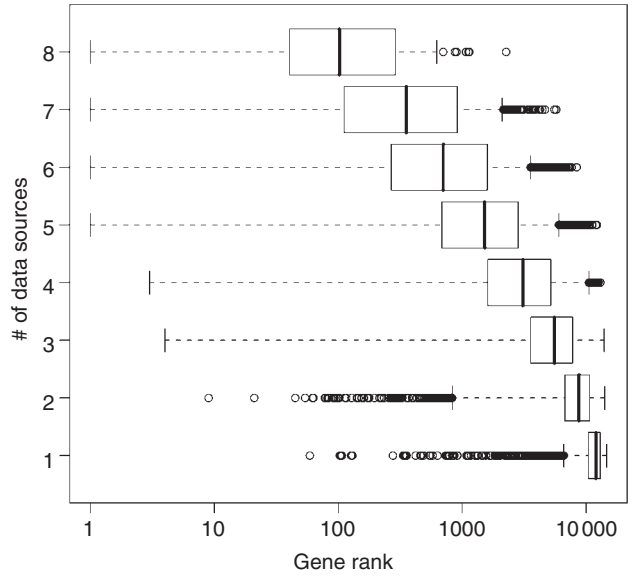

Fig. 4. The relationship between the rank of a gene and the number of data sources in which it is annotated, using the max and int4 methods. Ranks are plotted on a log scale. Box and whisker plots were constructed as described for Figure 3.

54-fold average enrichment. Six of the 16 test genes (37.5%) ranked in the 98th percentile, while four (25%) ranked in the 99th percentile (Table 2).

For comparison, we selected for each disease the relevant records from the OMIM database. For all tests the int4 method was used (Table 1). The test for candidate genes of myocardial infarction was omitted because the OMIM record for this disease is only ~100 words in length, which would be insufficient for reliably scoring a large number of ontology terms. Of the remaining 17 genes tested, 15 sufficient information to be ranked. The median rank was 879 with an average 43-fold enrichment. The best performance was observed for *CFB*, with 293-fold enrichment. Three of the 17 test genes (17.6%) ranked in the 98th percentile of all ranked genes, while 2 of 17 (11.8%) ranked in the 99th percentile. Only one gene, *SEMA5A*, had a dramatically improved rank relative to that obtained using a corpus of published review articles. Thus, the ranks for the test genes using OMIM records, while still clearly an improvement over random expectation, are in most cases inferior to those obtained using review articles.

We examined whether the length of the input text could help explain the difference in performance between the two types of input text. The length of each corpus was measured as the number of words excluding stop words and non-word characters. There was no significant correlation between the length of the corpus and the rank obtained for each test gene (Spearman's $\rho = -0.21$, $P = 0.27$).

3.3 Analysis of bias

CAESAR is dependent on available annotations to rank genes. Therefore, the preferential ranking of well-annotated genes is a potential source of bias in the results. We addressed this issue in two ways, by measuring the effect of both breadth and depth of annotation on gene rank. We first measured the correlation between gene rank and the breadth of annotation, or the number of sources for which a gene is annotated, across each integration method. Using the default methods (max and int4), there is a strong correlation ($\rho = -0.75$), as shown in Figure 4. By comparison, again using the max method, int2 ($\rho = -0.15$) and int3 ($\rho = -0.06$) showed little correlation, while int1 showed modest correlation ($\rho = -0.47$).

We next addressed the correlation between gene rank and annotation depth by considering the number of GO annotations (biological process + molecular function) per gene. For each data-mining method, and using int4 for data integration, we calculated the mean number of GO terms for genes ranked within the top 98th percentile (max: 7.2 ± 4.1; avg: 6.2 ± 3.7; sum: 9.8 ± 5.3) and found this to be significantly higher than the mean number of GO terms across all ranked genes (4.6 ± 2.9) for all three data methods (two-tailed, unpaired *t*-tests, *P*-values $< 2 \times 10^{-16}$).

Data sources used by CAESAR include diverse available sources of gene-centric information; however, non-independence among data sources could also potentially bias the results. To address this issue, we measured the average correlation between the ranked gene lists for each tested trait using the review article corpus (Table 3). The majority of the sources show a mild, yet significant, correlation. No two data sources show a correlation greater than $\rho = 0.43$. Several pairs of sources show very weak negative correlations.

Table 3. Independence of CAESAR data sources

	GAD	GObp	GOmf	PPI	IPro	MGD	Path	Tissue
GAD	–	−0.04	−0.04	0.08	0.06	0.10	0.11	−0.03
GObp	$2e^{-6}$	–	0.43	−0.06	0.12	−0.11	−0.10	−0.06
GOmf	$5e^{-6}$	$2e^{-16}$	–	−0.07	0.16	−0.15	−0.08	−0.11
PPI	$2e^{-16}$	$2e^{-13}$	$2e^{-16}$	–	0.08	0.18	0.21	−0.04
IPro	$1e^{-10}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	0.08	0.13	−0.10
MGD	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	0.27	−0.13
Path	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	−0.18
Tissue	$2e^{-4}$	$2e^{-10}$	$2e^{-16}$	$1e^{-6}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–

Top: Spearman rank correlations among pairs of sources. Each value represents the maximum correlation found for a given pair across data for all 11 tested complex traits using default parameters. Bottom: Significance of each correlation. GAD = genetic association database data, GObp = GO biological process data, GOmf = GO molecular function data, PPI = protein–protein interaction data, IPro = InterPro data, MGD = mouse genome database data, path = KEGG pathway data, tissue = Swiss-prot/TrEMBL tissue data.

4 DISCUSSION

The extraordinary amount of biological information available in the published literature and in publicly available databases about complex human diseases, on the one hand, and genes and their protein products, on the other, is well suited to the *in silico* identification of candidate genes for disease. The approach is enabled by ontologies that provide a semantic mapping between the natural language description

of diseases and traits, and the functional annotation of genes and their products. It is further enabled by the availability of well-curated pathway and protein-interaction datasets, and a wide variety of functional information about not only the genes themselves, but also their homologs in model organisms. The approach implemented in CAESAR can, in principle, be applied to any complex trait in any organism for which similar information resources exist.

CAESAR relies on human expert knowledge in order to function effectively, but it does not require that the user actually possess all of this knowledge. At a minimum, the user needs to select a relevant corpus, but much more user intervention is possible. The user may manually modify the scores from the text-mining step and/or introduce genes in addition to those that were extracted from the corpus. The final rankings may be modified based on user perceptions of the importance of particular data sources. The user may also restrict the algorithm to consider only certain genomic regions or particular sets of genes. While it is not advisable to eliminate human judgment and oversight of the candidate gene selection process, due to the volume and the complexity of the information involved, semi-automated methods such as CAESAR may well outperform an unaided expert. At the very least, CAESAR provides a quantitative starting point for which the assumptions are clear and the user's biases are minimized.

The success of CAESAR in any given instance is due both to factors that are, at least to some extent, under the user's control and those that are not. The user's choice of a corpus that accurately reflects the biology of the trait is clearly of critical importance. In our experiments, we found that review articles generally, though not always, yielded better results than OMIM records. The explanation for this difference is not clear; it does not appear to be due to differences in corpus length.

Other factors under the user's control are algorithmic, e.g. how to calculate a score for a gene within a data source and to rank genes across multiple data sources. The variety of simple methods used here can, in some cases, lead to substantially different rankings. One example is *NPSR1*, which had ranks of 749 and 2751 using int1 and int2, respectively. Four different data sources (GO bp, GO mf, IPro and tissue) report information on *NPSR1*, and the scores vary from high to low. Int1, which calculates the maximum, favors genes with a high score in one data source regardless of the others, whereas the low scores are detrimental to the final rank using int2, which calculates the average. Each of the methods can be justified (see Method section), and it is not clear a priori which should be superior.

Overall, we found that the best results on the test set were obtained using a corpus of review articles, the maximum method for combining scores for a gene within a data source, and the int4 method for data integration across multiple sources. However, other combinations of parameters were superior for particular test genes. On the basis of our test results, we have selected the 'max' data-mining and 'int4' data-integration methods to be the default settings for CAESAR. The OMIM record, if available, is used as the input text by default, though our results suggest that one or more review articles should be used instead, or in addition, when possible.

A number of factors affecting CAESAR's success are outside of the user's control. One is the depth of biological knowledge about the complex trait under study and the extent to which this knowledge has been recorded. Another is the extent to which ontologies can be used to mediate between trait-centric and gene-centric information sources. For example, anatomical ontologies are available for mammals, but not yet for all organisms. Even where an ontology exists, certain terms may not exist, have listed synonyms, or be sufficiently well defined.

The process of extracting gene names from unstructured text is also error-prone (Hirschman *et al.*, 2005), especially when using older bodies of text containing outdated gene names and symbols. Gene extraction is complicated further by the fact that genes often share symbols with other genes and non-gene acronyms.

Perhaps most importantly, CAESAR depends on the availability of functional information. Approximately half of the unique entries in our reference set remained unranked for any trait due to lack of annotation, including two of the test genes, *LOC387715* and *LOC439999*. As the total number of ranked genes depends on the number of ontology terms that are mapped from the corpus, the success of CAESAR for a given trait depends on the information content of the corpus. One tested trait, myocardial infarction did not have a sufficiently informative OMIM record. Therefore, CAESAR is limited to genes and traits for which there is sufficient information in the form of annotations and text descriptions, respectively. To the extent that this reflects incomplete knowledge of genes and traits, it is a limitation shared by all candidate gene approaches. The lack of gene-centric information, at least, can be partially overcome by including additional data sources from map-based studies, systematic functional genomic screens and other model systems in which homologs may have been characterized.

Given the importance of including a wide variety of functional information, CAESAR could be enhanced by the inclusion of additional data sources. A particularly valuable source would be data from transcription profiling experiments, which would provide information on a large proportion of genes that are lacking information from other sources. Inclusion of this data will be challenging, however, as the datasets available are diverse and heterogeneous, and it is not clear how best to score the relevance of a particular expression pattern to a trait.

Inclusion of additional data sources could potentially raise the issue of non-independence among them. Although no two data sources used in this study are highly correlated, most of them have a significant weak correlation. CAESAR does not currently correct for non-independence during the data-integration step.

A variety of *in silico* methods for candidate gene selection have previously been reported, though most have been designed and tested to prioritize positional candidates. Gene-Seeker (van Driel *et al.*, 2003) selected candidates in a given genomic region through web-based data mining of expression and phenotype databases. This approach enriched for disease genes

in 10 monogenic disorders, providing at best 25- and 7-fold enrichment on average. POCUS (Turner *et al*., 2003) exploited functional similarities between genes at two or more loci to predict candidates, requiring no user input beyond the genomic regions of interest. It provided 12-, 29- and 42-fold enrichment on average for three test loci of increasing size and at best provided 81-fold enrichment. Perez-Iratxeta *et al*. (2002) used literature mining to associate pathology with GO terms and then used these terms to rank candidate genes. The authors created artificial loci containing an average of 300 genes for testing and found 10-fold enrichment on average and, at best, 38-fold enrichment. The correct disease gene was present in their enriched set for ~50% of the loci. Freudenberg and Propping (2002) computed similarity-based clusters of known disease genes based on phenotypic sharing between diseases. Their method selected the correct disease gene in roughly two-thirds of the cases, on average resulting in 10-fold enrichment, and in the top one-third of the cases resulting in 33-fold enrichment. Franke *et al*. (2006) developed a functional network of human genes to select candidate genes found in pathways with known disease genes. They constructed artificial loci that contained on average 100 genes, and found 20- and 10-fold enrichment on average in 27 and 34% of tested genes, respectively.

More recently, SUSPECTS (Adie *et al*., 2006) and ENDEAVOUR (Aerts *et al*., 2006) have been developed for application to more complex traits. Both of these systems prioritized genes using a combination of annotation and sequence features based on similarity to a training set. SUSPECTS was able to identify a test gene in artificial loci on average within the top 13% of candidates, a 7-fold enrichment. In half the cases, the test gene was in the top 5% of candidates, a 20-fold enrichment. ENDEAVOUR tested both monogenic and polygenic (complex) disorders using a test set of 200 genes. Over all tested disorders, ENDEAVOUR provided 9-fold enrichment on average and 200-fold enrichment at best. Considering polygenic disorders only, ENDEAVOUR provided 5-fold enrichment on average and 18-fold enrichment at best.

The measure of success for an approach such as CAESAR ultimately depends on the specific application. Our goal has been the enrichment of candidates within the top 2% of ranked genes, which represents roughly the top 1% of genes in the human genome. Given the number of functionally annotated human genes, this corresponds to 250–300 genes, which is a reasonable number included in a high-resolution SNP association study for a complex disease in human populations. Our results suggest that approximately one-third to one-half of the genes previously associated with complex human disease would be included in this enriched candidate set. With a complex trait, for which the true effectors are only partially known, it is difficult to quantify the number of true and false positives. Nonetheless, assuming all genes outside of our test set are negatives, we can calculate sensitivity as TP/(TP+FN) and specificity as TN/(TN+FP), where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives. Considering positives to be the top 2% of ranked genes, we obtained an overall sensitivity of 39% and specificity of 98% for our test set. Other measures of success may be relevant for different applications, such as prioritizing SNPs for follow-up work from a genome-wide association study. By standard measures, CAESAR compares favorably with other methods, even though we use a test set of genes associated with complex rather than monogenic or oligogenic diseases. The highest (293) and average (67) fold enrichment obtained with CAESAR are greater than those reported for other systems.

CAESAR makes use of a relatively small trait-specific corpus, comprised of one to several review articles, and a large body of gene-centric information. A similar approach could be useful for other applications involving semantic mediation between larger corpora or sets of corpora.

In conclusion, CAESAR can successfully mine large amounts of biological information to guide the selection of candidate genes for complex diseases in humans. Applications include selection of candidate genes for association or re-sequencing studies, prioritization of candidates for functional genomics experiments, or evaluation of results from linkage and genome-wide association studies. The approach may be extended to select candidates for complex traits in other organisms for which similar informatic resources are available. No computational system can select candidate genes with certainty; however, when used as a guide, CAESAR is a useful tool for candidate gene prioritization.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health (DK72193 to K.L.M.), the National Science Foundation (0227314 to T.J.V.) and a Burroughs Wellcome Career Award in the Biomedical Sciences (K.L.M.). Funding to pay the Open Access publication charges was provided by 0227314, DK72193, and the UNC Office of the Vice Chancellor for Research and Economic Development. Funding to pay the Open Access publication charges was provided by 0227314, DK72193 and the UNC Office of the Vice Chancellor for Research and Economic Development.

Conflict of Interest: none declared.

REFERENCES

Adie,E. *et al*. (2006) Suspects: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.

Aerts,S. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol*., **24**, 537–544.

Alfarano,C. *et al*. (2005) The biomolecular interaction database and related tools 2005 update. *Nucleic Acids Res*., **33**, D418–D424.

Apweiler,R. *et al*. (2000) Interpro-an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.

Bairoch,A. *et al*. (2005) The universal protein resource (Uniprot). *Nucleic Acids Res*., **33**, D154–D159.

Becker,K. *et al*. (2004) The genetic association database. *Nat. Genet*., **36**, 431–432.

Begovich,A. *et al*. (2004) A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am. J. Hum. Genet*., **75**, 330–337.

Birney,E. *et al*. (2006) Ensembl 2006. *Nucleic Acids Res*., **34**, D556–D561.

Blake,J. *et al*. (2003) MGD: the mouse genome database. *Nucleic Acids Res*., **31**, 193–195.

Bottini,N. *et al*. (2004) A functional variant of lymphoid tyrosine phosphatase is associated with type 1 diabetes. *Nat. Genet*., **36**, 337–338.

Camon,E. *et al.* (2003) The gene ontology annotation (GOA) project: implementation of GO in swiss-prot, trembl and interpro. *Genome Res.*, **13**, 662–672.

Dean,M. (2003) Approaches to identify genes for complex human diseases: lessons from mendelian disorders. *Hum. Mutat.*, **22**, 261–274.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18**, S110–S115.

Gharani,N. *et al.* (2004) Association of the homeobox transcription factor, ENGRAILED 2, 3, with autism spectrum disorder. *Mol. Psychiatry*, **5**, 474–484.

Gold,B. *et al.* (2006) Variation in factor B (BF) and complement component 2 (C2) genes is associated with age-related macular degeneration. *Nat. Genet.*, **38**, 458–462.

Grant,S. *et al.* (2006) Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.*, **38**, 320–323.

Grupe,A. *et al.* (2006) A scan of chromosome 10 identifies a novel locus showing strong association with late-onset alzheimer disease. *Am. J. Hum. Genet.*, **78**, 78–88.

Guo,D. *et al.* (2004) A functional variant of SUMO4, a new I kappa B alpha modifier, is associated with type 1 diabetes. *Nat. Genet.*, **36**, 837–841.

Hamosh,A. *et al.* (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **33**, D514–D517.

Harris,M. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.

Helgadottir,A. *et al.* (2006) A variant of the gene encoding leukotrine A4 hydrolase confers ethnicity-specific risk of myocardial infarction. *Nat. Genet.*, **38**, 68–74.

Hirschman,L. (2005) Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC Bioinformatics*, **6**, S11.

Kanehisa,M. *et al.* (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.

Kelso,J. *et al.* (2003) eVOC: a controlled vocabulary for unifying gene expression data. *Genome Res.*, **13**, 1222–1230.

Klein,R. *et al.* (2005) Complement factor H polymorphism in age-related macular degeneration. *Science*, **308**, 385–389.

Kochi,Y. *et al.* (2005) A functional variant in FCRL3, encoding fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.*, **37**, 478–485.

Laitinen,T. *et al.* (2004) Characterization of a common susceptibility locus for asthma-related traits. *Science*, **304**, 300–304.

Maglott,D. *et al.* (2005) Entrez gene: gene-centric information at NCBI. *Nucleic Acids Res.*, **33**, D54–D58.

Maraganore,D. *et al.* (2005) High-resolution whole-genome association study of parkinson's disease. *Am. J. Hum. Genet.*, **77**, 685–693.

Monsuur,A. *et al.* (2005) Myosin IXB variant increases the risk of celiac disease and points toward a primary intestinal barrier defect. *Nat. Genet.*, **37**, 1341–1344.

Newton-Cheh,C. and Hirschhorn,J. (2005) Genetic association studies of complex traits: design and analysis issues. *Mutat. Res.*, **573**, 54–69.

Peltonen,L. and McKusick,V. (2001) Genomics and medicine: dissecting human disease in the postgenomic era. *Science*, **291**, 1224–1229.

Perez-Iratxeta,C. *et al.* (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316–319.

Peri,S. *et al.* (2004) Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.*, **32**, D497–D501.

Pimm,J. *et al.* (2005) The epsin 4 gene of chromosome 5q, which encodes the clathrin-associated protein enthoprotin, is involved in the genetic susceptibility to schizophrenia. *Am. J. Hum. Genet.*, **76**, 902–907.

Rivera,A. *et al.* (2005) Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum. Mol. Genet.*, **14**, 3227–3236.

Salton,G. *et al.* (1975) A Vector Space Model for Automatic Indexing. *Commun. ACM*, **18**, 613–620.

Smith,C. *et al.* (2005) The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.*, **6**, R7.

Thomas,D. (2006) Are we ready for genome-wide association studies? *Cancer Epidemiol. Biomarkers Prev.*, **15**, 595–598.

Turner,F. *et al.* (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.

Ueda,H. *et al.* (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, **423**, 503–511.

van Driel,M. *et al.* (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet.*, **11**, 57–63.

Vella,A. *et al.* (2005) Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.*, **75**, 773–779.

Wheeler,D. *et al.* (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **22**, D173–D180.