BMC Bioinformatics

DATABASE Open Access

Gentrepid V2.0: a web server for candidate disease gene prediction

Sara Ballouz[1,2,7], Jason Y Liu[1], Richard A George[1], Naresh Bains[1], Arthur Liu[1], Martin Oti[3], Bruno Gaeta[2], Diane Fatkin[4,5] and Merridee A Wouters[6*]

Abstract

Background: Candidate disease gene prediction is a rapidly developing area of bioinformatics research with the potential to deliver great benefits to human health. As experimental studies detecting associations between genetic intervals and disease proliferate, better bioinformatic techniques that can expand and exploit the data are required.

Description: *Gentrepid* is a web resource which predicts and prioritizes candidate disease genes for both Mendelian and complex diseases. The system can take input from linkage analysis of single genetic intervals or multiple marker loci from genome-wide association studies. The underlying database of the *Gentrepid* tool sources data from numerous gene and protein resources, taking advantage of the wealth of biological information available. Using known disease gene information from OMIM, the system predicts and prioritizes disease gene candidates that participate in the same protein pathways or share similar protein domains. Alternatively, using an *ab initio* approach, the system can detect enrichment of these protein annotations without prior knowledge of the phenotype.

Conclusions: The system aims to integrate the wealth of protein information currently available with known and novel phenotype/genotype information to acquire knowledge of biological mechanisms underpinning disease. We have updated the system to facilitate analysis of GWAS data and the study of complex diseases. Application of the system to GWAS data on hypertension using the ICBP data is provided as an example. An interesting prediction is a ZIP transporter additional to the one found by the ICBP analysis. The webserver URL is https://www.gentrepid.org/.

Keywords: Candidate disease gene prediction, Candidate disease genes, Mendelian diseases, Complex diseases, Genome-wide association studies, Genotype, Phenotype, Candidate gene identification, Genetic-association studies, Hypertension

Background

The identification of genes implicated in human disease enables an understanding of disease mechanisms and is essential for the development of diagnostics and therapeutics. While genetic approaches such as linkage mapping or genome-wide association studies (GWAS) can successfully identify genomic regions linked to a particular disease, identification of the disease-causing genes is still a difficult task. For linkage analysis, genomic regions are large, as these analyses look at inheritance over several generations in a known pedigree which may have few recombination events. Genetic association studies, which do not focus on recent observable ancestry but rather look at historic recombination, have shorter disease-associated regions [1]. However, when applied at the genome-wide scale (eg. GWAS), the large number of markers used requires stringent correction for multiple testing, limiting the number of reliably identified candidate genes. Lower scoring markers may still indicate potential disease genes but may also be false positives. Analysis of this valuable but noisy data would benefit from a candidate disease gene prioritization approach [2].

A variety of publicly available systems employ various methodologies to map phenotype to genotype in order to predict or rank candidate disease genes [3,4]. Many of these systems are available as web services. Because oligogenic diseases are associated with disruption of genes that have similar functions [5], the most prevalent

* Correspondence: m.wouters@deakin.edu.au
[6]School of Medicine, Deakin University, Geelong, VIC 3217, Australia
Full list of author information is available at the end of the article

method employed is gene clustering [6]. Genes are clustered based on shared features, such as common domains, similar functional annotation, involvement in the same protein complex or signaling pathways, co-expression, or combinations of these [6-11].

Here we present the *Gentrepid* web server, a public candidate disease gene prediction system that associates genes with specified phenotypes using genetic and biomolecular data (Figure 1). *Gentrepid* draws on two gene clustering methods to make candidate gene predictions; the Common Pathway Scanning (CPS) and Common Module Profiling (CMP) approaches [10]. Both methods identify links between genes in loci associated with a disease phenotype. CPS is based primarily on protein interaction data, whereas CMP is based on sequence data. *Gentrepid* can be assisted by phenotype-associated genes as seeds (*seeded* mode), or can work in the absence of disease gene knowledge using only phenotype-associated loci (*ab initio* mode). The system can be applied to both Mendelian and complex diseases [4,6,10].

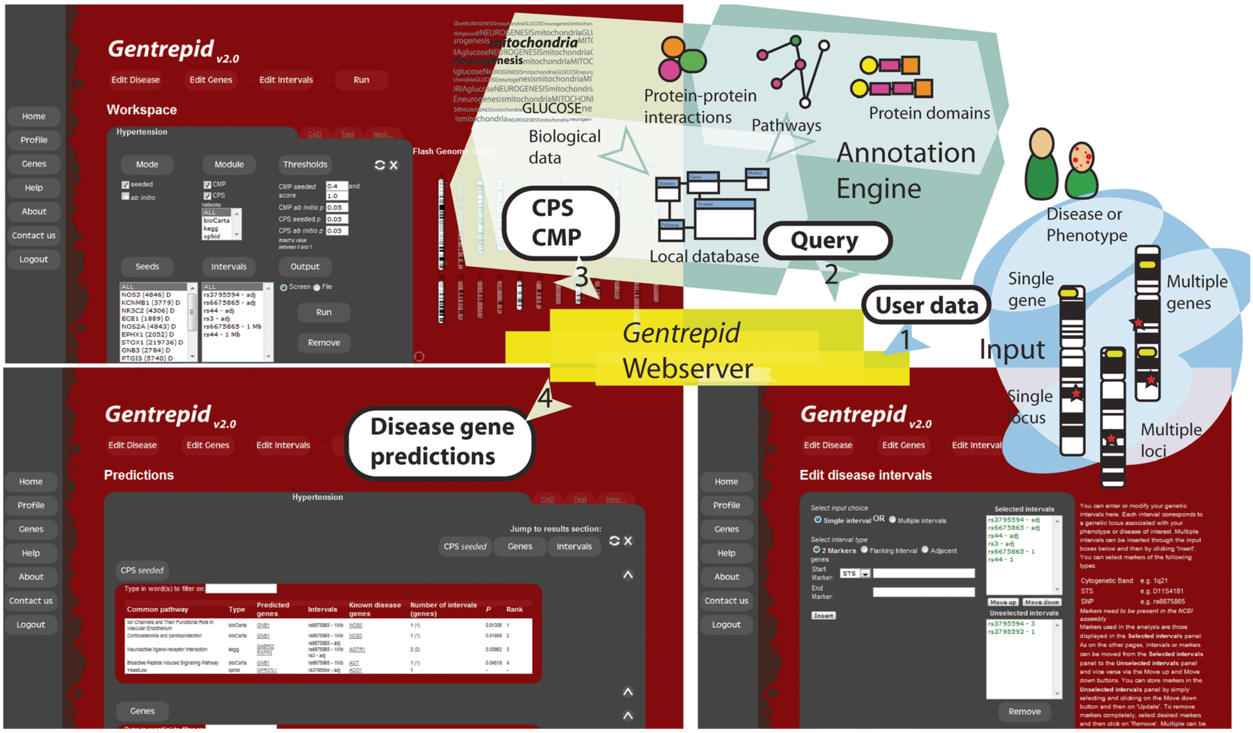
Construction and content

Underlying databases and data sources

The relational database underlying *Gentrepid* was created using the PostgreSQL database management system with current biological data on human proteins and genes, including pathways, interaction data and domains. The core of the database is built on gene annotation and sequence data of the NCBI reference assembly (build 36) from RefSeq [12], with corresponding transcript (mRNA) and protein information. Protein-protein interaction (PPI) data are gathered from the Interlogous Interaction Database (I2D, formerly known as the Online Predicted Human Interaction Database, or OPHID [13]). I2D contains literature-derived data from multiple databases including BIND [14], MINT [15] and HPRD [16]. Pathway data are retrieved from BioCarta [17] and KEGG [18], which contain information on signalling and metabolic pathways respectively. Single nucleotide polymorphisms (SNPs), sequence tagged sites (STSs) and cytogenic data which are included as marker positions for input of genetic loci, are pooled from dbSNP (build 131 at the time of the analyses) [19], dbSTS and Map Viewer from NCBI [12]. Human disease gene information is extracted from the Online Mendelian Inheritance in Man (OMIM) [20] Morbid Map flat file.

Data generation and processing

For all proteins, several sequence features are pre-calculated including signal peptide cleavage predictions



Gentrepid v2.0

Edit Disease · Edit Genes · Edit Intervals · Run

Home · Profile · Genes · Help · About · Contact us · Logout

Workspace

Hypertension · CAD · Test · New...

Mode: ☑ seeded ☐ *ab initio*

Module: ☑ CMP ☑ CPS

Networks: ALL, bioCarta, kegg, ophid

Thresholds:
CMP *seeded* score 0.4 and 1.0
CMP *ab initio p* 0.05
CPS seeded *p* 0.05
CPS *ab initio p* 0.05
Insert a value between 0 and 1

Seeds: ALL, NOS3 (4846) D, KCNMB1 (3779) D, NR3C2 (4306) D, ECE1 (1889) D, NOS2A (4843) D, EPHX1 (2052) D, STOX1 (219736) D, GNB3 (2784) D, PTGIS (5740) D

Intervals: ALL, rs3795594 - adj, rs6675865 - adj, rs44 - adj, rs3 - adj, rs6675865 - 1 Mb, rs44 - 1 Mb

Output: ● Screen ○ File

Run · Remove

Gentrepid v2.0

Edit Disease · Edit Genes · Edit Intervals

Predictions

Hypertension · CAD · Test · New...

Jump to results section: *CPS seeded* · Genes · Intervals

CPS seeded

Type in word(s) to filter on:

Common pathway	Type	Predicted genes	Intervals	Known disease genes	Number of intervals (genes)	*P*	Rank
Ion Channels and Their Functional Role in Vascular Endothelium	bioCarta	GNB1	rs6675865 - 1Mb	NOS3	1 (1)	0.01309	1
Corticosteroids and cardioprotection	bioCarta	GNB1	rs6675865 - 1Mb	NOS3	1 (1)	0.01958	2
Neuroactive ligand-receptor interaction	kegg	GABRD RXFP2	rs6675865 - adj rs6675865 - 1Mb rs3 - adj	AGTR1	2 (2)	0.03962	3
Bioactive Peptide Induced Signaling Pathway	bioCarta	GNB1	rs6675865 - 1Mb	AGT	1 (1)	0.04518	4
YeastLow	ophid	GPR37L1	rs3795594 - adj	ADD1	1	-	-

Genes

Gentrepid v2.0

Edit Disease · Edit Genes · Edit Interval

Edit disease intervals

Select input choice
● Single interval OR ○ Multiple intervals

Select interval type
● 2 Markers ○ Flanking Interval ○ Adjacent genes

Start Marker: STS
End Marker:
Insert

Selected intervals: rs3795594 - adj, rs6675865 - adj, rs44 - adj, rs3 - adj, rs6675865 - 1, rs44 - 1

Move up · Move down

Unselected intervals: rs3795594 - 3, rs3795593 - 1

Remove

You can enter or modify your genetic intervals here. Each interval corresponds to a genetic locus associated with your phenotype or disease of interest. Multiple intervals can be inserted through the input boxes below and then by clicking 'Insert'. You can select markers of the following types:

Cytogenetic Band	e.g. 1q21
STS	e.g. D11S4181
SNP	e.g. rs6675865

Markers need to be present in the NCBI assembly
Markers used in the analysis are those displayed in the **Selected intervals** panel. As on the other pages, intervals or markers can be moved from the **Selected intervals** panel to the **Unselected intervals** panel and vice versa via the Move up and Move down buttons. You can store markers in the **Unselected intervals** panel by simply selecting and clicking on the Move down button and then on 'Update'. To remove markers completely, select desired markers and then click on 'Remove'. Multiple can be

Figure 1 *Gentrepid* architecture and web interface overview. Users input data as a disease locus or loci, and a phenotype, if available **(1)**. Selected loci and known genes (if applicable) are displayed on the overview page **(2)**, from which a CPS or CMP analysis can be initiated **(3)**. The interactive *Gentrepid* web server matches candidate genes against known disease genes using pre-calculated similarity scores from the annotation engine and displays the results **(4)**, or uses the *ab initio* search for enrichment of common gene annotations.

by SignalP [21], coiled-coil domain predictions by Marcoil [22] and Multicoil [23], and predictions of transmembrane helices by TMHMM [24]. These data are used to provide more information about phenotype-related candidate proteins identified in the analyses. The OMIM [20] morbid map file is remediated to standardize phenotype names. In addition, simple phenotype clustering is implemented using similar text/character strings. For example, the multiple entries for Alzheimer's disease subtypes are merged into one phenotype based on the text string "Alzheimer". The purpose of this simple phenotype clustering is to capture as much relevant genetic information on the phenotype as possible [25]. Users have the ability to deconstruct phenotype clusters during the prediction process if desired. For the CMP algorithm, domain annotation of the proteome is performed by parsing all protein sequences against the Pfam library of Hidden Markov models [26] using HMMer [27]. Pairwise similarity scores between common domains of proteins are calculated using the Smith-Waterman algorithm implemented in SSEARCH [28].

Quality control and updates

Data is sourced from publicly available databases that are constantly updated. In order to maintain an up-to-date version of the system, data is updated every 6 months, or when a major release of other data sources is made available. Once the data is updated, the system is benchmarked to check for consistency by running validation tests [10].

Algorithms

Prior disease knowledge and the ab initio approach

Gentrepid functions under two input modes, *seeded* and *ab initio*. *Seeded* mode is assisted by phenotype-associated genes from OMIM [20] as seeds. These seed genes help define the phenotype-specific data that *Gentrepid* bases its predictions upon. This approach was shown to work successfully [10]. *Gentrepid* additionally allows candidate disease genes to be prioritized across multiple disease-associated loci in the absence of known disease genes (*ab initio* mode). In this mode, common domains and common pathways linking candidate disease genes from different loci are determined, and the degree of overrepresentation of specific domain combinations or pathways across disease-associated loci is used to make predictions. Candidate genes are then ranked according to the probability that these overrepresented domain combinations or pathways would occur randomly. The *ab initio* approach is especially useful when little is known about the disease phenotype, and can potentially discover novel biological mechanisms underpinning the disease.

Common pathway scanning

The Common Pathway Scanning (CPS) approach is a Systems Biology method that is based on the finding that genes for a specific phenotype are more likely to interact with other phenotype-specific genes. More specifically, these phenotype-specific proteins are involved in the same pathway or complex [5,9,29]. Utilising this knowledge, potential disease genes are predicted by searching implicated disease loci for proteins that are part of the same pathway or complex.

The CPS algorithm in *seeded* mode uses the phenotype-specific disease genes to associate pathways with the phenotype [10]. In *seeded* mode, genes within the candidate locus are checked for their occurrence in pathways associated with the seed gene. For each disease, pathways are ranked by the significance of the association, as determined by the Fisher's exact test statistic. The p-value is calculated by partitioning genes in the phenotype-associated loci based on whether they are associated with the pathway in question or not [4]. Candidate genes are filtered and prioritized based on the lowest p-value score of the pathway they share. The default significance threshold is $p_{path} < 0.05$.

The CPS algorithm in *ab initio* mode requires multiple associated loci as input. In this mode, the candidate loci are searched for genes participating in common pathways. These common pathways are then ranked according to the disease-relevance of their associated pathways in the same manner as in *seeded* mode.

Common module profiling

The second approach, Common Module Profiling (CMP), distinguishes *Gentrepid* from other candidate disease gene prediction approaches, as it is based on the use of sequence-based domain similarity. Predicated on the principle that similar diseases are often caused by families of genes with related functions [30], it uses domain-based comparative sequence analysis to identify proteins with potential functional similarity. A domain-based comparison of proteins has several advantages over comparison of the full-length amino acid sequence of proteins. Because structure is conserved over sequence, domain-based sequence comparison searches have been shown to be more accurate than full-length sequence searches [31], improving the sensitivity of the candidate gene search. Dissection of a protein into domains provides a more fine-grained approach to identifying aspects of protein function affected in the disease. Using the Pfam library of Hidden Markov models [26], domains can be assigned to approximately 69% of human proteins which allows functional inference for around 54% of human genes. While some other web tools reference protein domains [7,11], only *Gentrepid* uses domain sequence comparison rather than keyword matching. As a result, relationships between similar but differently labelled domains can also be captured.

The CMP algorithm in *seeded* mode uses pairwise alignments of Pfam domains common to the candidate gene and the known disease gene. Candidate genes are scored based on their domain similarity to a known disease gene, and prioritized according to this score [10].

In *ab initio* mode, the CMP algorithm predicts and prioritises genes in a different manner. In this method, domain combinations are tested for over-representation in the associated loci compared to the genome as a whole through upper and lower significance tests [10]. Results are prioritized based on χ^2_{min} which has been shown to give a better estimate of the significance [4].

Webserver

Data input

Data input is a relatively straightforward step-wise process. All that is required is at least one phenotype-associated locus and a known disease gene, or multiple phenotype-associated loci. *Gentrepid* allows disease intervals to be specified by STSs, SNPs or cytogenetic bands in one of two ways: two markers specifying a start and stop site can be entered; or a central marker and an interval width can be used (Figure 1, bottom right panel). Both new diseases and new disease genes not currently included in the underlying *Gentrepid* database can be added by the user and incorporated in the analysis. New phenotypes may be created by the user if desired, in which case existing OMIM data will not be accessed. New disease genes are entered into the system by selecting a HUGO gene name or Entrez ID. Additionally, any of *Gentrepid*'s pre-loaded phenotype-associated genes can be deselected for the current project. As some phenotypes described in *Gentrepid* reference more than one OMIM ID, genes that an individual researcher believes to be anomalous can be removed from the project. Usage of *Gentrepid* is outlined in a help page reachable from the start page.

GWA input upgrade

The system was originally designed for input from linkage-mapping loci determined for Mendelian diseases. Such investigations generally involve relatively few loci, making it feasible to manually enter each locus separately into the web server. This applies similarly to the investigation of relatively small numbers of loci associated with highly significant SNPs in GWAS of complex diseases. However, candidate disease gene prediction tools such as *Gentrepid* are particularly useful when applied to larger numbers of loci. For example, when lower significance thresholds are applied to the GWAS SNPs in order to discover candidates from weaker associations [4]. To facilitate this, in version 2.0 of the system, users can enter multiple loci simultaneously by pasting a list into the text box.

Data retrieval

Once a phenotype and an interval, or multiple intervals, have been entered, *Gentrepid* displays the known disease genes and the genes in the selected regions, both graphically and in tabular format (Figure 1, top left panel). The graphical display utilizes Flash GViewer, a freely available flash program developed by the GMOD project (http://gmod.org/wiki/Flash_GViewer). It shows an overview of the genome, highlighting the individual genes and intervals on the chromosomes that have been selected for the project. From this page, users can choose to perform a CPS or CMP candidate disease gene analysis.

For the CPS analysis, the user can select which preloaded genes and intervals from OMIM to include in the analysis, as well as which pathway databases to include. By default all genes, intervals and pathway databases are used. The analysis is performed, and a list of candidate disease gene predictions is displayed in table format, along with the interval, the known disease genes located in the same pathway upon which the prediction was made, and the pathway source database (Figure 1, bottom left panel).

For the CMP analysis, users have the option of selecting genes and intervals to include as for CPS, and may choose the cut-off scores for the domain comparisons. The results page displays a table of predicted candidate disease genes and their locations, along with the disease gene(s) with similar domains.

For both CPS and CMP, further information about genes of interest can be found on the gene information pages, which can be accessed by clicking on the hyperlinked gene names on the results page. Gene pages are compilations of in-house bioinformatic analyses of gene products, with protein isoforms displayed under individual tabs. Protein domains, coils, and transmembrane helices are all displayed as an image, with links to the text-based results of the Pfam [27], SignalP [21], Marcoil [22], Multicoil [23], and TMHMM [24] output. The mRNA sequence can also be accessed, and if applicable, current protein isoform sequences. Both are in FASTA format. The gene page links to relevant entries in the Entrez Gene [12] and GeneCards [32] webservers, as well as the UCSC Genome Browser [33].

User account-based project management

Gentrepid can be accessed in two ways: immediately through Quicksearch or via login after registering for an account. Both are accessible via the homepage. With Quicksearches, results are generated and immediately displayed. Quicksearch results can be saved locally by the user, but they are not stored on the webserver. In order to save results on the webserver for future use, a login account is required. Login accounts are available free of charge to users from academic institutions or non-profit research institutes. They allow users to save their input and access their project at a later date or from a different client, enabling user mobility and collaboration. With an account, users can also create, edit or delete multiple projects. User data are kept in individual password-protected PostgreSQL databases and are not accessible by other accounts. Data transfer to and from the website is also encrypted and protected by a security certificate. Projects remain available until explicitly removed by the user, and can be accessed through tabs along the top panel. Data input and retrieval are analogous to the Quicksearch version.

Utility and discussion

The system can be used on both linkage and genome-wide association data. Linkage data from family studies associates much larger regions of the genome with the disease or phenotype. Similar but not identical to an enrichment analysis, *Gentrepid*'s modules can be employed to sift through these regions and select genes that warrant further study. In the case of GWAS data, the tool can be used to search the genetic data holistically for affected pathways or molecular mechanisms. GWAS use SNPs in linkage disequilibrium (LD) to pinpoint phenotype-associated areas of the genome. When employing *Gentrepid*, genes within the vicinity of the associated SNP, not necessarily in LD, are analysed. The associated SNPs may be identifying haplotype blocks which contain control regions affecting the distal protein coding gene in *cis*. Other systems such as GRAIL [34] allow users a similar flexibility by requesting either an interval, a gene list or a set of SNPs.

The usefulness of each module for candidate gene prediction seems to depend on the heritability of the phenotype. For Mendelian diseases, CPS is the more effective of the two modules. Because of the small number of loci typically involved in Mendelian disease, it works best when there is some prior knowledge of the genes underlying the phenotype. For complex diseases, where it is believed multiple genes are involved, both modules, common pathways and common functional domains, appear to be effective. Provided enough significant loci are supplied, significant predictions can be made for complex diseases either with or without prior knowledge [4].

Previous validation

Validation studies of the *Gentrepid* approach have been published previously for both Mendelian diseases [10] as well as complex diseases [6]. To summarize, for Mendelian diseases benchmark tests on a set of 170 disease genes for 29 diseases showed that the CMP and CPS methods have a combined sensitivity of 0.52 and a specificity of 0.97, and reduce the candidate list by 13-fold [10]. These tests were performed using artificially constructed loci of 50, 100 or 150 genes around the target disease genes, numbers which are typical of linkage-based disease loci.

Complex diseases are more difficult to benchmark due to limited knowledge of the underlying genetics. In a comparison using loci previously determined by linkage analysis against 11 highly significant Type 2 Diabetes GWAS SNPs, *Gentrepid* had a sensitivity of 0.18 and a specificity of 0.96 while reducing the candidate list by 19-fold [6]. In a second benchmark on seven complex diseases using GWAS SNP data, *Gentrepid* was capable of extracting known disease genes and predicting plausible novel disease genes in known and novel loci. Depending on the size of the search space used, the system had a sensitivity that ranged between 0.09 and 1, specificity between 0.55 and 1, and enrichment ratios up to 25-fold [4].

In addition to these benchmark tests, we have used *Gentrepid* in-house to successfully predict a novel gene for autosomal recessive spondylocostal dysostosis [35] and analyse mutations in *MESP2*, *LNFG* and *HES7* [36]. An advantage of *Gentrepid*'s biological clustering approach is that, as it does not use machine learning algorithms, it does not rely on training data sets and thus avoids concomitant problems such as model overfitting [37].

Case study: hypertension and blood pressure

Hypertension (HTN) is a medical condition where blood pressure in the arteries is significantly elevated resulting in increased risk for cardiovascular disease, kidney disease or stroke. To demonstrate the application of the *Gentrepid* system in the analysis of GWAS data, we ran our analyses on a set of 29 significantly associated loci from a meta-GWAS performed by the International Consortium for Blood Pressure (ICBP) [38-40] where the phenotypes investigated were hypertension (HTN), systolic blood pressure (SBP), and diastolic blood pressure (DBP). The study reported 45 likely candidate and confirmed disease genes (Additional file 1: Table S1). We collated a set of 23 HTN-implicated genes from OMIM as seeds (Additional file 1: Table S2), using only genes known prior to the publication of GWAS so as not to skew our results. The seed genes are involved in pathways that regulate blood pressure and blood volume such as: calcium signalling, the renin-angiotensin system, and hormone metabolism [41-43]. Using the 29 significant loci reported, we generated a gene search space based on SNP-gene proximity labelled the *adjacent* approach where we pool genes adjacent to the associated SNP both upstream and downstream on the + and - strands [4]. We also generated a second gene search space based on SNP-gene distance, labelled the *bystander* approach. We used a 1Mbp window centred on each SNP and pooled genes within that interval. We then ran *Gentrepid* in *seeded* mode, with the 23 seed genes; and in *ab initio* mode, where no additional genotype/phenotype information is used. We report the predictions made by *Gentrepid*. We compared these predictions to the reported candidates from the study; which themselves, may or may not be the causal gene. Predictions different to the reported genes in the ICBP study [38] are annotated as alternate predictions.

From the 29 loci implicated by the ICBP study [38], the *adjacent* mapping generated a search space of 77 genes, as some loci mapped to fewer than 4 genes. From these 77 genes, *Gentrepid* returned 19 gene predictions for 15 of the 29 loci (Table 1). Of these 19 predictions, 12 were gene candidates reported by the ICBP study to be the likely disease genes. *Gentrepid* thus made 7 alternate gene predictions in 7 loci. Of these alternate predictions, many are in generic signalling pathways (eg. MAPK signalling, $p = 0.08$). Although the mitogen-activated protein kinase (MAPK) cascade is involved in

Table 1 Comparison of reported genes and candidate disease genes predicted by *Gentrepid* using the gene search space from the *adjacent* mapping

Reported SNP	Reported gene(s) by study	Predicted candidate gene(s)	Method	Common property used in prediction	Seed gene	Score	Relative rank
rs419076	*MDS1*, *EVI1*	*EVI1*	CMP-ab	zf-C2H2	-	S = 5.17	3
rs13107325	*SLC39A8*	*SLC39A8*	CMP-ab	Zip	-	S = 150.96*	1
		NFKB1	CPS-s	Corticosteroids & cardioprotection	*NOS3*	P = 0.09	1
rs13139571	*GUCY1A3*, *GUCY1B3*	*GUCY1A3*, *GUCY1B3*	CPS-ab	Purine metabolism	-	P = 0.23	7
			CPS-s	Long-term depression	*NOS2A*, *NOS3*	P = 0.39	7
rs4373814	*CACNB2*	*SLC39A12*	CMP-ab	Zip	-	S = 150.96*	1
rs932764	*PLCE1*	*PLCE1*	CPS-ab	Calcium signaling pathway	-	P = 0.31	11
			CPS-s	Calcium signaling pathway	*AGTR1*, *NOS2A*, *NOS3*	P = 0.31	5
rs7129220	*ADM*	*AMPD3*	CPS-ab	Purine metabolism	-	P = 0.23	9
rs2521501	*FURIN*, *FES*	*FES*	CMP-ab	Pkinase_Tyr	-	S = 13.97	2
rs17608766	*GOSR2*	*WNT9B*	CPS-ab	Basal cell carcinoma	-	P = 0.05*	3
rs6015450	*GNAS*, *EDN3*, *ZNF831*, *MRPS16P*	*C20orf174*	CMP-ab	zf-C2H2	-	S = 5.17	3
rs3774372	*ULK4*	*CTNNB1*	CPS-ab	Cell to Cell Adhesion Signaling	-	P = 0.01*	1
rs1458038	*FGF5*	*FGF5*	CPS-ab	MAPK signaling pathway	-	P = 0.08	5
		PRDM8	CMP-ab	zf-C2H2	-	S = 5.17	3
rs1813353	*CACNB2*	*CACNB2*	CPS-ab	MAPK signaling pathway	-	P = 0.08	5
rs17249754	*ATP2B1*	*ATP2B1*	CPS-ab	Calcium signaling pathway	-	P = 0.31	11
			CPS-s	Calcium signaling pathway	*AGTR1*, *NOS2A*, *NOS3*	P = 0.31	5
rs1378942	*CYP1A2*, *CSK*	*CYP1A2*	CMP-s	p450	*CYP3A5*	S = 0.15	1
		CSK	CPS-s	PPIN	*ACSM3*, *ADD1*	-	-
rs12940887	*ZNF652*	*ZNF652*	CMP-ab	zf-C2H2	-	S = 5.17	3

All predictions are reported. CMP/CPS s: *seeded*. ab :*ab initio*. Score (S) or p-value (P) reported for genes predicted. Significant scores or p-values are marked with an *. A dash (–) indicates no prediction. Relative rank is the rank of the gene in the method used. For PPIN, no rank has been assigned.

various cellular functions including vascular oxidative stress, it is debated whether this is solely a symptom of hypertension [44]. The study suggests that variant alleles of the MAPK signalling pathway predispose individuals to hypertension.

The *1Mbp bystander* approach generated a search space of 386 genes. *Gentrepid* returned predictions for 25 of the 29 implicated loci, with a total of 108 gene predictions (results not shown). Of the 25 loci with predictions, 15 gene predictions were congruent with those reported by the ICBP study. Alternate predictions of note include the nuclear factor NF-kappa-B (*NFKB1*), predicted via the "Corticosteroids and cardioprotection" pathway ($p = 9.99e\text{-}5$). Corticosteroids exert a variety of actions by binding to the glucocorticoid receptor and may play a role in increased water excretion in the kidneys to reduce blood volume and atrial pressure [45]. In addition to the ZIP domain ion transporter *SLC39A8* reported by the ICBP, *Gentrepid* CMP predicted the solute carrier *SLC39A12* that, like *SLC39A8*, has a Zinc transporter domain (Zip, Pfam PF02535). The ICBP study postulated that *SLC39A8* may play a role in disease through its transport of cadmium [46,47].

As demonstrated, both SNP/gene mapping approaches are useful to obtain likely candidates. We recommend the *adjacent* approach as a first step, and then the 1Mbp *bystander* approach can be used to generate possible candidates further from the associated locus.

Limitations and future directions

Gentrepid, as with all tools in the field, is limited by the current knowledge database. In an effort to compensate for the missing data, *Gentrepid* uses multiple modules (CPS, CMP) which gather data from multiple sources (protein-protein interactions, pathways, protein domains). Other protein function data that has been employed include tissue-specific expression levels [48-50], post-translational modifications [51], phylogenetic lineage [52], and other ontological classifications [53]. With the publication of regulatory annotation data such as those from the ENCODE project [54], a future direction for the tool would be to use such regulatory information as a likely candidate disease gene prediction module. Genes targeted by common transcription factors, or miRNAs, amongst other elements that affect gene expression, have also been disease associated and would be a useful update to the system [55,56].

Conclusions

The *Gentrepid* web server facilitates the prediction and prioritization of candidate disease genes for both Mendelian and complex diseases using two complementary approaches, namely Common Pathway Scanning and Common Module Profiling. The ability to apply different approaches separately enables the application of different prioritization strategies to different categories of disease—for instance, the optimal approach for the prioritization of candidate disease genes for Mendelian diseases might differ from that for complex diseases, or for cancer. The alternate predictions made by *Gentrepid* for hypertension and blood pressure traits are interesting candidates that require further validation. The system has shown to be capable of both replicating known or reported candidates and also making novel plausible predictions, demonstrating the usefulness of *Gentrepid* in *de novo* analysis and reanalysis of GWAS data. In the future, the addition and integration of yet more data types will further increase the utility of *Gentrepid* in candidate disease gene prediction and prioritization, for all types of diseases.

Availability and requirements

Gentrepid is available at https://www.gentrepid.org. It requires no special or additional data sources, other than the input data (genetic loci) from the user.

Additional file

Additional file 1: Table S1. Novel implicated loci, and reported candidate genes from the ICBP study. **Table S2.** OMIM hypertension associated genes used as seeds for the *seeded* disease gene approach.

Abbreviations

BP: Blood pressure; CPS: Common Pathway Scanning; CMP: Common Module Profiling; DBP: Diastolic blood pressure; GWAS: Genome-wide association studies; HTN: Hypertension; LD: Linkage disequilibrium; PPI: Protein-protein interactions; OMIM: Online Mendelian Inheritance in Man; SBP: Systolic blood pressure; SNP: Single nucleotide polymorphism; STS: Sequence tag sites.

Competing interest

The authors declare that they have no competing interest.

Authors' contributions

SB updated and validated V2.0. JYL developed the database, made the webpages and wrote the database queries. NB supervised construction of the database. AL set-up the webserver. RAG and MAW supervised construction of the website. SB, MO and MAW wrote the manuscript. MO worked on the CMP predictions. BG, DF and MB participated in the design of the study. MAW conceived the study, participated in its design and reviewed the results from the data analysis. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank industrial training students who contributed to the project including Rosemary Sadsad (database), Ignatius Pang (annotations) and Boer Xu (Gviewer implementation).

Author details

[1]Structural and Computational Biology Department, Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia. [2]School of Computer Science and Engineering, University of New South Wales, Kensington, NSW 2052, Australia. [3]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands. [4]School of Medical Sciences, University of New South Wales, Kensington, NSW 2052, Australia. [5]Molecular Cardiology and Biophysics Division, Victor Chang Cardiac Research Institute, Darlinghurst, NSW 2010, Australia. [6]School of

Medicine, Deakin University, Geelong, VIC 3217, Australia. [7]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, 500 Sunnyside Boulevard, 11797, Woodbury, NY, USA.

Received: 22 November 2012 Accepted: 13 August 2013
Published: 16 August 2013

References

1. Cardon LR, Bell JI: **Association study designs for complex diseases.** *Nat Rev Genet* 2001, **2**(2):91–99.
2. Cantor RM, Lange K, Sinsheimer JS: **Prioritizing GWAS results: a review of statistical methods and recommendations for their application.** *Am J Hum Genet* 2010, **86**(1):6–22.
3. Tranchevent LC, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y: **A guide to web tools to prioritize candidate genes.** *Brief Bioinform* 2011, **12**(1):22–32.
4. Ballouz S, Liu J, Oti M, Gaeta B, Fatkin D, Bahlo M, Wouters M: **Analysis of genome-wide association study data using the protein knowledge base.** *BMC Genet* 2011, **12**(1):98.
5. Badano JL, Katsanis N: **Beyond mendel: an evolving view of human genetic disease transmission.** *Nat Rev Genet* 2002, **3**(10):779–789.
6. Teber E, Liu J, Ballouz S, Fatkin D, Wouters M: **Comparison of automated candidate gene prediction systems using genes implicated in type 2 diabetes by genome-wide association studies.** *BMC Bioinforma* 2009, **10**(Suppl 1):S69.
7. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics* 2006, **22**(6):773–774.
8. Alarcón-Riquelme M: **Role of *RUNX* in autoimmune diseases linking rheumatoid arthritis, psoriasis and lupus.** *Arthritis Res Ther* 2004, **6**:169–173.
9. Franke L, Van Bakel H, Fokkens L, De Jong ED, Egmont-Petersen M, Wijmenga C: **Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes.** *Am J Hum Genet* 2006, **78**(6):1011–1025.
10. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction.** *Nucleic Acids Res* 2006, **34**(19):e130.
11. Tranchevent LC, Barriot R, Yu S, Van Vooren S, Van Loo P, Coessens B, De Moor B, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic Acids Res* 2008, **36**(suppl_2):W377–W384.
12. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**(suppl_1):D61–D65.
13. Brown KR, Jurisica I: **Online predicted human interaction database.** *Bioinformatics* 2005, **21**(9):2076–2082.
14. Bader GD, Betel D, Hogue CWV: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31**(1):248–250.
15. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, Cesareni G: **MINT: a Molecular INTeraction database.** *FEBS Lett* 2002, **513**(1):135–140.
16. Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, *et al*: **Development of human protein reference database as an initial platform for approaching systems biology in humans.** *Genome Res* 2003, **13**(10):2363–2371.
17. Nishimura D: **BioCarta.** *Biotech Software & Internet Report* 2001, **2**(3):117–120.
18. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M: **The KEGG resource for deciphering the genome.** *Nucleic Acids Res* 2004, **32:**D277–D280.
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29**(1):308–311.
20. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2002, **30**(1):52–55.
21. Emanuelsson O, Brunak S, Von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools.** *Nat Protoc* 2007, **2**(4):953–971.
22. Delorenzi M, Speed T: **An HMM model for coiled-coil domains and a comparison with PSSM-based predictions.** *Bioinformatics* 2002, **18**(4):617–625.
23. Wolf E, Kim PS, Berger B: **MultiCoil: a program for predicting two-and three-stranded coiled coils.** *Protein Sci* 1997, **6**(6):1179–1189.
24. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**(3):567–580.
25. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18:**S110–S115.
26. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30**(1):276–280.
27. Eddy S: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**(9):755–763.
28. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci U S A* 1988, **85**(8):2444–2448.
29. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *J Med Genet* 2006, **43**(8):691–698.
30. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes.** *Nature* 2001, **409**(6822):853–855.
31. George RA, Heringa J: **Protein domain identification and improved sequence similarity searching using PSI-BLAST.** *Proteins* 2002, **48**(4):672–681.
32. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D: **GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support.** *Bioinformatics* 1998, **14**(8):656–664.
33. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12**(6):996–1006.
34. Raychaudhuri S, Plenge RM, Rossin EJ, Ng ACY, Purcell SM, Sklar P, Scolnick EM, Xavier RJ, Altshuler D, Daly MJ, *et al*: **Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions.** *PLoS Genet* 2009, **5**(6):e1000534.
35. Sparrow DB, Guillén-Navarro E, Fatkin D, Dunwoodie SL: **Mutation of hairy-and-enhancer-of-split-7 in humans causes spondylocostal dysostosis.** *Hum Mol Genet* 2008, **17**(23):3761–3766.
36. Sparrow DB, Sillence D, Wouters MA, Turnpenny PD, Dunwoodie SL: **Two novel missense mutations in hairy-and-enhancer-of-split-7 in a family with spondylocostal dysostosis.** *Eur J Hum Genet* 2010, **18**(6):674–679.
37. Dietterich T: **Overfitting and undercomputing in machine learning.** *ACM computing surveys (CSUR)* 1995, **27**(3):326–327.
38. Ehret GB, Munroe PB, Rice KM, Bochud M, Johnson AD, Chasman DI, Smith AV, Tobin MD, Verwoert GC, Hwang S-J: **Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk.** *Nature* 2011, **478**(7367):103–109.
39. Levy D, Ehret GB, Rice K, Verwoert GC, Launer LJ, Dehghan A, Glazer NL, Morrison AC, Johnson AD, Aspelund T, *et al*: **Genome-wide association study of blood pressure and hypertension.** *Nat Genet* 2009, **41**(6):677–687.
40. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S: **Genome-wide association study identifies eight loci associated with blood pressure.** *Nat Genet* 2009, **41**(6):666–676.
41. Torielli L, Tivodar S, Montella RC, Iacone R, Padoani G, Tarsini P, Russo O, Sarnataro D, Strazzullo P, Ferrari P, *et al*: **α-Adducin mutations increase Na/K pump activity in renal cells by affecting constitutive endocytosis: implications for tubular Na reabsorption.** *Am J Physiol Renal Physiol* 2008, **295**(2):F478–F487.
42. Fyhrquist F, Saijonmaa O: **Renin-angiotensin system revisited.** *J Intern Med* 2008, **264**(3):224–236.
43. Carey RM, Siragy HM: **Newly recognized components of the renin-angiotensin system: potential roles in cardiovascular and renal regulation.** *Endocr Rev* 2003, **24**(3):261–271.
44. Grossman E: **Does increased oxidative stress cause hypertension?** *Diabetes Care* 2008, **31**(Supplement 2):S185–S189.
45. Guijarro C, Egido J: **Transcription factor-kappa B (NF-kappa B) and renal disease.** *Kidney Int* 2001, **59**(2):415–424.
46. He L, Wang B, Hay EB, Nebert DW: **Discovery of ZIP transporters that participate in cadmium damage to testis and kidney.** *Toxicol Appl Pharmacol* 2009, **238**(3):250–257.
47. Liu Z, Li H, Soleimani M, Girijashanker K, Reed JM, He L, Dalton TP, Nebert DW: **Cd 2+ versus Zn 2+ uptake by the ZIP8-dependent symporter: kinetics, electrogenicity and trafficking.** *Biochem Biophys Res Commun* 2008, **365**(4):814–820.

48. Liu X, Yu X, Zack D, Zhu H, Qian J: **TiGER: a database for tissue-specific gene expression and regulation.** *BMC Bioinforma* 2008, **9**(1):271.
49. Edgar R, Domrachev M, Lash AE: **Gene expression omnibus: NCBI gene expression and hybridization array data repository.** *Nucleic Acids Res* 2002, **30**(1):207–210.
50. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M: **ArrayExpress-a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Res* 2007, **35**(suppl 1):D747–D750.
51. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28**(1):45–48.
52. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN: **The COG database: an updated version includes eukaryotes.** *BMC Bioinforma* 2003, **4**(1):41.
53. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R: **The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology.** *Nucleic Acids Res* 2004, **32**(suppl_1):D262–D266.
54. Consortium EP: **A user's guide to the encyclopedia of DNA elements (ENCODE).** *PLoS Biol* 2011, **9**(4):e1001046.
55. Meola N, Gennarino V, Banfi S: **MicroRNAs and genetic diseases.** *PathoGenetics* 2009, **2**(1):7.
56. Kleinjan DJ, Coutinho P: **Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease.** *Brief Funct Genomic Proteomic* 2009, **8**(4):317–332.

doi:10.1186/1471-2105-14-249
Cite this article as: Ballouz *et al.*: ***Gentrepid*** **V2.0: a web server for candidate disease gene prediction.** *BMC Bioinformatics* 2013 **14**:249.