Prioritization of Positional Candidate Genes Using Multiple Web-Based Software Tools

Tobias A. Thornblad,[1] Kate S. Elliott,[2] Jeremy Jowett,[3] and Peter M. Visscher[1]

[1] *Genetic Epidemiology, Queensland Institute of Medical Research, Brisbane, Australia*
[2] *Wellcome Trust Centre for Human Genetics, University of Oxford, United Kingdom*
[3] *International Diabetes Institute, Caulfield, Victoria, Australia*

The prioritization of genes within a candidate genomic region is an important step in the identification of causal gene variants affecting complex traits. Surprisingly, there have been very few reports of bioinformatics tools to perform such prioritization. The purpose of this article is to investigate the performance of 3 positional candidate gene software tools available, *PosMed*, *GeneSniffer* and *SUSPECTS*. The comparison was made for 40, 20 and 10 Mb regions in the human genome centred around known susceptibility genes for the common diseases breast cancer, Crohn's disease, age-related macular degeneration and schizophrenia. The known susceptibility gene was not always ranked highly, or not ranked at all, by 1 or more of the software tools. There was a large variation between the 3 tools regarding which genes were prioritized, and their rank order. *PosMed* and *GeneSniffer* were most similar in their prioritization gene list, whereas *SUSPECTS* identified the same candidate genes only for the narrowest (10 Mb) regions. Combining 2 or all of the candidate gene finding tools was superior in terms of ranking positional candidates. It is possible to reduce the number of candidate genes from a starting set in a region of interest by combining a variety of candidate gene finding tools. Conversely, we recommend caution in relying solely on single positional candidate gene prioritization tools. Our results confirm the obvious, that is, that starting with a narrower positional region gives a higher likelihood that the true susceptibility gene is selected, and that it is ranked highly. A narrow confidence interval for the mapping of complex trait genes by linkage can be achieved by maximizing marker informativeness and by having large samples. Our results suggest that the best approach to classify a minimum set of candidate genes is to take those genes that are prioritized by multiple prioritization tools.

Comprehending genetic variation is imperative in order to understand the causes of hereditary diseases. Currently, the phenotypes of nearly 4000 genetic disorders are known, and the molecular basis is unknown for approximately one half of them (McKusick-Nathans Institute for Genetic Medicine). Genetic disorders may be transmitted by a single defective gene (Mendelian or monogenic disorders), from the culmination of many single defective genes giving rise to the same disorder, or through many genes with additive or multiplicative effects (complex disorders). Genome-wide linkage analyses using pedigrees have been successfully used to identify mutations affecting Mendelian disorders, but less so for complex traits (Abecasis et al., 2000; Morton, 2003).

The resolution of primary genome screens for genetic linkage is usually restricted to no less than 10 to 30 cM, corresponding in humans to approximately 10 to 30 megabase (Mb) pairs of DNA. A region of interest of this size can harbor several hundreds of genes. It is therefore of great importance to be able to prioritize genes before commencing time-consuming and costly wet laboratory work. The traditional approach to reduce the number of candidate genes is by fine-mapping studies using more markers and pedigrees (e.g., Glazier et al., 2002), followed by population-wide linkage disequilibrium studies (e.g., Abecasis et al., 2000).

Another approach to identifying candidate genes involved in complex disorders is by using available knowledge about biological pathways to prioritize positional candidate genes (Morton, 2003). However, when measured in the number of articles and journals that are published, the knowledge of biological systems is increasing at a considerable rate, making it impossible for a researcher, even on a specialized topic, to be up to date with all the relevant literature. A vast number of literature-mining tools have been made in order to identify relevant papers, and more advanced text-mining tools can be used to make novel hypotheses by combining information from multiple papers (Jensen et al., 2006). An example of its usefulness is presented by text-mining scientific literature for terms from an

Received 20 June, 2007; accepted 27 June, 2007.

Address for correspondence: Tobias Thornblad, Genetic Epidemiology, Queensland Institute of Medical Research, 300 Herston Road, Brisbane 4029, Australia. E-mail: tobias.thornblad@gmail.com

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

anatomical ontology, and then integrating the results with gene expression data (Tiffin et al., 2005). Ontologies define the terms within a specific subject area, and the gene ontology (GO), which describes molecular function, process, and location of action of a protein in a generic cell, is often used in text-mining, although Tiffin et al. did not use GO (Blaveri et al., 2001; Tiffin et al., 2005). Another type of useful input in text-mining is InterPro entries, which are described by at least one signature each, corresponding to a biologically meaningful domain, repeat, family, or pretrans-splicing molecule. Recently, InterPro entries were mapped to GO terms, so that a term applies to the number of proteins that matches that entry (Mulder et al., 2003).

The purpose of this study was to examine available software tools that prioritize positional candidate genes, and to investigate if the use of multiple tools could enhance the ranking of candidates. In contrast to recently published papers (e.g., Tiffin et al., 2006), we used only three web-based software tools, which were chosen because of their similar inputs to make the comparison as fair as possible, and to give an overview of the results as unambiguous as possible. The purpose of this article was not only to investigate the performance of the individual programs, but also to illustrate the usefulness of these web-based tools by applying them to the analysis of complex disease. We chose to work with four genetically complex disorders in the chromosomal regions where we already knew what genes that were involved. We subsequently formed the hypothesis that those known genes would be indicated as the most significant for each disease, under the assumption that the investigated tools worked perfectly. We then investigated whether the three readily available web-based tools could predict and accurately rank the genes that we expected in regions spanning 40, 20 and 10 Mb centred around the disease-causing loci. To compare and contrast the three software tools, we have produced lists of candidate genes using the different programs, and then looked for overlapping sets of genes. We have also compiled a mean rank of the candidate genes that were recognized by all three programs.

Methods

In our study we utilized three web-based software tools, *SUSPECTS*, *PosMed,* and *GeneSniffer*. The combination of three software tools provides a more complex selection of candidate genes, as opposed to using only one of the three programs, as all of the programs use different approaches to assign ranks to the available genes in each region of interest. Furthermore, it provides an approach to reduce the total number of candidate genes by selecting the overlapping genes ranked in the top 20 for all three programs for the same interval. The comparison between the three software tools was general, and should be interpreted mainly as an effort to contrast and highlight possibilities and limitations of using a combination of multiple programs. The essential focus of the study was on the accurate identification of candidate genes, rather than nominating one of the programs as superior to the others in finding these.

Software Tools

PROSPECTR and SUSPECTS (http://www.genetics.med.ed.ac.uk/SUSPECTS/). SUSPECTS ranks the candidate genes in the region of interest in the order of their likelihood of involvement in a specified disease (Adie et al., 2005). SUSPECTS retrieves a list of genes, a match set, implicated in the specified complex disease from the Human Gene Mutation Database (HGMD) and the Genetic Association Database (GAD; Becker et al., 2004; Stenson et al., 2003). Each of the candidate genes in the specified genetic interval is then compared to the match set. The comparison is scored in four different ways, which are then weighted and averaged to produce a final rank. The first score is given by PROSPECTR, which is based on sequence features of the genes. PROSPECTR classifies genes to be likely or unlikely to be involved in hereditary disease, based on previously shown results that these genes are more likely to have longer 3’ UTRs, signal peptides, and a higher percentage of bases conserved during evolution (Adie et al., 2005). The candidate genes are then given scores based on a comparison to the expression profiles found in the match set, using Spearman’s rho rank-order correlation. The remainder of the scores are given according to the number of shared Interpro domains, with the match set, and the number of GO annotations that are semantically similar, at a significant level, to the annotations found in the match set (Lord et al., 2003).

PosMed (http://omicspace.riken.jp/PosMed/search). *PosMed* has been developed by the Genome–Phenome Superbrain Project and Phenome Informatics Team at the Genomic Sciences Center, RIKEN, Japan. The inputs used in *PosMed* are phenotypic keywords and a genetic interval, and are thus the same as for *SUSPECTS*. These inputs are subsequently run as two different full-text searches. The first full-text search is carried out against the biological literature of MEDLINE . The system computes a *p* value, which determines the order of rank, for each candidate gene with a significance based on the number of hits in documents by the specified keyword and the name of the gene. Hence, ranking a candidate gene as more likely to be involved in a specific disease if it is mentioned in a higher number of medical articles, associated to that disorder, than a candidate gene with a lower significance. The second full-text search is done against OMIM and MGI locus records for which the system subsequently displays the number of hits (Blake et al., 2006).

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

GeneSniffer (http://www.GeneSniffer.org/). The third web-based tool, used in this article for finding and ranking candidate genes utilized, is *GeneSniffer*, developed by one of the authors, Kate Elliot. *GeneSniffer* also uses the inputs; phenotypic keywords and a genomic interval. The system then downloads the appropriate webpages from the NCBI's Gene, OMIM, and PubMed databases, and from Jackson's MGI database, for each of the genes in the genetic interval (Blake et al., 2006; Pruitt et al., 2000). The resulting database is subsequently interrogated, using computer-intensive database mining, for the phenotypic specific keywords. BLAST then identifies homologs for each of the genes in the region of interest. These are then scored for their OMIM, Gene, PubMed, and Jackson entries. The degree of homology then determines how the scores for each gene should be weighted, and a cumulative score, based on the number of hits, is used to rank the candidate genes according to their predicted involvement in a disease.

Diseases

A set of four genetically complex trait disorders, and the corresponding genes known to be associated with the diseases was compiled (described in Table 1). A brief description of the diseases follows. The locations of the associated genes were used as guidelines of where genetic intervals were to be examined. The three intervals constructed from each of these positions spanned 40, 20, and 10 Mb (Table 1) for each of the diseases, on the chromosome where each of the associated genes are found.

Breast cancer [BRCA1]. Cancer evolves from a series of mutations of genes involved in the regulation of cell differentiation and proliferation. Inherited mutations in the breast cancer susceptibility gene 1, BRCA1 (17q21), have, over the last 10 years, been found to be associated with a role in disrupting DNA damage signalling, repair, and cell cycle checkpoints (Katiyar et al., 2006).

Crohn's disease [CARD15]. Crohn's disease lacks a simple Mendelian transmission pattern and involves several susceptibility genes, which make the genetic factors involved highly complex. The presence of a susceptibility gene has been demonstrated for the development of CD on chromosome 16 (16q21). C-terminal Caspase Recruitment Domain 15, CARD15 (initially called NOD2), has been associated with a 2-3 fold increase in the risk for CD in the presence of one polymorphism, while the presence of two other polymorphisms represent an increase of 20-40 fold (Mendoza & Taxonera, 2005).

Macular degeneration [CFH]. One of the major causes of blindness in the elderly is characterized by progressive destruction of the retina's central region, the macula, causing central field visual loss. It has been shown that an intronic and common polymorphism in the Complement Factor H gene, CFH, is strongly associated with age-related macular degeneration. CFH, located on chromosome 1q32, is a key regulator in the complement system of innate immunity, which protects against infection, and attacks diseased and dysplastic cells and normally spares healthy cells (Klein et al., 2005).

Schizophrenia [NRG1]. Neuregulin 1, NRG1 (8p21-12), is involved in neurodevelopment, regulation of glutamate,and other neurotransmitter receptor expression, and synaptic plasticity, and has been shown to be a susceptibility gene for schizophrenia (Stefansson et al., 2002; Tosato et al., 2005).

Comparison of *SUSPECTS*, *PosMed* and *GeneSniffer*

A comparison of the ability to find the known susceptibility gene and assign it a high rank was carried out for each program in all regions. The overlap in selecting the same candidate genes across the programs was examined to attempt to reduce the total number of

Table 1

The Inputs Used for Finding Gene Candidates

Disease (gene)	Width of region (Mb)	Genomic location	No. genes in region	Keyword for prioritization
Breast cancer (BRCA1)	40	17p11.2-q23.3	765	Breast cancer
	20	17q11.2-q22	478	
	10	17q12-q21.32	313	
Crohn's disease (CARD15)	40	16p11.2-q22.1	456	Crohn disease
	20	16q11.1-q21	143	
	10	16q11.2-q12.2	70	
Macular degeneration (CFH)	40	1q25.2-q41	347	MD
	20	1q31.1-q32.1	175	
	10	1q31.2-q32.1	67	
Schizophrenia (NRG1)	40	8p23.1-q11.22	272	Schizophrenia
	20	8p21.3-11.21	161	
	10	8p21.2-p12	62	

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

candidates, and to reduce the likelihood that the candidate gene selected was a false positive. The overlap was investigated by comparing the top 20 candidate genes for each region. The top 20 were chosen because this was deemed to be a manageable number of genes to examine for subsequent functional studies. Venn diagrams (presented in the figures section) were created for each of the three regions for all the four disorders, in order to visualize the overlap in identification of candidate genes. The reduction of the number of candidates by combining the three programs was compared to the total number of genes in the starting set. A mean total rank was constructed by choosing candidate genes present in the same regions for the three programs, and subsequently assigning an average ranking score for these. Two mean total ranks were calculated for each disorder, one calculated in the 10Mb region, and another ‘overall mean total rank’ using all three regions. In the case of the mean total rank for the 10 Mb region, the candidate genes ranked among the top 20 for each of the programs in the 10 Mb region were selected, and each assigned an average ranking score and then tabulated.

Results

Breast Cancer

From the starting set of 765 genes for 40 Mb, four were chosen as being among the top 20 of the ranked genes in that region by all three software tools. Another five genes were selected by at least two of the programs as being part of the top 20, resulting in a reduction of 765 genes to a total of 9 candidate genes, as illustrated by Figure 1a (a reduction of 98.2%). When the region is narrowed down to 20 Mb, the same four genes are selected by all three programs, and another seven are chosen by at least two, which corresponds to a reduction of 478 genes to 11 (97.7%), illustrated in Figure 1b. The same four candidate genes are selected when the region spans 10 Mb, but the genes chosen by at least two of the software tools, however, decreases to five, as illustrated in Figure 1c. This corresponds to a reduction of candidates genes from 313 to 9 (97.2%). Both PosMed and GeneSniffer rank ERBB2 and BRCA1 as the two most likely gene candidates (rank 1 and 2) at all ranges. Interestingly, SUSPECTS does not rank BRCA1 at any of these top positions for 40 or 20 Mb, and ERBB2, a gene also known to be involved in cancer biology, does not reach the top at all. Table 2 presents the top

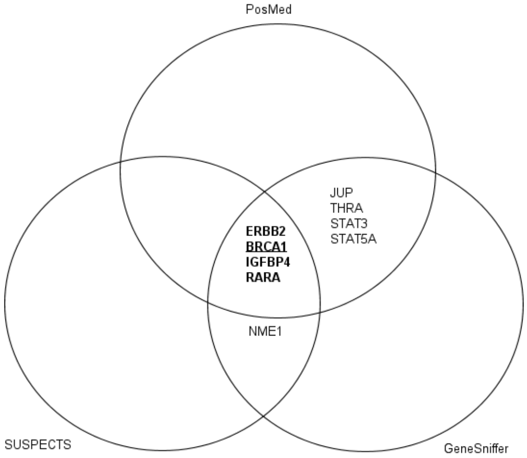

Figure 1a

Candidate genes for breast cancer (40 Mb).

Note: Candidate genes for breast cancer in the genomic region 17p11.2-q23.3, spanning a DNA region of ~ 40 Mb pairs containing 765 genes. The figure illustrates the concurrence of the same candidate genes predicted by the three software packages, *PosMed*, *SUSPECTS* and *GeneSniffer*. The candidate genes in bold were predicted by all three software tools. The ‘true’ gene is underlined.

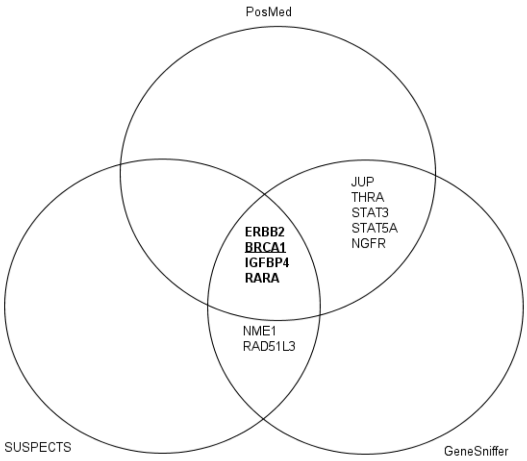

Figure 1b

Candidate genes for breast cancer (20 Mb).

Note: Candidate genes for breast cancer in the genomic region 17q11.2-q22, spanning a DNA region of ~ 20 Mb pairs containing 478 genes.

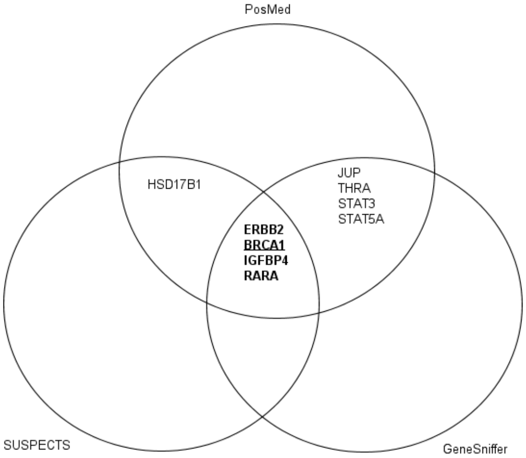

Figure 1c

Candidate genes for breast cancer (10 Mb).

Note: Candidate genes for breast cancer in the genomic region 17q12-q21.32, spanning a DNA region of ~10 Mb pairs containing 313 genes.

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

Table 2

Mean Ranks for All the Software Tools, for All Regions, and the 10 Mb Regions

	Mean total rank for all regions		Mean total rank for the 10 Mb region	
	Candidate gene	Rank	Candidate gene	Rank
Breast cancer	BRCA1	2.89	BRCA1	2
	ERBB2	5.0	ERBB2	3.67
	IGFBP4	7.89	IGFBP4	6.33
	RARA	10.78	RARA	8.33
			THRA	11
Crohn's disease	—*	—*	CARD15	2
	—*	—*	MMP2	3.33
Macular degeneration	—*	—*	CFH	5.33
	—*	—*	PTPRC	6.33
Schizophrenia	CHRNA2	4.89	CHRNA2	2.33
			PNOC	4
			FZD3	4.33
			NRG1	7

Note: The mean total rank is compiled from the candidate genes which are present in the 10, 20 and 40 Mb lists of PosMed, SUSPECTS and GeneSniffer. The mean total rank for the 10 Mb region includes those candidate genes that are present in the 10 Mb region in the lists of all three software tools.

*indicates that no candidate genes were ranked in all regions for all three programs at the same time.

The total range of values used for calculating the mean total rank can be found in Additional File 1.

ranked candidate genes for all regions as a mean total rank, which is also presented for the 10 Mb region for the three software tools. For the whole lists of candidate genes ranked as the top 20 for all regions, please contact the author.

Crohn's Disease

The starting set for Crohn's disease contained 456 genes for 40 Mb. Two of these available genes were chosen by all three software tools as being likely candidates, and another 11 were selected by at least two of the programs, as illustrated in Figure 2a. This decrease in the number of candidates corresponded to a 97.2% reduction. Narrowing the region down to 20 Mb resulted in a reduction from 143 candidate genes to a total of 11 (92.3%), where five of those were selected by all three programs, as described in Figure 2b. The number of candidate genes was reduced significantly when looking at a region of 10 Mb (70 genes). The total number of candidates for that region was five (two selected by all programs), resulting in a reduction of 92.9%. Not a single one of the genes is ranked as being among the top 20 by the three programs for all

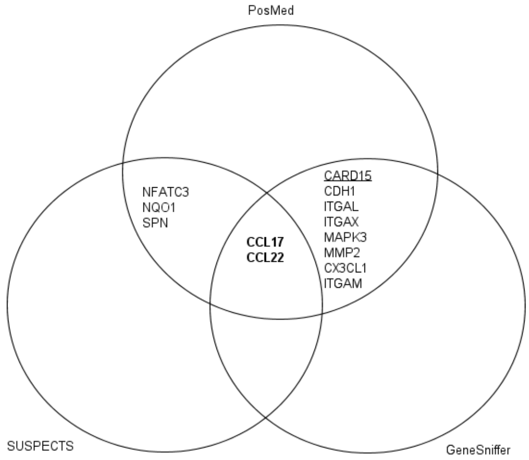

Figure 2a

Candidate genes for Crohn's disease (40 Mb).

Note: Candidate genes for Crohn's disease in the genomic region 16p11.2-q22.1, spanning a DNA region of ~ 40 Mb pairs containing 456 genes.

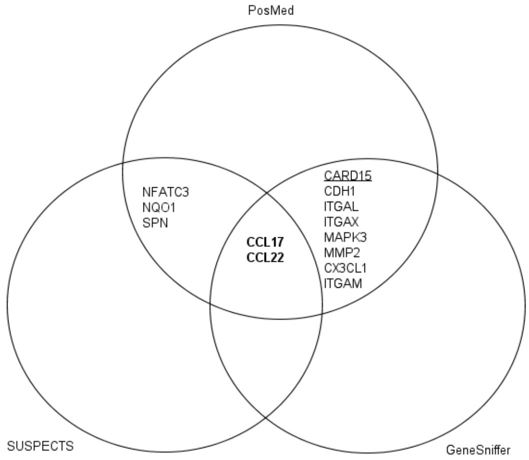

Figure 2b

Candidate genes for Crohn's disease (20 Mb).

Note: Candidate genes for Crohn's disease n the genomic region 16q11.1-q21, spanning a DNA region of ~ 20 Mb pairs containing 143 genes.

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

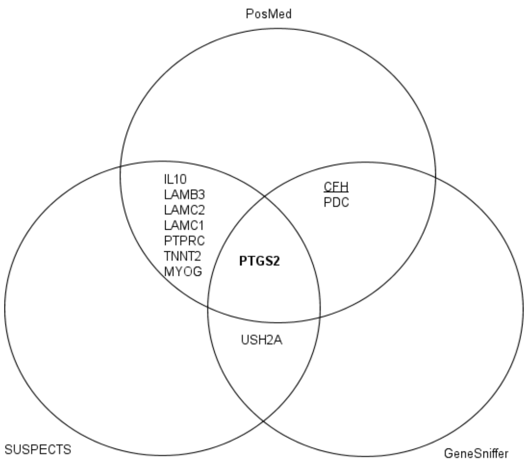

Figure 3a

Candidate genes for macular degeneration (40 Mb)

Note: Candidate genes for macular degeneration in the genomic region 1q25.2-q41, spanning a DNA region of ~ 40 Mb pairs containing 347 genes.

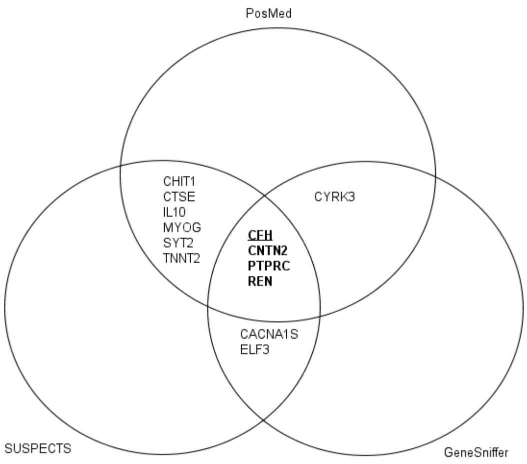

Figure 3b

Candidate genes for macular degeneration (20 Mb).

Note: Candidate genes for macular degeneration in the genomic region 1q31.1-q32.1, spanning a DNA region of ~ 20 Mb pairs containing 175 genes.

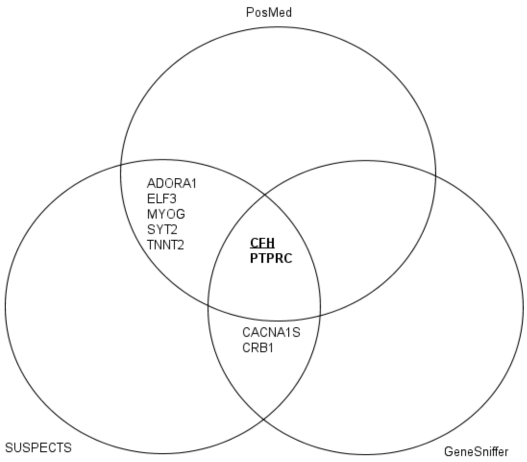

Figure 3c

Candidate genes for macular degeneration (10 Mb).

Note: Candidate genes for macular degeneration in the genomic region 1q31.2-q32.1, spanning a DNA region of ~ 10 Mb pairs containing 67 genes.

intervals. As for the result of breast cancer above, SUSPECTS does not rank the expected gene, in this case CARD15, to be among the top 20 candidates until the interval is reduced to 20 and 10 Mb. Table 2 presents the top ranked candidate genes for the 10 Mb region for the three software tools. For the whole lists of candidate genes ranked as the top 20 for all regions, please contact the author.

Macular Degeneration

The region of interest for macular degeneration contained a set of 347 genes for 40 Mb, from which one gene was selected by all three programs, and 10 genes were chosen by at least two of the programs, which is illustrated in Figure 3a (a reduction of 96.8%). In the 20 Mb region, four genes are selected as likely candidates by all the three programs and another nine by at least two of them (described in Figure 3b), corresponding to a reduction of 92.6%, as the starting set includes 175 genes in this region. Two of the candidate genes chosen in the 20 Mb set for all three programs remain when the region is narrowed down to 10 Mb. For likely candidate genes selected by at least two of the software tools, a total number of seven are chosen (as described in Figure 3c), reducing the amount of genes of interest in the region from 67 to nine (86.6%). As for previous results (above), SUSPECTS did not rank the expected gene, in this case CFH, to be in the top 20 as a top result, until the region was narrowed down to 20 or 10 Mb. Table 2 presents the top ranked candidate genes for the 10 Mb region for the three software tools. For the whole lists of candidate genes ranked as the top 20 for all regions, please contact the author.

Schizophrenia

From the starting set of 272 genes in the 40 Mb region, three were chosen by all three software tools and another 11 were chosen by at least two of the programs (as described in Figure 4a), reducing the amount of genes by 94.9%. In the 20 Mb region, the starting set comprises 161 genes, of which four are chosen by all three programs, and 12 more by at least two of the programs (90%) as described in Figure 4b. When the region was narrowed down to 10 Mb, all the three software tools chose four candidate genes, and another two were selected by two of the programs as described in Figure 4c, resulting in a reduction of 90.3% from the starting set of 62 genes in the region. As for previous results, SUSPECTS did not rank the expected gene, in this case NRG1, to be in the top 20 as a top result, until the region was narrowed down to 10 Mb. Table 2 presents the top

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

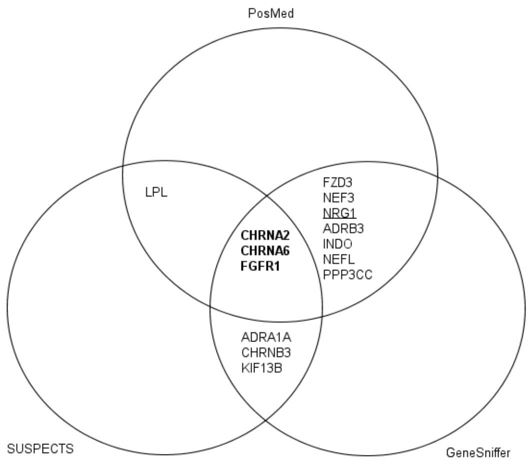

Figure 4a

Candidate genes for schizophrenia (40 Mb)

Note: Candidate genes for schizophrenia in the genomic region 8p23.1-q11.22. spanning a DNA region of ~ 40 Mb pairs containing 272 genes.

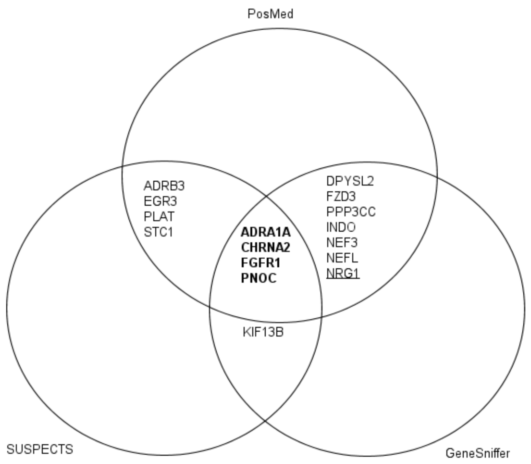

Figure 4b

Candidate genes for schizophrenia (20 Mb).

Note: Candidate genes for schizophrenia in the genomic region 8p21.3-p11.21, spanning a DNA region of ~ 20 Mb pairs containing 161 genes.

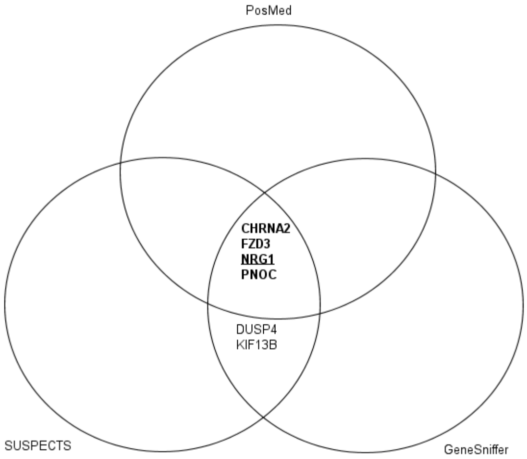

Figure 4c

Candidate genes for schizophrenia (10 Mb).

Note: Candidate genes for schizophrenia in the genomic region 8p21.2-p12, spanning a DNA region of ~ 10 Mb pairs containing 62 genes.

ranked candidate genes for all regions as a mean total rank, which is also presented for the 10 Mb region for the three software tools. For the whole lists of candidate genes ranked as the top 20 for all regions, please contact the author.

Discussion

We have compared the currently available software tools used to prioritize positional candidate genes, using a set of four disorders for which a susceptibility locus was known. These are 'best case scenario' examples, because both the gene and the phenotype were known. When success is defined as the known gene being among the top-20 prioritized set, the identification of the 'correct' gene was 100% of the cases for both *PosMed* and *GeneSniffer*, with *SUSPECTS* succeeding in 67% of the cases. When looking at single intervals, *SUSPECTS* succeeded in one of four cases in the 40 Mb region, three of four in the 20 Mb region, and in all the cases in the 10 Mb region. One possible explanation for the large overlap between *PosMed* and *GeneSniffer* may be that they both use text-mining, in contrast to *SUSPECTS*, making the selection of probable candidate genes biased towards the genes co-mentioned with the phenotypic keyword in the most abstracts. Another conjecture is that the scoring of the characteristics of genes implicated in disease is given too much weight in *SUSPECTS*, resulting in a failure to identify the expected candidate genes.

In the case of breast cancer, a mean total rank for all regions results in a list of four genes all known to be implicated in the disorder. This may be a combination resulting from both the vast knowledge about breast cancer, and the fact that these genes coincidentally lie in very close proximity to each other. The same approach to compile a mean rank for all regions was not successful in Crohn's disease or macular degeneration. Interestingly, in the case of schizophrenia, all programs identify the same candidate gene for all regions, and it is not the one expected. The candidate gene identified, CHRNA2, has been reported as not increasing the susceptibility to schizophrenia, according to one study by Blaveri et al. (2001), but may play a particular role in the nicotinic system in smoking for schizophrenic families, according to another study by Faraone et al. (2004). By limiting the mean total ranking of the three software tools to the 10 Mb region, three of four of the expected candidate genes are ranked as the most probable. This restriction of intervals yields an identification of more candidate genes. The additional candidate gene, THRA, identified for breast cancer, has similarities to the THRA1 gene, which has been shown to be a

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

strong candidate tumour suppressor gene in primary breast cancer (Sourvinos et al., 1997). In the mean total rank for the 10 Mb region of Crohn's disease, the candidate gene MMP2 is identified by all three programs as a likely candidate. The activity of MMP2 has been reported to be associated with quiescent Crohn's disease (Kossakowska et al., 1999). In the case of macular degeneration, the candidate PTPRC (formerly known as CD45), comes up for all three programs in the 10 Mb region. This gene is expressed in microglia, in the human retina, which may be associated with age-related macular degeneration according to a study by Chen et al. (2002). In schizophrenia, three additional candidates were prioritized in the 10 Mb region, one of which is the expected NRG1 gene. The PNOC gene, also identified as a candidate in this region, has been reported as unlikely to increase the susceptibility for schizophrenia (Blaveri et al., 2001). In contrast, the FZD3 gene has been reported to predispose to schizophrenia in a Chinese Han population (Zhang et al., 2004); however, this predisposition was not significant in subsequent studies in a British population (Wei & Hemmings, 2004; Zhang et al., 2004).

In practice, researchers might prefer to use more refined keywords to search for genes and some tools. For example, *GeneSniffer* allows for more elaborate text searches, using multiple disease-relevant keywords. Although this provides a greater depth of searching, our study suggests that caution should be applied. The context of keywords within abstracts may need further examination, for the selected candidate genes, to determine whether they reflect chance or negative associations among the prevalent interesting and positive findings. Researchers might also prefer to use programs that allows for inclusion of additional data sources that would be of interest in the prioritization strategy. Recently, a tool with this feature was developed (Aerts et al., 2006). However, this program, called Endeavour, does not use the same inputs as the three software tools described in this article, which is the reason for not including it in our study. Another approach to make the gene candidate prioritization as complete as possible, is to combine data from several sources to create new functional human gene networks (e.g., Franke et al., 2006).

Conclusions

It is possible to reduce the number of candidate genes from a starting set in a region of interest by combining a variety of candidate gene finding tools. We recommend caution in relying solely on single positional candidate gene prioritization tools. Our results confirm the obvious, that is, that starting with a narrower positional region gives a higher likelihood that the true susceptibility gene is selected and that it is ranked highly. A narrow confidence interval for the mapping of complex trait genes by linkage can be achieved by maximizing marker informativeness and by having large samples (Visscher & Goddard, 2004). Our results suggest that the best approach to classify a minimum set of candidate genes is to take those genes that are prioritized by multiple prioritization tools.

Acknowledgments

The authors wish to thank Dr Norio Kobayashi (Genomic Sciences Centre, RIKEN, Japan) for providing information about *PosMed*, and Georgia Chenevix-Trench and Dale Nyholt for helpful discussions. We thank Chemgenex Pharmaceuticals for allowing use of *GeneSniffer*. This work was supported by Australian NHMRC grant 389892.

References

Abecasis, G. R., Cardon, L. R., & Cookson, W. O. (2000). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, *66*, 279–292.

Adie, E. J., Adams, R. R., Evans, K. L., Porteous, D. J. & Pickard, B. S. (2005). Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, *6*, 55.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nature Biotechnology*, *24*, 719.

Becker, K. G., Barnes, K. C., Bright, T. J., & Wang, S. A. (2004). The genetic association database. *Nature Genetics*, *36*, 431–432.

Blake, J. A., Eppig, J. T., Bult, C. J., Kadin, J. A., Richardson, J. E. ; Mouse Genome Database Group. (2006). The Mouse Genome Database (MGD): Updates and enhancements. *Nucleic Acids Research*, *34(Database issue)*, D562–567.

Blaveri, E., Kalsi, G., Lawrence, J., Quested, D., Moorey, H., Lamb, G., Kohen, D., Shiwach, R., Chowdhury, U., Curtis, D., McQuillin, A., Gramoustianou, E. S., & Gurling, H. M. (2001). Genetic association studies of schizophrenia using the 8p21-22 genes: Prepronociceptin (PNOC), neuronal nicotinic cholinergic receptor alpha polypeptide 2 (CHRNA2) and arylamine N-acetyltransferase 1 (NAT1). *European Journal of Human Genetics*, *9*, 469–472.

Chen, L., Yang, P., & Kijlstra, A. (2002). Distribution, markers, and functions of retinal microglia. *Ocular immunology and inflammation, 10,* 27–39.

Faraone ,S. V., Su, J., Taylor, L., Wilcox, M., Van Eerdewegh, P., Tsuang, M. T. (2004). A novel permutation testing method implicates sixteen nicotinic acetylcholine receptor genes as risk factors for smoking in schizophrenia families. *Human Heredity 57, 59*–68.

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

Franke, L., Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., & Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *American Journal of Human Genetics*, *78*, 1011-1025. Epub 2006 Apr 1025.

Glazier, A. M., Nadeau, J. H.,& Aitman, T. J .(2002). Finding genes that underlie complex traits. *Science*, *298*, 2345–2349.

Jensen, L. J., Saric, J., & Bork, P. (2006). Literature mining for the biologist: From information retrieval to biological discovery. *Nature Review, Genetics, 7,* 119–129.

Katiyar, P., Ma,Y., Fan, S., Pestell, R. G., Furth, P. A., & Rosen, E. M. (2006). Regulation of progesterone receptor signaling by BRCA1 in mammary cancer. *Nuclear Receptor Signaling*, *4,* e006.

Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., Bracken, M. B., Ferris, F. L., Ott, J., Barnstable, C., & Hoh, J. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science, 308,* 385–389.

Kossakowska, A. E., Medlicott, S. A., Edwards, D. R., Guyn, L., Stabbler, A. L., Sutherland, L. R., & Urbanski, S. J. (1999). Elevated plasma gelatinase A (MMP-2) activity is associated with quiescent Crohn's Disease. *Annals of the New York Academy of Sciences*, *878,* 578–580.

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Semantic similarity measures as tools for exploring the gene ontology. *Pacific Symposium on Biocomputing*. 601–612.

McKusick-Nathans Institute for Genetic Medicine, J.H.U.B., MD and National Center for Biotechnology Information. Accessed using PosMed: May 25, 2006 (breast cancer), 12 June 2006 (Crohn's disease), June 14, 2006 (schizophrenia) and July 11, 2006 (macular degeneration). GeneSniffer accessed the database August 20, 2006 (all of the disorders above).

Mendoza, J. L., & Taxonera, C. (2005). Clinical value of gene NOD2/CARD15 mutations in Crohn's disease, *Revista Espanola de Enfermades Digestivas*, *97,* 541–546.

Morton, N. (2003). Genetic epidemiology, genetic maps and positional cloning, *Philosophical Transactions of the Royal Society of London, Series B, Biological Science, 358*, 1701–1708.

Mulder, N. J., Aweiler, R., Attwood, T. K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R. R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S. E., Pagni, M., Peyruc, D., Ponting, C. P., Selengut, J. D., Servant, F., Sigrist, C. J., Vaughan, R., & Zdobnov, E. M. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Research*, *31,* 315–318.

NCBI Entrez Gene Database — PubMed. Accessed using PosMed: May 25, 2006 (breast cancer); June 12, 2006 (Crohn's disease); June 14, 2006 (schizophrenia); July 11, 2006 (Macular degeneration). GeneSniffer accessed the database August 20, 2006 (all of the disorders previously listed).

Online Mendelian Inheritance in Man, OMIM (TM). National Library of Medicine (Bethesda, MD). Accessed using PosMed: May 25, 2006 (breast cancer); June 12, 2006 (Crohn's disease); June 14, 2006 (schizophrenia); July 11, 2006 (macular degeneration). GeneSniffer accessed the database August 20, 2006 (all of the disorders previously listed).

Pruitt, K. D., Katz, K. S., Sicotte, H., & Maglott, D. R. (2000). Introducing RefSeq and LocusLink: Curated human genome resources at the NCBI. *Trends in Genetics*, *16,* 44–47.

Sourvinos, G., Kiaris, H., Tsikkinis, A., Vassilaros, S., & Spandidos, D. A. (1997). Microsatellite instability and loss of heterozygosity in primary breast tumours. *Tumour Biology, 18,* 157–166.

Stefansson, H., Sigurdsson, E., Steinthorsdottir, V., Bjornsdottir, S., Sigmundsson, T., Ghosh, S., Brynjolfsson, J., Gunnarsdottir, S., Ivarsson, O., Chou, T. T., Hjaltason, O., Birgisdottir, B., Jonsson, H., Gudnadottir, V. G., Gudmundsdottir, E., Bjornsson, A., Ingvarsson, B., Ingason, A., Sigfusson, S., Hardardottir, H., Harvey, R. P., Lai, D., Zhou, M., Brunner, D., Mutel, V., Gonzalo, A., Lemke, G., Sainz, J., Johannesson, G., Andresson, T., Gudbjartsson, D., Manolescu, A., Frigge, M. L., Gurney, M. E., Kong, A., Gulcher, J. R., Petursson, H., & Stefansson, K. (2002). Neuregulin 1 and susceptibility to schizophrenia. *American Journal of Human Genetics, 71,* 877–892. Epub 2002, July 23.

Stenson, P. D., Ball, E. V., Mort, M., Phillips, A. D., Shiel, J. A., Thomas, N. S., Abeysinghe, S., Krawczak, M., & Cooper, D. N. (2003). Human Gene Mutation Database (HGMD): 2003 update. *Human Mutation, 21,* 577–581.

Tiffin, N., Adie, E., Turner, F., Brunner, H. G., van Driel, M. A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-Iratxeta, C., Andrade-Navarro, M. A., Adeyemo, A., Patti, M. E., Semple, C. A., & Hide, W. (2006). Computational disease gene identification: A concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Research*, *34*, 3067–3081.

Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B., & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research*, *33*, 1544–1552.

Tosato, S., Dazzan, P., & Collier, D. (2005). Association between the neuregulin 1 gene and schizophrenia: A systematic review. *Schizophrenia Bulletin, 31*, 613–617.

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press

Visscher, P. M., & Goddard, M. E. (2004). Prediction of the confidence interval of QTL location. *Behavior Genetics, 34*, 477–482.

Wei, J. & Hemmings, G. P. (2004). Lack of a genetic association between the frizzled-3 gene and schizophrenia in a British population. *Neuroscience Letter*, *366*, 336–338.

Zhang, Y., Yu, X., Yuan, Y., Ling, Y., Ruan, Y., Si, T., Lu, T., Wu, S., Gong, X., Zhu, Z., Yang, J., Wang, F., & Zhang, D. (2004). Positive association of the human frizzled 3 (FZD3) gene haplotype with schizophrenia in Chinese Han population. *American Journal of Medical Genetics, Part B, Neuropsychiatric Genetics. 129*, 16–19.

https://doi.org/10.1375/twin.10.6.861 Published online by Cambridge University Press