Method

Open Access

POCUS: mining genomic sequence annotation to predict disease genes

Frances S Turner[¤], Daniel R Clutterbuck[¤] and Colin AM Semple

Address: MRC Human Genetics Unit, Crewe Road, Western General Hospital, Edinburgh EH4 2XU, UK.

¤ These authors contributed equally to this work.

Correspondence: Colin AM Semple. E-mail: Colin.Semple@hgu.mrc.ac.uk

Published: 10 October 2003

Genome ***Biology*** 2003, **4:**R75

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2003/4/11/R75

Received: 18 July 2003
Revised: 19 August 2003
Accepted: 17 September 2003

Abstract

Here we present POCUS (prioritization of candidate genes using statistics), a novel computational approach to prioritize candidate disease genes that is based on over-representation of functional annotation between loci for the same disease. We show that POCUS can provide high (up to 81-fold) enrichment of real disease genes in the candidate-gene shortlists it produces compared with the original large sets of positional candidates. In contrast to existing methods, POCUS can also suggest counterintuitive candidates.

Background

Over the past two decades, linkage analysis and positional cloning have been remarkably successful in the identification of human genes responsible for mendelian diseases. Success has been more modest for the more common, complex diseases, because numerous genes with weaker genotype-phenotype correlations are involved [1]. Reports of linkage for one complex disease to many different loci are common in the literature. Unfortunately, the loci implicated are often very large, necessitating the laborious and expensive investigation of hundreds of positional candidate genes. Furthermore, the number of loci implicated per disease is expected to increase as the emerging high-density single-nucleotide polymorphism (SNP) map of the human genome is exploited [2].

Some examples of oligogenic diseases are already documented, where alleles at more than one gene contribute to the same disease, but the molecular basis of such phenomena is often poorly understood, partly because of the lack of functional data for many of the genes involved. It is generally assumed that oligogenicity reflects disruptions in proteins that participate in a common complex or pathway [3]. Where this assumption is accurate, one might expect genes involved in the same disease to share commonalities in their functional annotation, relative to other genes in the genome. At present many human genes lack detailed functional annotation and so these commonalities may often be elusive. Nevertheless, in the wake of the near-complete sequence of the human genome, there is an opportunity to investigate disease susceptibility loci on a large scale, in terms of the likely or known functions of the annotated genes present within them. Three computational methods have recently been developed to exploit this opportunity.

Perez-Iratxeta *et al.* [4] developed a sophisticated treatment of the biomedical literature that associates pathological conditions with particular Gene Ontology (GO) terms, which then allowed candidate disease genes to be ranked according to the number of these terms they share. Freudenberg and Propping [5] produced clusters of known disease genes based on a measure of phenotypic similarity between diseases. Candidate genes were then scored according to the GO terms

shared with known disease genes in the clusters. Van Driel *et al.* [6] developed a web tool that integrates data from mapping, expression and phenotypic databases and allows genes meeting user-defined criteria to be retrieved. All three methods aim to mimic the process that takes place when researchers prioritize positional candidate genes for further study, but also to increase the speed, objectivity and consistency of this process. Essentially, in all three methods, one extrapolates from what is already known about a disease or the genes underlying it to find other promising candidate disease genes. All three methods also implicitly assume that the disease genes we have yet to discover will be consistent with what is already known about a disease and/or its genetic basis. Unfortunately, the literature on genetic susceptibility to disease is rich in unexpected findings and it is not unusual for novel disease genes to be counterintuitive, given the literature on the disease in question.

Here we present POCUS (prioritization of candidate genes using statistics), a novel approach to the computational prediction of disease genes, and report results that have emerged from its application to known disease loci. Our method requires no prior knowledge of the disease under study other than the location of two or more susceptibility loci. Similarly, we make no assumptions about the protein functions, expression, or any other characteristics associated with individual genes involved in the disease. We assume only that two or more of those genes share some aspect of their expression pattern or of the function of the encoded protein. Examination of known disease genes suggests that in most cases this assumption is reasonable. In addition, the method provides an assessment of when the analysis is likely to have failed or been successful with a given set of loci. The basis of our method is the identification of unexpected enrichment of any annotated functional class of genes at a given set of susceptibility loci relative to the genome at large.

Results

Genes underlying complex diseases tend to share functional annotation

A list of 29 Online Mendelian Inheritance in Man (OMIM) diseases was compiled for which three or more contributing genes were known (and were also present within Ensembl), and the degree to which genes for the same disease shared InterPro domain and GO identifiers (IDs) was assessed. Of the 163 disease genes in the 29 sets, 131 (80%) shared an ID with another gene for the same disease, and 102 (63%) share two or more IDs with other genes for the same disease (Table 1). Across the diseases surveyed, 26 of 29 (90%) have contributing genes that share at least one ID. Thus there appears to be a strong tendency for disease genes to share annotation (but see Materials and methods for the potential bias in these data). This extends previous work that showed a correlation between gene functions and features (such as age of onset) of the diseases they predispose to [7].

Prediction of disease genes by over-representation of shared identifiers

The disease genes for each of the 29 diseases were used to produce artificial locus sets containing the disease genes and many flanking genes, and each set was analyzed using POCUS to find genes possessing IDs over-represented among two or more loci. The three locus sizes tested were 100, 500 and 1,000 IDs, and these corresponded to 20, 94 and 187 genes per locus on average, respectively. Assuming a rough average gene density of one every 100 kilobases (kb) [8] the range of physical locus sizes examined was therefore 2-19 megabases (Mb), which is within the range for susceptibility regions identified in positional cloning studies. Various threshold scores were tested, but here we present data only for the most successful (0.8) and a more liberal value (0.5) for comparison. The results refer to the positive control sets of disease genes at the 0.8 threshold unless stated otherwise. POCUS was found to perform differently on the locus sets for different diseases. The method was successful (correctly identifying two or more disease genes) for 65%, 19% and 15% of positive control sets, respectively, at the three locus sizes. Four diseases in particular (epidermolysis bullosa letalis, inflammatory bowel disease, colorectal cancer and cardiomyopathy) were found to be particularly amenable to analysis using POCUS, with many of their disease genes still correctly identified at the largest locus size (Table 1).

For any given locus three outcomes were possible, POCUS may have returned the disease gene (and often others) above the threshold, it may have returned only non-disease genes scoring above the threshold, or it may have returned no genes above the threshold. Figure 1 depicts the rates of each possible outcome per locus. It shows that in 49-75% of loci (depending on locus size), no genes scored above the threshold of 0.8, that is, POCUS was unable to detect any candidates but equally did not return any non-disease genes either. Correspondingly, in 6-15% of loci, only non-disease genes exceeded this threshold, and in the remainder of loci the disease gene was correctly identified (45%, 15% and 11% respectively at the three locus sizes). As Figure 1 shows, compared with 0.5, the more stringent threshold of 0.8, while resulting in a small loss of true positives (correctly identified disease genes), more efficiently reduced the number of false positives (non-disease genes) returned as candidates by POCUS. At the 0.8 threshold, the relative enrichment for disease genes within those genes above the threshold was 12-fold (95% confidence intervals (CI): 9.74-15.83), 29-fold (95% CI: 18.79-43.24) and 42-fold (95% CI: 25.36-69.45), respectively, at the three locus sizes. This means, for example, that any gene from a locus 1,000 IDs in size was 42 times more likely to be the disease gene if it was picked from those genes above the threshold than if it was chosen at random from the locus.

Only modest improvement in POCUS performance was observed with the inclusion of UniGene expression data. Although disease genes often shared the same expression ID

Table 1

The full results of POCUS analysis for 29 OMIM diseases, over locus sizes of 100, 500 and 1,000 IDs at a threshold of 0.8

Disease (representative OMIM number)	Number of genes*	Genes sharing†	Correctly identified‡			Non-disease genes§			Total number of genes¶			Enrichment¥		
			100	500	1,000	100	500	1,000	100	500	1,000	100	500	1,000
Parkinson's disease (168600)	3	0	0	0	0	0	0	1	62	264	571	1	1	0
Lupus erythematosus, systematic (152700)	3	0	0	0	0	1	0	0	43	262	547	0	1	1
Glaucoma, primary open angle, juvenile-onset (137750)	3	2	0	0	0	0	0	0	59	279	527	1	1	1
Bardet Biedl (209900)	4	0	0	0	0	3	0	0	75	390	776	0	1	1
Meningioma, familial (607174)	4	2	1	0	0	6	17	10	69	363	722	2.46	0	0
Acute myelogenous leukemia, familial (601626)	4	4	0	0	0	0	0	0	84	385	771	1	1	1
Basal cell carcinoma (605462)	4	3	2	0	0	0	0	0	83	418	810	20.8	1	1
Adrenoleukodystrophy, autosomal neonatal form (202370)	4	4	2	0	0	0	3	3	76	369	733	19	0	0
Epidermolysis bullosa letalis (226700)	4	4	3	2	2	2	1	4	73	380	728	11	63.3	60.67
Familial adenomatous polyposis (175100)	4	4	3	2	0	8	12	4	64	371	727	4.36	13.3	0
Ovarian carcinoma (167000)	4	4	0	0	0	0	0	0	70	360	733	1	1	1
Hypertension (145500)	5	2	0	0	0	3	0	0	95	472	920	0	1	1
Alzheimer's disease (104300)	5	4	3	0	0	0	0	0	135	460	875	27	1	1
Charcot-Marie-Tooth disease, types 1A-1F (118200)	5	4	3	0	0	0	0	0	98	449	937	19.6	1	1
Gastric cancer (137215)	5	4	0	0	0	0	0	0	87	483	932	1	1	1
Cystic fibrosis (219700)	5	5	1	0	0	2	0	0	99	458	900	6.6	1	1
Inflammatory bowel disease (266600)	5	5	2	0	2	3	0	3	99	506	1,013	7.92	1	81.04
Long-segment Hirschsprung disease (142623)	5	5	0	0	0	0	0	0	102	468	972	1	1	1
Leber congenital amaurosis (204000)	6	5	5	0	0	4	0	0	125	508	1,120	11.6	1	1
Maturity onset diabetes of the young (606391)	6	5	2	0	0	0	0	0	111	551	1,078	18.5	1	1
Prostate cancer (176807)	6	5	0	0	0	2	0	0	128	550	1,157	0	1	1
Colorectal cancer, hereditary nonpolyposis (114500)	6	6	5	6	6	4	14	17	115	560	1,095	10.6	28	47.61
Epiphyseal dysplasia, multiple types 1-5 (132400)	7	6	6	3	0	4	0	0	135	596	1,232	11.6	85.1	1
Muscular dystrophy, limb-girdle, autosomal recessive (601173)	7	5	2	0	0	1	0	0	124	634	1,282	11.8	1	1

Table 1 *(Continued)*

The full results of POCUS analysis for 29 OMIM diseases, over locus sizes of 100, 500 and 1,000 IDs at a threshold of 0.8

Diabetes mellitus, non-insulin dependent (125853)	8	6	2	0	0	0	0	0	155	719	1,420	19.4	1	1
Breast cancer (114480)	9	7	3	0	0	2	0	0	170	819	1,592	11.3	1	1
Retinitis pigmentosa (268000)	10	8	6	0	0	3	7	6	197	977	1,897	13.1	0	0
Cardiomyopathy, familial hypertrophic (192600)	11	11	9	7	4	7	7	15	194	1,011	2,029	9.92	46	38.83
Thyroid carcinoma, papillary (188550)	11	11	0	0	0	1	0	0	217	1,059	2,142	0	1	1

* The total number of genes for a disease. †The number of disease genes that share IDs. ‡The number of disease genes above the threshold at the three locus sizes. §The number of non-disease genes above the threshold at the three locus sizes. ¶The total number of genes present at the loci considered. ¥The enrichment of disease genes in genes above the threshold compared with the initial loci, zeros denote diseases where only non-disease genes were above the theshold.

(that is they were 'solely' or 'highly' expressed in the same EST library according to UniGene), the identifiers were generally too commonly shared in the genome at large to contribute significantly to gene scores. More rigorous expression data may significantly improve POCUS performance.

The diseases were grouped by number of disease genes to investigate the effects of locus number. Enrichment levels were calculated for two categories of disease gene sets, 3-5 loci and 6-11 loci, and enrichment was found to vary with the number of loci. At each locus size the 6-11-loci category was found to have higher levels of enrichment by a factor of 1.55-3. As larger loci on average contain more genes, one might expect more modest success at higher locus sizes and, indeed, this effect was evident in a decreasing true-positive rate (Figure 1). However, it is also notable that at higher locus sizes the enrichment factor increased over the three locus sizes (12-fold, 29-fold and 42-fold respectively). Thus at larger locus sizes POCUS was successful for fewer loci, but when successful it yielded a list of candidate genes that was more highly enriched. Of course, for any locus there were numerous non-disease genes, whereas there was only one disease gene. This was reflected by the ratios of disease genes to non-disease genes above the threshold, which were 0.70, 2.00, and 3.36, respectively. However, these estimates of false positives mask the high specificity that POCUS often achieved in its ranking of genes above the threshold, with 86-95% of correctly identified disease genes ranked first. Indeed, in 60%, 33%, and 12% of the loci (at the three locus sizes) with genes above the threshold, the disease gene was the only gene above the threshold.

A total of 32 disease genes were found to share no IDs with the other genes for the same disease, and were therefore undetectable using POCUS. The loci containing such undetectable genes were included in further analyses, as they mimic the inclusion of erroneously implicated loci - those that are later found to be artifacts of the positional cloning process. The protocol appears to be robust to the presence of such potentially misleading loci, with enrichment dropping from 12-fold, 29-fold and 42-fold to 10-fold, 25-fold and 39-fold (at the three locus sizes) when they are included in the sets analyzed by POCUS (data not shown).

In common with any well studied set of genes, the disease genes examined here were significantly better annotated relative to the average in the human genome. To be precise, our positive control set of genes possesses 10.71 IDs per gene (standard deviation (SD): 4.51) whereas the genome average is 5.39 IDs per gene (SD: 4.51). As POCUS relies on over-representation of IDs between loci, and as highly annotated genes are more likely to contain a given ID than poorly annotated genes, this could introduce bias into the protocol. In this case we might expect the positive control genes scoring above the threshold to be those with the highest number of IDs. In fact, the mean numbers of IDs possessed by positive control genes above the threshold are 11.25 (SD: 5.32), 12.55 (SD: 4.93), and 13 (SD: 4.44) at the three locus sizes, respectively. None of these means is significantly different from that of the positive control set in general (by two-tailed, unpaired *t* tests). Other data suggest that, in general, POCUS identified disease genes correctly because these genes have highly scoring IDs rather than a large number of moderately scoring IDs. This effect is visible in comparisons between the average scores per ID for positive control disease gene sets versus non-disease genes within the simulated loci. At the three locus sizes examined (100, 500, and 1,000 IDs), these scores are 0.0144 compared to 0.0014, 0.0043 compared to 0.0003, and 0.0025 compared to 0.0002, respectively. In each case, the positive control disease genes possess IDs that score

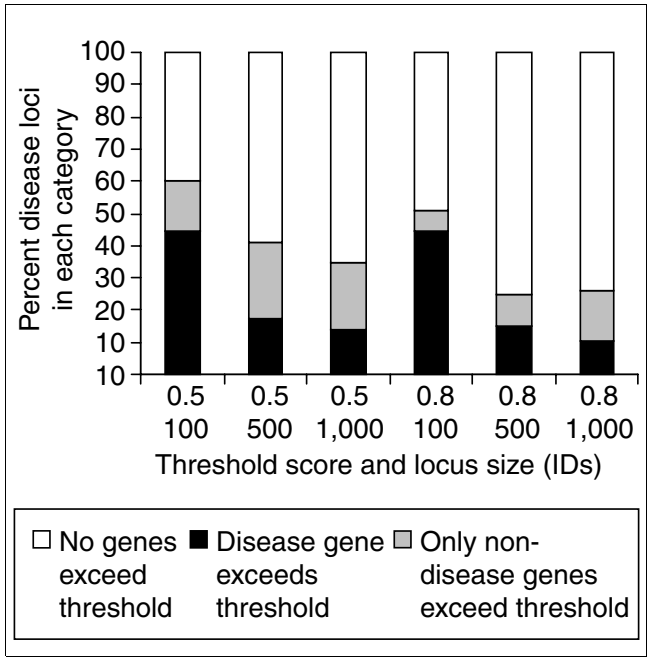

Figure 1
POCUS results per locus for positive control sets of disease genes: the percentage of loci for each of three outcomes is plotted against locus size (100, 500, and 1,000 IDs) at two threshold scores (0.5 and 0.8). The outcome 'No genes exceed threshold' corresponds to the rate of false negatives, 'Only non-disease genes exceed threshold' corresponds to the rate of false positives, and 'Disease gene exceeds threshold' corresponds to the rate of true positives.

around 10 times more than the IDs possessed by non-disease genes. Thus, the success of POCUS is largely attributable to patterns of ID sharing between disease genes that are unusual with respect to other genes arbitrarily selected from the genome. The score assigned to an ID depends on the rarity of the ID in the genome, the number of genes at different loci sharing it, and the size and number of loci examined. In practical terms, when genes do score above threshold they do so as a result of possessing high-scoring IDs (greater than 0.4). The stringency of POCUS is evident in the observation that at the largest locus size the genes scoring above threshold shared IDs seen only 10 or fewer times in the genome. Alternatively, less rare identifiers (occurring 10-50 times in the genome) could also score highly provided they were shared between five or more genes at different loci.

Less subtle bias could also have been introduced by IDs that are associated with a disease itself, but on examination the annotation of disease genes contained only four such IDs (GO:0007601 vision, GO:0007048 oncogenesis, GO:0008181 tumor suppressor, and GO:0008016 regulation of heart). All these IDs were omitted from our analyses to remove such bias. All annotation IDs for all disease genes are available as additional data from [9] and see Additional data files with the online version of this article.

A case study: neuroligins in autism

During the course of these analyses, and after the release of the version of the Ensembl database we used, it was shown that mutations in the neuroligin genes *NLGN3* and *NLGN4* are associated with autism [10]. These disease genes constitute a compelling test of POCUS for two reasons. First, at the time of writing *NLGN3* and *NLGN4* possessed no annotation IDs related to their roles in autism and could both be identified within Ensembl (see Materials and methods). Second, they represent one of the rare occasions when two novel disease genes have been identified simultaneously in the literature, which is important as POCUS requires two or more loci to examine. Following the procedure above, we produced artificial locus sets 100, 500, and 1,000 IDs in size, corresponding to the regions around *NLGN3* and *NLGN4*. At the smallest locus size, POCUS selected the two genes as the best candidates within their loci from a total of 34 genes (18 genes in one locus and 16 in the other). At the 500-ID locus size, POCUS selected each of the two genes as the second-best candidates within their respective locus, from 175 genes (80 in one locus and 95 in the other), and at the 1,000-ID locus size, each of the two genes were ranked seventh within their locus out of 383 genes (172 and 211 genes in the two loci). In spite of these successful rankings, at the two larger locus sizes the two genes failed to score above the 0.8 threshold - which is unsurprising given that this threshold was developed for the sets of three or more loci in the positive control set. This demonstrates that POCUS can be successful where the genes' functional annotation is not biased by the study of disease.

Comparison to other candidate gene prioritization techniques

As mentioned previously, in addition to the protocol presented here, three other techniques for candidate gene prioritization have recently been proposed by [4-6]. Although these different techniques are not incompatible, it is useful to compare the effectiveness of each approach. Freudenberg and Propping [5] emphasize their 'medium stringency' results on a selection of known disease genes used to test the method. These results show that for around a third of the cases (that is, diseases) examined the best results were achieved: the known disease gene was ranked within the top 3% of genes scored (321 of 10,672 genes), representing a 33-fold enrichment (95% CI: 2.08-530.34) of disease genes. More frequently though, in around two thirds of cases examined, the known disease gene was within the top 15% (1,600 of 10,672 genes) which is equivalent to sevenfold enrichment (95% CI: 0.42-106.58). By implication, in the remaining third of cases the known disease genes were not present in the top 15% and the method failed to identify the known disease gene. Thus, the method of Freudenberg and Propping [5], used at a reasonable threshold, resulted two thirds of the time in sevenfold enrichment (though half these cases achieved a maximum enrichment of 33-fold) and one third of the time failed altogether.

Perez-Iratxeta *et al.* [4] tested their method on a series of 30-Mb regions surrounding known disease genes and found that for around a quarter of these regions the known disease gene was within the eight top-ranked genes. If we assume an approximate average gene density of one every 100 kb [8], each region should have contained around 300 genes on average, which indicates 38-fold enrichment (95% CI: 2.57-547.71). For around half of the regions examined, the known disease gene was ranked within the top 30 genes, representing 10-fold enrichment (95% CI: 0.64-155.85), and failed to be in this shortlist for the other half of the regions examined. Thus, the method of Perez-Iratxeta *et al.* [4] provided a 38-fold maximum enrichment one quarter of the time, 10-fold enrichment half the time and failed altogether half the time. Van Driel *et al.* [6] gave a more modest assessment of their method using a test dataset of only 10 regions of variable sizes, each containing a known disease gene. For one of these regions the method reduced a list of 49 candidate genes to two, providing 25-fold enrichment (95% CI: 2.26-265.80). They observed that "on average, a list of 163 genes based on position alone was reduced to a more manageable list of 22 genes", which is equivalent to sevenfold enrichment (95% CI: 0.48-114.27). Because all 10 of the known disease genes tested were found by the method, the failure rate is unknown.

Although these assessments of existing techniques do not allow rigorous comparisons between them, it seems reasonable to conclude that they can perform with similar effectiveness. All three techniques appear to provide around 25-38-fold enrichment at best, but 7-10-fold enrichment more usually. In addition, the methods of Freudenberg and Propping [5] and Perez-Iratxeta *et al.* [4] fail to identify the correct disease gene 33-50% of the time, and it seems reasonable to assume that the method of Van Driel *et al.* [6] is not infallible. We have found that POCUS provided up to 81-fold maximum enrichment (Table 1). More commonly it achieved 12-fold (95% CI: 9.74-15.83), 29-fold (95% CI: 18.79-43.24), and 42-fold (95% CI: 25.36-69.45) enrichment at the three locus sizes. Over the same three locus sizes it failed to return the correct disease gene 58%, 86%, and 89% of the time, respectively. This was usually due to a failure to return any genes above threshold rather than solely non-disease genes. Thus, it would appear that POCUS, using currently available annotation, performs similarly to (and occasionally better than) existing methods when it returns candidate genes above threshold. At larger locus sizes, however, POCUS seems to be notably more conservative than existing methods, usually failing to return any candidate genes above threshold. In reality, POCUS and all of the existing methods could be used in combination as they are likely to be complementary to one another.

Discussion

We have shown that the genes predisposing to a given disease tend to share commonalities in their annotation and the extent of such commonalities is often sufficient to identify these genes using POCUS from regions containing hundreds, or even thousands, of other genes. Depending on the sizes of the loci, we correctly identified 11-45% of disease genes with a modest degree of false positives (0.70-3.36 non-disease genes per real disease gene). This represents a 12-42-fold enrichment for disease genes in the sets of candidates that our protocol returned. The protocol is conservative, with 52-73% of loci yielding no genes scoring highly enough to be regarded as good candidates. As expected, not all disease gene sets possess genes that share sufficient annotation for the protocol to be successful, but this may change as the extent and depth of annotation of human genes increases. POCUS is also robust to variation in the number of loci examined and to the inclusion of loci containing no detectable disease genes. This is important, as the disease literature is expected to contain susceptibility loci that are experimental artifacts. In addition, our case study of the *NLGN3* and *NLGN4* autism genes shows that POCUS can be successful in identifying real disease genes before there is any hint of the disease process in their functional annotation. We conclude that this protocol should prove useful to groups who wish to prioritize genes from susceptibility loci or quantitative trait loci (QTLs) for further study.

It should be stressed that some of these 29 diseases are classed within OMIM as oligogenic (where each contributing gene is not necessary but is sufficient) rather than complex (where no particular gene is necessary and no gene is sufficient). This is important, as greater functional similarity may be expected between genes contributing to an oligogenic disease compared with those contributing to a complex disease. For our purposes, however, it is only important that the predisposing genes for a disease share some degree of functional similarity (as this is the basis of POCUS), regardless of the mode of inheritance. It should also be noted that some unavoidable bias would be expected to exist in collections of known disease susceptibility genes such as those examined here. Where a disease gene has been successfully identified for a given phenotype, later work may have, directly or indirectly, relied on functional similarities to the original gene to discover further genes. This would lead to bias in the disease gene sets examined here, resulting in sets with greater functional similarity to one another than average.

Other groups have recently produced protocols, conceptually related to POCUS, for the identification of enriched functional annotation IDs within a single set of genes, usually derived from large-scale studies of gene expression using microarrays [11-13]. To our knowledge, however, there are no existing methods that perform such analyses between sets of genes. It is clear that, with minor modifications, POCUS could be used to prioritize candidate genes for any mapped trait in any sequenced organism, which could be helpful in the investigation of QTLs in model organisms. We intend to extend POCUS to consider any user-defined annotation, for example

a set of genes upregulated in a microarray gene expression experiment, to identify unexpected enrichment at susceptibility regions known from the literature.

Conclusions

POCUS is a novel protocol that appears to give comparable levels of enrichment for disease genes to existing methods but, in contrast, requires no prior knowledge of the etiology of the disease under study. Indeed, it is possible to identify candidates that are counterintuitive given the literature about the disease. We are developing combined approaches using POCUS as a complement to existing techniques. POCUS does require more than one susceptibility locus to be known, although in the study of complex diseases the bottleneck is usually not in finding susceptibility loci but in identifying the genes underlying them. With the completion of the human genome, and the advancing efforts to rapidly provide functional annotation for the genes [14], POCUS will become an even more potent tool for candidate gene prioritization.

Materials and methods

Functional annotation data

InterPro domain IDs [15] and GO terms [16] for Ensembl genes were obtained from the Ensembl human database (Release 12.31 [17]). InterPro domains are assigned as part of the Ensembl annotation pipeline. The GO data in Ensembl are inherited from the European Bioinformatics Institute GOA project [18] GO term assignments for gene products in the Swiss-Prot and TrEMBL databases [19]. GOA assignments are partly derived from four main sources: manual curation using the literature and automated assignment using either Swiss-Prot keywords, enzyme EC numbers or InterPro domains. In the GOA annotation of Swiss-Prot and TrEMBL (version 6.0) the contributions of these four sources were 2%, 40%, 7%, and 51% of assignments, respectively. The vast majority of GO terms in Ensembl are therefore derived from Swiss-Prot keywords and InterPro domains. As Swiss-Prot provides high-quality, manually curated functional annotation, and InterPro covers around three quarters of the proteins in Swiss-Prot and TrEMBL, the functional annotation in Ensembl represents much of the best-quality and best coverage functional annotation currently available for human genes. Expression data was obtained from the National Center for Biotechnology Information (NCBI) UniGene database [20] in the form of UniGene clusters reported to be 'highly' or 'solely' expressed in a given cDNA library. Such clusters were assigned to the appropriate Ensembl genes according to the NCBI LocusLink database [20].

Disease mapping data

We retrieved data for 29 disorders from the OMIM database [21] that had at least three or more contributing disease-susceptibility genes or modifier genes that were also present within the Ensembl database. Positive control sets of disease genes were derived from these data together with modifier genes from [3]. Specifically, the 131 disease genes that shared one or more ID with another gene for the same disease were regarded as positive control genes, as POCUS can only proceed from the basis of shared IDs.

The *NLGN4* gene was identified in Ensembl as ENSG00000146938 (NCBI RefSeq: NM_020742) but in spite of the *NLGN3* gene being present in the sequence databases for three years (AF217411 [22]) it was not predicted as an Ensembl gene. *NLGN3* was found to be represented by two Ensembl 'EST genes' (gene structures predicted according to matches between ESTs and genomic sequence) instead - ENSESTG00000021460 and ENSESTG00000021462 - and the region spanned by these two EST genes was taken as the genomic location of *NLGN3*. Neither of the Ensembl EST genes possessed any functional annotation, but GO and InterPro IDs were successfully retrieved from the GOA project pages. *NLGN3* was found to possess three InterPro domain matches (IPR000460 Neuroligin; IPR002018 Carboxylesterase, type B; IPR000379 Esterase/lipase/thioesterase, active site) and five GO terms (GO:0007155 cell adhesion; GO:0016020 membrane; GO:0016789 carboxylic ester hydrolase activity; GO:0003824 enzyme activity; GO:0016787 hydrolase activity). All IDs are the result of automated annotation (derived either from matches to InterPro domains or from Philibert *et al.* [22]) that has remained the same since before the publication of *NLGN3* and *NLGN4* as autism genes [10]. The annotation IDs obtained for *NLGN4* from Ensembl (IPR002018 Carboxylesterase, type B; IPR000379 Esterase/lipase/thioesterase, active site; GO:0005177 neuroligin; GO:0007155 cell adhesion; GO:0016020 membrane; GO:0016787 hydrolase; GO:0016789 carboxylic ester hydrolase) were similar to those for *NLGN3*. The two genes shared six IDs in total.

Identification of disease genes

POCUS aims to detect significant enrichment of IDs between loci relative to the genome at large, and then to use this information to score the genes within these loci. The behavior of IDs across the genome was modeled using simulated locus sets. Locus sizes were measured in numbers of IDs to account for variation in gene density and annotation around the genome. We sampled 200,000 nonoverlapping simulated loci sets (up to 15 loci per set) for each of the three locus sizes examined: 100 IDs, 500 IDs, and 1,000 IDs. For example, at locus size 500 IDs we recorded 61,548,000 sharing events among five locus sets. A sharing event is defined as any occasion on which genes from more than one locus possess a particular ID. The number of times each event has occurred is then counted and divided by the total number of events observed in the simulations. The events are then ordered by the resulting frequencies, rarest first. Using this ordered list we calculate a cumulative probability (p) for each possible event, by summing the frequencies of all the events with equal or lower frequencies.

This cumulative probability provides the probability of observing a sharing event equally frequent or less frequent than the frequency of the observed event (f). In effect this is the probability per event; a set of loci will, however, result in a number of sharing events (n). The probability (P_f) of any of these n events having a frequency of f or less is given by the following formula:

$$P_f = 1 - (1 - p)^n$$

Essentially, P_f is the probability that for the given set of loci the observed event would happen by chance. The formula given is mathematically identical to the Bonferroni inequality correction for multiple tests [23]. A score for each ID is then taken to be $1 - P_f$. We then calculate an *ad hoc* score for each gene as the sum of the scores for each ID shared by that gene.

The relative enrichment ratio for disease genes achieved by POCUS was estimated as the proportion of disease genes within the input loci divided by the proportion of disease genes returned above the threshold. Such enrichment measures are essentially comparisons of two proportions and consequently the confidence intervals were estimated using the method of Newcombe [24] implemented as the 'confidence interval calculator' (available as a spreadsheet from [25]). For instance, at the 100 ID locus size, the 131 disease genes sharing IDs were situated within loci containing 3,144 genes in total (Table 1); that is, there was a starting concentration of disease genes equal to 0.042. Processing these gene sets with POCUS produced shortlists of 116 candidate genes containing 60 genuine disease genes (a final concentration of 0.517), representing 12.4-fold enrichment. Using the method mentioned above [24], the 95% CI for the comparison of these two ratios (131/3144 and 60/116), the range of enrichment is 9.737-15.827 fold.

During these calculations we have assumed for convenience that all sharing events are independent; this assumption is often incorrect, however. Because of the hierarchical nature of GO classifications, the presence of lower terms will be dependent on higher terms. For example, all genes with the term for 'insulin receptor complex' (GO:0005899) will also have the terms 'integral to plasma membrane' (GO:0005887), 'intregral to membrane' (GO:0016021), and 'membrane' (GO:0016020). In many of these cases, however, the higher terms are so common across the genome that they will not substantially contribute to the final score of the gene. In addition, our results show that our protocol can score genes above the threshold on the basis of a single shared ID.

Additional data files

The following files are available with the online version of this article: a list of all Ensembl genes with their chromosome numbers (chromosomes 23 and 24 denote X and Y, respectively) and their positions in base pairs, with associated InterPro domains and GO terms (Additional data file 1); a list of the gene expression libraries in which the genes analyzed are found (Additional data file 2); a list of all the diseases analyzed with abbreviations used (Additional data file 3); a list of all the disease genes analyzed (Additional data file 4); Perl scripts calculating the probability of observing each possible pattern of sharing of identifiers for regions of a specified size (Additional data file 5) and calculating the probability of sharing (for a range of numbers of loci) from the simulations results from Additional data file 5 (Additional data file 6); and a fuller description of these Perl scripts (Additional data file 7). Full results of all analyses at thresholds of 0.5 and 0.8 (formatted as a MYSQL database or as plain text) are available on request.

Acknowledgements

This work benefited from the financial support of the UK Medical Research Council. We are grateful to Andrew Carothers for discussion and advice on statistical issues. Martin S. Taylor provided useful comments on the manuscript.

References

1. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405:**847-856.
2. Roses AD: **Pharmacogenetics and the practice of medicine.** *Nature* 2000, **405:**857-865.
3. Badano JL, Katsanis N: **Beyond Mendel: an evolving view of human genetic disease transmission.** *Nat Rev Genet* 2002, **3:**779-789.
4. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining.** *Nat Genet* 2002, **31:**316-319.
5. Freudenberg J, Propping P: **A similarity-based method for genome-wide prediction of disease-relevant human genes.** *Bioinformatics* 2002, **18(Suppl 2):**S110-S115.
6. Van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG: **A new web-based data mining tool for the identification of candidate genes for human genetic disorders.** *Eur J Hum Genet* 2003, **11:**57-63.
7. Jimenez-Sanchez G, Childs B, Valle D: **Human disease genes.** *Nature* 2001, **409:**853-855.
8. International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
9. **MRC HGU Semple Lab** [http://www.hgu.mrc.ac.uk/Users/Colin.Semple/lab_data.html]
10. Jamain S, Quach H, Betancur C, Rastam M, Colineaux C, Gillberg IC, Soderstrom H, Giros B, Leboyer M, Gillberg C, *et al.*: **Mutations of the X-linked genes encoding neuroligins *NLGN3* and *NLGN4* are associated with autism.** *Nat Genet* 2003, **34:**27-29.
11. Castillo-Davis CI, Hartl DL: **GeneMerge-post-genomic analysis, data mining, and hypothesis testing.** *Bioinformatics* 2003, **19:**891-892.
12. Grosu P, Townsend JP, Hartl DL, Cavalieri D: **Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks.** *Genome Res* 2002, **12:**1121-1126.
13. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**R60.
14. King OD, Foulger RE, Dwight SS, White JV, Roth FP: **Predicting gene function from patterns of annotation.** *Genome Res* 2003, **13:**896-904.
15. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al.*: **The InterPro Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31:**315-318.
16. Hill DP, Blake JA, Richardson JE, Ringwald M: **Extension and integration of the gene ontology (GO): combining GO vocabu-**

laries with external vocabularies. *Genome Res* 2002, **12:**1982-1991.
17. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, *et al.*: **Ensembl 2002: accommodating comparative genomics.** *Nucleic Acids Res* 2003, **31:**38-42.
18. **European Bioinformatics Institute GOA project** [http://www.ebi.ac.uk/GOA/index.html]
19. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
20. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, *et al.*: **Database resources of the National Center for Biotechnology.** *Nucleic Acids Res* 2003, **31:**28-33.
21. 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:**52-55.
22. Philibert RA, Winfield SL, Sandhu HK, Martin BM, Ginns EI: **The structure and expression of the human neuroligin-3 gene.** *Gene* 2000, **246:**303-310.
23. Bland JM, Altman DG: **Multiple significance tests: the Bonferroni method.** *Brit Med J* 1995, **310:**170.
24. Newcombe RG: **Improved confidence intervals for the difference between binomial proportions based on paired data.** *Stat Med* 1998, **17:**2635-2650.
25. **The confidence interval calculator** [http://129.78.28.173/pedro/CIcalculator.xls]