Seven existing candidate disease gene prioritisation methods were used to analyse the starting set of candidate genes. Because of the large number of implicated loci, we were unable to use methods Prioritiser (4) and POCUS (12) which are designed to analyse candidate genes lying in a few loci. The seven methods used are:

**Disease Gene Prediction (DGP) (http://www.cmbi.ru.nl/geneseeker/) (6):**

Based on specific sequence property patterns that have been determined for known monogenic hereditary disease genes, DGP assigns probabilities to all genes (from Ensembl version 15.33.1), indicating their likelihood to mutate. Properties analysed include protein length, degree of conservation, phylogenetic extent and paralogy pattern. Previous assessment of the performance of this method used a model in which 75% of the test set was used as learning set and 25% used as a test set. On average, 70% of the gene set was correctly predicted with 67% precision. In this study, genes with a DGP score over 0.5 were selected as potential disease-causing genes.

**PROSPECTR (http://www.genetics.med.ed.ac.uk/prospectr/) (1):**

PROSPECTR is an alternating decision tree which has been trained to differentiate between genes likely to be involved in disease and those that are not, according to sequence-based features such as gene length, protein length and conservation. PROSPECTR assigns a score ranging from 0 to 1, and genes over the threshold of 0.5 are classified as likely to be involved in hereditary disease, whilst those scoring less than 0.5 are unlikely.

Methods that select subsets of genes based only on DNA and protein features will select the same subset from the common starting set, regardless of the disease under consideration.

**eVOC method (11):**

This method selects candidate genes by determining disease relevance of their expression profile, as determined from cDNA libraries, Serial Analysis of Gene Expression (SAGE) and Massively Parallel Signature Sequencing (MPSS) data (5). The eVOC anatomical system ontology, a controlled vocabulary for unifying gene expression data (available at [http://www.sanbi.ac.za/resources/tools-](http://www.sanbi.ac.za/resources/tools-))
for-downloading/), is used to integrate clinical and molecular data through a combination of text- and data-mining. The method first associates eVOC anatomy terms with the disease name, according to their co-occurrence in PubMed abstracts (http://research.i2r.a-star.edu.sg/DRAGON/DE/), then ranks the selected terms. Genes are selected according to their annotation with top-ranking terms. Thus candidate genes are selected according to an expression profile that matches anatomical sites most frequently associated with the disease within the biomedical literature. In a test of 20 known disease-causing genes, the gene was presented in the selected subset of candidates for 19/20 candidates (95%), with an average reduction in candidate set size to 64.2% (±10.7%) of the original set size. Table 1 shows top ranking terms identified for metabolic syndrome and individual phenotypes, and search terms used to identify PubMed abstracts for analysis. Where the same set of anatomy terms are associated with different disease terms, the same subset of genes will be selected according to those anatomy terms.

GeneSeeker (http://www.cmbi.ru.nl/geneseeker/) (13, 14):
GeneSeeker is a web-based tool that filters positional candidate disease genes based on expression and phenotypic data from both human and mouse databases. In a test using 10 syndromes, GeneSeeker reduced the candidate gene lists from an average of 163 position-based candidate genes to an average of 22 candidates based on position and expression or phenotype. The search terms used for metabolic syndrome and the five phenotypes are shown in Table 1.

Suspects (http://www.genetics.med.ed.ac.uk/suspects/) (2):
SUSPECTS mines annotation data from GO, Interpro and expression libraries, and incorporates sequence analysis by the method PROSPECTR. Candidate genes are scored on how significantly similar their annotation is to a set of genes already implicated in the same disorder (the “training set”). For metabolic syndrome, there are no confirmed disease genes, so as a training set we used the set of genes described by Matsunaga et al. (7) derived by a sophisticated knowledge-based analysis involving text-mining of PubMed abstracts for genes associated with metabolic syndrome. The training sets used for the phenotypes are shown in table 1. SUSPECTS returns a ranked list of genes, and we selected the top 20% of these genes to populate the candidate lists.
**G2D (http://www.ogic.ca/projects/g2d_2/) (8-10):**

This system scores all terms in Gene Ontology (GO) according to their relevance to each disease or phenotype starting from MEDLINE queries featuring the name of the disease/phenotype. The terms used for these queries are shown in table 1, column 2. This is done by relating symptoms to GO terms through chemical compounds, combining fuzzy binary relations between them previously inferred from the whole MEDLINE and RefSeq databases. In a test with 100 diseases chosen at random from OMIM (15), using bands of 30 Mb (the average size of linkage regions (9)), G2D detected the disease gene in 87 cases. In 39% of these it was among the best 3 candidates, and in 47% among the best 8% of candidates (10). In this instance, GO term scores were then used to score each gene in the candidate set, assigning the maximum mapped GO score to each candidate. This returns a ranked list, and we selected the top-scoring third of the scored genes to populate the gene lists with most likely candidates.

**Endeavour (http://homes.esat.kuleuven.be/~bioiuser/endeavour/endeavour.php) (3):**

This method ranks candidate genes according to their similarity to a training set of genes. Endeavour accesses MEDLINE abstracts, LocusLink data, Gene Ontology (GO) data, InterPro and Bind protein-protein interaction data, KEGG pathway data, microarray and EST-based expression data, TFBS cis-regulatory modules and sequence similarity by BLAST, and returns a ranked list of candidates. When applied to a test set of 200 candidates, the algorithm gave ten known monogenic disease-causing genes an average rank of $13 \pm 5$ out of 200 test genes, and polygenic genes were ranked at position $40 \pm 10$. We selected the top 25% of genes scored by Endeavour to populate the candidate lists. The training set used for metabolic syndrome was that generated by Matsunaga et al. (7), as described for the method SUSPECTS; and those used for the individual phenotypes are shown in Table 1.

**References:**


