Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/50007

Please be advised that this information was generated on 2017-07-06 and may be subject to change.

Published online June 6, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 10 ***3067–3081***
doi:10.1093/nar/gkl381

SURVEY AND SUMMARY

Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes

Nicki Tiffin*, Euan Adie[1], Frances Turner[2], Han G. Brunner[3], Marc A. van Driel[4], Martin Oti[5], Nuria Lopez-Bigas[6], Christos Ouzounis[7], Carolina Perez-Iratxeta[8], Miguel A. Andrade-Navarro[8,9], Adebowale Adeyemo[10,11], Mary Elizabeth Patti[12], Colin A. M. Semple[2] and Winston Hide

South African National Bioinformatics Institute, University of the Western Cape, Bellville, 7535, South Africa, [1]Medical Genetics Section, Department of Medical Sciences, The University of Edinburgh, Edinburgh, UK, [2]MRC Human Genetics Unit, Crewe Road, Western General Hospital, Edinburgh, EH42XU, UK, [3]Department of Human Genetics, University Medical Centre Nijmegen, PO Box 9101, 6500HB Nijmegen, The Netherlands, [4]Department of Molecular Biology, Nijmegen Center for Molecular Life Sciences, Radboud University, 6500 HB Nijmegen, The Netherlands, [5]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen, PO Box 9010, 6500GL Nijmegen, The Netherlands, [6]Research Unit on Biomedical Informatics (GRIB), Universitat Pompeu Fabra, Passeig Martim de la Barceloneta 37–49, 08003, Barcelona, Spain, [7]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, [8]Ontario Genomics Innovation Centre, Ottawa Health Research Institute, 501 Smyth, Ottawa, ON, Canada K1H 8L6, [9]Department of Cellular and Molecular Medicine, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada, [10]National Human Genome Center, Howard University, Genetic Epidemiology Unit, College of Medicine, 2216 6th Street, NW, Washington, DC 20059, USA, [11]University of Ibadan, College of Medicine, Ibadan, Nigeria and [12]Harvard Medical School, Joslin Diabetes Center, 1 Joslin Place, Boston, MA 02215, USA

Received January 3, 2006; Revised March 25, 2006; Accepted May 2, 2006

ABSTRACT

Genome-wide experimental methods to identify disease genes, such as linkage analysis and association studies, generate increasingly large candidate gene sets for which comprehensive empirical analysis is impractical. Computational methods employ data from a variety of sources to identify the most likely candidate disease genes from these gene sets. Here, we review seven independent computational disease gene prioritization methods, and then apply them in concert to the analysis of 9556 positional candidate genes for type 2 diabetes (T2D) and the related trait obesity. We generate and analyse a list of nine primary candidate genes for T2D genes and five for obesity. Two genes, *LPL* and *BCKDHA*, are common to these two sets. We also present a set of secondary candidates for T2D (94 genes) and for obesity (116 genes) with 58 genes in common to both diseases.

INTRODUCTION

Genome-wide empirical studies generate large sets of potential candidate genes. However, it remains difficult to identify the most likely disease-related genes. In this study, we review existing computational methods for disease gene identification and describe their differences and similarities, and then we apply these methods in a complementary fashion to the identification of candidate disease genes for the complex disease type 2 diabetes (T2D) and the related trait obesity. Our aim is to offer the prospective user an overview of the inputs, outputs and functionality of the methods, to illustrate the use of these methods applied to a single problem, and to present a new set of most likely candidate disease genes for the complex diseases T2D and obesity.

Familial studies indicate that many diseases are caused by ectopic loss or gain of gene function, and research to date has identified numerous genes implicated in simple (Mendelian) diseases. Existing methods have been successful in identifying single high relative risk disease genes, but have typically failed to identify genes underlying complex diseases or traits

*To whom correspondence should be addressed. Tel: +27 21 9592611; Fax: 27 21 9592512; Email: nicki@sanbi.ac.za
Present address:
Christos Ouzounis, Computational Genomics Unit & Institute of Agrobiotechnology, Centre for Research and Technology Hellas, Thessalonica, GR-57001, Greece

© 2006 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commerical use, distribution, and reproduction in any medium, provided the original work is properly cited.

that often present with a wide range of phenotypes and generally involve multiple aetiological mechanisms and contributing genes (1,2). In particular, the contribution of each of several genes to the complex disease state is likely to be small, and only the joint effect of several susceptibility genes (often in concert with predisposing environmental factors) leads to disease, making functional validation of complex disease-causing genes difficult (3).

Success in identifying causative genes for Mendelian diseases has raised hopes that such genes can be used as therapeutic targets for sophisticated new drugs, and that genetic profiling of affected individuals could improve accuracy of diagnostic and prognostic information and more effective drug use in the clinic (4). However, the positional cloning techniques successfully employed for Mendelian diseases have proved less effective for complex diseases, and genome-wide and high throughput techniques are generating ever larger sets of candidate disease genes for further analysis (5). For example, loci implicated in susceptibility to multifactorial disease by linkage analysis can be up to 30 Mb, often containing several hundreds of genes (6). Comprehensive empirical analysis of so many candidate genes is prohibitively time-consuming and expensive.

A pertinent role for bioinformatics research exists in the analysis of biological data for candidate genes, and subsequent selection of a subset of most likely disease gene candidates for empirical validation. Many computational methods have been developed to address this problem, and existing methodologies mine many different data sources containing sequence data, biological information, functional information and expression data for candidate genes. There is an ever-increasing wealth of such genomic data now available in public databases (7,8).

A variety of existing computational approaches to selection of candidate disease genes has been tested, peer-reviewed and presented independently. We illustrate the combined use of multiple methods that in principle achieve the same aim in a fair common ground. Unlike other comparative bioinformatics studies, such as the Critical Assessment of protein Structure Prediction (CASP) comparison of protein modelling algorithms (9), we do not aim to select a 'best' method due to the difference in focus of the methods presented. Additionally, it is currently impossible to prove that a predicted candidate gene is not associated with the disease of interest, making such an assessment unfeasible. Instead, we identify the most commonly selected candidate genes.

Because of the different inputs and algorithms used, genes with less comprehensive annotation may be selected by some methods and not others, so are not missed in the overall study. A prospective user presented with an array of diverse methods and outputs may employ the methods depending on the data available for their candidate gene list. In the case study presented here, we prioritize a subset of candidate genes according to their selection by multiple methods, with the premise that these genes are less likely to be artefacts or false positives due to idiosyncrasies of any single approach.

In this context, we present the independent methods and what we believe to be a useful illustration of application of these methods to the analysis of T2D and obesity. This will help prospective users of these methods to understand method inputs and outputs, and to select suitable methods to investigate their disease of interest.

Case study disease: T2D

Our aim is to harness in concert the individual strengths of the different approaches to provide an accurate, informative and appropriate selection of candidate disease genes for the complex disease T2D and the related trait obesity. Diabetes mellitus is an increasingly common disease with profound impact on health and longevity, and is estimated by the World Health Organization to affect more than 150 million people worldwide. Around 90% of cases are T2D, a disease closely linked with obesity in which the body is unable to respond appropriately to insulin produced by the pancreas. T2D is defined by elevations in plasma glucose levels (hyperglycemia), but encompasses a variety of metabolic abnormalities, including reduced responsiveness to insulin (insulin resistance) in key insulin target tissues such as muscle, adipose tissue, liver, kidney and brain; abnormal accumulation of lipids in non-adipose tissue, and abnormal pancreatic beta-cell function leading to insufficient insulin secretion (10,11). While previously considered a disease of adult life, T2D is increasingly observed in children and adolescents (http://www.who.int/topics/diabetes_mellitus/en/), an alarming trend which may relate to the increasing prevalence of obesity worldwide.

The contribution of heredity to T2D appears to differ considerably between different populations and in different environments, consistent with the profile of a complex disease. While family and twin studies, population studies and genetic admixture data offer good evidence that heredity plays an important role (12), it is clear that both pre- and post-natal environment and lifestyle are critical contributors to diabetes pathogenesis (13).

Several relatively uncommon monogenic forms of diabetes have been characterized. These frequently entail defects in pancreatic β-cells and insulin secretion [e.g. MODY diabetes (14)] or adipose tissue development [e.g. lipodystrophy syndromes (15)]. These single genetic changes can result in highly penetrant monogenic diabetes, suggesting that more subtle genetic changes in these pathways might play a role in determining susceptibility to T2D (16).

Genome-wide linkage-based approaches and association studies have been extensively employed in an attempt to identify T2D susceptibility genes, and some regions of replicated significant linkage have been documented, although consensus on all loci has not been reached (10). To date, the only candidate susceptibility gene identified by positional cloning is *CAPN10*, encoding the protease Calpain 10, and this finding has not been widely replicated [reviewed in Ref. (11)].

Obesity is a significant health burden that affects an increasingly large proportion of the global population, frequently co-occurs with T2D, and is the main risk factor for T2D at both individual and entire population levels due to the link between adiposity, abnormal function of adipocytes and insulin resistance (17). Susceptibility to obesity is strongly influenced by multiple genetic factors in addition to the effects of environmental factors, and causative genes for this complex disease have proved elusive (18–20).

In our study, we additionally apply computational methods to the identification of candidate genes for obesity, with the speculation that co-occurrence of T2D and obesity in such a large proportion of the affected population may indicate a

commonality of predisposing genetic background for the two complex diseases. However, not all obese people develop T2D, and not all T2D patients are obese. This implies that we may find a set of genes that are involved with both diseases, and in addition genes that are associated with only one of the two.

T2D, and the associated disease obesity, offers an appropriate case study for a collaborative bioinformatic analysis of candidate disease genes. Both are high impact diseases that are reaching epidemic proportions and present a severe health burden worldwide. There is a vast wealth of clinical (phenotype) data and positional cloning data available for T2D in the biomedical literature, and the high co-occurrence of obesity with T2D allows a possible additional insight as to genetic background of these related diseases. Finally, extensive research into the genetic causes of T2D has to date yielded few convincing answers, and we hope that a combined effort by the bioinformatics disease gene community may offer some new leads for T2D researchers.

Existing methods

The researchers contributing to this collaboration have developed, tested and published their methodologies previously. Here, we provide summaries of these methods with reference to seminal publications and further information for each method.

GeneSeeker (http://www.cmbi.ru.nl/geneseeker/) (21,22): GeneSeeker is a web tool that filters positional candidate disease genes based on expression and phenotypic data from both human and mouse. It queries several online databases directly through the web, guaranteeing that the most recent data are used at all times and removing the need for local repositories. 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. Though particularly well suited for syndromes in which the disease gene shows altered expression patterns in the affected tissues, it can also be applied to more complex diseases.

Analysis of candidate gene expression using eVOC annotation (23): This method performs candidate disease gene selection using the eVOC (a controlled vocabulary for unifying gene expression data) anatomy ontology. It selects candidate disease genes according to their expression profiles, using the eVOC anatomical system ontology as a bridging vocabulary to integrate clinical and molecular data through a combination of text- and data-mining. The method first makes an association between each eVOC anatomy term and the disease name according to their co-occurrence in PubMed abstracts, and then ranks the identified anatomy terms and selects candidate genes annotated with the top-ranking terms. Candidate disease genes are thus selected according to their expression profiles within tissues associated with the disease of interest. In a test of 20 known disease associated genes, the gene was present in the selected subset of candidate genes for 19/20 cases (95%), with an average reduction in size of the candidate gene set to 64.2% (±10.7%) of the original set size.

Disease Gene Prediction (DGP) (http://cgg.ebi.ac.uk/services/dgp/) (24): The genes that are already known to be involved in monogenic hereditary disease have been shown to follow specific sequence property patterns that would make them more likely to suffer pathogenic mutations. Based on these patterns, DGP is able to assign probabilities to all the genes that indicate their likelihood to mutate solely based on their sequence properties. In particular, the properties analysed by DGP are protein length, degree of conservation, phylogenetic extent and paralogy pattern. The performance of this method has been assessed previously on a test dataset by building a model with a part of the data (learning set: 75%) and testing with the rest (test set: 25%). On average 70% of the disease genes in the test set were predicted correctly with 67% precision (24). Genes involved in complex diseases, similarly to monogenic disease genes, need to have mutations or variations in the gene sequence that impair or modify the function or expression of the protein they encode, leading to a disease phenotype. Thus, we believe that, although DGP has been designed for the prediction of mendelian diseases, it can also be useful for the identification of complex-disease genes as it will identify those genes with higher likelihood of suffering mutations.

PROSPECTR and SUSPECTS (http://www.genetics.med.ed.ac.uk/suspects/) (25,26): It can be shown that genes implicated in disease share certain patterns of sequence based features like larger gene lengths and broader conservation through evolution. PROSPECTR is an alternating decision tree which has been trained to differentiate between genes likely to be involved in disease and genes unlikely to be involved in disease. By using sequence-based features like gene length, protein length and the percent identity of homologs in other species as input a score (ranging from 0 to 1) can be obtained for any gene of interest. Genes with scores over a certain threshold, 0.5, are classified as likely to be involved in some form of human hereditary disease while genes with scores under that threshold are classified as unlikely to be involved in disease. The score itself is a measure of confidence in the classification. PROSPECTR requires only basic sequence information to classify genes as likely or unlikely to be involved in disease.

SUSPECTS builds on this by incorporating annotation data from Gene Ontology (GO), InterPro and expression libraries. Candidate genes are scored using PROSPECTR and also on how significantly similar their annotation is to a set of genes already implicated in the same disorder (the 'training set'). This enables SUSPECTS to rank genes according to the likelihood that they are involved in a particular disorder rather than human hereditary disease in general. SUSPECTS leverages the structure of the GO, requiring GO terms to be closely enough related semantically speaking to be considered significant (27). As a rank-based system, it requires potential candidates to share GO terms with other disease genes to a greater extent than the other genes in the same region of interest.

Performance of both PROSPECTR and SUSPECTS was tested separately with a set of oligogenic and complex disorders including Alzheimer's disease, hypertension, autism and systemic lupus erythematosus. At least two implicated genes for each disease were available. For each implicated gene, a region of interest was created containing the implicated gene itself (the 'target gene') and every gene within

7.5 Mb on either side. On average each region of interest contained 155 genes. Associated training sets were then created for SUSPECTS containing the remaining implicated genes for each disorder.

Using PROSPECTR, on average the target gene was in the top 31.23% of the resulting ranked lists of candidates and in the top 5% of those lists 20 times out of 156 (13%). In comparison, on average the target gene was in the top 12.93% of the ranked list from SUSPECTS, which took both the region of interest and the relevant training set as input in each case. The target gene was in the top 5% of the ranked list 87 times out of 156 (56%) (25,26).

G2D (http://www.ogic.ca/projects/g2d_2/) (6,28): This system scores all terms in GO according to their relevance to each disease starting from MEDLINE queries featuring the name of the disease. 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. Then, to identify candidate genes in a given a chromosomal region, G2D (genes to diseases) performs BLASTX searches (29) of the region against all the (GO annotated) genes in RefSeq. All hits in the region with an *E*-value $<10e^{-10}$ are registered and sorted according to the GO-score of the RefSeq gene they hit (the average of the scores of their GO annotations). Note that hits in the genome might correspond to known or unknown genes, or to a pseudogene. In a test with 100 diseases chosen at random from OMIM (Online Mendelian Inheritance in Man) (30), using bands of 30 Mb [the average size of linkage regions (6)], G2D detected the disease gene in 87 cases. In 39% of these it was among the best three candidates, and in 47% among the best 8 candidates (28).

POCUS (http://www.hgu.mrc.ac.uk/Users/Colin.Semple/) (31): POCUS exploits the tendency for genes predisposing to the same disease to have identifiable similarities, such as shared GO annotation, shared InterPro domains or a similar expression profile. Therefore where genes within different susceptibility regions for the same disease share GO or InterPro annotation and/or are co-expressed, these genes may be considered good candidates. Although genes may be selected as candidates on the basis of sharing only a single GO term or InterPro domain, genes lacking this annotation completely will not be selected. Some polygenic/complex diseases may be caused by different genes that are not functionally related. In such cases this method would not be expected to select the disease genes as candidates, but may still, by chance, find functional similarities between some other genes in the regions (especially where there are many regions or the regions contain many genes). Each observed similarity between genes in different regions is given a score. The score is based on the probability of seeing such specific (or more specific) similarities between genes in different randomly chosen regions of the genome containing many genes. Where such a specific (or more specific) similarity would not be seen by chance in >5% of sets of randomly chosen region analysed, the similar genes are considered to be good candidates. Therefore in cases where disease genes are not functionally related (or where there is no data to suggest the disease genes are functionally related) POCUS will select no candidate genes in 95% of cases. This means that POCUS is far more conservative than the other methods discussed. Where many large regions are analysed almost any similarity between genes in different regions will have a considerable probability of being seen by chance. Therefore this method is not likely to be successful when many large regions are analysed, so analysis should be restricted to the most tightly defined and best-supported regions available.

The performance of POCUS was tested by using it to look for known disease genes. Test susceptibility regions were created containing known disease genes and the surrounding genes (31). Test susceptibility regions were created for 120 diseases for which more than one associated gene appears in the OMIM database. POCUS was then used to analyse the set of test regions corresponding to each disease. The performance was measured by the percentage of known disease genes selected as candidates from the test regions. The enrichment for disease genes in the selected genes compared to the whole susceptibility region was also considered. Enrichment was calculated as Enrichment = (disease genes selected/non-disease genes selected) / (disease genes in region/all genes in region). Where the test regions contained 20 genes in total the percentage of disease genes found was 41.7% and enrichment was 10.5-fold. For 100 genes the equivalent figures were 25.8% and 36.9, respectively, and for 200 genes 14.9% and 46.3. It is important to note that these results were obtained with no prior knowledge of disease pathogenesis. However, POCUS can also take into account prior knowledge of the disease, either in the form of known disease genes or preferred genes that are weighted during the analysis. Preferred genes could be genes expressed in the affected tissue or genes selected by other programs as being likely candidates.

MATERIALS AND METHODS

Collation of the starting set of candidate T2D genes

We use the results of association studies as a primary filter of all known genes, in a first step in the selection of likely T2D and obesity candidate genes. We have used all loci clearly identified by association studies (Table 1) in order to include all possible likely candidate genes. Genetic loci of candidates are detailed in Supplementary Table S1, allowing researchers to focus on their regions of interest. We have selected multiple genetic loci, designated by cytogenetic band, that have been implicated in T2D by linkage and association studies, and are described in the biomedical literature (Table 1). Our aim is to cover the majority of loci thus associated with T2D. We have assembled all genes from these regions into our starting set of candidate genes as defined by Ensembl (www.ensembl.org, v.31) (Supplementary Table S1). Our search is focused on finding T2D candidates, and the overlapping set of candidates for both T2D and obesity. This starting set of 9556 candidate T2D genes has therefore been used as the same candidate gene set for identifying obesity genes, under the assumption that any genes that are causative for both diseases would be represented within the T2D candidate gene group. We also discuss briefly candidates from this set that are selected as candidates for obesity only.

Table 1. Loci associated with T2D through linkage and association studies

Locus	Reference
1q21–25	(74–82)
1p31	(83)
2p11	(84)
2p22–2p13	(75,79,85)
2p25	(83)
2q12	(86)
2q24	(87)
2q33–2q37	(81,88–91)
3p12–3p13	(83,92)
3p24-22	(74,84,93)
3p26	(83)
3q11	(75)
3q27–29	(79,87)
4q27–4q28	(74,92,94)
4q32–34	(84,95)
5q13	(82,93)
5q31–5q32	(74,82)
6p21–6p22	(88)
6q12	(75)
6q15–6q27	(74,78,85,96–99)
7p15	(82)
7p21–7p22	(90)
7q22	(78)
7q36	(89)
8p21–8p22	(81,82,87)
8p11–8p12	(81)
8q11	(95)
8q24	(82)
9p13–p24	(84,88,92)
9q31	(78)
9q33	(83)
10p13	(93)
10q23	(82)
10q26	(79,83,92)
11p12–p14	(90,96)
11q23	(78)
12p11	(74)
12q15–12q21	(93,100)
12q24	(85,94,101)
14q11–14q13	(76,83,90)
14q23–14q24	(98)
14q32	(95)
15q11	(86)
15q13–q21	(90)
16p12–16q11	(84,85)
17p13–17q22	(84,85,94)
18p11	(74,75,81,83,86,89)
18q21–18q23	(83,89)
19p13	(94)
19q13	(86)
20p11–20p13	(79,94,95,97,98)
20q12–13	(79,88,90,94,97,102)
Xq23–27.3	(93,98)

GeneSeeker (21,22)

GeneSeeker was run in batch mode with the following expression/phenotypic profiles: Obesity: '(fat OR adipose OR hypothalamus OR pituitary OR gut) NOT (eye OR bone OR skin OR hair)'; T2D: '(insulin OR glucose OR pancreas OR fat OR adipose OR liver OR kidney OR gut) OR (muscle AND glucose) NOT (eye OR bone OR skin OR hair)'. For obesity, hypothalamus and pituitary were included to allow for hormonal influences. For T2D, the settings were chosen based on prior knowledge and a brief literature review. The term 'brain' was intentionally left out of the 'NOT' section of both queries to avoid spuriously excluding valid genes that may also be expressed in the brain, given the broad expression profile of this organ (e.g. the muscle glucose transporter GLUT4 is also expressed in the brain). Housekeeping genes were not filtered out since glucose metabolism is a fundamental cellular process. The remaining settings were left at their defaults (10 cM maximum Oxford-grid distance, no databases excluded). The resulting candidate genes were validated against their respective loci using Ensembl BioMart, since GeneSeeker uses (Oxford-grid) chromosomal synteny for orthology determination and not per-gene orthology.

Analysis of candidate gene expression using eVOC annotation (23)

Association of each eVOC anatomy term with the disease name is measured according to co-occurrence in PubMed abstracts, to generate a 'frequency of association' value for each term. The method then determines the frequency with which each term has been used to annotate genes in the Ensembl database (www.ensembl.org), to give the 'frequency of annotation' value. The ranking score s is calculated for each term using these values. Four top-scoring eVOC terms are selected from the ranked list and these terms are compared with eVOC terms annotated to candidate disease genes. The system allows one mismatched term between the terms identified by text-mining and the terms used to annotate candidate genes. Thus, genes selected as most likely candidates from the candidate gene list are those annotated with at least three eVOC terms that match the four top-scoring disease-associated eVOC terms.

DGP (24)

A decision tree-based model was built based on sequence properties (i.e. protein length, phylogenetic extent, degree of conservation and paralogy). This model was then applied to all the genes in the disease loci analysed in order to obtain a probability score for these proteins to be involved in hereditary disease. Note that this probability score is indicative of the probability of the genes to suffer mutations that impair the functionality of the protein encoded to cause a disease phenotype. It does not assume any particular phenotype and it does not account for specific phenotype features.

PROSPECTR and SUSPECTS (25,26)

The genes in each locus were scored by SUSPECTS first using a training set made up of genes already implicated in obesity and then, separately, using a training set made up of genes already implicated in T2D (Supplementary Table S2). The top 10th percentile of each results set was then taken to represent a group of genes enriched for good candidates. This proportion, providing a balance of sensitivity and specificity, was chosen on the basis of tests using positive controls as described in Adie *et al*. (26). All genes were also scored by PROSPECTR based on their sequence features. Genes with PROSPECTR scores >0.65 (~8% of the total) were selected as possible candidates.

G2D (6,28)

G2D makes predictions of candidates on chromosomal regions by defining and scoring a number of BLASTX matches of that region against a scored database of genes. For the sake of the comparison presented in this work, the results had to be mapped to genes and not genomic locations, therefore the BLASTX hits that did not overlap with any current ENSEMBL gene prediction were filtered out (these can be obtained using the G2D web server (6,28) for a particular genomic region and disease). The final result is an ordered list of candidates for each chromosomal region and disease with a score that depends on their GO annotation. We have added a second score to the candidates, the *R*-score. This is the relative score of a sequence according to the distribution of GO scores of the RefSeq set used to characterize the region (the sequence ranking according to its GO-score minus one divided by the total number of sequences in the RefSeq set). *R*-score values close to zero indicate a strong possible relation of the sequence to the disease under consideration according to the current knowledge. The *R*-score allows comparing candidates for a given disease across different genomic regions linked to it; that is, one can see for which of the multiple genomic regions analysed G2D obtained better candidates (28).

POCUS (31)

The POCUS method is sensitive to noise. The inclusion of poorly defined susceptibility regions, or regions with a questionable association with the disease can result in failure to select similarities between disease genes, as such similarities are obscured by the background noise (see 'Existing methods, POCUS'). Therefore the analysis was confined to the best supported and most tightly defined regions. These were 3p22–p24, 3q27–q terminal, 10q26, 11p12–p14, 14q11–14q13, 15q13 and 18q22–p23. Genes scoring above a threshold of 0.95 were considered good candidates. This stringent threshold is a direct reflection of the degree of statistical support for the candidate genes returned by POCUS, according to performance on positive controls (known disease genes) unrelated to the present data. At this threshold, spurious, non-disease genes are expected to be nominated as candidates for <5% of diseases analysed. Using more liberal thresholds results in only a small increase in true positives (correctly identified disease genes) but an accompanying large increase in false positives (non-disease genes) returned as candidates by POCUS (31).

Analysis of candidate gene lists generated

For each gene, the methods that select that gene have been collated. The candidate gene list is sorted by the number of methods that have selected each gene. The likelihood of any gene being selected by all seven methods is very small, as the method POCUS was run on only a subset of the starting set of candidate genes, and also the extent of annotation of genes in the initial candidate gene set is highly varied. In our study, the maximum number of methods that selected a candidate gene was six out of a total of seven. The most frequently selected genes are ranked as the top group selected by six of the seven methods (referred to as the 6/7 set), and the second group is a larger set of genes selected by five out of the seven methods (referred to as the 5/7 set). The candidates in the 6/7 set are assessed individually through review of literature and available data. For the larger groups (the 5/7 set), we have investigated the pathways in which these genes are most frequently implicated, and also present an approach to rank the candidates according to the characteristics of the methods that select each gene, as described below.

Pathways represented by candidate genes

Mootha *et al*. (32) studied the regulation of sets of genes in tissues from diabetic patients. The authors collated groups of genes that are known to be involved in eleven different pathways (Table 2, and Supplementary Table S3), in order to determine in which pathways their sets of genes are most active (32). We have used this pathway data to identify which pathways are most commonly represented in our sets of candidate genes for T2D and obesity, and in the overlapping set of candidate genes.

Ranking of the 5/7 sets of T2D and obesity candidate genes

To accommodate the potential bias that is introduced when several methods utilize very similar data sources and are therefore more likely to select the same genes, we have

Table 2. Comparison of gene sets selected by 5/7 methods, with pathways defined by Mootha *et al*. (32)

Pathway with total number of assigned genes	No. of genes	T2D	% of gene set	% of pathway	Obesity	% of gene set	% of pathway	T2D and Obesity	% of gene set	% of pathway
Fatty acid metabolism	24	1	1.1	4.2	4	3.4	16.7	1	1.7	4.2
Gluconeogenesis	32	3	3.2	9.4	2	1.7	6.3	2	3.4	6.3
Glycolysis	31	1	1.1	3.2	2	1.7	6.5	1	1.7	3.2
Glycogen metabolism	24	2	2.1	8.3	0	0.0	0.0	0	0.0	0.0
Insulin signaling	50	1	1.1	2.0	2	1.7	4.0	1	1.7	2.0
Ketogenesis	9	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Pyruvate metabolism	435	11	11.7	2.5	16	13.8	3.7	8	13.8	1.8
Reactive oxygen species homeostasis	7	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Krebbs cycle	16	1	1.1	6.3	1	0.9	6.3	1	1.7	6.3
Oxidative phosphorylation	87	1	1.1	1.1	0	0.0	0.0	0	0.0	0.0
Mitochondria	457	14	14.9	3.1	17	14.7	3.7	9	15.5	2.0
Total Candidate Genes selected		94			116			58		

No. of genes, the total number of genes assigned to each pathway; T2D/Obesity, the number of T2D or obesity 5/7 candidates found in this pathway; T2D and Obesity, the number of genes in common to T2D and obesity found in this pathway; % of gene set, the percentage of T2D and/or obesity candidates that are found in this pathway; % of pathway, the percentage of total genes assigned to this pathway found in the T2D and/or obesity candidate set.

Table 3. (A) Sources of input data for each method and (B) number of genes in the starting candidate set, and number of genes selected by each method

	Methods GeneSeeker	eVOC system	DGP	PROSPECTR	SUSPECTS	G2D	POCUS
(A) Input							
PubMed abstracts	X	X				X	
eVOC annotation		X					
Sequence data			X	X	X	X	
GO annotation					X	X	X
Protein data	X				X	X	X
Expression libraries	X				X		X
Orthologous mouse genes	X						
OMIM	X						
(B) Number of genes selected							
Starting set of candidates	9556	9556	9556	9556	9556	9556	562
T2D	642	2504	1174	791	956	N/A[a]	2
Obesity	281	3046	1174	791	995	N/A[a]	2

[a]G2D ranks all candidate genes, rather than selecting a subset of candidates.

developed a system to rank the candidate genes according to 'which' methods, rather than simply 'how many' methods have successfully selected them. The input data sources for each method are defined (Table 3), and for each gene selected we have counted the total number of independent data sources that are represented across all methods that select that gene. Genes with more independent data sources involved in their selection are thus given a higher ranking. Additionally, we have taken into consideration the number of candidate genes that each method has selected: a method that selects a large percentage of the starting gene set will have lower weighting than a method that is more selective and chooses few candidate genes. This weighting is calculated as the inverse of the fraction of the initial candidate set that is selected and is averaged across the methods used to select each gene. A higher score therefore predicts a more likely candidate:

$$\text{Score} = (\text{number of data sources})^{*}\text{average}\ (\text{total genes}/\text{selected genes})$$

RESULTS

From the starting set of 9556 genes, each of the methods selected a subset of most likely candidate genes for T2D and obesity, ranging in size from 2 to 3046 genes (Table 3). The selected genes were grouped according to the number of methods that selected each gene (Supplementary Table S1). Nine genes are selected as potential T2D genes by 6/7 methods (a summary of these candidate genes is shown in Table 4). We selected these as our primary candidate gene set, with the premise that genes selected by the most independent methods are least likely to be false positives or artefacts of the type of approach used. Four of the nine are located in the mitochondrion (*BCKDHA*, *OAT*, *ACAA2*, *ECHS1*). Some are involved in the metabolism of fatty acids (*ACAA2*, *ECHS1*, *LPL*), lipids (*LPL* and *ACAA2*), amino acids (*BCKDHA* and *OAT*), and glycogen and glucose (*PRKCSH* and *PGM1*). *CSF1R* is a tyrosine kinase transmembrane receptor for the cytokine colony stimulating factor 1 (CSF1), and is involved in macrophage differentiation, function and production; and *TGFBR2* is a Ser/Thr kinase transmembrane receptor for transforming growth factor-beta (TGFβ), with a role in transcriptional regulation.

A total of five genes are selected as most likely candidate genes for obesity by 6/7 methods. Two of these, *LPL* and *BCKDHA*, overlap with the set of most likely T2D candidates, and the additional three selected as candidates for obesity only are *CAT*, *NEU1* and *VLDLR* (Table 4).

Five methods select 94 candidates for T2D and 116 candidates for obesity. Of these, 58 genes are common to both candidate sets (Supplementary Table S1).

DISCUSSION

Candidate genes selected by 6/7 methods

Insulin resistance is one of the principal features of T2D that appears early in the disease state, and is strongly associated with mitochondrial dysfunction. Defects in muscle and liver glucose metabolism have been implicated in T2D and insulin resistance [reviewed in Ref. (33)]. Increasingly, however, aberrant processing of fatty acids and their metabolites within the mitochondria has been implicated in the development of insulin resistance. Defects in fatty acid metabolism may cause accumulation in the cell of fatty acyl CoAs and diacylglycerol, with subsequent inhibition of insulin-stimulated glucose transport activity leading to insulin resistance [reviewed in Ref. (33,34)]. In both obesity and T2D, there is a reduction in glucose oxidation and storage, TCA cycle activity and β-oxidation and electron transport enzyme activity (35).

Extensive evidence shows that mitochondrial function is also required for normal glucose-stimulated insulin secretion from pancreatic β-cells via the production of ATP by oxidative phosphorylation, thus linking mitochondrial pathways to another component of diabetes pathology (34,36). Microarray studies have also shown that expression of nuclear-encoded mitochondrial genes involved in oxidative phosphorylation is dysregulated in tissues obtained from both humans with diabetes (32,37–39), and from mouse models of insulin deficient diabetes (40). Moreover, oxidative phosphorylation activity and mitochondrial bioenergetic capacity are both impaired in diabetic humans (41,42). Even in insulin-resistant but normoglycemic individuals (prediabetes), expression of oxidative genes and upstream regulators is decreased (39)

Table 4. Properties of candidate genes selected by 6/7 methods for both T2D and obesity, for T2D only and for obesity only

Name	HUGO ID	Ensembl_ID	Locus	LOD score	Disease affiliation	Gene function
T2D and obesity						
Lipoprotein lipase precursor	*LPL*	ENSG00000175445	8p21.3	2.55 (73)	Hyperlipoproteinemia type I	Hydrolysis of triglycerides
Branched-chain alpha-keto acid dehydrogenase	*BCKDHA*	ENSG00000142046	19q13.2	1.50 (77)	Maple syrup urine disease	Amino acid metabolism
T2D only						
Ornithine aminotransferase	*OAT*	ENSG00000065154	10q26.13	2.88 (83), 1.10 (74), 1.69 (70)	Ornithine aminotransferase deficiency.	Amino acid metabolism
Macrophage colony stimulating factor I	*CSF1R*	ENSG00000182578	5q32	1.22 (73)	Myeloid malignancy	Receptor for colony stimulating factor 1, a cytokine which controls the production, differentiation, and function of macrophages.
Glucosidase II beta subunit precursor	*PRKCSH*	ENSG00000130175	19p13.2	1.81 (85)	Polycystic liver disease	Beta-subunit of glucosidase II, an N-linked glycan-processing enzyme in the endoplasmic reticulum
TGFβ receptor type II	*TGFBR2*	ENSG00000163513	3p24.1, 3p22	1.27 (65), 2.20 (75)	Colon cancer, Marfan syndrome type II	Receptor for TGFβ, with a protein kinase domain that activates transcription factors
Phospho-glucomutase 1	*PGM1*	ENSG00000079739	1p31.3	1.50 (74)	OMIM 171900	Catalyses the transfer of a phosphate group between the 1- and 6-positions of glucose.
Acetyl-CoA acyltransferase 2	*ACAA2*	ENSG00000167315	18q21.1	2.62 (80), 1.10 (74)	OMIM 604770	Mitochondrial fatty acid metabolism (β-oxidation)
Enoyl-CoA hydratase	*ECHS1*	ENSG00000127884	10q26.3	1.69 (70), 2.88 (83), 1.10 (74)	OMIM 602292	Mitochondrial fatty acid metabolism (β-oxidation)
Obesity only						
Catalase	*CAT*	ENSG00000121691	11p13	3.08 (81), 2.89 (87)	Acatalasia,	Catalysis of hydrogen peroxide to water and oxygen in the peroxisome
Sialidase 1	*NEU1*	ENSG00000184494	6p21.33	2.07 (79)	Neuraminidase deficiency	Cleavage of terminal sialic acid residues from substrates such as glycoproteins and glycolipids in the lysosome
VLDLR	*VLDLR*	ENSG00000147852	9p24.2	2.38 (83)	OMIM 192977	Transport of triacylglycerol from the liver to adipose tissue

and ATP production is reduced by 30% compared with controls (43). T2D does not show exclusively maternal inheritance, and appropriately our studied focused on nuclear-encoded rather than mitochondrial T2D candidates. Taken together, these data suggest that mitochondrial dysfunction, perhaps mediated via changes in nuclear-encoded mitochondrial gene expression and/or function, may be an early phenotype associated with insulin resistance and diabetes risk.

Thus dysregulated fatty acid metabolism and dysregulation of the oxidative phosphorylation pathway have been identified as potential causes of diabetes. Our integrated computational approach, however, has preferentially identified seven candidate T2D genes (*PRKCSH*, *PGM1*, *LPL*, *ECHS1*, *ACAA2, BCKDHA* and *OAT*) that are active in oxidative metabolic pathways (Figure 1). These pathways contribute via fatty acid, carbohydrate and amino acid metabolism to the generation of precursors, such as acyl CoA and acetyl CoA. These precursors feed into the citric acid cycle, which subsequently leads into the oxidative phosphorylation pathway (Figure 1). Dysregulation of the selected genes in the upstream metabolic pathways may thus be contributing to the observed reductions in TCA flux and ATP synthesis (43).

Evidence suggests that circulating cytokines and growth factors secreted by fat tissue modulate responsiveness of liver and muscle to insulin (44). TGFBR2 is a Ser/Thr kinase transmembrane receptor for the growth factor TGFβ. It is intriguing that previous studies have implicated TGFβ in diabetes, diabetic nephropathy and pancreatic dysfunction, for example analysis of *TGFβ* gene polymorphisms in a cohort of Chinese T2D patients has identified the *TGFβ* polymorphism T869C (Leu10Pro) as being associated with diabetic nephropathy in these T2D patients (45). Similarly, CSF1R is a tyrosine kinase transmembrane receptor for the CSF1, and is involved in macrophage differentiation, function and production. While not previously linked to diabetes, CSF1 is expressed in adipose tissue and is regulated by tumour necrosis factor-α (TNFα,) (46), a cytokine that is expressed in macrophages and adipose tissue and may contribute to obesity-associated insulin resistance and diabetes (44,47). Potential links between macrophages and adipocyte function are of particular interest in this regard, as macrophage

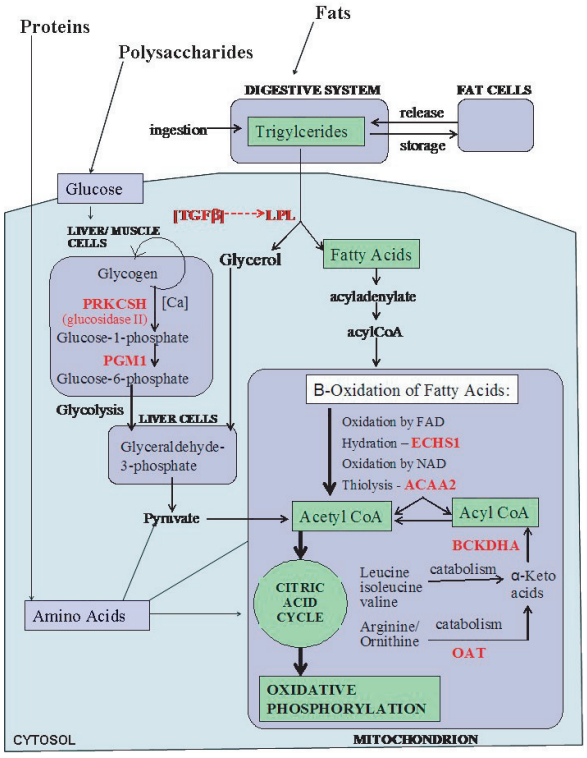

Figure 1. Involvement of candidate T2D genes in cellular metabolism pathways, selected candidates are shown in red.

infiltration of adipose tissues is common in obesity, and inflammation is likely to be a key component of insulin resistance and diabetes risk (44,48).

The link between malnutrition and immunosuppression is well recognized, and conversely it appears that overnutrition and obesity may result in activation of proinflammatory pathways. Multiple lines of evidence suggest that such immune mediators, including TNFα, and by implication CSF1, may be important regulators of insulin resistance, mitochondrial function, ectopic lipid storage and beta cell dysfunction—offering further support for a possible role for CSF1 in T2D and/or obesity (49–51).

Three genes were selected as likely candidates for obesity only. Catalase, the product of the *CAT* gene, acts in the peroxisome to convert hydrogen peroxide to water and oxygen, as critical step in control of reactive oxygen species which may accumulate in excess with oxidative stress and/or mitochondrial dysfunction. *NEU1* encodes sialidase 1, a lysosomal enzyme which cleaves terminal sialic acid residues from substrates, such as glycoproteins and glycolipids. Very low density lipoprotein (VLDL) is a lipoprotein subclass assembled in the liver from cholesterol and apolipoproteins, and carries triglycerides from the liver to adipose tissue. VLDL levels are shown to be significantly elevated in T2D patients (52). It is likely that defects in the VLDL receptor (VLDLR) could disrupt VLDL processing and similarly contribute to T2D and obesity (53,54). Furthermore, *in vivo* studies suggest that abrogated expression of VLDLR can protect mice from obesity (55).

Some of the identified genes are known to interact, suggesting that they belong to common pathways that are implicated in T2D and adding weight to their selection as most likely T2D candidate genes. *TGF*β downregulates *LPL* expression via Sp1 and Sp3 transcription factor binding sites (56), and *LPL* activation has in turn been found to regulate *ACAA2* and *ECHS1* in a rat model (57). Of interest, regulation of *LPL* by *PPAR*γ, a well-documented T2D susceptibility gene, has also been extensively described (58–60). In addition, a strong relationship exists between *LPL* and *VLDLR*, with direct binding of *LPL* to *VLDLR*. LPL lipolytic activity converts VLDL to smaller particles for endocytosis by VLDLR, enhancing VLDL binding to the receptor (53). Coordinated regulation of LPL and VLDLR expression levels has also been demonstrated in response to diet, and may contribute to obesity by promoting shunting of dietary fat from skeletal muscle to adipose tissue (54). Likewise increased LPL activity increases the propensity for obesity and insulin-resistance in a mouse model (61,62).

Our methods select *LPL* and *BCKDHA* as the most likely candidates in common for both obesity and T2D. The interaction of *LPL* with multiple T2D- and obesity-related genes supports the selection of this candidate for both diseases. Activity of LPL is central to metabolism of trigylcerides and fatty acid metabolism, and is associated with *VLDLR* and *PPAR*γ. To a lesser degree, *LPL* is associated with regulation of *ACAA2* and *ECHS1*, also identified as potential candidates in this study. There is also some evidence that *TGF*β interacts with *LPL* (56,63,64). Of particular interest, a study by Wang *et al*. (65) concludes that the lipoprotein lipase *Pvu* II polymorphism is significantly associated with coronary artery disease severity and with type 2 diabetes in coronary artery disease patients, independent of changes in circulating lipid levels, and a transgenic mouse study highlights a role for β-cell-derived LPL in regulation of glucose metabolism and insulin secretion in pancreatic islets (66). *BCKDHA* appears to be less extensively networked (according to current information), but plays an important role in amino acid metabolism and the processing of α-keto acids in the mitochondrion into intermediate molecules for the citric acid cycle. It is possible that disruption of these pathways might lead to altered concentrations of intermediates with subsequent insulin resistance and increase in adipose storage. Microarray analysis has also shown dysregulated expression of both *LPL* and *BCKDHA* in skeletal muscle of Mexican–American patients with T2D (39).

Candidate genes selected by 5/7 methods

Five methods select 94 candidates for T2D and 116 for obesity. Of these, 58 genes are common to both candidate sets (Supplementary Table S1). We analysed these gene sets in three different ways:

First, we compared these gene lists to the pathway datasets compiled by Mootha *et al*. (32) (Supplementary Table S3). For T2D, obesity and the overlapping candidate sets, the pathway sets containing the largest number of the 5/7 candidate genes are gluconeogenesis, pyruvate metabolism and mitochondrial genes. When comparing the percentage of the genes from each pathway that appear in the candidate gene sets, the most extensively represented pathway sets

are those of fatty acid metabolism and gluconeogenesis for T2D, obesity and the overlapping set; glycogen metabolism for T2D, and glycolysis for obesity and the overlapping candidate set (Table 2). As with candidates selected by 6/7 methods, the genes most commonly represented are those from processes related to and preceding oxidative phosphorylation, including processes in the mitochondrion.

Second, we compared the genes selected by 5/7 methods to candidate T2D genes determined by empirical SNP and microarray analysis in previous studies (Supplementary Table S4). In their SNP analysis, Florez *et al*. (11) identified three polymorphic genes in common with our T2D and obesity candidates, including *PCSK2* and *INSR* for both diseases and *GYS1* as an additional T2D candidate. In array analyses, both *GYS1* and *GPI* were dysregulated in skeletal muscle from humans with T2D (38,39). Again, these genes generally function in metabolite-processing pathways that lead into oxidative phosphorylation: *GYS1* (muscle glycogen synthase, ENSG00000104812) has a role in glycogen metabolism and has been extensively implicated in diabetes (OMIM no. 138570); *PCSK2* (neuroendocrine convertase 2 precursor, ENSG00000125851) encodes a proinsulin-processing enzyme that plays a key role in regulating insulin biosynthesis (OMIM no. 162151); *INSR* (ENSG00000171105) encodes the insulin receptor, the first step in a complex insulin signalling network, and has been implicated in diabetes (OMIM no. 147670); and *GPI* (Glucose-6-phosphate isomerase, ENSG00000105220) encodes a phosphoglucose isomerase protein involved in the glycolytic pathway (OMIM no. 172400).

Third, an issue for consideration in the analysis of selected candidate genes is the similarity/dissimilarity in the seven methods applied in this study. These methods employ an array of input data types to inform the selection of candidates, and there is significant overlap in input data type between some methods, and less between others (Table 3). This introduces potential bias to the selection process: selection of a candidate gene by several methods using the same input data types may be less valuable than selection of a candidate by several methods that use disparate data sources. Equally, a candidate gene selected as one of a large subset of the candidates carries less weight than a candidate selected as one of a small subset of candidates. In order to demonstrate how this type of bias may be accounted for in the analysis of our candidate gene sets, we have developed a scoring system that takes these variables into consideration and ranks the candidate set accordingly. This approach could be of use to a researcher wishing to determine empirically the involvement of a subset of genes from the larger 5/7 candidate sets, giving an approximate rating of the merit of the selected candidates. The ranked sets of T2D and obesity 5/7 genes list the candidates from most to least likely, depending on the subset of methods that selected them (Supplementary Table S5)

We have presented an overview of seven independent computational methods for identifying candidate disease genes, and we have applied them to the analysis of candidate genes for the complex disease T2D and the related complex disease obesity. Rather than selecting a 'best' method, we have used the methods to complement each other in selecting most likely candidates, and to offer the prospective user a better understanding of the inputs, outputs and functionality of each available method. A survey of these methods working in concert also offers the bioinformatics community an opportunity to assess the efficacy of current computational approaches to disease gene identification, and inform future directions for research. We show here that applying a concert of computational approaches to finding disease genes can select genes with functions and characteristics appropriate

Table 5. Analysis of previously proposed T2D and obesity disease genes from loci studied

Symbol	Ensembl ID	Locus	G2D	eVOC	SUSPECTS	DGP	GeneSeeker	POCUS	PROSPECTR	Total
T2D										
GYS1[a]	ENSG00000104812	19q13	X	X	X	X	X			5
IRS1[a]	ENSG00000169047	2q33–2q37			X	X				2
CAPN10[a]	ENSG00000142330	2q33–2q37				X				1
LMNA[b]	ENSG00000160789	1q21.2–q21.3				X				1
ENPP1[b]	ENSG00000197594	6q22–q23	X	X						2
HNF4A[b]	ENSG00000101076	20q12–q13.1	X			X				2
PRKAB1 (71)	ENSG00000111725	12q24	X	X			X			3
PRKAB2 (71)	ENSG00000131791	1q21	X	X						2
ATF6 (72)	ENSG00000118217	1q23	X	X						2
Total			6	5	2	5	2	0	0	
Obesity										
ACDC[a]	ENSG00000181092	3q27–29			X	X			X	3
ADRB2[a]	ENSG00000169252	5q31–5q32			X	X	X			3
ADRB3[a]	ENSG00000188778	8p11–8p12	X				X			2
LEPR[a]	ENSG00000116678	1p31	X	X		X				3
NR3C1[a]	ENSG00000113580	5q31–5q32	X	X	X	X	X			5
TNF[b](73)	ENSG00000111956	6p21.3	X	X		X				3
APOA4[b]	ENSG00000110244	11q23				X				1
LDLR[b]	ENSG00000130164	19p13.3	X	X	X	X			X	5
Total			5	4	4	7	3	0	2	

[a]Genes used in PROSPECTR training set (68–70).
[b]Genes from the Genetic Association Database, located in regions studied and with more than one independent non-negative association study (http://geneticassociationdb.nih.gov/).

to T2D and obesity aetiology, based on the assumption that novel disease genes may share some characteristics with identified disease genes (67), and identify pathways that warrant further investigation. We show that this approach is appropriate for the investigation of complex disease aetiology due to selection of multiple most likely disease genes, and can be used to identify pathways and regulatory networks of aetiological significance. In the identification of genes underlying complex disorders, such as T2D and obesity, the challenge is to reduce the large lists of candidate genes from genome-wide studies without discarding the potentially numerous aetiological genes. However, the majority of 'known' disease genes identified to date that are used to populate artificially assembled test datasets are for monogenic disorders, and it is consequently difficult to assess the performance of existing methods in the complex disease scenario.

The methods we present here have each been previously rigorously tested independently using such test datasets, and their performance in these circumstances has been well documented (22–25,28,31). The logical progression is to apply such methods to real disease scenarios and identify most likely candidate genes. These are studies, however, for which there is no 'known' answer, and confirmation of 'correct' candidate selection consequently becomes unfeasible. Also, most methods generate fairly large lists of most likely candidates rather than single genes or small subsets. We have aimed to overcome some of the limitations of these methods applied in isolation by applying all methods to a single problem and then comparing the sets of candidate genes selected by each method. In this way, the genes selected for further analysis are less likely to be artefacts or false positives resulting from the limitations of a single approach.

In this study we have focused on the candidates selected by the most number of methods as being the most likely candidates, as discussed below. This approach could also be used to highlight potential candidates that are selected by a subset of methods and are missed by the other methods, depending on the type of data available. For example, DGP uses sequence characteristics to select the candidate gene *GATM*, which is involved in amino acid metabolism in the mitochondrion and would be a sensible candidate gene for T2D: additional annotation of this gene (e.g. GO and eVOC annotation) is not sufficient for selection by any other methods.

An understanding of the different approaches used allows us to refine the lists of selected candidates further. For example, genes selected by G2D are ranked by a scoring system that prioritizes most likely candidates according to a calculated *R*-score (6,28), and by selecting T2D candidates that are ranked, for example, in the top 1000 genes of the candidate set, we can prioritize five of the nine candidates in order (from most to least likely: *LPL*, *BCKDHA*, *PRKCSH*, *TGFBR2* and *ACAA2*) in the 6/7 candidate gene list. Similarly, we can use the *R*-score to prioritize and order 34/94 candidates from the 5/7 candidate set, and for candidate obesity genes (see Supplementary Table S7).

Our lists of most frequently selected genes (by 6/7 methods) contain a predominance of genes involved in the processing of metabolites destined to feed into glycolytic and mitochondrial oxidative phosphorylation pathways. We have compared the selected genes with empirical data for genes dysregulated in T2D, and find some overlap of pathways and function between the gene sets. Our selected candidate genes appear to be realistic candidates for a role in T2D and obesity aetiology.

We have identified a selection of genes that have previously been proposed as disease genes for T2D (six genes) and obesity (eight genes) and fall within the loci that we have analysed (68–73), (http://geneticassociationdb.nih.gov/). We determine which of these genes are identified as likely candidates by our cohort of methods (Table 5). From this analysis, it can be seen that many of the genes have been identified as candidates by multiple methods (e.g. *GYS1* has been identified by 5/7 methods as a T2D candidate gene, and *NR3C1* and *LDLR* have been identified by 5/7 methods as obesity candidate genes). None of these genes have been proven conclusively to be common T2D or obesity causative genes, and true positive common disease genes for T2D and obesity have yet to be unequivocally identified (68–70). However, *ATF6*, selected by 2/7 methods, may have a nominal association with T2D in Pima Indians and has been recommended for further research (72), and *TNF* was recently found to be upregulated in morbidly obese T2D patients (73). Equally, there is no true negative control set: even for characterized candidate genes for which no dysregulation has been identified in the disease state we are limited by current knowledge regarding how genes may be dysregulated, although recent research shows that the candidates *PRKAB1* and *PRKAB2*, selected by 3/7 and 2/7 methods, respectively, are not associated with T2D (71). Selection of novel genes outside the 'known' set of genes by our methods does not therefore indicate failure of our approach, as candidate genes identified for T2D to date have not been convincingly proven or disproven.

We recognize that prioritizing candidate genes by commonality of selection by methods relies on assumptions that all methods are equally 'good' and that they perform with equivalent specificity and sensitivity. These assumptions may introduce bias to our selected gene sets. Some methods have significant overlap with other methods in their input data, and output in terms of the number of candidate genes selected from a common starting set can vary greatly according to method. For example, a gene that has been extensively studied for a long time will have a large amount of associated literature and has a better chance of being selected by all methods that use PubMed abstracts as a datasource. Genes that have well-characterized protein products are more likely to be selected by methods that access InterPro data on protein domains, or genes that are ubiquitously expressed are more likely to be identified as good candidates by methods that rely on gene expression data. Even among methods that mine overlapping sources of data, there may be subtle differences in the way data are used and combined with other sources which can affect the bias introduced by multiple methods; for example POCUS mines GO terms shared between different genetic loci whereas G2D deduces associations between symptoms and GO terms from MEDLINE through an annotated protein database (RefSeq).

Thus a candidate selected by two methods with very diverse data inputs may carry more weight than a candidate selected by two methods using the same input data; and a gene belonging to a large subset of selected candidates will carry less weight than a member of a much smaller candidate

subset. In our primary set of most likely T2D candidates, selected by 6/7 methods, the impact of this set size and input data bias are minimal because all candidates were selected by the same subset of six methods, so are equally biased with regard to input data sources and size of candidate sets selected. Also, with the small numbers of candidates in these sets it is possible to consider each candidate gene independently and in greater detail. However, in the larger 5/7 sets, this bias has been taken into consideration and we have described an approach for approximate ranking of these candidate gene sets according to the similarities and dissimilarities of their selection methods. We have also described the metabolic pathways that are most commonly represented in the candidate sets, and finally compared the 5/7 candidate genes with those previously identified by microarray analysis as dysregulated in T2D patient samples.

Further potential for bias in computational approaches to disease gene selection arises from the varying extent of gene annotation and available data for genes in the starting set of candidates. This is particularly evident when the starting set of candidate genes has been compiled according to locus: each locus contains genes across the full spectrum from totally uncharacterized to well characterized and annotated. The greater the extent of annotated data for a given candidate gene, the more likely it is to be selected by data-mining techniques. This could be the explanation for the high prevalence of disease-associated genes selected in the 6/7 candidate set (Table 4): previously identified disease genes are likely to be well studied and consequently extensively annotated. In contrast, a very recently identified gene has less associated literature and annotation, and is less likely to be selected by these methods regardless of its suitability. This is balanced to some extent, however, by the methods that use gene sequence data input, including DGP and PROSPECTR, regardless of additional gene annotation. G2D also uses the raw genomic sequence as input, and a gene can be detected as a candidate even if it has not been previously characterized. Moreover, while it is true that it will tend to point to functions/features already associated with the disease in the literature, it can also point to relations never made before in an explicit way. This is again illustrative of the strength of using all methods in a complementary fashion to balance each other's capacities.

The methods select diverse gene sets according to their different input data and analyses, and use existing genetic data to propose novel disease gene candidates. Even for the method SUSPECTS, which uses these proposed candidate genes within a training set, other genes were selected as more appropriate due to novel interpretation of the existing data. Each method has been previously rigorously tested within the limits of current knowledge for well characterized diseases, and there is no way to know whether T2D candidates will follow the same general 'rules' as for other diseases. The nature of the bioinformatics approach is to generate likely candidates by extensive analysis of known characteristics of genes, and is inevitably restricted by existing information be it GO annotation, current expression data or DNA sequence. However we show here that existing information *can* be synthesized to select novel candidates, and this is the aim of our methods. We do not seek to prove that these are true disease genes—this can only be undertaken empirically.

We believe it is important to progress from intellectual analysis of individual bioinformatics disease gene identification approaches using artificially assembled test datasets, to application of these techniques in the most appropriate way to real and pressing disease scenarios; and this has been the aim of our collaboration. By making such a transition with a collaborative and focused effort, using a concert of complementary methods, we hope to generate new, appropriate leads from existing data to assist researchers in prioritizing candidate disease genes for further empirical analysis.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

N.T. is funded by the South African Medical Research Council, the University of Western Cape and South African National Bioinformatics Institute Unit for Capacity Development funding, and the National Bioinformatics Network, South Africa. M.O. is supported by the BioRange programme of The Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through The Netherlands Genomics Initiative (NGI). N.L.-B. is recipient of a Ramon y Cajal contract of the Spanish Ministry of Science and Education (MEC) and acknowledges funding from the International Human Frontier Science Program Organization. C.O. acknowledges support by the BioSapiens project, funded by the EC, contract number LSHG-CT-2003-503265. M.A.A.-N is the recipient of a Canada Research Chair in Bioinformatics. M.A.A.-N. and C.P.-I. acknowledge funding from the Canadian Foundation for Innovation, Ontario Innovation Trust, and Canadian Institutes of Health Research. M.E.P. is supported by NIH grants DK062948 and DK060837. C.A.M.S. is funded by the United Kingdom Medical Research Council. W.H. is supported in part by the Ludwig Institute of Cancer Research. The Open Access publication charges for this article were waived by the Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Glazier,A.M., Nadeau,J.H. and Aitman,T.J. (2002) Finding genes that underlie complex traits. *Science*, **298**, 2345–2349.
2. Tabor,H.K., Risch,N.J. and Myers,R.M. (2002) Opinion: candidate-gene approaches for studying complex genetic traits: practical considerations. *Nature Rev. Genet*., **3**, 391–397.
3. Hoh,J. and Ott,J. (2004) Genetic dissection of diseases: design and methods. *Curr. Opin. Genet. Dev*., **14**, 229–232.
4. Yang,Q., Khoury,M.J., Botto,L., Friedman,J.M. and Flanders,W.D. (2003) Improving the prediction of complex diseases by testing for multiple disease-susceptibility genes. *Am. J. Hum. Genet*., **72**, 636–649.
5. Risch,N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
6. Perez-Iratxeta,C., Bork,P. and Andrade,M.A. (2002) Association of genes to genetically inherited diseases using data mining. *Nature Genet*., **31**, 316–319.
7. Kanehisa,M. and Bork,P. (2003) Bioinformatics in the post-sequence era. *Nature Genet*., **33**, 305–310.
8. Stein,L.D. (2003) Integrating biological databases. *Nature Rev. Genet*., **4**, 337–345.

9. Kinch,L.N., Wrabl,J.O., Krishna,S.S., Majumdar,I., Sadreyev,R.I., Qi,Y., Pei,J., Cheng,H. and Grishin,N.V. (2003) CASP5 assessment of fold recognition target predictions. *Proteins*, **53**, 395–409.
10. van Tilburg,J., van Haeften,T.W., Pearson,P. and Wijmenga,C. (2001) Defining the genetic contribution of type 2 diabetes mellitus. *J. Med. Genet.*, **38**, 569–578.
11. Florez,J.C., Hirschhorn,J. and Altshuler,D. (2003) The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu. Rev. Genomics Hum. Genet.*, **4**, 257–291.
12. Gloyn,A.L. and McCarthy,M.I. (2001) The genetics of type 2 diabetes. *Best. Pract. Res. Clin. Endocrinol. Metab.*, **15**, 293–308.
13. Ong,K.K. and Dunger,D.B. (2004) Birth weight, infant growth and insulin resistance. *Eur. J. Endocrinol.*, **151**, U131–U139.
14. McCarthy,M.I. and Froguel,P. (2002) Genetic approaches to the molecular understanding of type 2 diabetes. *Am. J. Physiol Endocrinol. Metab.*, **283**, E217–E225.
15. Garg,A. (2000) Lipodystrophies. *Am. J. Med.*, **108**, 143–152.
16. O'Rahilly,S., Barroso,I. and Wareham,N.J. (2005) Genetic factors in type 2 diabetes: the end of the beginning? *Science*, **307**, 370–373.
17. Mokdad,A.H., Ford,E.S., Bowman,B.A., Dietz,W.H., Vinicor,F., Bales,V.S. and Marks,J.S. (2003) Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA*, **289**, 76–79.
18. O'Rahilly,S., Farooqi,I.S., Yeo,G.S. and Challis,B.G. (2003) Minireview: human obesity-lessons from monogenic disorders. *Endocrinology*, **144**, 3757–3764.
19. Swarbrick,M.M. and Vaisse,C. (2003) Emerging trends in the search for genetic variants predisposing to human obesity. *Curr. Opin. Clin. Nutr. Metab. Care*, **6**, 369–375.
20. Rankinen,T., Perusse,L., Rauramaa,R., Rivera,M.A., Wolfarth,B. and Bouchard,C. (2004) The human gene map for performance and health-related fitness phenotypes: the 2003 update. *Med. Sci. Sports Exerc.*, **36**, 1451–1469.
21. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A. and Brunner,H.G. (2003) A new web-based data mining tool for the identification of candidate genes for human genetic disorders. *Eur. J. Hum. Genet.*, **11**, 57–63.
22. van Driel,M.A., Cuelenaere,K., Kemmeren,P.P., Leunissen,J.A., Brunner,H.G. and Vriend,G. (2005) GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, **33**, W758–W761.
23. Tiffin,N., Kelso,J.F., Powell,A.R., Pan,H., Bajic,V.B. and Hide,W.A. (2005) Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, **33**, 1544–1552.
24. Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, **32**, 3108–3114.
25. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.
26. Adie,E.A., Adams,R.R., Evans,K.L., Porteous,D.J. and Pickard,B.S. (2006) SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, **22**, 773–774.
27. Lord,P.W., Stevens,R.D., Brass,A. and Goble,C.A. (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.
28. Perez-Iratxeta,C., Wjst,M., Bork,P. and Andrade,M.A. (2005) G2D: a tool for mining genes associated with disease. *BMC Genet.*, **6**, 45.
29. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
30. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al.* (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **33**, D39–D45.
31. Turner,F.S., Clutterbuck,D.R. and Semple,C.A. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, **4**, R75.
32. Mootha,V.K., Lindgren,C.M., Eriksson,K.F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstrale,M., Laurila,E. *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genet.*, **34**, 267–273.
33. Shulman,G.I. (2000) Cellular mechanisms of insulin resistance. *J. Clin. Invest.*, **106**, 171–176.
34. Lowell,B.B. and Shulman,G.I. (2005) Mitochondrial dysfunction and type 2 diabetes. *Science*, **307**, 384–387.
35. Simoneau,J.A., Veerkamp,J.H., Turcotte,L.P. and Kelley,D.E. (1999) Markers of capacity to utilize fatty acids in human skeletal muscle: relation to insulin resistance and obesity and effects of weight loss. *FASEB J.*, **13**, 2051–2060.
36. Maechler,P. and Wollheim,C.B. (2001) Mitochondrial function in normal and diabetic beta-cells. *Nature*, **414**, 807–812.
37. Antonetti,D.A., Reynet,C. and Kahn,C.R. (1995) Increased expression of mitochondrial-encoded genes in skeletal muscle of humans with diabetes mellitus. *J. Clin. Invest.*, **95**, 1383–1388.
38. Sreekumar,R., Halvatsiotis,P., Schimke,J.C. and Nair,K.S. (2002) Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabetes*, **51**, 1913–1920.
39. Patti,M.E., Butte,A.J., Crunkhorn,S., Cusi,K., Berria,R., Kashyap,S., Miyazaki,Y., Kohane,I., Costello,M., Saccone,R. *et al.* (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of *PGC1* and *NRF1*. *Proc. Natl Acad. Sci. USA*, **100**, 8466–8471.
40. Yechoor,V.K., Patti,M.E., Saccone,R. and Kahn,C.R. (2002) Coordinated patterns of gene expression for substrate and energy metabolism in skeletal muscle of diabetic mice. *Proc. Natl Acad. Sci. USA*, **99**, 10587–10592.
41. Simoneau,J.A. and Kelley,D.E. (1997) Altered glycolytic and oxidative capacities of skeletal muscle contribute to insulin resistance in NIDDM. *J. Appl. Physiol.*, **83**, 166–171.
42. Kelley,D.E., He,J., Menshikova,E.V. and Ritov,V.B. (2002) Dysfunction of mitochondria in human skeletal muscle in type 2 diabetes. *Diabetes*, **51**, 2944–2950.
43. Petersen,K.F., Dufour,S., Befroy,D., Garcia,R. and Shulman,G.I. (2004) Impaired mitochondrial activity in the insulin-resistant offspring of patients with type 2 diabetes. *N. Engl. J. Med.*, **350**, 664–671.
44. Lazar,M.A. (2005) How obesity causes diabetes: not a tall tale. *Science*, **307**, 373–375.
45. Wong,T.Y., Poon,P., Chow,K.M., Szeto,C.C., Cheung,M.K. and Li,P.K. (2003) Association of transforming growth factor-beta (TGF-beta) *T869C* (Leu 10Pro) gene polymorphisms with type 2 diabetic nephropathy in Chinese. *Kidney Int.*, **63**, 1831–1835.
46. Umezawa,A., Tachibana,K., Harigaya,K., Kusakari,S., Kato,S., Watanabe,Y. and Takano,T. (1991) Colony-stimulating factor 1 expression is down-regulated during the adipocyte differentiation of H-1/A marrow stromal cells and induced by cachectin/tumor necrosis factor. *Mol. Cell. Biol.*, **11**, 920–927.
47. Sartipy,P. and Loskutoff,D.J. (2003) Expression profiling identifies genes that continue to respond to insulin in adipocytes made insulin-resistant by treatment with tumor necrosis factor-alpha. *J. Biol. Chem.*, **278**, 52298–52306.
48. Cai,D., Yuan,M., Frantz,D.F., Melendez,P.A., Hansen,L., Lee,J. and Shoelson,S.E. (2005) Local and systemic insulin resistance resulting from hepatic activation of IKK-beta and NF-kappaB. *Nature Med.*, **11**, 183–190.
49. Kolb,H. and Mandrup-Poulsen,T. (2005) An immune origin of type 2 diabetes? *Diabetologia*, **48**, 1038–1050.
50. Wellen,K.E. and Hotamisligil,G.S. (2005) Inflammation, stress, and diabetes. *J. Clin. Invest.*, **115**, 1111–1119.
51. Uysal,K.T., Wiesbrock,S.M., Marino,M.W. and Hotamisligil,G.S. (1997) Protection from obesity-induced insulin resistance in mice lacking TNF-alpha function. *Nature*, **389**, 610–614.
52. Hiukka,A., Fruchart-Najib,J., Leinonen,E., Hilden,H., Fruchart,J.C. and Taskinen,M.R. (2005) Alterations of lipids and apolipoprotein CIII in very low density lipoprotein subspecies in type 2 diabetes. *Diabetologia*, **48**, 1207–1215.
53. Takahashi,S., Sakai,J., Fujino,T., Hattori,H., Zenimaru,Y., Suzuki,J., Miyamori,I. and Yamamoto,T.T. (2004) The very low-density lipoprotein (VLDL) receptor: characterization and functions as a peripheral lipoprotein receptor. *J. Atheroscler. Thromb.*, **11**, 200–208.
54. Roberts,C.K., Barnard,R.J., Liang,K.H. and Vaziri,N.D. (2002) Effect of diet on adipose tissue and skeletal muscle VLDL receptor

and LPL: implications for obesity and hyperlipidemia. *Atherosclerosis*, **161**, 133–141.
55. Goudriaan,J.R., Tacken,P.J., Dahlmans,V.E., Gijbels,M.J., van Dijk,K.W., Havekes,L.M. and Jong,M.C. (2001) Protection from obesity in mice lacking the VLDL receptor. *Arterioscler. Thromb. Vasc. Biol.*, **21**, 1488–1493.
56. Irvine,S.A., Foka,P., Rogers,S.A., Mead,J.R. and Ramji,D.P. (2005) A critical role for the Sp1-binding sites in the transforming growth factor-beta-mediated inhibition of lipoprotein lipase gene expression in macrophages. *Nucleic Acids Res.*, **33**, 1423–1434.
57. Doi,M., Kondo,Y. and Tsutsumi,K. (2003) Lipoprotein lipase activator NO-1886 (ibrolipim) accelerates the mRNA expression of fatty acid oxidation-related enzymes in rat liver. *Metabolism*, **52**, 1547–1550.
58. Nagashima,K., Lopez,C., Donovan,D., Ngai,C., Fontanez,N., Bensadoun,A., Fruchart-Najib,J., Holleran,S., Cohn,J.S., Ramakrishnan,R. *et al.* (2005) Effects of the PPARgamma agonist pioglitazone on lipoprotein metabolism in patients with type 2 diabetes mellitus. *J. Clin. Invest.*, **115**, 1323–1332.
59. Laplante,M., Sell,H., MacNaul,K.L., Richard,D., Berger,J.P. and Deshaies,Y. (2003) PPAR-gamma activation mediates adipose depot-specific effects on gene expression and lipoprotein lipase activity: mechanisms for modulation of postprandial lipemia and differential adipose accretion. *Diabetes*, **52**, 291–299.
60. Schoonjans,K., Peinado-Onsurbe,J., Lefebvre,A.M., Heyman,R.A., Briggs,M., Deeb,S., Staels,B. and Auwerx,J. (1996) PPARalpha and PPARgamma activators direct a distinct tissue-specific transcriptional response via a PPRE in the lipoprotein lipase gene. *EMBO J.*, **15**, 5336–5348.
61. Kim,J.K., Fillmore,J.J., Chen,Y., Yu,C., Moore,I.K., Pypaert,M., Lutz,E.P., Kako,Y., Velez-Carrasco,W., Goldberg,I.J. *et al.* (2001) Tissue-specific overexpression of lipoprotein lipase causes tissue-specific insulin resistance. *Proc. Natl Acad. Sci. USA*, **98**, 7522–7527.
62. Duivenvoorden,I., Teusink,B., Rensen,P.C., Romijn,J.A., Havekes,L.M. and Voshol,P.J. (2005) Apolipoprotein C3 deficiency results in diet-induced obesity and aggravated insulin resistance in mice. *Diabetes*, **54**, 664–671.
63. Friedman,G., Ben Yehuda,A., Ben Naim,M., Matsa,D., Stein,O. and Stein,Y. (1995) Effect of transforming growth factor-beta on lipoprotein lipase in rat mesenchymal heart cell cultures. *Biochim. Biophys. Acta*, **1254**, 140–146.
64. Butterwith,S.C. and Gilroy,M. (1991) Effects of transforming growth factor beta 1 and basic fibroblast growth factor on lipoprotein lipase activity in primary cultures of chicken (*Gallus domesticus*) adipocyte precursors. *Comp. Biochem. Physiol. A*, **100**, 473–476.
65. Wang,X.L., McCredie,R.M. and Wilcken,D.E. (1996) Common DNA polymorphisms at the lipoprotein lipase gene. Association with severity of coronary artery disease and diabetes. *Circulation*, **93**, 1339–1345.
66. Pappan,K.L., Pan,Z., Kwon,G., Marshall,C.A., Coleman,T., Goldberg,I.J., McDaniel,M.L. and Semenkovich,C.F. (2005) Pancreatic beta-cell lipoprotein lipase independently regulates islet glucose metabolism and normal insulin secretion. *J. Biol. Chem.*, **280**, 9023–9029.
67. Lopez-Bigas,N., Blencowe,B. and Ouzounis,C.A. Highly consistent patterns for inherited human diseases at the molecular level. *Bioinformatics*, **22**, 269–277.
68. Stumvoll,M., Goldstein,B.J. and van Haeften,T.W. (2005) Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*, **365**, 1333–1346.
69. Bell,C.G., Walley,A.J. and Froguel,P. (2005) The genetics of human obesity. *Nature Rev. Genet.*, **6**, 221–234.
70. Lohmueller,K.E., Pearce,C.L., Pike,M., Lander,E.S. and Hirschhorn,J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.*, **33**, 177–182.
71. Sun,M.W., Lee,J.Y., de Bakker,P.I., Burtt,N.P., Almgren,P., Rastam,L., Tuomi,T., Gaudet,D., Daly,M.J., Hirschhorn,J.N. *et al.* (2006) Haplotype structures and large-scale association testing of the 5′ AMP-activated protein kinase genes *PRKAA2*, *PRKAB1*, and *PRKAB1* with type 2 diabetes. *Diabetes*, **55**, 849–855.
72. Thameem,F., Farook,V.S., Bogardus,C. and Prochazka,M. (2006) Association of amino acid variants in the activating transcription factor 6 gene (*ATF6*) on 1q21-q23 with type 2 diabetes in pima indians. *Diabetes*, **55**, 839–842.
73. Chacon,M.R., Richart,C., Gomez,J.M., Megia,A., Vilarrasa,N., Fernandez-Real,J.M., Garcia-Espana,A., Miranda,M., Masdevall,C., Ricard,W. *et al.* (2006) Expression of TWEAK and its receptor Fn14 in human subcutaneous adipose tissue. Relationship with other inflammatory cytokines in obesity. *Cytokine*, **33**, 129–137.
74. Ng,M.C., So,W.Y., Cox,N.J., Lam,V.K., Cockram,C.S., Critchley,J.A., Bell,G.I. and Chan,J.C. (2004) Genome-wide scan for type 2 diabetes loci in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21–q25. *Diabetes*, **53**, 1609–1613.
75. Ng,M.C., So,W.Y., Lam,V.K., Cockram,C.S., Bell,G.I., Cox,N.J. and Chan,J.C. (2004) Genome-wide scan for metabolic syndrome and related quantitative traits in Hong Kong Chinese and confirmation of a susceptibility locus on chromosome 1q21–q25. *Diabetes*, **53**, 2676–2683.
76. Hsueh,W.C., St Jean,P.L., Mitchell,B.D., Pollin,T.I., Knowler,W.C., Ehm,M.G., Bell,C.J., Sakul,H., Wagner,M.J., Burns,D.K. *et al.* (2003) Genome-wide and fine-mapping linkage studies of type 2 diabetes and glucose traits in the Old Order Amish: evidence for a new diabetes locus on chromosome 14q11 and confirmation of a locus on chromosome 1q21–q24. *Diabetes*, **52**, 550–557.
77. Das,S.K., Hasstedt,S.J., Zhang,Z. and Elbein,S.C. (2004) Linkage and association mapping of a chromosome 1q21–q24 type 2 diabetes susceptibility locus in northern European Caucasians. *Diabetes*, **53**, 492–499.
78. Hanson,R.L., Ehm,M.G., Pettitt,D.J., Prochazka,M., Thompson,D.B., Timberlake,D., Foroud,T., Kobes,S., Baier,L., Burns,D.K. *et al.* (1998) An autosomal genomic scan for loci linked to type II diabetes mellitus and body-mass index in Pima Indians. *Am. J. Hum. Genet.*, **63**, 1130–1138.
79. Vionnet,N., Hani,E., Dupont,S., Gallina,S., Francke,S., Dotte,S., De Matos,F., Durand,E., Lepretre,F., Lecoeur,C. *et al.* (2000) Genomewide search for type 2 diabetes-susceptibility genes in French whites: evidence for a novel susceptibility locus for early-onset diabetes on chromosome 3q27-qter and independent replication of a type 2-diabetes locus on chromosome 1q21–q24. *Am. J. Hum. Genet.*, **67**, 1470–1480.
80. Xiang,K., Wang,Y., Zheng,T., Jia,W., Li,J., Chen,L., Shen,K., Wu,S., Lin,X., Zhang,G. *et al.* (2004) Genome-wide search for type 2 diabetes/impaired glucose homeostasis susceptibility genes in the Chinese: significant linkage to chromosome 6q21–q23 and chromosome 1q21–q24. *Diabetes*, **53**, 228–234.
81. Elbein,S.C., Hoffman,M.D., Teng,K., Leppert,M.F. and Hasstedt,S.J. (1999) A genome-wide search for type 2 diabetes susceptibility genes in Utah Caucasians. *Diabetes*, **48**, 1175–1182.
82. Wiltshire,S., Hattersley,A.T., Hitman,G.A., Walker,M., Levy,J.C., Sampson,M., O'Rahilly,S., Frayling,T.M., Bell,J.I., Lathrop,G.M. *et al.* (2001) A genomewide scan for loci predisposing to type 2 diabetes in a U.K. population (the Diabetes UK Warren 2 Repository): analysis of 573 pedigrees provides independent replication of a susceptibility locus on chromosome 1q. *Am. J. Hum. Genet.*, **69**, 553–569.
83. Aulchenko,Y.S., Vaessen,N., Heutink,P., Pullen,J., Snijders,P.J., Hofman,A., Sandkuijl,L.A., Houwing-Duistermaat,J.J., Edwards,M., Bennett,S. *et al.* (2003) A genome-wide search for genes involved in type 2 diabetes in a recently genetically isolated population from the Netherlands. *Diabetes*, **52**, 3001–3004.
84. Lindgren,C.M., Mahtani,M.M., Widen,E., McCarthy,M.I., Daly,M.J., Kirby,A., Reeve,M.P., Kruglyak,L., Parker,A., Meyer,J. *et al.* (2002) Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: the Botnia study. *Am. J. Hum. Genet.*, **70**, 509–516.
85. Demenais,F., Kanninen,T., Lindgren,C.M., Wiltshire,S., Gaget,S., Dandrieux,C., Almgren,P., Sjogren,M., Hattersley,A., Dina,C. *et al.* (2003) A meta-analysis of four European genome screens (GIFT Consortium) shows evidence for a novel region on chromosome 17p11.2–q22 linked to type 2 diabetes. *Hum. Mol. Genet.*, **12**, 1865–1873.
86. van Tilburg,J.H., Sandkuijl,L.A., Franke,L., Strengman,E., Pearson,P.L., van Haeften,T.W. and Wijmenga,C. (2003)

Genome-wide screen in obese pedigrees with type 2 diabetes mellitus from a defined Dutch population. *Eur. J. Clin. Invest.*, **33**, 1070–1074.
87. Busfield,F., Duffy,D.L., Kesting,J.B., Walker,S.M., Lovelock,P.K., Good,D., Tate,H., Watego,D., Marczak,M., Hayman,N. *et al.* (2002) A genomewide search for type 2 diabetes-susceptibility genes in indigenous Australians. *Am. J. Hum. Genet.*, **70**, 349–357.
88. Luo,T.H., Zhao,Y., Li,G., Yuan,W.T., Zhao,J.J., Chen,J.L., Huang,W. and Luo,M. (2001) A genome-wide search for type II diabetes susceptibility genes in Chinese Hans. *Diabetologia*, **44**, 501–506.
89. Li,W.D., Dong,C., Li,D., Garrigan,C. and Price,R.A. (2004) A quantitative trait locus influencing fasting plasma glucose in chromosome region 18q22-23. *Diabetes*, **53**, 2487–2491.
90. Mori,Y., Otabe,S., Dina,C., Yasuda,K., Populaire,C., Lecoeur,C., Vatin,V., Durand,E., Hara,K., Okada,T. *et al.* (2002) Genome-wide search for type 2 diabetes in Japanese affected sib-pairs confirms susceptibility genes on 3q, 15q, and 20q and identifies two new candidate Loci on 7p and 11p. *Diabetes*, **51**, 1247–1255.
91. Hanis,C.L., Boerwinkle,E., Chakraborty,R., Ellsworth,D.L., Concannon,P., Stirling,B., Morrison,V.A., Wapelhorst,B., Spielman,R.S., Gogolin-Ewens,K.J. *et al.* (1996) A genome-wide search for human non-insulin-dependent (type 2) diabetes genes reveals a major susceptibility locus on chromosome 2. *Nature Genet.*, **13**, 161–166.
92. Duggirala,R., Blangero,J., Almasy,L., Dyer,T.D., Williams,K.L., Leach,R.J., O'Connell,P. and Stern,M.P. (1999) Linkage of type 2 diabetes mellitus and of age at onset to a genetic location on chromosome 10q in Mexican Americans. *Am. J. Hum. Genet.*, **64**, 1127–1140.
93. Ehm,M.G., Karnoub,M.C., Sakul,H., Gottschalk,K., Holt,D.C., Weber,J.L., Vaske,D., Briley,D., Briley,L., Kopf,J. *et al.* (2000) Genomewide search for type 2 diabetes susceptibility genes in four American populations. *Am. J. Hum. Genet.*, **66**, 1871–1881.
94. Rotimi,C.N., Chen,G., Adeyemo,A.A., Furbert-Harris,P., Parish-Gause,D., Zhou,J., Berg,K., Adegoke,O., Amoah,A., Owusu,S. *et al.* (2004) A genome-wide search for type 2 diabetes susceptibility genes in West Africans: the Africa America Diabetes Mellitus (AADM) Study. *Diabetes*, **53**, 838–841.
95. Permutt,M.A., Wasson,J.C., Suarez,B.K., Lin,J., Thomas,J., Meyer,J., Lewitzky,S., Rennich,J.S., Parker,A., DuPrat,L. *et al.* (2001) A genome scan for type 2 diabetes susceptibility loci in a genetically isolated population. *Diabetes*, **50**, 681–685.
96. Nawata,H., Shirasawa,S., Nakashima,N., Araki,E., Hashiguchi,J., Miyake,S., Yamauchi,T., Hamaguchi,K., Yoshimatsu,H., Takeda,H. *et al.* (2004) Genome-wide linkage analysis of type 2 diabetes mellitus reconfirms the susceptibility locus on 11p13–p12 in Japanese. *J. Hum. Genet.*, **49**, 629–634.
97. Ghosh,S., Watanabe,R.M., Valle,T.T., Hauser,E.R., Magnuson,V.L., Langefeld,C.D., Ally,D.S., Mohlke,K.L., Silander,K., Kohtamaki,K. *et al.* (2000) The Finland–United States investigation of non-insulin-dependent diabetes mellitus genetics (FUSION) study. I. An autosomal genome scan for genes that predispose to type 2 diabetes. *Am. J. Hum. Genet.*, **67**, 1174–1185.
98. Silander,K., Scott,L.J., Valle,T.T., Mohlke,K.L., Stringham,H.M., Wiles,K.R., Duren,W.L., Doheny,K.F., Pugh,E.W., Chines,P. *et al.* (2004) A large set of Finnish affected sibling pair families with type 2 diabetes suggests susceptibility loci on chromosomes 6, 11, and 14. *Diabetes*, **53**, 821–829.
99. Sale,M.M., Freedman,B.I., Langefeld,C.D., Williams,A.H., Hicks,P.J., Colicigno,C.J., Beck,S.R., Brown,W.M., Rich,S.S. and Bowden,D.W. (2004) A genome-wide scan for type 2 diabetes in african-american families reveals evidence for a locus on chromosome 6q. *Diabetes*, **53**, 830–837.
100. Bektas,A., Suprenant,M.E., Wogan,L.T., Plengvidhya,N., Rich,S.S., Warram,J.H., Krolewski,A.S. and Doria,A. (1999) Evidence of a novel type 2 diabetes locus 50 cM centromeric to NIDDM2 on chromosome 12q. *Diabetes*, **48**, 2246–2251.
101. Mahtani,M.M., Widen,E., Lehto,M., Thomas,J., McCarthy,M., Brayer,J., Bryant,B., Chan,G., Daly,M., Forsblom,C. *et al.* (1996) Mapping of a gene for type 2 diabetes associated with an insulin secretion defect by a genome scan in Finnish families. *Nature Genet.*, **14**, 90–94.
102. Klupa,T., Malecki,M.T., Pezzolesi,M., Ji,L., Curtis,S., Langefeld,C.D., Rich,S.S., Warram,J.H. and Krolewski,A.S. (2000) Further evidence for a susceptibility locus for type 2 diabetes on chromosome 20q13.1–q13.2. *Diabetes*, **49**, 2212–2216.