Title

FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease

Journal

Genome Biology, 9(12)

ISSN

1474-760X

Authors

Chen, Rong
Morgan, Alex A
Dudley, Joel
et al.

Publication Date

2008

DOI

10.1186/gb-2008-9-12-r170

Peer reviewed

Research

FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease

Rong Chen[*†‡], Alex A Morgan[*†‡], Joel Dudley[*†‡], Tarangini Deshpande[§], Li Li[†], Keiichi Kodama[*†‡], Annie P Chiang[*†‡] and Atul J Butte[*†‡]

Addresses: [*]Stanford Center for Biomedical Informatics Research, 251 Cmpus Drive, Stanford, CA 94305, USA. [†]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA 94305, USA. [‡]Lucile Packard Children's Hospital, 725 Welch Road, Palo Alto, CA 94304, USA. [§]NuMedii Inc., Menlo Park, CA 94025, USA.

Correspondence: Atul J Butte. Email: abutte@stanford.edu

Published: 5 December 2008

Genome ***Biology*** 2008, **9:**R170 (doi:10.1186/gb-2008-9-12-r170)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2008/9/12/R170

Received: 17 June 2008
Revised: 26 September 2008
Accepted: 5 December 2008

Abstract

Background: Candidate single nucleotide polymorphisms (SNPs) from genome-wide association studies (GWASs) were often selected for validation based on their functional annotation, which was inadequate and biased. We propose to use the more than 200,000 microarray studies in the Gene Expression Omnibus to systematically prioritize candidate SNPs from GWASs.

Results: We analyzed all human microarray studies from the Gene Expression Omnibus, and calculated the observed frequency of differential expression, which we called differential expression ratio, for every human gene. Analysis conducted in a comprehensive list of curated disease genes revealed a positive association between differential expression ratio values and the likelihood of harboring disease-associated variants. By considering highly differentially expressed genes, we were able to rediscover disease genes with 79% specificity and 37% sensitivity. We successfully distinguished true disease genes from false positives in multiple GWASs for multiple diseases. We then derived a list of functionally interpolating SNPs (fitSNPs) to analyze the top seven loci of Wellcome Trust Case Control Consortium type 1 diabetes mellitus GWASs, rediscovered all type 1 diabetes mellitus genes, and predicted a novel gene (*KIAA1109*) for an unexplained locus 4q27. We suggest that fitSNPs would work equally well for both Mendelian and complex diseases (being more effective for cancer) and proposed candidate genes to sequence for their association with 597 syndromes with unknown molecular basis.

Conclusions: Our study demonstrates that highly differentially expressed genes are more likely to harbor disease-associated DNA variants. FitSNPs can serve as an effective tool to systematically prioritize candidate SNPs from GWASs.

Background

A major goal of biomedical research is to identify genes that contribute to the molecular pathology of specific diseases. This process has been accelerated by two types of high-throughput studies: genome-wide association studies (GWASs) and gene expression microarray studies. A GWAS

scans a genome for single nucleotide polymorphisms (SNPs) associated with disease, whereas microarrays identify genes that are differentially expressed between disease and control samples. These methods have been integrated into molecular profiling to identify expression quantitative trait loci and to build pathways that are involved in various diseases, including type 2 diabetes [1,2], atherosclerosis [3], dystrophic cardiac calcification [4], metabolic disorders [5], and cardiovascular disorders [6]. To lower the cost, GWASs are frequently designed as a two-stage study [7]; first is a stage involving identification of candidate SNPs, and then a validation stage is conducted, in which the effect of the candidate SNPs in a larger population is determined. However, in a recent two-stage GWAS of prostate cancer, most of the SNPs determined to be significant were not even ranked in the top 1,000 SNPs in the identification stage [7], which suggests that existing candidate SNP prioritization methods, which are largely based on known functional annotations, are inadequate.

There are many candidate gene and SNP prioritization methods, including the use of sequence information [8,9], protein-protein interaction networks [10,11], literature and ontology [12,13], and various combination of these methods [14]. For a detailed description of the available tools, the reader is referred to comprehensive reviews [15,16]. Gene expression is often taken into consideration when prioritizing candidate genes or SNPs, but this is most often within the context of the specific disease, such as disease-related anatomical regions and tissue specificity [17-20], conserved co-expression [21], coherent expression profile with known disease-associated genes [22], or several expression datasets in model organisms [23]. These disease-specific gene expression prioritization methods are somewhat informative, but they are cumbersome, requiring extensive manual work. Given that there are more than 200,000 microarray studies included in the National Center for Biotechnology Information's Gene Expression Omnibus (GEO) [24] and more than 10,000 disease-associated DNA variants in the Genetic Association Database (GAD) [25] and Human Gene Mutation Database (HGMD) [26], we hypothesize that a more general (and therefore more systematic) link exists between a gene's expression and the likelihood that it is associated with disease.

Recognizing the wealth of gene expression data in public repositories, we propose an integrative genomics method to systematically prioritize DNA markers that aims to accelerate the identification of novel causative genes and variants. Here, we analyzed every available human microarray study in GEO; we calculated the frequency of differential expression for every gene; and we found that the more often a gene was differentially expressed, the more likely it was that it contained disease-associated variants. Based on this discovery, we derived a list of functionally interpolating SNPs (fitSNPs) from differential gene expression, and we showed how fitSNPs could have been used to successfully prioritize genes from type 1 and type 2 diabetes mellitus GWASs, as well as previously identified Online Mendelian Inheritance in Man (OMIM) loci with unknown molecular basis.

Results

Highly differentially expressed genes are more likely to harbor disease-associated variants

In order to determine whether differentially expressed genes are genetically associated with disease, we downloaded all 476 curated human GEO datasets to serve as our human gene expression set. The probes from these GEO datasets, which include groups of microarrays organized by experimental variable (for example, time, tissue, agent, temperature, and so on), were annotated with the latest National Center for Biotechnology Information Entrez Gene annotations using AILUN [27]. We conducted 4,877 group-versus-group comparisons using significance analysis of microarrays (SAM) [28] and obtained a list of 19,879 genes that were differentially expressed with q value under 0.05 in one or more experiments. We then created a list of curated human disease-associated genes by combining GAD [25] and HGMD [26], resulting in a list of 3,221 genes with disease-associated variants.

We compared our list of differentially expressed genes with the list of genes with disease-associated variants, and we found that 99% of disease-associated genes were differentially expressed in one or more GEO datasets, with 14% specificity (Additional data file 1). The likelihood of having variants associated with disease was 12 times higher among differentially expressed genes than among constantly expressed genes ($P < 0.0001$, Fisher's exact test), whereas the likelihood of having a nonsynonymous coding SNP was 1.6 times higher among differentially expressed genes than among constantly expressed genes.

In order to characterize better the relationship between DNA variance and expression in all human genes, we tested whether genes differentially expressed in multiple microarray studies are more likely to have disease-associated variants. For each gene, a differential expression ratio (DER) was calculated as the count of GEO datasets in which it was differentially expressed (q value ≤ 0.05) divided by the count of GEO datasets in which it was measured. The calculation was restricted to genes that were measured in at least 5% of all GEO datasets.

The precision of rediscovering a disease gene was 16% for genes with a DER greater than 0. This precision improved gradually to 28% when the DER was greater than 0.62, and then increased dramatically to 100% when the DER was greater than 0.72 (Figure 1). As a control, a similar graph is also plotted in Figure 1 for constantly expressed genes with a DER less than the cutoffs used. The more GEO datasets in which a gene was constantly expressed, the less likely it was

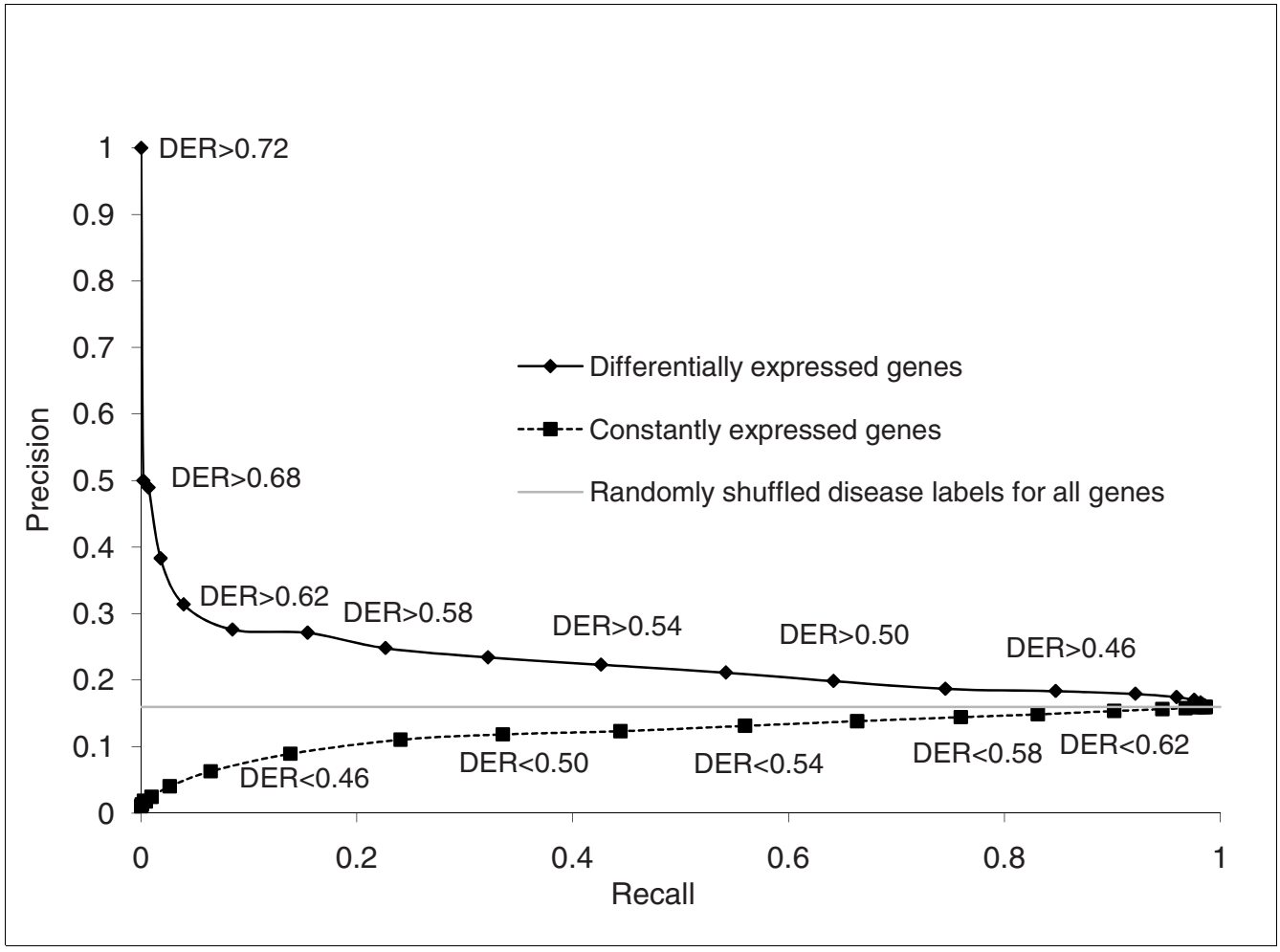

Figure 1
Use of differentially and constantly expressed genes to rediscover disease genes. The DER was calculated as the count of GEO datasets in which a gene was differentially expressed divided by the count of GEO datasets in which it was measured. For any cutoff *x*, differentially expressed genes were defined as genes with DER > *x*, whereas constantly expressed genes were defined as genes with DER <*x*. The precision/recall graphs show that the likelihood of harboring disease mutations for a gene increases when its DER value increases. For the control, we shuffled disease labels 10,000 times among all genes and obtained a predicted precision of 16%. DER, differential expression ratio; GEO, Gene Expression Omnibus.

to contain disease-associated variants. As an additional control, we randomly shuffled disease labels for all genes 10,000 times, and the precision of rediscovering disease genes remained at the predicted 16%. Compared with constantly expressed or randomly shuffled disease genes, the more often a gene was differentially expressed, the more likely it was that it contained DNA variants associated with diseases.

In a receiver operating characteristic curve constructed to rediscover disease genes using the DER values, a DER value ≥ 0.55 exhibited the best performance, with 79% specificity and 37% sensitivity. As shown in Figure 2, genes with DER ≥ 0.55 were 2.25 times more likely to harbor disease-associated variants than others ($P < 0.0001$, Fisher's exact test). Varying the threshold, we achieved 56% specificity and 65% sensitivity at DER ≥ 0.50, and 93% specificity and 16% sensitivity at DER ≥ 0.60.

DER distinguishes true type 1 diabetes mellitus genes from false positive genes in GWASs

The likelihood of harboring disease-associated variants in genes with high DER values could be used to prioritize candidate SNPs from GWASs. To lower the cost, GWASs are often designed as a two-stage experiment: identifying candidate SNPs and then validating them in a larger population. Most often, functionally important genes are manually selected from the loci around positive SNPs for sequencing or high-quality genotyping in a larger population. This prior knowledge based gene prioritization method is not only time consuming but is also likely to miss novel genes. Indeed, associations for a large number of candidate genes from identification stage of GWASs were found to be false positives in the validation stage. A test to distinguish true disease genes from these false-positive genes will demonstrate the prioritizing power of DER in GWASs.

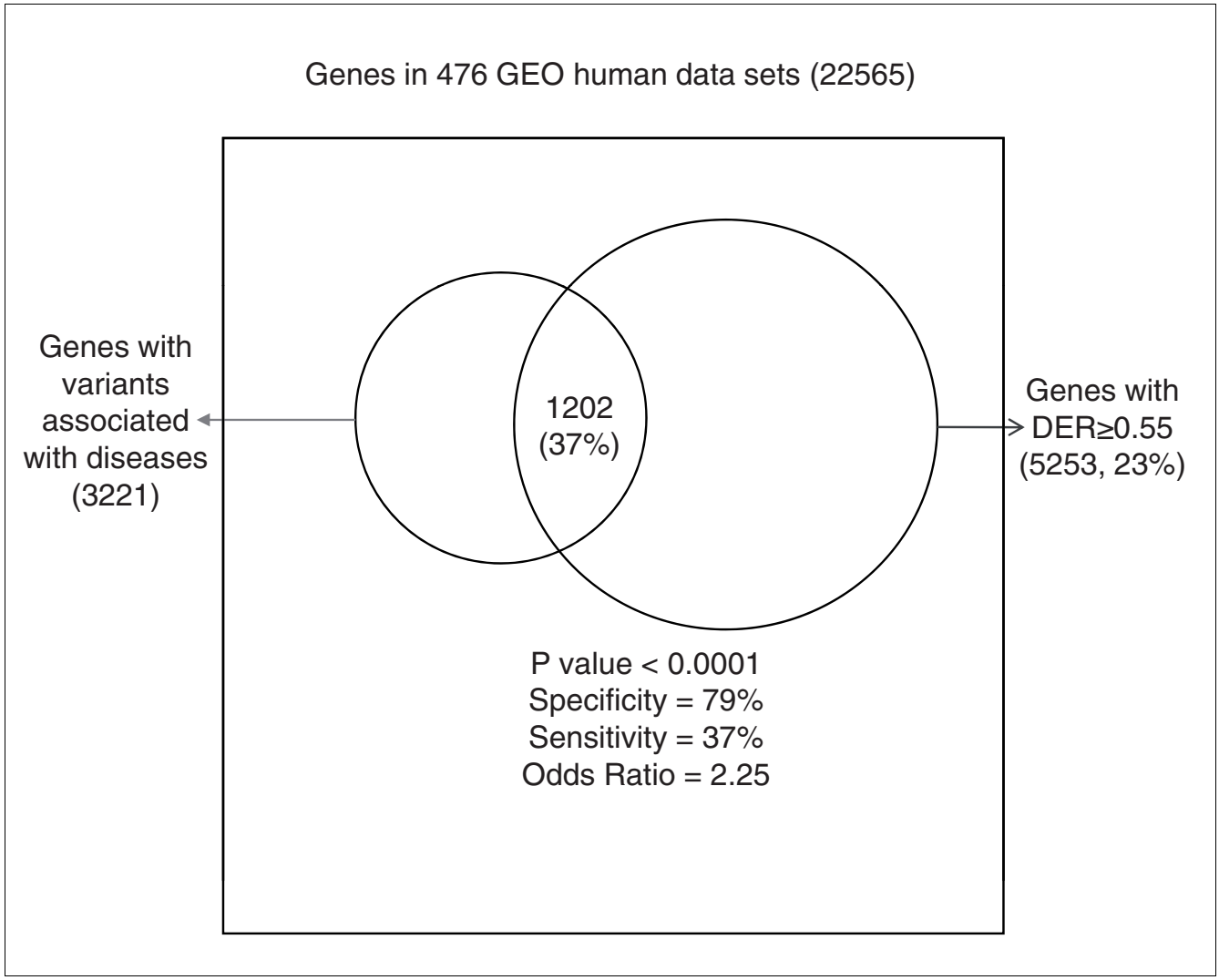

Figure 2
Performance of rediscovering disease genes by DER. Genes with DER ≥ 0.55 were predicted to be disease genes, and compared with genes with disease-associated DNA variants listed in GAD and HGMD. *P* values were calculated using Fisher's exact test. DER, differential expression ratio; GAD, Genetic Association Database; GEO, Gene Expression Omnibus; HGMD, Human Gene Mutation Database.

We first evaluated the performance in type 1 diabetes mellitus (T1DM). Within the top seven T1DM loci (6p21, 12q24, 12q13, 16p13, 18p11, 12p13, and 4q27) identified from the Wellcome Trust Case Control Consortium (WTCCC) GWAS [29], 21 genes were reported with genotyping results in two follow-up studies [30,31]. Table 1 lists their DER values along with their validation results. As shown in Figure 3, the DER values of T1DM genes were significantly higher than those for false-positive genes ($P = 0.003$, *t*-test), with clear separation of the 25th to 75th percentile ranges. Among the ten genotyped candidate genes with DER ≥ 0.55, all but *ITPR3* were validated as true T1DM genes. Of the 11 genotyped genes with DER < 0.55, all but three (*HLA-DPB1*, *C12orf30*, and *KIAA0350*) were found to be unassociated with T1DM. We successfully distinguished true T1DM genes from false positives with 89% specificity and 75% sensitivity ($P = 0.02$, Fisher's exact test). If we only genotype genes with DER ≥ 0.50, then we identify all true T1DM genes, with a 56% false discovery rate.

DER distinguishes true type 2 diabetes mellitus genes from false-positive genes in GWASs

To validate the robustness of this method, we applied it to another disease, namely type 2 diabetes mellitus (T2DM), which had been studied in six large-scale GWASs [29,32-36] and tens of targeted association studies in more than 20 populations. We extracted all significant T2DM genes described in the abstracts, and limited the list to those with significant association in at least three different populations, and derived 15 widely accepted T2DM genes (Table 2). We also retrieved SNPs that were reported to exhibit significant association in the identification stage but no association in the validation stage in a large-scale T2DM GWAS [32]. We annotated these

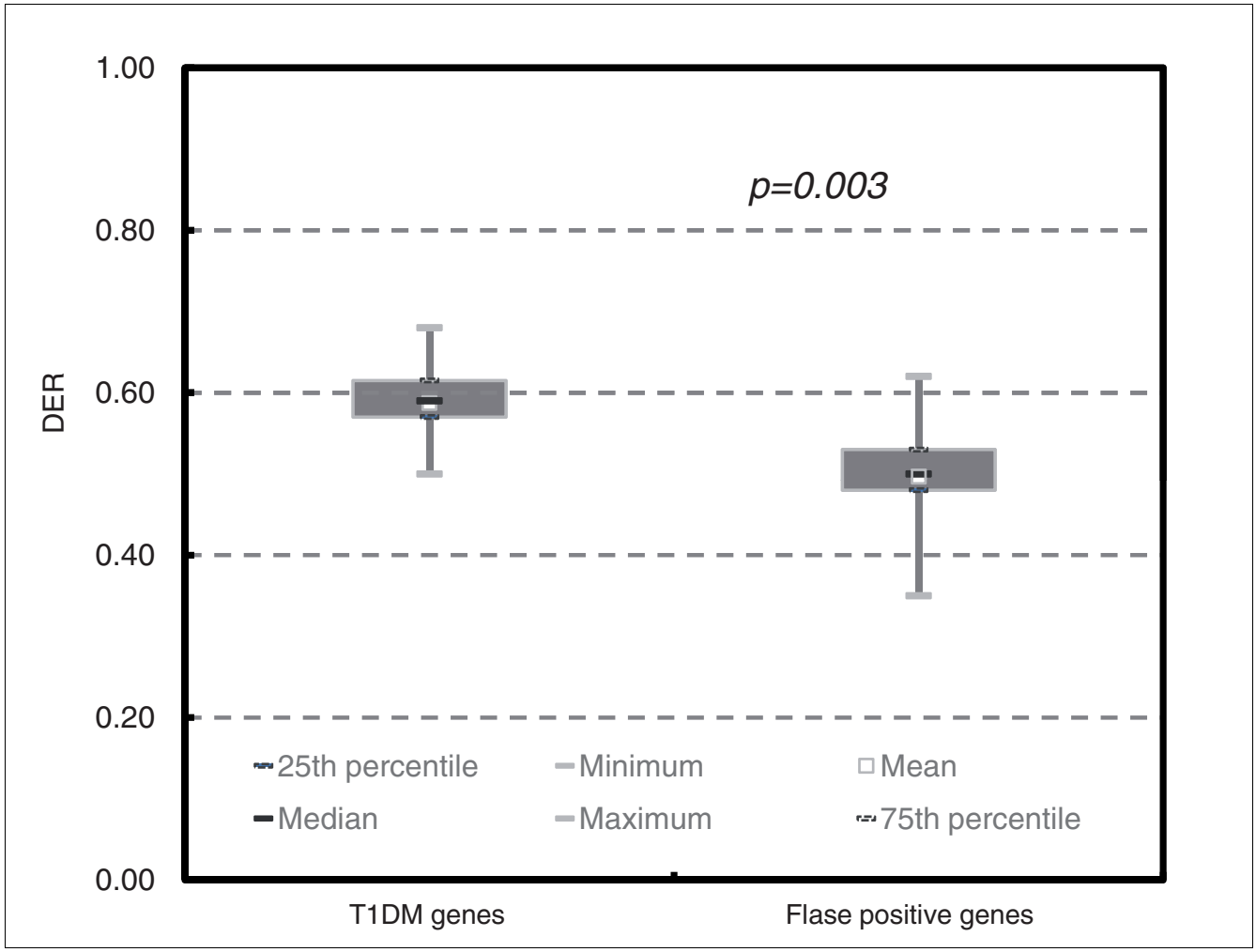

Figure 3
Distinguishing T1DM genes from false positives in the top seven loci from GWASs using DER. Genes in the top seven loci from the WTCCC T1DM GWASs are reported with validation results. False-positive genes were shown as positive in the initial scan but found to be unassociated with T1DM in the follow-up validation studies. T1DM genes had significantly higher DER values than did false positive genes (P = 0.003). The mean DER values for T1DM and false-positive genes were 0.59 and 0.50, respectively. DER, differential expression ratio; GWAS, genome-wide association study; T1DM, type 1 diabetes mellitus; WTCCC, Wellcome Trust Case Control Consortium.

negative SNPs with their associated genes using Entrez dbSNP, and we removed those without gene annotations, and derived 13 negative genes. As shown in Table 2, DER ≥ 0.55 successfully distinguished T2DM genes from negative genes with 85% specificity and 60% sensitivity (P = 0.02, Fisher's exact test).

FitSNPs predicts T1DM genes directly from the top seven WTCCC T1DM loci

The robustness of DER to distinguish disease genes from false positives in T1DM and T2DM GWASs led us to hypothesize that it may also be used to predict disease genes directly from the loci identified from GWASs. To facilitate the visualization of DER values along with GWAS results on the human genome, we created a tool called functionally interpolating SNPs (fitSNPs) [37]. It is a list of human SNPs with DER values assigned according to their associated genes. It can be easily loaded into the University of California Santa Cruz (UCSC) genome graph [38] and visualized on the human genome along with a wealth of preloaded or user-defined genomic data, such as GWAS results. We called the tool 'functionally interpolating SNPs' because it not only infers the likelihood of disease association for all human SNPs but also suggests potential diseases to guide functional studies. In the Gene page of the FitSNPs server, clicking the DER value for any gene will display all biologic and clinical conditions in which it was found to be differentially expressed, with statistical comparisons and filter/sort functions [39].

We therefore examined each of the top seven WTCCC T1DM loci on the UCSC genome browser to evaluate whether we could predict T1DM genes using fitSNPs. The hypothesis is

Table 1

DER values for T1DM and false positive genes in the top 7 WTCCC T1DM loci

Loci	Gene[a]	Associated?[b]	DER	Correct?[c]
4q27	*TENR*	No	0.54	True negative
4q27	*IL2*	No	0.48	True negative
4q27	*IL21*	No	0.46	True negative
6p21	*HLA-DQB1*	Yes	0.68	True positive
6p21	*HLA-DRB1*	Yes	0.61	True positive
6p21	*HLA-B*	Yes	0.59	True positive
6p21	*HLA-A*	Yes	0.59	True positive
6p21	*HLA-DPB1*	Yes	0.54	False negative
6p21	*TAP2*	Yes	0.58	True positive
6p21	*CFB*	Yes	0.59	True positive
6p21	*MICA*	No	0.5	True negative
6p21	*MICB*	No	0.53	True negative
6p21	*MASIL*	No	0.35	True negative
6p21	*UBD*	No	0.48	True negative
6p21	*ITPR3*	No	0.62	False positive
12p13	*CLEC2D*	No	0.51	True negative
12q13	*ERBB3*	Yes	0.63	True positive
12q24	*C12orf30*	Yes	0.52	False negative
12q24	*SH2B3*	Yes	0.58	True positive
16p13	*KIAA0350*	Yes	0.5	False negative
18p11	*PTPN2*	Yes	0.64	True positive

[a]The positive candidate genes from WTCCC GWAS with reported validation results. [b]Validated to be associated or unassociated with T2DM in the high-quality genotyping. [c]The predicted result using DER ≥ 0.55. DER, differential expression ratio; GWAS, genome-wide association study; T1DM, type 1 diabetes mellitus; WTCCC, Wellcome Trust Case Control Consortium.

that a gene with a significantly higher DER value than other genes in the vicinity will probably explain the observed disease association from the locus.

In 12q13, *ERBB3* is the only gene with high scores in both the WTCCC T1DM GWAS and fitSNPs, and this gene was indeed found to contain rs2292239, which is the only confirmed T1DM marker within this region. In 18p11, *PTPN2* is the only gene suggested by fitSNPs (DER = 0.64), and it was confirmed to explain the association with T1DM for this region. In 16p13, we predicted *SOCS1* to be the most significant gene (DER = 0.64), and the follow-up study showed that it contains the validated marker rs243329 ($-\log_{10}P$ = 4.19). However, we missed *KIAA3350* (DER = 0.5) from 16p13, which has a confirmed association with T1DM and a higher $-\log_{10}P$ than *SOCS1*. In 12p13, no gene has a high score in both GWAS and fitSNPs, which is consistent with the fact that no association was found in the follow-up parent-child trio study [31].

Within 12q24, *SH2B3* and *ALDH2* have high scores in both T1DM and fitSNPs, and indeed *SH2B3* was confirmed to contain a mutation in R262W that explains the association with T1DM in this region in the follow-up study [31]. The association of *SH2B3* with T1DM is somewhat fortuitous because it was originally excluded based on data quality. Only upon recovering additional, poorly clustered nonsynonymous SNPs was it screened for association. This highlights an inadequate prioritization approach, which currently is based on existing functional annotations. This gene prioritization problem is addressed by fitSNPs because it is not biased by existing functional annotations. It is not clear whether there was any follow-up study on mitochondrial aldehyde dehydrogenase 2 (the protein encoded by *ALDH2*), which detoxifies aldehydes generated by alcohol metabolism and lipid peroxidation in the mitochondrial matrix. The association of inactive *ALDH2* genotype with maternal inheritance of T1DM, previously reported in a Japanese population [40], suggests that it may also play a role in T1DM.

Within 4q27, *IL2*, *IL21*, and *TENR* were selected for deep sequencing in the T1DM follow-up study because of the association of T1DM susceptibility with *IL2* in nonobese diabetic mice. However, no T1DM marker had been found in these three genes, and the T1DM association of 4q27 remains unexplained. Figure 4 shows the fitSNPs DER values along with T1DM GWAS $-\log_{10}P$ at 4q27 on the UCSC genome browser [38]. We found that *KIAA1109*'s DER value (0.63) is much greater than those for all other genes in 4q27, including *IL2* (0.48), *IL21* (0.46), and *TENR* (0.54). It is flanked by two most significant T1DM GWAS SNPs, and is highly likely to be associated with T1DM. The $-\log_{10}P$ curve within *KIAA1109* was missing because it was not listed in the genotyping array used in the WTCCC T1DM GWAS (Affymetrix 500K SNP array; Affymetrix Inc., Santa Clara, CA, USA).

Interestingly, the 4q27 region has also been found to be associated with celiac disease [41] and rheumatoid arthritis [42], suggesting that it might be a general risk factor for multiple autoimmune diseases. It has been reported that rs13119723 in *KIAA109* has the most significant association with celiac disease outside the HLA region ($P = 2 \times 10^{-7}$) [41]. By examining our annotated microarray database of disease versus normal gene expression datasets [43], we found that *KIAA1109* was significantly downregulated in peripheral blood cells in juvenile rheumatoid arthritis in two independent studies [44,45]. Additionally, the GNF SymAtlas lists it as being highly expressed in T cells [46]. Therefore, *KIAA1109* is a valuable gene for further investigation in T1DM and other autoim-

Table 2

DER values for T2DM and false positive genes from GWAS

Locus or SNP	Gene	Associated in populations	DER	Correct?[a]
2q37.3	*CAPN10*	Finish. [55], Korean [56], Mexican. [55], Tunisian [57]	0.57	True positive
3p25	*PPARG*	Caucasian. [58], Finish. [59], German. [60], Indian Sikhs. [61], Japanese. [62], Mexican. [63]	0.53	False negative
3q27.2	*IGF2BP2*	Asian. [64], Caucasian. [33], Chinese [65], Danish. [66], French. [67], German. [60], Hispanic. [68], Indian Sikhs. [61], Japanese. [69], Norwegian. [70]	0.54	False negative
6p22.3	*CDKAL1*	Asian. [64], Ashkenazi Jewish. [71], Caucasian. [33], Chinese [65], German. [60], Hispanic. [68], Japanese. [69], Norwegian. [70]	0.55	True positive
8q24.11	*SLC30A8*	Asian. [64], African. [68], Caucasian. [33], Chinese [65], Hispanic. [68], Japanese. [69], Norwegian. [70]	0.42	False negative
9p21	*CDKN2A*	Asian. [64], Caucasian. [34], Chinese [65], Danish. [66], French [72], Japanese. [69]	0.59	True positive
9p21	*CDKN2B*	Asian. [64], Caucasian. [33], Chinese [65], Danish. [66], French [72], Japanese. [69], Norwegian. [70]	0.49	False negative
10q23	*HHEX*	Asian. [64], Caucasian. [33], Chinese [65], Danish. [66], German. [60], Japanese. [69], Norwegian. [70]	0.58	True positive
10q23	*IDE*	Caucasian. [73], Chinese [65], Danish. [66], Japanese. [74], Korean. [75]	0.61	True positive
10q24	*KIF11*	Caucasian, Chinese [65], Danish. [66], Japanese. [74]	0.54	False negative
10q25.3	*TCF7L2*	African. [76], Ashkenazi Jewish. [71], Asian. [64], Caucasian. [33], Chinese. [77], German. [60], Hispanic. [78], Indian Sikhs. [61], Japanese. [79], Spanish, UK white. [80]	0.64	True positive
11p15.1	*KCNJ11*	Arab. [81], Caucasian. [33], Czech [82], Japanese. [69]	0.39	False negative
11p15.5	*KCNQ1*	Singaporean. [35], European. [35], Japanese. [35]	0.6	True positive
16q12.2	*FTO*	Asian. [64], Caucasian. [34], Indian Sikhs. [61], German. [60], Japanese. [83], Norwegian. [70]	0.55	True positive
20q12	*HNF4A*	Amish. [84], Ashkenazim [85], Danish. [86], Finish. [87], Swedish. [87], Mexican. [88], Norwegian. [89], UK Caucasian. [90]	0.63	True positive
rs11078674	*NLGN2*	No	0.53	True negative
rs2866016	*TSPAN5*	No	0.53	True negative
rs12629276	*RFTN1*	No	0.54	True negative
rs8101509	*ZNF649*	No	0.4	True negative
rs945384	*FAM69B*	No	0.53	True negative
rs2050831	*VPS13A*	No	0.63	False positive
rs6670163	*RYR2*	No	0.55	False positive
rs859101	*SLC44A3*	No	0.5	True negative
rs11084127	*ZNF615*	No	0.46	True negative
rs7950175	*KIRREL3*	No	0.48	True negative
rs13064991	*SLC6A20*	No	0.45	True negative
rs6541240	*TTC13*	No	0.51	True negative
rs2278419	*ZNF350*	No	0.45	True negative

[a]The predicted result using DER $\geq$ 0.55. DER, differential expression ratio; GWAS, genome-wide association study; T2DM, type 2 diabetes mellitus.

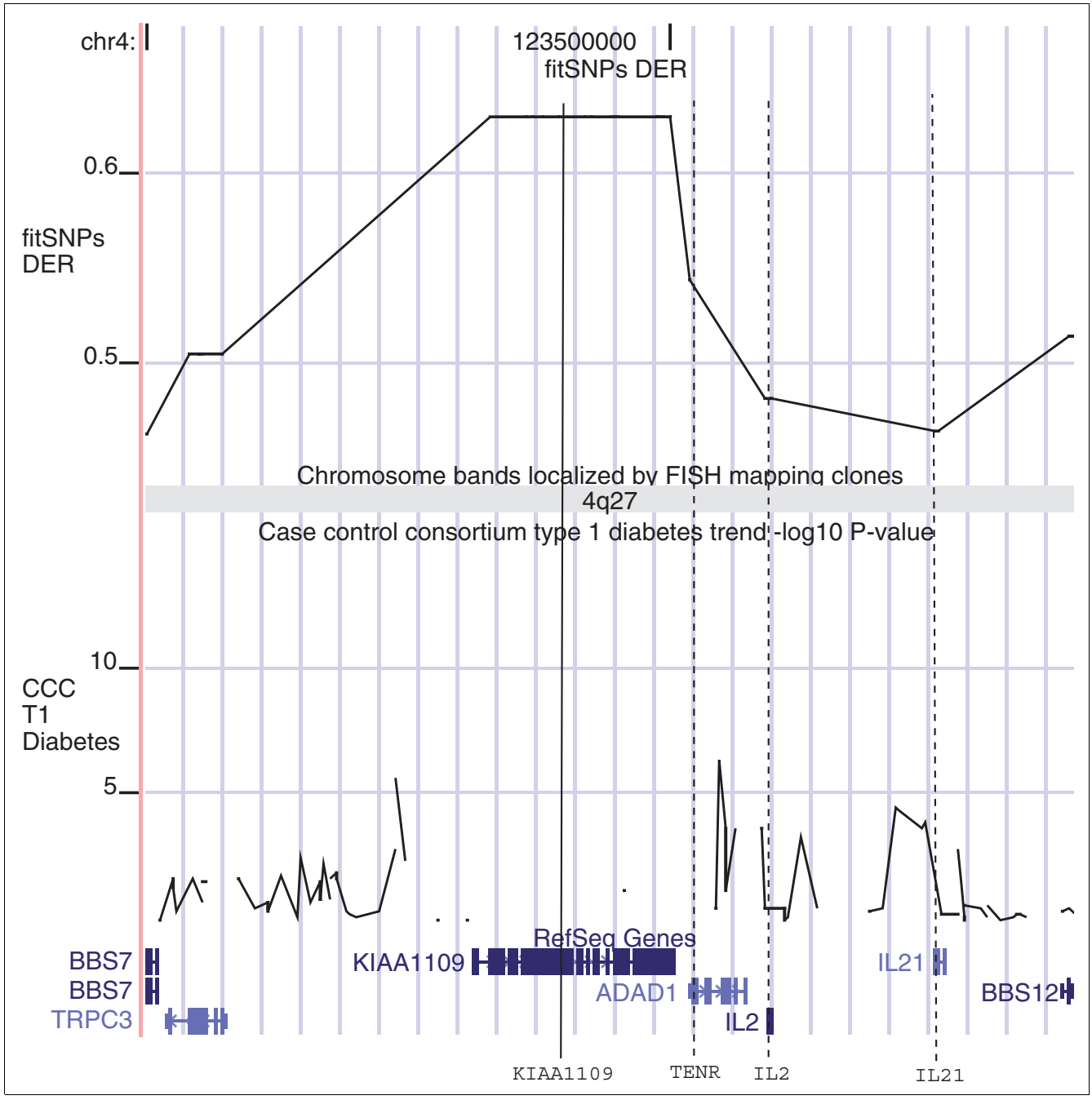

Figure 4
Interpreting T1DM GWAS findings at 4q27 using fitSNPs. The region 4q27 has been identified as a risk factor area for T1DM, celiac disease, and rheumatoid arthritis. *IL2*, *IL21*, and *TENR* were selected based on prior knowledge for sequencing in the follow-up studies, but no association was found. *KIAA1109* has a much higher fitSNPs DER value than all other genes in the region, and is flanked by two significant T1DM GWAS SNPs ($-\log_{10}P > 5$). We predicted that this gene may explain the T1DM association in this region. The GWAS $-\log_{10}P$ curve for *KIAA1109* is missing because it was not listed in the Affymetrix 500 K SNP array used for the GWAS. DER, differential expression ratio; fitSNPs, functionally interpolating single nucleotide polymorphisms; GWAS, genome-wide association study; SNP, single nucleotide polymorphism; T1DM, type 1 diabetes mellitus.

mune diseases, and we predict that it is likely to explain the T1DM association in 4q27.

Comparing DER values among different types of disease genes

The success of these three validation studies demonstrates that fitSNPs could be used not only to prioritize different loci from GWASs but also to prioritize genes from each locus. Before applying fitSNPs to all diseases, one important question is whether genes associated with different type of diseases have different DER values. We downloaded lists of disease genes for Mendelian diseases (highly penetrant diseases caused by a single mutation), complex diseases, and cancer, which were compiled by Ran Blekhman and coworkers [47]. As shown in Table 3, no significant DER difference were observed between Mendelian and complex disease genes (0.53 versus 0.54; $P = 0.2$, *t*-test). Cancer genes exhibited significantly higher DER values (0.56) than did both Mendelian ($P < 0.0001$, *t*-test) and complex disease genes ($P = 0.001$, *t*-test). Furthermore, all types of disease genes exhibited significantly higher DER values than did nondisease genes ($P < 0.0001$, *t*-test). These findings suggest that fitSNPs could be used to prioritize disease genes for both Mendelian and complex diseases, and would be even more effective in prioritizing cancer genes.

FitSNPs predicts disease genes in OMIM loci with unknown molecular basis

FitSNPs could be used not only to prioritize disease genes from GWASs for multiple disease types, but also to predict disease associations for genes with high DER values. There are 5,253 human genes with DER ≥ 0.55. Of these, 23% have known variants for various diseases according to GAD and HGMD. The remaining 4,052 genes have not yet been shown to associate with any diseases through mutations or polymorphisms, making them promising leads. To systematically predict disease associations for them, we searched OMIM and found that 830 diseases and syndromes have been linked to cytogenetic locations but not specific genes. From these cytogenetic locations, we predicted 3,331 highly differentially expressed genes with DER ≥ 0.55 in 610 diseases. From this group, 2,586 genes, which are currently not associated with any disease according to GAD and HGMD, were predicted to be associated with 597 diseases [48].

For example, systemic lupus erythemetosus (SLE) is an autoimmune disease with multiple organ involvement and a genetic predisposition. Renal disease occurs in 40% to 75% of SLE patients and up to 90% of childhood SLE patients, and significantly contributes to morbidity and mortality. A genome scan was performed with more than 300 microsatellite markers in the 75 pedigrees that had SLE with nephritis, and linkage was identified at 2q34-q35 with $P = 0.000001$ (*SLEN2*; OMIM %607966). To date, no gene in 2q34-q35 has been associated with *SLEN2*. The DER for the gene *OBSL1* (obscurin-like 1; DER = 0.71) is significantly greater than that for all other genes (Figure 5). Actually, it has the second highest DER value among all human genes without known disease-associated variants. By examining our annotated microarray database of disease versus normal gene expression datasets [43], we found that *OBSL1* was significantly differentially expressed in juvenile idiopathic arthritis (GEO series 8650) and several kidney diseases, such as kidney cancer (GEO dataset 9) and kidney transplant rejection (GEO dataset 724). Therefore, we suggest that *OBSL1* might be associated with *SLEN2*. Similarly, we suggest that the 2,586 genes predicted with DER values are top candidate genes for the 597 syndromes in question.

Discussion

We analyzed 476 human GEO datasets and calculated the frequency of differential expression for every gene, which we called the differential expression ratio (DER). The enrichment analysis on a comprehensive list of curated disease genes revealed a positive association between DER values and the likelihood of harboring disease-associated mutations. We were able to rediscover all disease genes with 79% specificity and 37% sensitivity using a simple threshold of DER ≥ 0.55. These highly differentially expressed genes were 2.25 times

Table 3

DER value comparisons among Mendelian, complex, cancer, all disease genes and nondisease genes

P value[a]	Mendelian (mean = 0.53, n = 931)	Complex (mean = 0.54, n = 70)	Cancer (mean = 0.56, n = 324)	All diseases (mean = 0.53, n = 3,178)	Nondisease (mean = 0.50, n = 16,698)
Mendelian		0.2	<0.0001	0.4	<0.0001
Complex			0.001	0.3	<0.0001
Cancer				<0.0001	<0.0001
All diseases					<0.0001
Nondisease					

**P* values were calculated using *t*-test. DER, differential expression ratio.

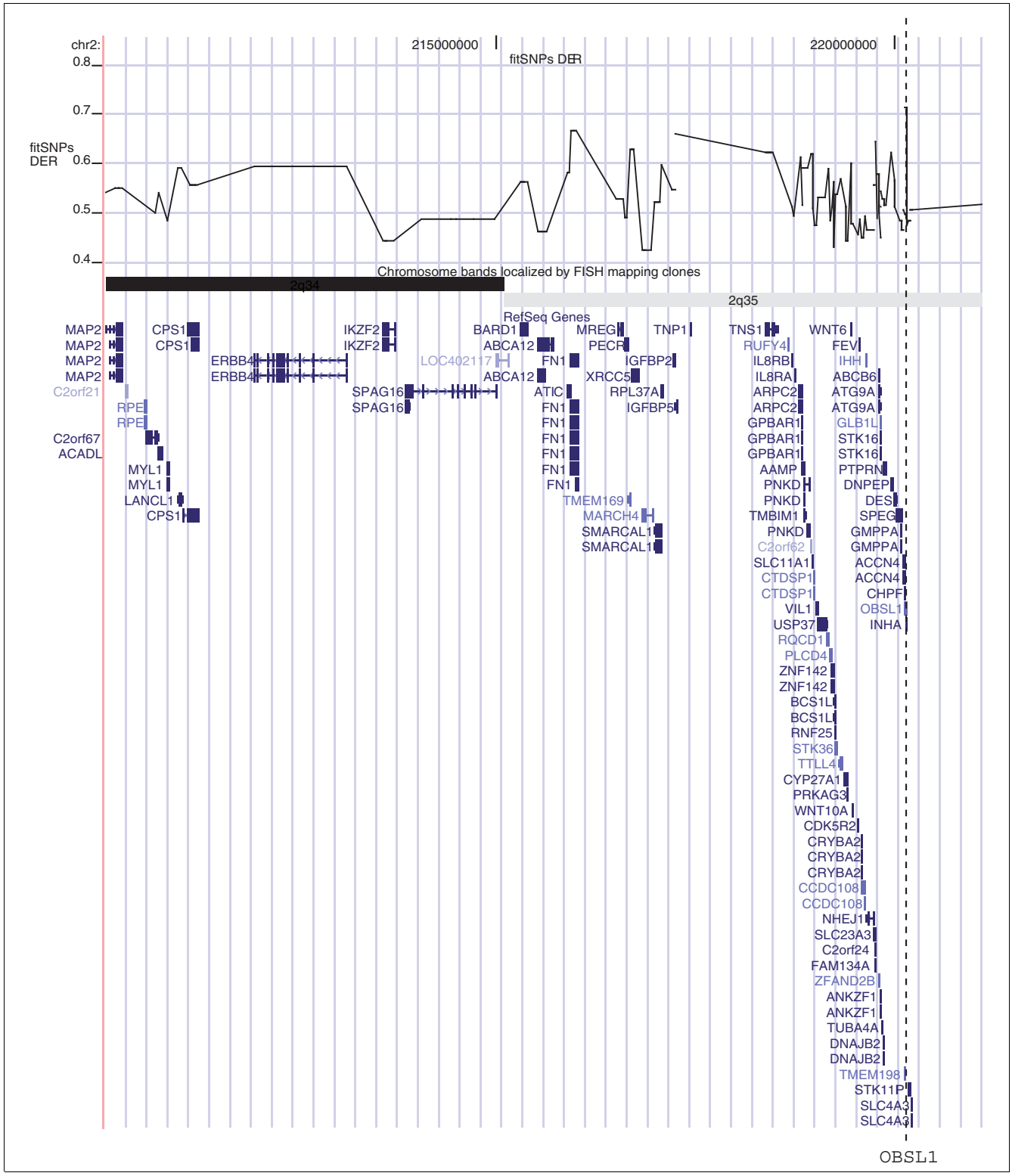

Figure 5
Prediction that *OBSL1* is associated with systemic lupus erythematosus with nephritis through 2q34-q35. Systemic lupus erythemetosus with nephritis (*SLEN2*; OMIM %607966) was identified to be associated with 2q34-q35 but without identification of specific genes. *OBSL1* has a much higher DER value (0.71) than those of all other genes from 2q34-q35. It was also found to be differentially expressed in juvenile idiopathic arthritis, kidney cancer, and kidney transplant rejection. Therefore, we suggest that it should be sequenced for its potential association with *SLEN2*.

more likely to harbor disease-associated variants than others. The positive association between DER and our precision to rediscover disease genes was consistently observed across ranges of DER values, in spite of variable adjustments, including adjusting the *q* value cutoff from 0.005 to 0.2, and the removal of genes measured in fewer than 0% to 30% microarray studies. Additionally, we analyzed disease genes from three different human genetic association databases, namely GAD, HGMD, and OMIM, individually and observed the same DER-related increase in precision. We also used the absolute GEO dataset counts instead of the DER to rediscover disease genes and observed the same pattern. The majority of 476 GEO datasets are genome-wide experiments; 98% of GEO datasets contained more than 5,000 probes, and 89% contained more than 10,000 probes, which are unlikely to be targeted arrays. These results demonstrated a robust association between differential expression and disease variants.

Based on the observed associations, we created a tool called fitSNPs to prioritize disease genes from candidate GWAS loci. First, we successfully distinguished true disease genes from false positives (positive SNPs from initial scan subsequently found to be negative during validation) for T1DM GWASs with 89% specificity and 75% sensitivity, and T2DM GWASs with 85% specificity and 60% sensitivity. We then directly rediscovered true T1DM genes by analyzing the top seven loci of WTCCC GWAS initial scan results using fitSNPs. Furthermore, in an unexplained locus (4q27), fitSNPs predicted that a novel gene, *KIAA1109*, may explain the association for T1DM and several autoimmune diseases. We also examined the findings of a segmental copy number variation (CNV) study [49], which was performed using a whole-genome tiling-path bacterial artificial chromosome array to detect a gain or loss of more than 40 kilobases in 93 human samples. The results were uploaded into the UCSC genome browser as a custom track. Using the custom track, we found a CNV in *KIAA1109*, suggesting that CNV might play a role in T1DM.

Although there are existing gene prioritization methods, this is the first to describe the use of differential expression to systematically prioritize candidate genes or SNPs. We acknowledge that no single gene prioritization method is perfect and suggest that fitSNPs can also be used in a complementary manner with other prioritization methods. Given that there are more than 100 published GWASs, we believe that fitSNPs can serve as an effective tool to systematically prioritize candidate SNPs from them.

In theory, FitSNPs can also be used to design SNP arrays for GWASs. It has been shown that tagSNPs could lower costs by 53% while capturing 80% of common SNPs in the African population [50]. In comparison, a DER of 0.48 achieved similar sensitivity; 57% of genes in the genome have a DER value larger than 0.48. They comprise 74% of genes known to have disease-associated variants. A GWAS focusing on these genes could lower experimental costs by 43% while covering at least 74% of disease genes. Therefore, fitSNPs could reduce GWAS costs in a way comparable to that of tagSNPs, but with the additional advantages of gene prioritization and direct linkage to functional experiments. Furthermore, fitSNPs could be combined with tagSNPs in the design of GWASs to further reduce costs and to expedite the discovery of causative genes and DNA variants.

To facilitate the use of fitSNPs, we developed a web server [51] that retrieves DER values, and a comprehensive list of validated and predicted disease associations for all human genes and their underlying microarray study results.

Conclusion

This study demonstrates that highly differentially expressed genes are more likely to harbor disease-associated variants. FitSNPs successfully distinguished true disease genes from false positives of GWASs for multiple diseases, and can serve as a powerful and convenient tool to prioritize disease genes from GWASs. We further proposed 2,586 genes to sequence for 597 syndromes with unknown molecular basis. With the wealth of genomic, genetic, and disease databases in public international repositories, we are now able to investigate systematically the molecular and genetic mechanisms of diseases, make predictions, and validate them using commercial kits and core facilities. To maximize their value, these molecular measurements must be placed within the context of physiology. A public repository of de-identified clinical measurements will greatly accelerate this process [52].

Materials and methods

GEO datasets

The GEO contains gene expression profiles for more than 200,000 individual microarray samples. They are assembled into biologically meaningful and comparable GEO datasets with manually annotated experimental details, such as variables that were studied in the experiment. All samples within a GEO dataset were measured on the same platform with the same background processing and normalization, and their values were directly comparable. We downloaded, processed, and annotated all GEO datasets from GEO, and obtained 476 human GEO datasets, in which both the GEO platform and the GEO dataset were annotated as human.

Differentially expressed genes

Each GEO dataset was categorized into subsets annotated with one of the 24 types, including disease state, genotype/variation, strain, infection, development stage, age, time, agent, dose, tissue, cell type, cell line, metabolism, stress, growth protocol, protocol, gender, individual, isolate, shock, species, specimen, temperature, and others. We performed all possible subset-versus-subset comparisons in each comparison type in every GEO dataset, ignoring subsets with fewer than three samples. For every comparison, we identi-

fied all differentially expressed probes using two class unpaired analysis in the R package of SAM (SAMR) with version 1.25 [28]. We used all default parameters with standard *t*-statistics: nperms = 50, fold > 0, and delta < 0.4. All differentially expressed probes with $q \leq 0.05$ were recorded and annotated with the latest Entrez Gene IDs using AILUN [27]. For 4,552 out of 4,877 comparisons, at least one gene exhibited a significant difference.

DER

The DER was calculated for each Entrez Gene ID as the count of GEO datasets in which it was differentially expressed divided by the count of GEO datasets in which it was measured. Genes measured in fewer than 5% of GEO datasets were removed.

Disease genes

Human genes with known disease-associated variants were downloaded from HGMD Professional (Biobase) and GAD. HGMD gene symbols were related to Entrez Gene IDs using AILUN [27]. Entrez Gene IDs were retrieved from GAD entries with validated disease associations, and compared with the latest Entrez Gene ID list to replace or remove outdated Gene IDs and nonhuman genes.

Differentially expressed genes versus disease genes

For a cutoff from 0 to 1 with an increment of 0.02, differentially expressed genes with DER values greater than the cutoff were compared with the list of disease genes to calculate the precision and recall. Constantly expressed genes with DER less than the cutoff were similarly evaluated. For the control, the label of disease genes was also shuffled 10,000 times within all human genes and compared with differentially expressed genes.

Comparison of DER values for T1DM genes with those of false positives in GWASs

For each of top seven loci described in the WTCCC T1DM GWAS [31], all genes with described validation results were manually extracted from the paper and supplementary materials [30,31]. The DER values were compared between T1DM and non-T1DM genes in accordance with the validation result.

Comparison of DER values for T2DM genes with those of false positives in GWASs

T2DM genes were extracted from six T2DM GWASs [29,32-36] and tens of association studies. Of them, genes associated with T2DM in three or more populations were recorded as true T2DM genes. False-positive SNPs were extracted from Table S7 of the report of a T2DM GWAS [32], which were found to be positive in the stage 1 GWAS but found to be unassociated with T2DM during the validation phase, with p value from permutation larger than 0.05. They were annotated with Entrez Gene IDs using Entrez dbSNP [53]. All SNPs without gene annotations were removed.

FitSNPs

FitSNPs [39] is a list of human Entrez Gene IDs with DER values. Genes with disease-associated variants and corresponding diseases were retrieved from HGMD [26] and GAD [25]. To facilitate integration between fitSNPs and GWASs on the human genome, all reference SNPs were downloaded from dbSNP [53] and assigned DER scores according to their associated genes. For SNPs mapping to multiple genes, the highest DER value was selected. FitSNPs can be loaded into the UCSC genome graph, in accordance with the instructions in the GWAS page of the FitSNPs server [37]. It will automatically show up as a custom track in the UCSC genome browser that can be compared with a wealth of genomic data, including multiple GWAS study results.

Predicting T1DM genes from the top seven loci of GWASs

Both DER values and WTCCC T1DM GWAS $-\log_{10}P$ were visualized in UCSC genome browser [54] for the top seven loci. Genes with DER value ≥ 0.55 and $-\log_{10}P > 5$ were predicted to be T1DM genes, and compared with the validation findings.

Mapping diseases without known molecular basis to lead genes

All diseases in OMIM morbid map with a percentage preceding their MIM numbers were considered to be Mendelian disorders without known molecular association. Cytogenetic locations of these diseases and all human genes were retrieved from the OMIM morbid map and the Human Gene Nomenclature Committee, respectively. Highly differentially expressed genes with DER ≥ 0.55 were identified from the cytogenetic location for each disease. Within them, genes that have not been known to have disease-associated variants were predicted to be associated with a corresponding disease.

Abbreviations

CNV: copy number variation; DER: differential expression ratio; fitSNPs: functionally interpolating single nucleotide polymorphisms; GAD: Genetic Association Database; GEO: Gene Expression Omnibus; GWAS: genome-wide association study; HGMD: Human Gene Mutation Database; OMIM: Online Mendelian Inheritance in Man; SAM: significance analysis of microarrays; SLE: systemic lupus erythemetosus; SNP: single nucleotide polymorphism; T1DM: type 1 diabetes mellitus; T2DM: type 2 diabetes mellitus; UCSC: University of California Santa Cruz; WTCCC: Wellcome Trust Case Control Consortium.

Authors' contributions

RC designed and performed the experiments, and wrote the manuscript. AB provided the overall project guidance. AC provided critical review and edited the manuscript. KK collected T2DM gene lists and gave advice on the validation. AM,

JD, TD, and LL gave advice on the experiments. All authors read and approved the manuscript.

Additional data files

The following additional data are available with the online version of this paper. Additional data file 1 is an enrichment comparison between genes differentially expressed in one or more microarray studies and genes with disease-associated variants.

Acknowledgements

This work was supported by Lucile Packard Foundation for Children's Health, US National Library of Medicine (K22 LM008261), National Institute of General Medical Sciences (R01 GM079719), Howard Hughes Medical Institute, and Pharmaceutical Research and Manufacturers of America Foundation. We thank Alex Skrenchuk from Stanford University for computer support, and Rohan Mallelwar and Ajit Thosar from Optra Systems for website development.

References

1. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir G, Reynisdottir I, Gudbjartsson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir U, Gretarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, Lamb JR, *et al.*: **Genetics of gene expression and its effect on disease.** *Nature* 2008, **452:**423-428.
2. Keller MP, Choi Y, Wang P, Davis DB, Rabaglia ME, Oler AT, Stapleton DS, Argmann C, Schueler KL, Edwards S, Steinberg HA, Neto EC, Kleinhanz R, Turner S, Hellerstein MK, Schadt EE, Yandell BS, Kendziorski C, Attie AD: **A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility.** *Genome Res* 2008, **18:**706-716.
3. Wang SS, Schadt EE, Wang H, Wang X, Ingram-Drake L, Shi W, Drake TA, Lusis AJ: **Identification of pathways for atherosclerosis in mice: integration of quantitative trait locus analysis and global gene expression data.** *Circ Res* 2007, **101:**e11-e30.
4. Meng H, Vera I, Che N, Wang X, Wang SS, Ingram-Drake L, Schadt EE, Drake TA, Lusis AJ: **Identification of Abcc6 as the major causal gene for dystrophic cardiac calcification in mice through integrative genomics.** *Proc Natl Acad Sci USA* 2007, **104:**4530-4535.
5. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, Macneil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy MF, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE: **Variations in DNA elucidate molecular networks that cause disease.** *Nature* 2008, **452:**429-435.
6. Schadt EE, Lum PY: **Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Reverse engineering gene networks to identify key drivers of complex disease phenotypes.** *J Lipid Res* 2006, **47:**2601-2613.
7. Thomas G, Jacobs KB, Yeager M, Kraft P, Wacholder S, Orr N, Yu K, Chatterjee N, Welch R, Hutchinson A, Crenshaw A, Cancel-Tassin G, Staats BJ, Wang Z, Gonzalez-Bosquet J, Fang J, Deng X, Berndt SI, Calle EE, Feigelson HS, Thun MJ, Rodriguez C, Albanes D, Virtamo J, Weinstein S, Schumacher FR, Giovannucci E, Willett WC, Cussenot O, Valeri A, *et al.*: **Multiple loci identified in a genome-wide association study of prostate cancer.** *Nat Genet* 2008, **40:**310-315.
8. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6:**55.
9. Lopez-Bigas N, Ouzounis CA: **Genome-wide identification of genes likely to be involved in human genetic disease.** *Nucleic Acids Res* 2004, **32:**3108-3114.
10. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *J Med Genet* 2006, **43:**691-698.
11. Wu X, Jiang R, Zhang MQ, Li S: **Network-based global inference of human disease genes.** *Mol Syst Biol* 2008, **4:**189.
12. Hristovski D, Peterlin B, Mitchell JA, Humphrey SM: **Using literature-based discovery to identify disease candidate genes.** *Int J Med Inform* 2005, **74:**289-298.
13. Perez-Iratxeta C, Wjst M, Bork P, Andrade MA: **G2D: a tool for mining genes associated with disease.** *BMC Genet* 2005, **6:**45.
14. Tranchevent LC, Barriot R, Yu S, Vooren SV, Loo PV, Coessens B, Moor BD, Aerts S, Moreau Y: **ENDEAVOUR update: a web resource for gene prioritization in multiple species.** *Nucleic Acids Res* 2008, **36:**W377-W384.
15. Oti M, Brunner HG: **The modular nature of genetic diseases.** *Clin Genet* 2007, **71:**1-11.
16. Zhu M, Zhao S: **Candidate gene identification approach: progress and challenges.** *Int J Biol Sci* 2007, **3:**420-427.
17. Gaulton KJ, Mohlke KL, Vision TJ: **A computational system to select candidate genes for complex human traits.** *Bioinformatics* 2007, **23:**1132-1140.
18. Masseroli M, Galati O, Pinciroli F: **GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists.** *Nucleic Acids Res* 2005, **33:**W717-W723.
19. Tiffin N, Kelso JF, Powell AR, Pan H, Bajic VB, Hide WA: **Integration of text- and data-mining using ontologies successfully selects disease gene candidates.** *Nucleic Acids Res* 2005, **33:**1544-1552.
20. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases.** *Nucleic Acids Res* 2005, **33:**W758-W761.
21. Ala U, Piro RM, Grassi E, Damasco C, Silengo L, Oti M, Provero P, Di Cunto F: **Prediction of human disease genes by human-mouse conserved coexpression analysis.** *PLoS Comput Biol* 2008, **4:**e1000043.
22. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S: **TOM: a web-based integrated approach for identification of candidate disease genes.** *Nucleic Acids Res* 2006, **34:**W285-W292.
23. Ma X, Lee H, Wang L, Sun F: **CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data.** *Bioinformatics* 2007, **23:**215-221.
24. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles: database and tools update.** *Nucleic Acids Res* 2007, **35:**D760-D765.
25. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36:**431-432.
26. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN: **Human Gene Mutation Database (HGMD): 2003 update.** *Hum Mutat* 2003, **21:**577-581.
27. Chen R, Li L, Butte AJ: **AILUN: reannotating gene expression data automatically.** *Nat Methods* 2007, **4:**879.
28. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci USA* 2001, **98:**5116-5121.
29. Wellcome Trust Case Control Consortium: **Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.** *Nature* 2007, **447:**661-678.
30. Nejentsev S, Howson JM, Walker NM, Szeszko J, Field SF, Stevens HE, Reynolds P, Hardy M, King E, Masters J, Hulme J, Maier LM, Smyth D, Bailey R, Cooper JD, Ribas G, Campbell RD, Clayton DG, Todd JA: **Localization of type 1 diabetes susceptibility to the MHC class I genes HLA-B and HLA-A.** *Nature* 2007, **450:**887-892.
31. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, Lowe CE, Szeszko JS, Hafler JP, Zeitels L, Yang JH, Vella A, Nutland S, Stevens HE, Schuilenburg H, Coleman G, Maisuria M, Meadows W, Smink LJ, Healy B, Burren OS, Lam AA, Ovington NR, Allen J, Adlem E, Leung HT, *et al.*: **Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes.** *Nat Genet* 2007, **39:**857-864.
32. Sladek R, Rocheleau G, Rung J, Dina C, Shen L, Serre D, Boutin P, Vincent D, Belisle A, Hadjadj S, Balkau B, Heude B, Charpentier G, Hudson TJ, Montpetit A, Pshezhetsky AV, Prentki M, Posner BI, Balding DJ, Meyre D, Polychronakos C, Froguel P: **A genome-wide association study identifies novel risk loci for type 2 diabetes.** *Nature* 2007, **445:**881-885.
33. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H,

Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ, Hughes TE, Groop L, Altshuler D, Almgren P, Florez JC, Meyer J, Ardlie K, Bengtsson Bostrom K, Isomaa B, Lettre G, Lindblad U, Lyon HN, Melander O, Newton-Cheh C, Nilsson P, Orho-Melander M, Rastam L, Speliotes EK, Taskinen MR, Tuomi T, *et al.*: **Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels.** *Science* 2007, **316:**1331-1336.

34. Scott LJ, Mohlke KL, Bonnycastle LL, Willer CJ, Li Y, Duren WL, Erdos MR, Stringham HM, Chines PS, Jackson AU, Prokunina-Olsson L, Ding CJ, Swift AJ, Narisu N, Hu T, Pruim R, Xiao R, Li XY, Conneely KN, Riebow NL, Sprau AG, Tong M, White PP, Hetrick KN, Barnhart MW, Bark CW, Goldstein JL, Watkins L, Xiang F, Saramies J, *et al.*: **A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.** *Science* 2007, **316:**1341-1345.
35. Unoki H, Takahashi A, Kawaguchi T, Hara K, Horikoshi M, Andersen G, Ng DP, Holmkvist J, Borch-Johnsen K, Jorgensen T, Sandbaek A, Lauritzen T, Hansen T, Nurbaya S, Tsunoda T, Kubo M, Babazono T, Hirose H, Hayashi M, Iwamoto Y, Kashiwagi A, Kaku K, Kawamori R, Tai ES, Pedersen O, Kamatani N, Kadowaki T, Kikkawa R, Nakamura Y, Maeda S: **SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations.** *Nat Genet* 2008, **40:**1098-1102.
36. Yasuda K, Miyake K, Horikawa Y, Hara K, Osawa H, Furuta H, Hirota Y, Mori H, Jonsson A, Sato Y, Yamagata K, Hinokio Y, Wang HY, Tanahashi T, Nakamura N, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Takeda J, Maeda E, Shin HD, Cho YM, Park KS, Lee HK, Ng MC, Ma RC, *et al.*: **Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus.** *Nat Genet* 2008, **40:**1092-1097.
37. **FitSNPs for GWAS** [http://fitSNPs.stanford.edu/fitSNPs.php]
38. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.
39. **FitSNPs Gene** [http://fitSNPs.stanford.edu/gene.php]
40. Suzuki Y, Matsuura N, Suzuki S, Muramatsu T, Taniyama M, Ohta S, Higuchi S, Tsukahara M, Atusmi Y, Matsuoka K: **Aldehyde dehydrogenase 2 genotype in type 1 diabetes mellitus.** *Diabetes Res Clin Pract* 2003, **60:**139-141.
41. van Heel DA, Franke L, Hunt KA, Gwilliam R, Zhernakova A, Inouye M, Wapenaar MC, Barnardo MC, Bethel G, Holmes GK, Feighery C, Jewell D, Kelleher D, Kumar P, Travis S, Walters JR, Sanders DS, Howdle P, Swift J, Playford RJ, McLaren WM, Mearin ML, Mulder CJ, McManus R, McGinnis R, Cardon LR, Deloukas P, Wijmenga C: **A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21.** *Nat Genet* 2007, **39:**827-829.
42. Zhernakova A, Alizadeh BZ, Bevova M, van Leeuwen MA, Coenen MJ, Franke B, Franke L, Posthumus MD, van Heel DA, Steege G van der, Radstake TR, Barrera P, Roep BO, Koeleman BP, Wijmenga C: **Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases.** *Am J Hum Genet* 2007, **81:**1284-1288.
43. Kodama K, Butte AJ, Creusot RJ, Su L, Sheng D, Hartnett M, Iwai H, Soares LR, Fathman CG: **Tissue- and age-specific changes in gene expression during disease induction and progression in NOD mice.** *Clin Immunol* 2008, **129:**195-201.
44. Barnes MG, Aronow BJ, Luyrink LK, Moroldo MB, Pavlidis P, Passo MH, Grom AA, Hirsch R, Giannini EH, Colbert RA, Glass DN, Thompson SD: **Gene expression in juvenile arthritis and spondyloarthropathy: pro-angiogenic ELR+ chemokine genes relate to course of arthritis.** *Rheumatology (Oxford)* 2004, **43:**973-979.
45. Fall N, Barnes M, Thornton S, Luyrink L, Olson J, Ilowite NT, Gottlieb BS, Griffin T, Sherry DD, Thompson S, Glass DN, Colbert RA, Grom AA: **Gene expression profiling of peripheral blood from patients with untreated new-onset systemic juvenile idiopathic arthritis reveals molecular heterogeneity that may predict macrophage activation syndrome.** *Arthritis Rheum* 2007, **56:**3793-3804.
46. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101:**6062-6067.
47. Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M: **Natural selection on genes that underlie human disease susceptibility.** *Curr Biol* 2008, **18:**883-889.
48. **FitSNPs prediction.** [http://fitsnps.stanford.edu/prediction.php]
49. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, Lam WL: **A comprehensive analysis of common copy-number variations in the human genome.** *Am J Hum Genet* 2007, **80:**91-104.
50. International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.
51. **FitSNPs** [http://fitSNPs.stanford.edu]
52. Butte AJ: **Medicine. The ultimate model organism.** *Science* 2008, **320:**325-327.
53. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29:**308-311.
54. **UCSC genome browser** [http://genome.ucsc.edu/cgi-bin/hgGateway]
55. Horikawa Y, Oda N, Cox NJ, Li X, Orho-Melander M, Hara M, Hinokio Y, Lindner TH, Mashima H, Schwarz PE, del Bosque-Plata L, Horikawa Y, Oda Y, Yoshiuchi I, Colilla S, Polonsky KS, Wei S, Concannon P, Iwasaki N, Schulze J, Baier LJ, Bogardus C, Groop L, Boerwinkle E, Hanis CL, Bell GI: **Genetic variation in the gene encoding calpain-10 is associated with type 2 diabetes mellitus.** *Nat Genet* 2000, **26:**163-175.
56. Kang ES, Kim HJ, Nam M, Nam CM, Ahn CW, Cha BS, Lee HC: **A novel 111/121 diplotype in the Calpain-10 gene is associated with type 2 diabetes.** *J Hum Genet* 2006, **51:**629-633.
57. Kifagi C, Makni K, Mnif F, Boudawara M, Hamza N, Rekik N, Abid M, Rebai A, Granier C, Jarraya F, Ayadi H: **Association of calpain-10 polymorphisms with type 2 diabetes in the Tunisian population.** *Diabetes Metab* 2008, **34:**273-278.
58. Beamer BA, Yen CJ, Andersen RE, Muller D, Elahi D, Cheskin LJ, Andres R, Roth J, Shuldiner AR: **Association of the Pro12Ala variant in the peroxisome proliferator-activated receptor-gamma2 gene with obesity in two Caucasian populations.** *Diabetes* 1998, **47:**1806-1808.
59. Lindi VI, Uusitupa MI, Lindstrom J, Louheranta A, Eriksson JG, Valle TT, Hamalainen H, Ilanne-Parikka P, Keinanen-Kiukaanniemi S, Laakso M, Tuomilehto J: **Association of the Pro12Ala polymorphism in the PPAR-gamma2 gene with 3-year incidence of type 2 diabetes and body weight change in the Finnish Diabetes Prevention Study.** *Diabetes* 2002, **51:**2581-2586.
60. Herder C, Rathmann W, Strassburger K, Finner H, Grallert H, Huth C, Meisinger C, Gieger C, Martin S, Giani G, Scherbaum WA, Wichmann HE, Illig T: **Variants of the PPARG, IGF2BP2, CDKAL1, HHEX, and TCF7L2 genes confer risk of type 2 diabetes independently of BMI in the German KORA studies.** *Horm Metab Res* 2008, **40:**722-726.
61. Sanghera DK, Ortega L, Han S, Singh J, Ralhan SK, Wander GS, Mehra NK, Mulvihill JJ, Ferrell RE, Nath SK, Kamboh MI: **Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2 (Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk.** *BMC Med Genet* 2008, **9:**59.
62. Horiki M, Ikegami H, Fujisawa T, Kawabata Y, Ono M, Nishino M, Shimamoto K, Ogihara T: **Association of Pro12Ala polymorphism of PPARgamma gene with insulin resistance and related diseases.** *Diabetes Res Clin Pract* 2004, **66(suppl 1):**S63-S67.
63. Black MH, Fingerlin TE, Allayee H, Zhang W, Xiang AH, Trigo E, Hartiala J, Lehtinen AB, Haffner SM, Bergman RN, McEachin RC, Kjos SL, Lawrence JM, Buchanan TA, Watanabe RM: **Evidence of interaction between PPARG2 and HNF4A contributing to variation in insulin sensitivity in Mexican Americans.** *Diabetes* 2008, **57:**1048-1056.
64. Ng MC, Park KS, Oh B, Tam CH, Cho YM, Shin HD, Lam VK, Ma RC, So WY, Cho YS, Kim HL, Lee HK, Chan JC, Cho NH: **Implication of genetic variants near TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, and FTO in type 2 diabetes and obesity in 6,719 Asians.** *Diabetes* 2008, **57:**2226-2233.
65. Wu Y, Li H, Loos RJ, Yu Z, Ye X, Chen L, Pan A, Hu FB, Lin X: **Common variants in CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8 and HHEX/IDE genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population.** *Diabetes* 2008, **57:**2834-2842.
66. Grarup N, Rose CS, Andersson EA, Andersen G, Nielsen AL, Albrechtsen A, Clausen JO, Rasmussen SS, Jorgensen T, Sandbaek A, Lauritzen T, Schmitz O, Hansen T, Pedersen O: **Studies of association of variants near the HHEX, CDKN2A/B, and IGF2BP2 genes with type 2 diabetes and impaired insulin release in 10,705 Danish subjects: validation and extension of genome-wide**

association studies. *Diabetes* 2007, **56:**3105-3111.
67. Duesing K, Fatemifar G, Charpentier G, Marre M, Tichet J, Hercberg S, Balkau B, Froguel P, Gibson F: **Evaluation of the association of IGF2BP2 variants with type 2 diabetes in French Caucasians.** *Diabetes* 2008, **57:**1992-1996.
68. Palmer ND, Goodarzi MO, Langefeld CD, Ziegler J, Norris JM, Haffner SM, Bryer-Ash M, Bergman RN, Wagenknecht LE, Taylor KD, Rotter JI, Bowden DW: **Quantitative trait analysis of type 2 diabetes susceptibility loci identified from whole genome association studies in the Insulin Resistance Atherosclerosis Family Study.** *Diabetes* 2008, **57:**1093-1100.
69. Omori S, Tanaka Y, Takahashi A, Hirose H, Kashiwagi A, Kaku K, Kawamori R, Nakamura Y, Maeda S: **Association of CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8, and KCNJ11 with susceptibility to type 2 diabetes in a Japanese population.** *Diabetes* 2008, **57:**791-795.
70. Hertel JK, Johansson S, Raeder H, Midthjell K, Lyssenko V, Groop L, Molven A, Njolstad PR: **Genetic analysis of recently identified type 2 diabetes loci in 1,638 unselected patients with type 2 diabetes and 1,858 control participants from a Norwegian population-based cohort (the HUNT study).** *Diabetologia* 2008, **51:**971-977.
71. Bronstein M, Pisante A, Yakir B, Darvasi A: **Type 2 diabetes susceptibility loci in the Ashkenazi Jewish population.** *Hum Genet* 2008, **124:**101-104.
72. Duesing K, Fatemifar G, Charpentier G, Marre M, Tichet J, Hercberg S, Balkau B, Froguel P, Gibson F: **Strong association of common variants in the CDKN2A/CDKN2B region with type 2 diabetes in French Europids.** *Diabetologia* 2008, **51:**821-826.
73. Florez JC, Wiltshire S, Agapakis CM, Burtt NP, de Bakker PI, Almgren P, Bengtsson Bostrom K, Tuomi T, Gaudet D, Daly MJ, Hirschhorn JN, McCarthy MI, Altshuler D, Groop L: **High-density haplotype structure and association testing of the insulin-degrading enzyme (IDE) gene with type 2 diabetes in 4,206 people.** *Diabetes* 2006, **55:**128-135.
74. Furukawa Y, Shimada T, Furuta H, Matsuno S, Kusuyama A, Doi A, Nishi M, Sasaki H, Sanke T, Nanjo K: **Polymorphisms in the IDE-KIF11-HHEX gene locus are reproducibly associated with type 2 diabetes in a Japanese population.** *J Clin Endocrinol Metab* 2008, **93:**310-314.
75. Kwak SH, Cho YM, Moon MK, Kim JH, Park BL, Cheong HS, Shin HD, Jang HC, Kim SY, Lee HK, Park KS: **Association of polymorphisms in the insulin-degrading enzyme gene with type 2 diabetes in the Korean population.** *Diabetes Res Clin Pract* 2008, **79:**284-290.
76. Lewis JP, Palmer ND, Hicks PJ, Sale MM, Langefeld CD, Freedman BI, Divers J, Bowden DW: **Association analysis in african americans of European-derived type 2 diabetes single nucleotide polymorphisms from whole-genome association studies.** *Diabetes* 2008, **57:**2220-2225.
77. Ng MC, Tam CH, Lam VK, So WY, Ma RC, Chan JC: **Replication and identification of novel variants at TCF7L2 associated with type 2 diabetes in Hong Kong Chinese.** *J Clin Endocrinol Metab* 2007, **92:**3733-3737.
78. Palmer ND, Lehtinen AB, Langefeld CD, Campbell JK, Haffner SM, Norris JM, Bergman RN, Goodarzi MO, Rotter JI, Bowden DW: **Association of TCF7L2 gene polymorphisms with reduced acute insulin response in Hispanic Americans.** *J Clin Endocrinol Metab* 2008, **93:**304-309.
79. Miyake K, Horikawa Y, Hara K, Yasuda K, Osawa H, Furuta H, Hirota Y, Yamagata K, Hinokio Y, Oka Y, Iwasaki N, Iwamoto Y, Yamada Y, Seino Y, Maegawa H, Kashiwagi A, Yamamoto K, Tokunaga K, Takeda J, Makino H, Nanjo K, Kadowaki T, Kasuga M: **Association of TCF7L2 polymorphisms with susceptibility to type 2 diabetes in 4,087 Japanese subjects.** *J Hum Genet* 2008, **53:**174-180.
80. Humphries SE, Gable D, Cooper JA, Ireland H, Stephens JW, Hurel SJ, Li KW, Palmen J, Miller MA, Cappuccio FP, Elkeles R, Godsland I, Miller GJ, Talmud PJ: **Common variants in the TCF7L2 gene and predisposition to type 2 diabetes in UK European Whites, Indian Asians and Afro-Caribbean men and women.** *J Mol Med* 2006, **84:**1005-1014.
81. Alsmadi O, Al-Rubeaan K, Wakil SM, Imtiaz F, Mohamed G, Al-Saud H, Al-Saud NA, Aldaghri N, Mohammad S, Meyer BF: **Genetic study of Saudi diabetes (GSSD): significant association of the KCNJ11 E23K polymorphism with type 2 diabetes.** *Diabetes Metab Res Rev* 2008, **24:**137-140.
82. Cejkova P, Novota P, Cerna M, Kolostova K, Novakova D, Kucera P, Novak J, Andel M, Weber P, Zdarsky E: **KCNJ11 E23K polymorphism and diabetes mellitus with adult onset in Czech patients.** *Folia Biol (Praha)* 2007, **53:**173-175.
83. Horikoshi M, Hara K, Ito C, Shojima N, Nagai R, Ueki K, Froguel P, Kadowaki T: **Variations in the HHEX gene are associated with increased risk of type 2 diabetes in the Japanese population.** *Diabetologia* 2007, **50:**2461-2466.
84. Damcott CM, Hoppman N, Ott SH, Reinhart LJ, Wang J, Pollin TI, O'Connell JR, Mitchell BD, Shuldiner AR: **Polymorphisms in both promoters of hepatocyte nuclear factor 4-alpha are associated with type 2 diabetes in the Amish.** *Diabetes* 2004, **53:**3337-3341.
85. Barroso I, Luan J, Wheeler E, Whittaker P, Wasson J, Zeggini E, Weedon MN, Hunt S, Venkatesh R, Frayling TM, Delgado M, Neuman RJ, Zhao J, Sherva R, Glaser B, Walker M, Hitman G, McCarthy MI, Hattersley AT, Permutt MA, Wareham NJ, Deloukas P: **Population-specific risk of type 2 diabetes (T2D) conferred by HNF4A P2 promoter variants: a lesson for replication studies.** *Diabetes* 2008, **57:**3161-3165.
86. Ek J, Rose CS, Jensen DP, Glumer C, Borch-Johnsen K, Jorgensen T, Pedersen O, Hansen T: **The functional Thr130Ile and Val255Met polymorphisms of the hepatocyte nuclear factor-4alpha (HNF4A): gene associations with type 2 diabetes or altered beta-cell function among Danes.** *J Clin Endocrinol Metab* 2005, **90:**3054-3059.
87. Beas-Zarate C, Morales-Villagran A, Tapia-Arizmendi G, Feria-Velasco A: **Effect of 3-acetylpyridine on serotonin uptake and release from rat cerebellar slices.** *Eur J Pharmacol* 1991, **198:**7-14.
88. Lehman DM, Richardson DK, Jenkinson CP, Hunt KJ, Dyer TD, Leach RJ, Arya R, Abboud HE, Blangero J, Duggirala R, Stern MP: **P2 promoter variants of the hepatocyte nuclear factor 4alpha gene are associated with type 2 diabetes in Mexican Americans.** *Diabetes* 2007, **56:**513-517.
89. Johansson S, Raeder H, Eide SA, Midthjell K, Hveem K, Sovik O, Molven A, Njolstad PR: **Studies in 3,523 Norwegians and meta-analysis in 11,571 subjects indicate that variants in the hepatocyte nuclear factor 4 alpha (HNF4A) P2 region are associated with type 2 diabetes in Scandinavians.** *Diabetes* 2007, **56:**3112-3117.
90. Weedon MN, Owen KR, Shields B, Hitman G, Walker M, McCarthy MI, Love-Gregory LD, Permutt MA, Hattersley AT, Frayling TM: **Common variants of the hepatocyte nuclear factor-4alpha P2 promoter are associated with type 2 diabetes in the U.K. population.** *Diabetes* 2004, **53:**3002-3006.