Published in final edited form as:

Int J Comput Biol Drug Des. 2014 ; 7(0): 183–194. doi:10.1504/IJCBDD.2014.061656.

Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes

Kai Wang,

Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Tennessee 37996, USA

Charles A. Phillips,

Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Tennessee 37996, USA

Gary L. Rogers,

National Institute for Computational Sciences, Oak Ridge, Tennessee 37831, USA

Fredrik Barrenas,

The Center for Individualized Medication, Linköping University Hospital, 58185, Linköping, Sweden

Mikael Benson, and

The Center for Individualized Medication, Linköping University Hospital, 58185, Linköping, Sweden

Michael A. Langston*

Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Tennessee 37996, USA

Kai Wang: kwang11@utk.edu; Charles A. Phillips: cphill25@utk.edu; Gary L. Rogers: grogers@eecs.utk.edu; Fredrik Barrenas: fredrik.barrenas@liu.se; Mikael Benson: mikael.benson@liu.se; Michael A. Langston: langston@eecs.utk.edu

Abstract

Differential expression has been a standard tool for analysing case-control transcriptomic data since the advent of microarray technology. It has proved invaluable in characterising the molecular mechanisms of disease. Nevertheless, the expression profile of a gene across samples can be perturbed in ways that leave the expression level unaltered, while a biological effect is nonetheless present. This paper describes and analyses differential Shannon entropy and differential coefficient of variation, two alternate techniques for identifying genes of interest. Ontological analysis across 16 human disease datasets demonstrates that these alternatives are effective at identifying disease-related genes not found by mere differential expression alone.

Copyright © 200x Inderscience Enterprises Ltd.

*Corresponding author.

This paper is a revised and expanded version of a paper entitled 'Differential Shannon entropy and differential coefficient of variation: alternatives and augmentations to differential expression in the search for disease-related genes' presented at the *International Conference on Intelligent Biology and Medicine (ICIBM 2013)*, Nashville, TN, USA, 11–13 August, 2013.

Because the two alternate techniques are based on somewhat different mathematical formulations, they tend to produce somewhat different gene lists. Moreover, each may pinpoint genes completely overlooked by the other. Thus, measures of entropy and variation can be used to replace or better yet augment standard differential expression computations.

Keywords

coefficient of variation; differential entropy; differential expression; gene ontology; ACOXL; ERBB4; ESR1; IRF4; disease-related gene

1 Introduction

One of the key goals of high-throughput transcriptomic experiments and data analysis is to identify genes relevant to disease, treatment or other factors. Such a gene is often viewed as one whose expression is altered in the presence of some relevant stimulus. Hence, mRNA microarray experiments are frequently designed with a case (treatment, stimulus) population and a control (unaffected) population. Differential expression has long been the standard technique to identify up-regulated or down-regulated genes in such experiments (Huber et al., 2002; Leung and Cavalieri, 2003; Smyth, 2004). It continues to be applied as new technologies such as RNA-Seq enter the picture (Anders and Huber, 2010; Oshlack et al., 2010). Although differential expression has proven to be an enormously valuable aid, it is limited in that it detects only altered *levels* of expression. Changes to the expression profile between case and control with little or no corresponding change in mean expression level will often remain undetected.

We discuss and analyse two variability-based techniques to identify genes of interest in case-control data. One is based on the notion of normalised Shannon entropy. The other harnesses the coefficient of variation. Both entropy and variation focus on expression variability on a normalised scale, while differential expression is based instead on simple mean differences. Over a collection of various biological datasets, we demonstrate that entropy and variation are able to identify genes relevant to disease but overlooked by differential expression.

Shannon entropy and coefficient of variation are well known in a great many application domains, from theoretical physics to computational chemistry to materials science. They have been applied in bioinformatics as well, most notably in statistical genetics and molecular biology. Entropy is derived from information theory (Shannon, 1948). It has been used, for example, to track gradual expression changes in cancer growth (Berretta and Moscato, 2010), as a measure of genetic diversity at levels ranging from gene expression to landscapes (Sherwin, 2010), as an estimate of the structural diversity of ecological species classifiers (Masisi et al., 2008), as a measure of the robustness of gene regulatory networks (Chen and Li, 2010), to accelerate feature elimination when classifying microarray expression data (Furlanello et al., 2003), and as a pre-processing filter on microarray expression data (Kohane et al., 2003). Coefficient of variation is a standard statistical measure. It has, for example, been used to assess variability of quantitative assays (Reed et al., 2002), as a predictor of risk sensitivity in animals (Weber et al., 2004), to analyse

synaptic plasticity (Faber and Korn, 1991), and to compare diversity among workforces (Bedeian and Mossholder, 2000). To our knowledge, neither technique has been applied to analyse transcriptomic data in the differential context in which we shall employ them here.

2 Materials and methods

We tested two differential metrics, one based on normalised Shannon entropy (SE) and the other on coefficient of variation (CV). We used publicly available mRNA gene expression data, and compared these results to those obtained with differential expression (DE). Both SE and CV measure the variability or dispersion of data in a normalised fashion. They are computed as follows. Let $x_1, x_2, \dots, x_n$ denote a list of n numbers, and let $x = \sum_{i=1}^n x_i$ signify their sum. Then the normalised Shannon entropy of this list is defined as:

$$SE = \frac{-\sum_{i=1}^n \frac{x_i}{x} \log_2 \frac{x_i}{x}}{\log_2 n}$$

and the coefficient of variation is defined as:

$$CV = \frac{\sigma}{|\bar{x}|}$$

where $\bar{x} = x/n$ is the sample mean and $\sigma = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1)}$ is the sample standard deviation. In the context of gene expression data, this list of numbers represents the expression levels of a particular probe set over different biological samples.

SE and CV are calculated separately for case and control. The case-control differences produce values we term differential Shannon entropy (DSE) and differential coefficient of variation (DCV). We note in passing that the phrase ‘differential entropy’ has been used previously in information theory to mean a continuous extension from the discrete version (Cover and Thomas, 2006), in a sense similar to differential calculus. That is not how we employ the term here, of course. There are many ways to calculate DE. For example, one can use a *t*-test with Bonferroni correction. Methodological differences tend to be minimal, however, and therefore for our purposes we simply took the magnitude of the difference between the mean expression level for case and the mean expression level for control. We applied our three metrics (DE, DSE and DCV) to 15 publicly available disease datasets obtained from the Gene Expression Omnibus (GEO) (Edgar et al., 2002), and to one in-house allergic rhinitis dataset. We chose case-control data in a form easy to analyse, striving for a reasonable variety of complex human diseases. All data are from mRNA microarray experiments. The GEO series identifiers are shown in Table 1. The number of probe sets ranges from roughly 12,000 to 61,000. The number of samples in case and control ranges from 4 to 63.

For the 100 probe sets with the highest differential values for each of the three techniques, we performed Gene Ontology (GO) enrichment analysis with the online hypergeometric test provided by DAVID Bioinformatics Resources 6.7 (Huang et al., 2009a, 2009b). We used the lowest p -value among all enriched categories as an estimate of the relative efficacy of each technique. The choice of 100 as a cut-off is arbitrary, of course. But for the purpose of demonstrating efficacy of the new techniques, any reasonable cut-off that allows analysis of probe sets in the top range of each metric suffices.

To help corroborate the potential usefulness of the new metrics, for each disease we obtained genes harbouring disease-associated single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) (Hindorf et al., 2012). Such studies avoid the discovery bias that may result from hypothesis-based genomic studies. Since the purpose of this analysis was to assess enrichment, not carry out a genetic study, we avoided setting cut-offs too stringently. We therefore included all SNPs annotated as intragenic or located less than 10kb upstream of the gene start; these SNPs were identified using a detection p -value of 10^{-5} or lower. We assessed the number of SNP-harboring genes among the top 100 genes identified by DE, DSE, and DCV using a one-tailed Fisher's exact test.

3 Results and discussion

Our results are centred on answering two questions. First, to what extent are DE, DSE and DCV surrogates for each other? In other words, do they tend to identify the same genes? Second, what is the relative effectiveness of each technique? Does each independently select genes relevant to the disease under study?

Figure 1 addresses the first question. Venn diagrams for each of the 16 diseases depict the overlap between the top 100 genes chosen by each technique. Of the genes found by the two new metrics, at most 36 of 100 would have been found by differential expression (in the Psoriasis dataset). Most of the diseases exhibit far fewer genes in common, 11 of 16 diseases having six or fewer genes shared between DSE-DCV and DE. Moreover, it is not the case that genes found by DSE and DCV are highly ranked but just barely missed by DE. See Table 2. DE thus seems to be an extremely weak predictor for genes found by DSE and DCV. As one might expect, however, the two variation-based approaches tend to produce more similar results. DSE and DCV show strong overlap. More than half their top 100 genes are shared in all but two diseases (type 2 diabetes and pancreatic tumour); more than two-thirds are shared in 12 of 16 diseases.

We address the second question by examining first the GO enrichments of the top 100 genes produced by each method, then the enrichment p -value by rank for each method. Figure 2 shows the lowest p -value among all GO categories for each disease. Differential expression tends to have the most significant p -values, although in five cases it is surpassed by DSE and/or DCV (asthma, colorectal cancer, lung adenocarcinoma, obesity, and ulcerative colitis). Figure 3 plots the p -value of the most enriched GO category by rank for the allergic rhinitis dataset. We selected sets of 100 genes at different rank positions for each method and performed a GO enrichment test for each set. In all three methods, the top sets of 100

genes had markedly lower p -value than all other sets of 100 genes. And all three methods show notably decreased functional enrichment after the first few hundred genes. It therefore appears that each metric is independently finding disease-relevant genes.

We next examined the enriched GO categories for the three different methods to check for relevance to specific diseases. We highlight three examples to illustrate the effectiveness of the variation-based techniques. In the asthma dataset, the most enriched GO category is 'secreted' for both DSE ($p = 3.37\text{E-}11$) and DCV ($p = 2.02\text{E-}09$). Yet the category is not enriched at all in the top 100 DE genes. In the Crohn's disease dataset, the category 'response to wounding' is the second most enriched for both DSE ($p = 3.43\text{E-}7$) and DCV ($p = 4.15\text{E-}5$), but is not enriched in the top 100 DE genes. And in the lung adenocarcinoma dataset, the category 'extracellular region' is highly enriched in the top 100 genes for all three metrics, DE ($p = 2.28\text{E-}8$), DSE ($p = 2.81\text{E-}11$) and DCV ($p = 9.36\text{E-}11$). Yet the latter two methods have no overlap at all with DE for this disease (see Figure 1).

Results of SNP enrichment identified by GWAS are shown in Table 3. Briefly, the top 100 genes identified by DE did not include a significant number of SNP-harboring genes for any disease. In contrast, the top 100 genes identified by DSE contain a significant number of SNP-harboring genes for breast cancer. Also, the top 100 genes identified by DCV include a significant number of SNP-harboring genes for chronic lymphocytic leukaemia. This suggests that DSE and DCV are capable of identifying some disease-related SNPs that DE would miss. In breast cancer, the SNP-harboring genes identified by DSE included ESR1 and ERBB4. The estrogen receptor ESR1 is a nuclear hormone receptor whose expression is a strong predictor for treatment response in breast cancer (Jensen and Jordan, 2003). However, strong expression of ERBB4 has been observed to have adverse (Bieche et al., 2003; Lodge et al., 2003) or beneficial effect (Pawlowski et al., 2000; Suo et al., 2002; Witton et al., 2003), which could be explained by different splice variants, or interactions with genetic variants in estrogen receptor genes (Junttila et al., 2005). The SNP-harboring genes identified by DCV in chronic lymphocytic leukaemia included ACOXL and IRF4. While the disease-causing mechanisms of these genes in chronic lymphocytic leukaemia are not well understood, the disease-associated genetic polymorphisms in IRF4 have been associated with differential expression (Havelange et al., 2011). Yet in this analysis, it was identified by DCV, not DE. The fact that these genes were detected by DSE and DCV, but not DE, suggests the usefulness of DSE and DCV in detecting

- genes whose expression is perturbed by disease in more complex ways
- combinations of genes with strongly interacting expression.

4 Conclusion

We have studied a pair of alternate techniques for identifying genes of interest in case-control microarray experiments: differential normalised Shannon entropy and differential coefficient of variation. Analysis of 16 human disease datasets has demonstrated that these techniques can identify genes not found by differential expression, and moreover these genes are likely to be disease-relevant based on GO enrichment. Thus, these two techniques can serve as viable alternatives or complements to standard differential expression in an

array analysis pipeline. While both techniques are measures of variability, each identifies a sufficient number of genes not found by the other to suggest that they are better used together, and that they complement not only differential expression but also each other.

Acknowledgments

This work has been supported in part by NIH awards R01-AA-018776 and 3P20MD000516-07S1, and by European Community Seventh Framework Programme grant 223367.

References

- Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol.* 2010; 11:R106. [PubMed: 20979621]
- Bedeian AG, Mossholder KW. On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods.* 2000; 3:285–297.
- Berretta R, Moscato P. Cancer biomarker discovery: the entropic hallmark. *PLoS One.* 2010; 5:e12262. [PubMed: 20805891]
- Bieche I, Onody P, Tozlu S, Driouch K, Vidaud M, Lidereau R. Prognostic value of ERBB family mRNA expression in breast carcinomas. *Int J Cancer.* 2003; 106:758–765. [PubMed: 12866037]
- Chen BS, Li CW. On the interplay between entropy and robustness of gene regulatory networks. *Entropy.* 2010; 12:1071–1101.
- Cover, TM.; Thomas, JA. *Elements of Information Theory.* Wiley-Interscience; Hoboken, NJ: 2006. p. xxiii. 748
- Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002; 30:207–210. [PubMed: 11752295]
- Faber DS, Korn H. Applicability of the coefficient of variation method for analyzing synaptic plasticity. *Biophys J.* 1991; 60:1288–1294. [PubMed: 1684726]
- Furlanello C, Serafini M, Merler S, Jurman G. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics.* 2003; 4:54. [PubMed: 14604446]
- Havelange V, Pekarsky Y, Nakamura T, Palamarchuk A, Alder H, Rassenti L, Kipps T, Croce CM. IRF4 mutations in chronic lymphocytic leukemia. *Blood.* 2011; 118:2827–2829. [PubMed: 21791429]
- Hindorf LA, MacArthur J, Wise A, Junkins H, Hall P, Klemm A, Manolio T. *A Catalog of Published Genome-Wide Association Studies.* 2012
- Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 2009a; 37:1–13. [PubMed: 19033363]
- Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009b; 4:44–57. [PubMed: 19131956]
- Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.* 2002; 18(Suppl. 1):S96–104. [PubMed: 12169536]
- Jensen EV, Jordan VC. The estrogen receptor: a model for molecular medicine. *Clin Cancer Res.* 2003; 9:1980–1989. [PubMed: 12796359]
- Junttila TT, Sundvall M, Lundin M, Lundin J, Tanner M, Harkonen P, Joensuu H, Isola J, Elenius K. Cleavable ErbB4 isoform in estrogen receptor-regulated growth of breast cancer cells. *Cancer Res.* 2005; 65:1384–1393. [PubMed: 15735025]
- Kohane, IS.; Kho, AT.; Butte, AJ. *Microarrays for an Integrative Genomics.* MIT Press; Cambridge, Mass: 2003. p. xviii. 306
- Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 2003; 19:649–659. [PubMed: 14585617]

- Lodge AJ, Anderson JJ, Gullick WJ, Haugk B, Leonard RC, Angus B. Type 1 growth factor receptor expression in node positive breast cancer: adverse prognostic significance of c-erbB-4. *J Clin Pathol.* 2003; 56:300–304. [PubMed: 12663644]
- Masisi, L.; Nelwamondo, F.; Marwala, T. The Use of Entropy to Measure Structural Diversity. *Stará Lesná, Slovakia*: 2008 Nov. p. 27–29.2008
- Oshlack A, Robinson MD, Young MD. From RNA-seq reads to differential expression results. *Genome Biol.* 2010; 11:220. [PubMed: 21176179]
- Pawlowski V, Revillion F, Hebbard M, Hornez L, Peyrat JP. Prognostic value of the type I growth factor receptors in a large series of human primary breast cancers quantified with a real-time reverse transcription-polymerase chain reaction assay. *Clin Cancer Res.* 2000; 6:4217–4225. [PubMed: 11106235]
- Reed GF, Lynn F, Meade BD. Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Diagnostic Laboratory Immunology.* 2002; 9:1235–1239. [PubMed: 12414755]
- Shannon CE. A mathematical theory of communication. *The Bell System Technical Journal.* 1948; 27:45.
- Sherwin WB. Entropy and information approaches to genetic diversity and its expression: genomic geography. *Entropy.* 2010; 12:1765–1798.
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004; 3(Article 3) <http://www.ncbi.nlm.nih.gov/pubmed/16646809>.
- Suo Z, Risberg B, Kallsson MG, Willman K, Tierens A, Skovlund E, Nesland JM. EGFR family expression in breast carcinomas. c-erbB-2 and c-erbB-4 receptors have different effects on survival. *J Pathol.* 2002; 196:17–25. [PubMed: 11748637]
- Weber EU, Shafir S, Blais AR. Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychol Rev.* 2004; 111:430–445. [PubMed: 15065916]
- Witton CJ, Reeves JR, Going JJ, Cooke TG, Bartlett JM. Expression of the HER1–4 family of receptor tyrosine kinases in breast cancer. *J Pathol.* 2003; 200:290–297. [PubMed: 12845624]

Biography

Biographical notes: Kai Wang is a PhD student in the Department of Electrical Engineering and Computer Science at the University of Tennessee. Currently he is doing research in graph theory and graph algorithms with applications in computational biology.

Charles A. Phillips is a Ph.D. student in the Department of Electrical Engineering and Computer Science at the University of Tennessee. His current research includes the graph theory and graph algorithms and their applications to computational biology.

Gary L. Rogers received his PhD in Computer Science from the University of Tennessee in 2011. He is working at the National Institute for Computational Sciences. His areas of interest include out of core algorithm design, parallel file systems, and high performance computing with a focus on bioinformatic analysis.

Fredrik Barrenas received his Ph.D. in Bioinformatics from Linköping University, Sweden, in 2012. He is currently a post doc at the University of Washington. His main interest is using biological networks to understand disease processes, with the goal of identifying new treatment strategies, markers for individual medication and drug targets.

Mikael Benson received his PhD in Medicine from the University of Göteborg, Faculty of Medicine, Sweden in 2000. Currently he is a Professor of Pediatrics at Linköping

University, Sweden. His main research interest is to integrate genomics and bioinformatics with functional and clinical studies to develop methods to individualise medicine.

Michael A. Langston received his PhD in Computer Science from Texas A&M University in 1981. He currently serves as Professor of Electrical Engineering and Computer Science at the University of Tennessee. His research interests include combinatorial optimisation, computational biology, discrete mathematics, fixed-parameter tractability, graph theory and high-performance computation.

Figure 1.

Commonality of genes over three metrics. In the Venn diagrams, one for each disease, each circle represents the top 100 genes selected by one of the three metrics under study. In most cases, few of the top DE genes are also ranked near the top for DSE or DCV. On the other hand, DSE and DCV have a majority of top-ranked genes in common for all but two diseases (see online version for colours)

Figure 2. Comparison of gene enrichment over three metrics. The p -value of the most enriched GO category for each disease is shown for DE, DSE and DCV (see online version for colours)

Figure 3.

Enrichment by rank over three metrics. For the allergic rhinitis dataset, the lowest GO enrichment p -value is shown for sets of 100 genes at various points in the rankings. The top 100 genes are shown at point 0, genes ranked 101–200 at point 1, genes ranked 2001–2100 at point 20 and so on. All three metrics show a sharp drop-off in functional enrichment after the top few hundred genes, suggesting that each metric is indeed identifying relevant genes. Other diseases exhibit similar results (see online version for colours)

Table 1

The most enriched categories for each metric on each disease. For each of the 16 diseases, the most highly enriched GO category by the hypergeometric test for each of the three metrics is shown, along with its enrichment *p*-value

Dataset	DE			DSE			DCV		
	Category	p-value	Category	p-value	Category	p-value			
Allergic rhinitis	Signal	8.52E-16	Signal	6.02E-13	Signal	6.16E-13			
Asthma (GSE4302)	Signal peptide	1.23E-05	Secreted	3.37E-11	Secreted	2.02E-09			
Breast cancer (GSE10810)	Response to endogenous stimulus	7.84E-13	Secreted	1.45E-06	Signal	2.17E-06			
Chronic lymphocytic leukaemia (GSE8835)	Immunoglobulin C1-set	4.40E-21	Disulphide bond	2.42E-15	Disulphide bond	1.42E-14			
Colorectal cancer (GSE9348)	Acetylation	9.82E-07	Signal	2.58E-06	Signal	5.14E-08			
Crohn's disease (GSE6731)	SM00407:IGc1	7.61E-12	Defence response	6.97E-08	Micro-some	3.92E-05			
Lung adenocarcinoma (GSE7670)	Immunoglobulin; major histocompatibility complex, conserved site	6.17E-10	Secreted	7.89E-12	Signal	8.25E-11			
Multiple sclerosis (GSE21942)	Translational elongation	5.2E-26	Immune system process	8.9E-6	Erythrocyte differentiation	4.1E-4			
Obesity (GSE12050)	Antimicrobial	8.99E-07	Defence response	4.30E-09	Oxygen transport	6.52E-08			
Pancreatic tumour (GSE 16515)	Digestion	3.14E-16	Digestion	4.9E-9	Response to chemical stimulus	2.7E-6			
Parkinson's disease (GSE20141)	Membrane associated complex	1.88E-06	Organic acid: sodium symporter activity	0.00335	Alternative splicing	9.31E-04			
Prostate cancer (GSE6919)	Translational elongation;	8.61E-38	Muscle protein	1.87E-18	Contractile ber part	6.95E-17			
Psoriasis (GSE13355)	Secreted	1.73E-15	Secreted	4.09E-10	Secreted	4.87E-11			
Schizophrenia (GSE 17612)	Acetylated amino end	3.10E-11	Secreted	2.07E-07	Extracellular region	1.41E-07			
Type 2 diabetes (GSE20966)	Pancreas	3.67E-20	Signal	2.32E-08	Signal	1.14E-04			
Ulcerative colitis (GSE6731)	SM00407:IGc1	8.29E-14	Signal	3.66E-15	Inflammatory response	2.36E-16			

Table 2

The DE ranking of genes ranked in the top 100 by DSE and DCV, but not ranked in the top 100 by DE. From this we conclude that DE would have missed great numbers of disease-related genes identified by DSE and DCV

Dataset	Gene count	Minimum position	Maximum position	Average position	Standard deviation
Allergic rhinitis	48,803	102	37,051	3037	5294
Asthma	54,675	116	43,008	8907	7728
Breast cancer	18,382	151	18,144	5340	4749
Chronic lymphocytic leukaemia	22,283	229	22,110	7954	6494
Colorectal cancer	54,675	109	53,921	20,255	14,808
Crohn's disease	12,625	144	12,283	4262	3400
Lime adenocarcinoma	22,283	236	17,572	5636	4049
Multiple sclerosis	54,675	141	23,661	7153	6399
Obesity	44,290	888	38,157	8064	9161
Pancreatic tumour	54,613	416	27,179	8377	5937
Parkinson's disease	54,675	1213	53,411	12,611	11,622
Prostate cancer	12,625	129	12,613	4567	3733
Psoriasis	54,675	102	49,861	5303	8532
Schizophrenia	54,675	308	54,518	21,731	17,796
Type 2 diabetes	61,294	328	61,274	30,388	18,305
Ulcerative colitis	12,625	122	12,257	3393	2952

Table 3

The enrichment p -values of SNPs identified as being related to each disease by GWAS or genes upstream (within 10kb) of such SNPs. Enrichment in the top 100 genes as ranked by each metric is shown. DE does not show significant enrichment for any disease. DSE and DCV each show enrichment for several diseases, and significant enrichment (p -value < 0.05) for at least one disease. The implication is that DSE and DCV can sometimes discriminate disease-related genes when DE alone does not

Dataset	DE	DSE	DCV
Allergic rhinitis	1	1	1
Asthma	1	0.112	0.114
Breast cancer	1	0.005	1
Chronic lymphocytic leukaemia	1	1	0.002
Colorectal cancer	1	1	1
Crohn's disease	1	0.063	0.063
Lung adenocarcinoma	1	1	1
Multiple sclerosis	1	0.199	1
Obesity	1	1	1
Pancreatic tumour	1	1	1
Parkinson's disease	0.127	1	1
Prostate cancer	1	1	1
Psoriasis	1	0.085	0.080
Schizophrenia	1	1	1
Type 2 diabetes	1	1	1
Ulcerative colitis	1	1	1