A new sample where the gene expression for Gene 1 was higher than the gene expression for Gene 2 would be classified as Group 1. 1 and 2 form a classifier based on their relative levels of expression. than Gene 2, and the reverse is true in Group 2. In this case, Genes arrays, separated by the black line. These groups could represent how the classification scheme works, consider the simulated gene most consistently switches between two groups. To understand search among all pairs of genes, and look for genes whose ranking et al. development of clinical tests from microarray experiments. Identifying biologically interpretable, robust and cheap classifiers and normalization steps to be effective (Quackenbush, 2006). gene expression measurements. This type of classifier is both hard to interpret and depends critically on the platform, pre-processing and normalization steps to be effective (Quackenbush, 2006). Identifying biologically interpretable, robust and cheap classifiers based on small subsets of genes would greatly speed progress in the development of clinical tests from microarray experiments. Top scoring pairs (TSPs) are pairs of genes that accurately classify patients into clinically relevant groups based on their ranks (Geman et al., 2004; Marchionni et al., 2008; Paik et al., 2004; van’t Veer et al., 2002). But standard microarray classifiers are based on complicated functions of many gene expression measurements. This type of classifier is both hard to interpret and depends critically on the platform, pre-processing and normalization steps to be effective (Quackenbush, 2006). Identifying biologically interpretable, robust and cheap classifiers based on small subsets of genes would greatly speed progress in the development of clinical tests from microarray experiments. Top scoring pairs (TSPs) are pairs of genes whose relative rankings can be used to accurately classify individuals into one of two classes. TSPs have two main advantages over many standard classifiers used in gene expression studies: (i) a TSP is based on only two genes, which leads to easily interpretable and inexpensive diagnostic tests and (ii) TSP classifiers are based on gene rankings, so they are more robust to variation in technical factors or normalization than classifiers based on expression levels of individual genes. Here I describe the R package, tspair, which can be used to quickly identify and assess TSP classifiers for gene expression data. Availability: The R package tspair is freely available from Bioconductor: http://www.bioconductor.org Contact: jtleek@jhu.edu

1 INTRODUCTION
Classification of patients into disease groups or subtypes is the most direct way to translate microarray technology into a clinically useful tool (Quackenbush, 2006). A small number of tests based on microarrays have even been approved for clinical use, for example, for diagnosing breast cancer subtypes (Ma et al., 2004; Marchionni et al., 2008; Paik et al., 2004; van’t Veer et al., 2002). But standard microarray classifiers are based on complicated functions of many gene expression datasets, with tens of thousands of genes. The TSP can be calculated both in R package tspair. This early success suggests that it may be possible to identify TSP classifiers for other important diseases and quickly develop new inexpensive diagnostic tests.

2 THE tspair PACKAGE
Calculating the TSP for a gene expression dataset is relatively straightforward, but computationally intensive. I have developed an R package tspair that can rapidly calculate the TSP for typical gene expression datasets, with tens of thousands of genes. The TSP can be calculated both in R or with an external C function, which allows both for rapid calculation and flexible development of the tspair package. The tspair package includes functions for calculating the statistical significance of a TSP by permutation test, and is fully compatible with Bioconductor expression sets. The R package is freely available from the Bioconductor web site (www.bioconductor.org).

3 AN EXAMPLE SESSION
Here I present an example session on a simple simulated dataset included in the tspair package. I calculate the TSP, assess the strength of evidence for the classifier with a permutation test, plot the output and show how to predict outcomes for a new dataset. The main function in the tspair package is tspcalc(). This function accepts either (i) a gene expression matrix or an expression set and a group indicator vector, or (ii) an expression set object and a column number, indicating which column of the annotation data to use as the group indicator. The result is a tsp object which gives...
the accuracy of the TSP.

from the TSP; the better the black line separates the colors the better
very simple and can be easily calculated either by hand or using the

is calculated for each null permutation,

of the null hypothesis that no TSP classifier is better than random

Indicates the TSP classification. The black line is the line where expression

1204

Fig. 2. A TSP plot. A TSP plot for the simulated data example in the
tspair package. The colors indicate the true groups, and the black line
indicates the TSP classification. The black line is the line where expression
for ‘Gene5’ equals the expression for ‘Gene338’; the classification boundary
is not data-driven, it is set in advance.

The function tspsig() can be used to calculate the significance
of a TSP classifier by permutation as described in Geman et al.
(2004). The class labels are permuted, a new TSP is calculated
for each permutation, and the null scores are compared with the
observed TSP score to calculate a P-value. Since the maximum score
is calculated for each null permutation, tspsig() performs a test
of the null hypothesis that no TSP classifier is better than random chance.

The tsp object with: 1 TSPs

Pair: TSP Score Tie-Breaker Indices
TSP 1: 0.64 NA 5 338

Once a TSP has been calculated, the tspplot() function can be
used to visualize the classifier. The resulting TSP figure (Fig. 2) plots
the expression for the first gene in the pair versus the expression
for the second gene in the pair. The true group difference is indicated by
the color of the points, and the score for the TSP classifier is shown
in the title of the plot. The black 45° line indicates the classification
from the TSP; the better the black line separates the colors the better
the accuracy of the TSP.

A major advantage of the TSP approach is that predictions are
very simple and can be easily calculated either by hand or using the
built-in functionality of the tspair package. In this example, the
expression value for ‘Gene5’ is greater than the expression value
for ‘Gene338’ much more often for the diseased patients. In a
new dataset, when the expression for ‘Gene5’ is greater than the
expression for ‘Gene338’ I predict that the patient will be diseased.
The tspair package can be used to predict the outcomes of
new samples based on new expression data. The new data can
take the form of a new expression matrix, or an expression set object.
The function predict() searches for the TSP gene names from
the original tspcalc() function call, and based on the row names
or featureNames of the new dataset identifies the genes to use
for prediction. If multiple TSPs are reported, the default is to predict
with the TSP achieving the top tie-breaking score (Tan et al., 2005),
but the user may also elect to use a different TSP for prediction.

In this example, the predict() function finds the genes with
labels ‘Gene5’ and ‘Gene338’ in the second dataset and calculates
the TSP predictions based on the values of these two genes. The new
data matrix need not be defined by a microarray, it could easily be
the result of RT-PCR or any other expression assay, imported into R
as a tab-delimited text file.

ACKNOWLEDGEMENTS

The author acknowledges the useful discussions with Giovanni
Parmigiani, Leslie Cope, Dan Naiman and Don Geman.

Funding: National Science Foundation (DMS034211); National
Institutes of Health (IUL1RR025005-01).

Conflict of Interest: none declared.

REFERENCES

Geman D. et al. (2004) Classifying gene expression profiles from pairwise mRNA
comparisons. Stat. Appl. Genet. Mol. Biol., 3.
Ma,X.J. et al. (2004) A two-gene expression ratio predicts clinical outcome in breast

cancer patients treated with tamoxifen. Cancer Cell, 6, 607–616.
Marchionini,L. et al. (2008) Systematic review: gene expression profiling assays in
early-stage breast cancer. Ann. Intern. Med., 148, 358–369.
Paik,S. et al. (2004) A multigene assay to predict recurrence of tamoxifen-treated,
node-negative breast cancer. N. Engl. J. Med., 350, 2817–2826.
Picou,N.D. et al. (2007) Highly accurate two-gene classifier for differentiating
gastrointestinal stromal tumors and leiomyosarcomas. Proc. Natl Acad. Sci. USA,
104, 3416-3419.
Quackenbush,J.(2006) Microarray analysis and tumor classification. N. Engl. J. Med.,
354, 2463–2472.
Tan,A.C. et al. (2005) Simple decision rules for classifying human cancers from gene
expression profiles. Bioinformatics, 21, 3906–3904.
van’t Veer,L.J. et al. (2002) Gene expression profiling predicts clinical outcome of
breast cancer. Nature, 415, 530–536.
Xu,L. et al. (2005) Robust prostate cancer marker genes emerge from direct integration
of inter-study microarray data. Bioinformatics, 21, 3905–3911.