Ranking, selecting, and prioritising genes with desirability functions

Stanley E. Lazic

2015-09-15

Introduction

Desirability functions were developed for multi-objective optimisation in the 1960s (Harrington, 1965; Derringer and Suich, 1980) and are used extensively in medicinal chemistry (Bickerton et al., 2012). They are co-opted here to select, rank, and prioritise genes, proteins, and metabolites. The `desirability` and `desire` packages on CRAN can be used for the traditional application of desirability functions. The `desiR` (pronounced "desire") package focuses on ranking, selection, and prioritisation.

Using this approach genes are characterised according to several criteria, such as fold-change and p-value for a particular comparison, whether a gene is in a particular pathway, or the correlation of a gene with a key or target gene. The importance of these criteria differ, they are measured on different scales, and have different units of measurement. Nevertheless, all of the variables can be mapped to a 0–1 desirability scale using several mathematical functions, where a value of one is maximally desirable and zero is unacceptable. A weighted average of the individual desirabilities is then calculated to provide an overall desirability. More specifically, applying the desirability approach involves three steps:

1. **Choose the relevant variables.** The choice of variables is determined by the biological question, along with technical considerations. For example, genes with a large fold-change are more desirable or interesting than those with smaller fold-changes.

2. **Map values for each variable onto a continuous 0–1 scale using the appropriate desirability function.** The choice of function depends on whether high, low, central, or extreme values are considered desirable. Categorical variables can also be included and examples of different functions are shown below.

3. **Calculate the overall desirability as a weighted combination of the individual desirabilities.** The results should usually be driven by one or two criteria (e.g. fold-change and p-value) that are given a high weight, with the other criteria acting as soft filters (e.g. average expression) with lower weights. Then, order genes by overall desirability. Genes are ranked along a single dimension and only one threshold at the final stage of the analysis is applied.

The desirability approach is both a generalisation and formalisation of currently used methods of selecting and prioritising genes. For example, genes that have an adjusted p-value of less than 0.05 and a fold-change greater than 2 may be deemed "differentially expressed". This approach has implicitly mapped the p-values and fold-change values to a binary 0/1 desirability scale and treated both criteria as equally important. The desirability approach formalises this procedure and expands the range of functions that are applied. Importantly, mapping continuous values to a binary scale is avoided because such dichotomisation is a poor use of the data (Cohen, 1983; Irwin and McClelland, 2003; Royston et al., 2006; Naggara et al., 2011; Kuss, 2013).

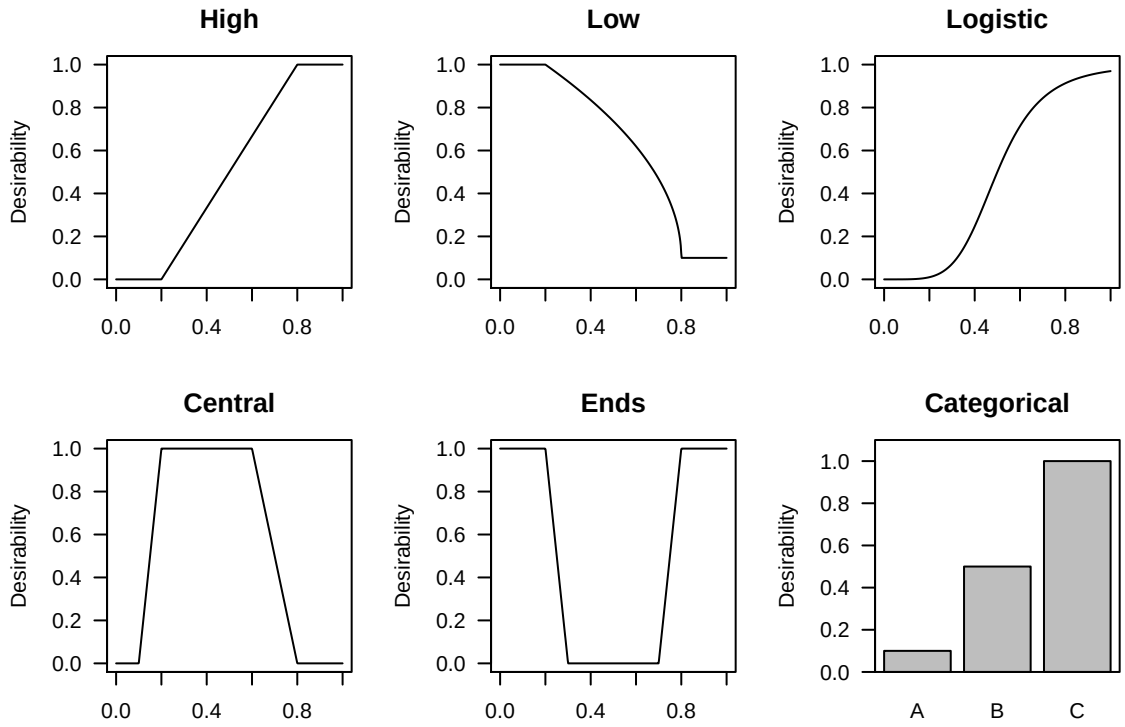
Examples of desirability functions

Desirability functions for continuous variables have four general shapes: high values are good; low values are good; middle, central, or target values are good; or extreme values are good. Examples of various functions are shown in the figure below. Categorical variables can be assigned a desirability value for each level or class. Most functions require two or more cut points to be specified, which determine where the function changes. The cut points can be based on relevant values (e.g. $p < 0.05$, fold-change > 2) or based on the shape of the distributions (e.g. top 20%). Most functions have a `scale` argument that controls how quickly the function changes between the maximum and minimum desirability. The default value is one. The final options are the maximum and minimum desirabilities, which default to one and zero, respectively. It may occasionally be beneficial to change these; for example, setting the minimum desirability to a small non-zero value such as 0.1. Otherwise, the overall desirability will be zero if any single desirability is zero, regardless of the values of the other functions.

```
library(desiR)
```

```
# generate values to convert to desirabilities
x <- seq(0,1, length.out=500)

# view mappings of x to different desirabilities
par(mfrow=c(2,3),
    mar=c(3,4.5,3,1),
    las=1)

# "high is good"
d1 <- d.high(x, cut1 = 0.2, cut2 = 0.8)
plot(d1 ~ x, ylab="Desirability", ylim=c(0,1), type="l",
     main="High")

# "low is good", with non-default scale and min.des arguments
d2 <- d.low(x, cut1 = 0.2, cut2 = 0.8, scale = 0.5, des.min = 0.1)
plot(d2 ~ x, ylab="Desirability", ylim=c(0,1), type="l",
     main="Low")

# "high is good" with a four parameter logistic function
d3 <- d.4pl(x, hill = 5, inflec = 0.5)
plot(d3 ~ x, ylab="Desirability", ylim=c(0,1), type="l",
     main="Logistic")

# "central or middle values are good"
d4 <- d.central(x, cut1 = 0.1, cut2 = 0.2, cut3 = 0.6, cut4 = 0.8)
plot(d4 ~ x, ylab="Desirability", ylim=c(0,1), type="l",
     main="Central")

# "extreme values are good"
d5 <- d.ends(x, cut1 = 0.2, cut2 = 0.3, cut3 = 0.7, cut4 = 0.8)
plot(d5 ~ x, ylab="Desirability", ylim=c(0,1), type="l",
     main="Ends")

# categorical: C is most desirable, followed by B
```

```
barplot(c(0.1, 0.5, 1), ylim=c(0,1.1),ylab="Desirability",
        main="Categorical", names.arg=c("A","B","C")); box()
```



Overall desirability

Individual desirabilities are combined with a weighted geometric mean: n desirabilities are multiplied together and taken to the root of n. A geometric mean is used because the overall desirability will be zero if any individual desirabilities are zero. The rationale is that if an individual desirability is zero, then the values of the other desirabilities—no matter how good—are irrelevant. For example, if a gene is not expressed, then the p-value for a comparison of interest is irrelevant. A weighted mean allows the different levels of importance of each criterion to be incorporated. In the equations below, n is the number of individual desirabilities (number of variables), d is the value of the individual desirabilities, and D is the overall desirability. The first equation is an unweighted overall desirability, the second, a weighted desirability, and the third, a weighted desirability written as the log of the desirabilities.

Geometric mean:

$$D = \left(\prod_{i=1}^{n} d_i \right)^{1/n}$$

Weighted geometric mean:

$$D = \left(\prod_{i=1}^{n} d_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i}$$

Weighted geometric mean (log-form):

$$D = \exp\left(\frac{\sum_{i=1}^{n} w_i \ln d_i}{\sum_{i=1}^{n} w_i}\right)$$

Microarray example

A microarray data set from Farmer et al. (2005; GEO: GDS1329) is used to illustrate the methods and options. The comparison of interest is between basal and luminal samples and the apocrine group is ignored. The data are the results from comparing the two groups with a linear model (from the `limma` package), along with a few additional variables. The data contains 1000 probesets randomly selected from the full data set.

```
# load the data (provided with the package)
data(farmer2005)

# look at first few rows
head(farmer2005)
```

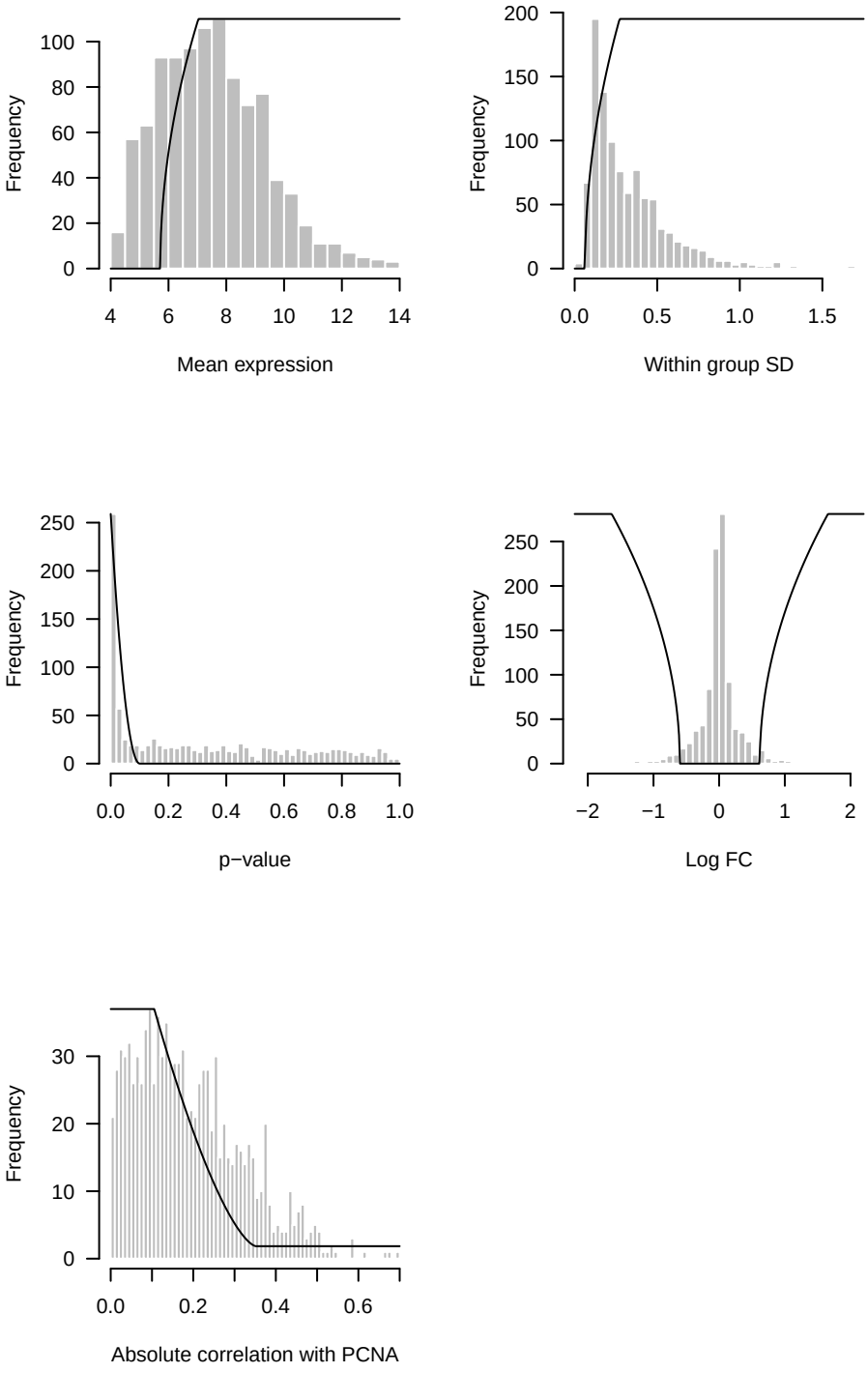
```
##      ProbeSet          GeneID       logFC  AveExpr      P.Value         SD
## 1   206373_at            ZIC1 -2.12762924 7.045365 1.039210e-07 1.49121932
## 2   208772_at ANKHD1-EIF4EBP3  0.32802634 9.097876 4.154784e-03 0.38040410
## 3 213344_s_at           H2AFX -0.03429972 6.981160 4.001303e-01 0.12449142
## 4   220810_at          CLCA3P  0.02606116 4.129490 1.532161e-01 0.04421642
## 5   204953_at          SNAP91 -0.14555023 5.267037 1.916564e-02 0.20066874
## 6   220590_at           ITFG2 -0.01703465 5.639281 6.255407e-01 0.10436762
##      PCNA.cor
## 1  0.49876807
## 2 -0.20420940
## 3 -0.14642003
## 4 -0.32368451
## 5 -0.01913128
## 6  0.16917274
```

Five variables are used to rank the genes and the distribution of these variables and the associated desirability functions are shown in the figure below. The first variable is the average expression across all samples (`AveExp`). If a gene is not expressed, then it is not of interest. A "high is good" function is therefore used. The second variable is the variability of expression across all samples (`SD`). Genes with constant expression (that is, low variability) are unlikely to be differentially expressed and are therefore uninteresting. A "high is good" function is therefore used. These two criteria are often used for non-specific gene filtering, which improves results by removing uninteresting genes (McClintick and Edenberg, 2006; Hackstadt and Hess, 2009; Bourgon et al., 2010). The third variable is the p-value (`P.Value`), with smaller values being more desirable, and so a "low is good" function is used. The fourth variable is the log2 fold-change (`LogFC`). Large effects in either direction are of interest and so an "extreme is good" function is used. The fold-change and p-value (either raw or adjusted) are often used to select genes that are differentially expressed. The final variable is the correlation of each gene with PCNA (proliferating cell nuclear antigen; `PCNA.cor`), which is a gene in this data set. PCNA is associated with cell proliferation, and for the purposes of this example, we are not interested in genes that are mainly markers of proliferation. Venet et al. (2011) showed that

randomly selected gene signatures are often good prognostic biomarkers, mainly because much of the breast cancer transcriptome is related to proliferation. Since the fraction of proliferating cells is known from biopsy samples, only genes unrelated to proliferation are of interest, that is, genes that provide *new* information. We therefore use a "low is good" function on the absolute correlation values to remove genes associated with proliferation.

The code below plots a histogram of each variable and the `des.line()` function shows how the values of the variables are mapped to the desirability scale. Density plots can be used in place of histograms if desired.

```
par(mfrow=c(3,2),
    las=1)
hist(farmer2005$AveExpr, breaks=30, col="grey", border="white", main="",
     xlab="Mean expression")
des.line(farmer2005$AveExpr, "d.high", des.args=c(cut1=5.75, cut2=7, scale=0.5))

hist(farmer2005$SD, breaks=30, col="grey", border="white", main="",
     xlab="Within group SD")
des.line(farmer2005$SD, "d.high", des.args=c(cut1=0.1, cut2=0.3, scale=0.5))

hist(farmer2005$P.Value, breaks=50, col="grey", border="white", main="",
     xlab="p-value")
des.line(farmer2005$P.Value, "d.low", des.args=c(cut1=0.0001, cut2=0.1, scale=2))

hist(farmer2005$logFC, breaks=50, col="grey", border="white", main="",
     xlab="Log FC")
des.line(farmer2005$logFC, "d.ends", des.args=c(cut1=log2(1/3), cut2=log2(1/1.5),
         cut3=log2(1.5), cut4=log2(3), scale=0.5))

hist(abs(farmer2005$PCNA.cor), breaks=50, col="grey", border="white", main="",
     xlab="Absolute correlation with PCNA")
des.line(farmer2005$P.Value, "d.low",  des.args=c(cut1=0.15, cut2=0.5,
         scale=1.5, des.min = 0.05))
```



Once suitable functions, cut points, and scales are selected, the final desirability values for each variable are calculated and saved as additional columns in the dataframe. The options for each desirability function below are identical to those in the figure above.

```
farmer2005$d1 <- d.high(farmer2005$AveExpr, cut1=5.75, cut2=7, scale=0.5)
farmer2005$d2 <- d.high(farmer2005$SD, cut1=0.1, cut2=0.3, scale=0.5)
```

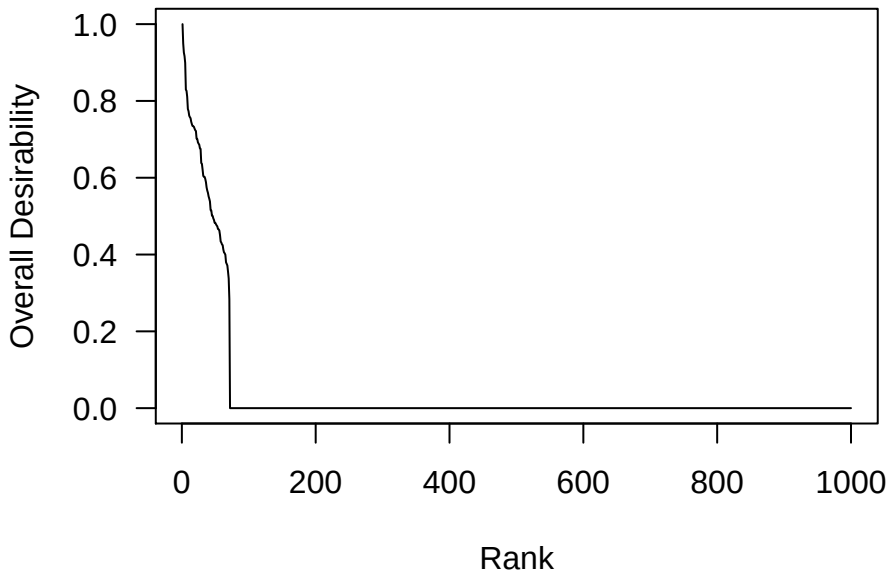
```
farmer2005$d3 <- d.low(farmer2005$P.Value, cut1=0.0001, cut2=0.1, scale=2)
farmer2005$d4 <- d.ends(farmer2005$logFC, cut1=log2(1/3), cut2=log2(1/1.5),
         cut3=log2(1.5), cut4=log2(3), scale=0.5)
farmer2005$d5 <- d.low(abs(farmer2005$PCNA.cor), cut1=0.15, cut2=0.5,
            scale=1.5, des.min = 0.05)
```

The next step is to combine the five individual desirabilities into an overall desirability and save the values to the dataframe. The weights for the individual desirabilities are chosen to reflect the importance of each variable. The actual values of the weights are unimportant, only the relative differences between them. For example, using the values 20 and 10, 2 and 1, or 1 and 0.5 are equivalent because the first weight is twice the value of the second, and therefore will contribute twice as much to the overall desirability. The values are scaled internally by the `d.overall()` function, so any set of values can be used, as long as they are positive. In this example, the p-value and fold-change are assigned a maximum weight of one so that the overall results are driven by these variables. The average expression and variability of expression are given weights of 0.1 (i.e. 10% as important as the p-value or fold-change). This means that the average expression and the variability in expression will have little influence on the overall desirability, unless they have a value of zero, in which case the overall desirability is zero. Finally, the correlation with PCNA is given a weight of 0.5, so it is half as important as the p-value and fold-change. Again, a value of zero will make the overall desirability zero, and intermediate values will have a modest influence on the ordering of the genes. If weights are not provided the default is to give all desirabilities an equal weight.

```
# order of weights needs to match the order of desirabilities
farmer2005$D <- d.overall(farmer2005$d1, farmer2005$d2, farmer2005$d3,
                          farmer2005$d4, farmer2005$d5,
                weights=c(0.1, 0.1, 1, 1, 0.5))
```

A plot of the ordered desirabilities shows that most values are zero, very few have a maximum value of one, and there are some with intermediate values that drop off quickly. We now have a ranked list of genes that can be taken forward for further experimentation.

```
par(las=1)
plot(rev(sort(farmer2005$D)), type="l", xlab="Rank", ylab="Overall Desirability")
```



We can also sort the dataframe by the overall desirability and look at the top ten genes. Putting the overall and individual desirabilities into the same dataframe as the raw data allows us to examine how the raw values map to the desirabilities and to understand why a known and expected gene did not have a high overall desirability. It is easy to see, for example, that the known gene received a very low desirability for the fold-change. If so, we might consider modifying the cut point or scale options.

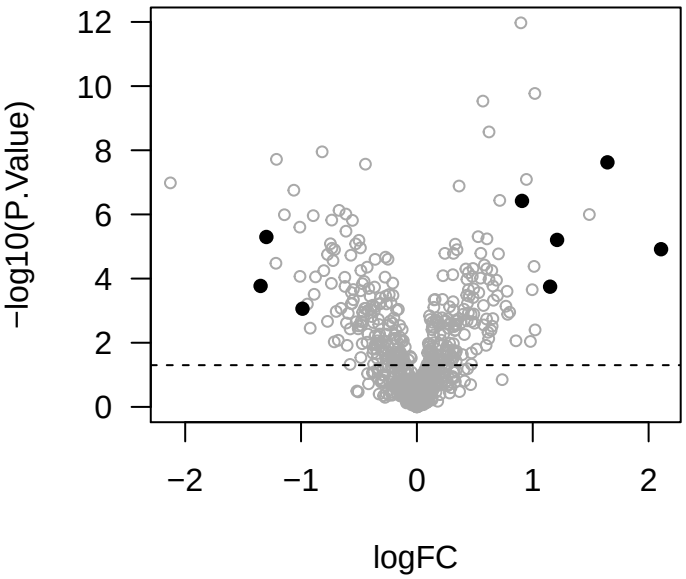
```
f2 <- farmer2005[order(farmer2005$D, decreasing=TRUE), ]
head(f2, 10)
```

```
##        ProbeSet  GeneID      logFC   AveExpr      P.Value        SD
## 936 216623_x_at    TOX3  2.1074670  7.164389 1.213388e-05 1.7101980
## 358   202270_at    GBP1 -1.3492447  9.128454 1.695831e-04 1.2389235
## 295 211712_s_at   ANXA9  1.6446407  7.474061 2.378950e-08 1.1147609
## 146 210519_s_at    NQO1  1.2100377  9.884189 6.233256e-06 0.9569021
## 734   201563_at    SORD  1.1496543  9.147931 1.787022e-04 1.0585756
## 609 221185_s_at    IQCG -1.2992278  7.937521 5.048276e-06 1.0197547
## 40  209631_s_at   GPR37 -0.9875021  6.257221 8.680605e-04 1.0063061
## 971   221732_at   CANT1  0.9071043 10.836180 3.787983e-07 0.6562165
## 859 205768_s_at SLC27A2  1.0200092  6.652243 3.976612e-03 1.1832011
## 158   202133_at   WWTR1 -0.8749355  9.579290 8.848779e-05 0.7753556
##        PCNA.cor        d1 d2        d3        d4        d5         D
## 936 -0.10107941 1.0000000  1 1.0000000 1.0000000 1.0000000 1.0000000
## 358  0.15155398 1.0000000  1 0.9986074 0.8742324 0.9936801 0.9498309
## 295 -0.23937398 1.0000000  1 1.0000000 1.0000000 0.6604473 0.9260548
## 146 -0.11169372 1.0000000  1 1.0000000 0.7906170 1.0000000 0.9166629
## 734 -0.07014069 1.0000000  1 0.9984250 0.7514598 1.0000000 0.8990537
## 609  0.28429472 1.0000000  1 1.0000000 0.8451422 0.5096346 0.8293267
## 40   0.07885628 0.6370062  1 0.9846825 0.6344601 1.0000000 0.8261894
## 971 -0.14974895 1.0000000  1 1.0000000 0.5675754 1.0000000 0.8107699
## 859 -0.22216881 0.8495847  1 0.9238960 0.6595807 0.7218820 0.7789294
## 158  0.18481265 1.0000000  1 1.0000000 0.5384914 0.8618480 0.7735338
```

It may be necessary to adjust the "tunable parameters" (functions, cut points, scales, and weights) in order to obtain better results. There is no "right" set of functions or values for the options, just like there is no correct adjusted p-value beyond which a gene is considered differentially expressed. The choice of cut points is arbitrary, but if too many variables have the maximum desirability, then there is little discrimination between the genes. For example, if 500 of the 1000 genes have a desirability of 1, then stricter values should be used to better distinguish between the top 500 genes.

The volcano plot below highlights the 8 genes with overall desirabilities greater then 0.8. Note how they are not necessarily those with the smallest p-values or most extreme fold-changes. This is even more apparent with the full data set.

```
# volcano plot
par(las=1)
plot(-log10(P.Value) ~ logFC, data=farmer2005, col="darkgrey", cex=0.75)

# overplot the top 8 genes
points(-log10(P.Value) ~ logFC, pch=16,
       data=farmer2005[farmer2005$D > 0.8, ])

# horizontal reference line at p = 0.05.
abline(h=-log10(0.05), lty=2)
```



What criteria can be used?

The desirability criteria can be grouped into (1) those that are external to the data set, and include information contained in bioinformatics databases as well as previous publications, and (2) those that are based on the data, and include p-values, fold-changes, and average expression levels. Categorical desirability functions are often suitable when incorporating information from databases, and continuous functions are better suited for experimental results. Several examples are provided below to stimulate ideas (most examples have not been tried), and the choice of criteria to include is determined by the biological question, along with any technical considerations. Any combination of

the examples can be used, and in most cases they would be in addition to the usual fold-change, p-value, average expression, and variance in expression desirabilities.

Database/external criteria:

- **In a gene set:** A gene set is a collection of genes that share certain attributes, such as belonging to a biochemical pathway or sharing the same molecular function. A collection of commonly used gene sets can be found in the Molecular Signature Data Base (MSigDB). It is possible to increase or decrease the desirability of genes according to whether they are in a relevant gene set. A categorical desirability function can be used, perhaps a value of 1 if a gene is in the set and 0.2 otherwise. The gene is not completely removed if it is not in the gene set, just reduced in importance. The gene ontology (GO) gene sets also have evidence codes that describe how the annotation of the gene was derived. One could give a higher desirability to annotations based on experimental evidence compared with computational predictions for example. Discovering new genes is less likely when previous information is used for ranking and selection, so incorporating external information will not be appropriate in all cases.
- **In a protein-protein interaction:** Some experiments may have one or more known proteins that are related to the disease or biological process under consideration. Differentially expressed genes that are in the same protein complex or interact directly with one of the key proteins can be given a desirability of 1, and genes that are one step away in the interaction network could be given a desirability of 0.8 for example. Other genes could be assigned a low desirability of 0.1.
- **In a list from previous publications:** Related publications often contain lists of differentially expressed genes, or those that were included in a gene signature because they were predictive of a clinical outcome. These genes are occasionally validated in subsequent experiments but may not be found in a database. It may be desirable therefore to give these genes a higher weight in the present experiment.
- **Sequence similarity with another species:** When using an animal model of a human disease, genes that are most similar to the human homolog can be prioritised. A continuous desirability function would be preferable.
- **Tissue specificity or expression:** If samples are from whole blood and the aim is to find biomarkers for a neurodegenerative disorder, then the expression level of genes in the target tissue (brain) can be prioritised. A continuous desirability function would be preferable and tissue specific expression information can be obtained from the Genotype-Tissue Expression project (GTex).
- **Secreted or cell surface protein:** For biomarker studies it may be preferable to prioritise genes whose protein product is secreted and therefore detectable in the blood or urine. This may be desirable because of the technologies available or preferred for a diagnostic assay.
- **Compounds available:** If the purpose of follow-up experiments is to validate the functional role of candidate genes with pharmacological methods, it may be beneficial to prioritise genes that are known targets of compounds. This information can be found in the Drug Gene Interaction Database (DGIdb).

Data-derived criteria:

- **Correlation with other genes:** The microarray study in the above example deprioritised genes that were correlated with a marker of cell division (PCNA). Other options are possible;

for example, if one is interested in discovering new genes or pathways it may be beneficial to deprioritise genes that are highly uncorrelated with one or more genes that are known to be relevant. It is likely that the highly correlated genes are part of the same pathway as the known genes.

- **Differentially expressed in multiple conditions:** A common aim is to find genes that are differentially expressed in several conditions compared with a common control. Genes are usually classified as being differentially expressed or not for each comparison and the results are often visualised with Venn diagrams. A problem with this approach is that the number of false negatives increases with the number of comparisons. For example, if the power to detect differential expression is 0.8 for three individual comparisons, there will only be about a 50% chance of correctly detecting this. Since the desirability functions use continuous mappings, genes just below the cut point are not misclassified.

- **Disease specificity**: Often the interest is in finding genes that are differentially expressed in one condition or group but not in another. For example, between healthy controls and Parkinson's disease patients but **not** between controls and Alzheimer's patients when the aim is to look for Parkinson-specific genes rather than genes associated with neurodegeneration or cell death in general. Therefore genes with high p-values and log fold-changes close to zero for the control–Alzheimer's comparison would be prioritised.

- **Consistency of expression:** In some cases it may be desirable to find genes that do not vary between groups or conditions. This is similar to the specificity example above, but with a slightly different focus. In the context of finding biomarkers, it would be useful to find genes that are not differentially expressed between males and females, do not have a circadian rhythm in expression, and are not affected by age or body mass index. Furthermore, the expression levels should be constant across different RNA extraction batches. In other words, genes that are robust to variations in biological and technological factors are considered desirable.

- **Based on parameters of linear models:** Many analyses use linear models from the 'limma' package (Smyth, 2005). The coefficients from complex models can be used as desirability criteria. For example, suppose four concentrations of a drug (including a no drug control condition) are treated as an ordered categorical factor and the interest is in finding genes that have a linear dose-response. The p-value and coefficient for the linear contrast will therefore be the main criteria, but many genes with a significant linear component will also have a strong quadratic (i.e. "U" or inverted "U" shapes) or cubic components. Such genes can be deprioritised with a "middle is good" desirability function on the higher order coefficients. More generally, specific patterns of expression can be selected.

Other uses for desirability functions

The methods described here naturally extend to other "-omics" data. In addition, desirability functions have been successfully used to rank compounds from high-content imaging screens, where the criteria are fluorescent, morphological, and texture readouts from cells (unpublished data). In addition, such methods have been used to combine diverse data across different experiments. Examples will be provided in future versions of the package.

Conclusion

The desirability approach to gene ranking and selection has several advantages. First, a diverse set of criteria can be numerically combined. Second, the importance of the different criteria can be taken into account. Third, distributions with unusual or non-Gaussian shapes and outliers are easily accommodated. Fourth, a continuous ranking of genes is returned, rather than just a list of genes that meet all criteria. Fifth, the reproducibility of research is improved because the decision criteria are captured by the functions and weights. Sixth, the criteria are explicit, and so can be shared with others, criticised, and modified as needed. Finally, the uncertainty in the values of the criteria are accounted for by avoiding unnecessary dichotomisation. Many criteria are measured with error, and classifying genes according to whether they are above or below a threshold is prone to misclassification error. For example, if a gene is truly expressed, but due to sampling error, has an average expression level just below the threshold, then it will be removed from further analyses. Since desirability functions use continuous mappings, genes near the threshold receive intermediate values rather than zero or one.

The main disadvantage of this approach is that the results are probabilistic—there are no p-values or confidence intervals associated with the overall desirability. The purpose of this approach however is not to declare that something new has been discovered, but to select and prioritise genes for further experimentation.

References

Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. *Nature Chemistry*. 2012 4(2):90-98.

Bourgon R, Gentleman R, Huber W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl Acad. Sci. USA*. 2010 107:9546-9551.

Cohen J. The cost of dichotomization. *Applied Psychological Measurement*. 1983 7(3):249–253.

Derringer G, Suich R. Simultaneous optimization of several response variables. *Journal of Quality Technology*. 1980 12(4):214-219.

Farmer P, Bonnefoi H, Becette V, Tubiana-Hulin M, Fumoleau P, Larsimont D, Macgrogan G, Bergh J, Cameron D, Goldstein D, Duss S, Nicoulaz AL, Brisken C, Fiche M, Delorenzi M, Iggo R. Identification of molecular apocrine breast tumours by microarray analysis. *Oncogene*. 2005 24(29):4660-4671.

Hackstadt AJ, Hess AM. Filtering for increased power for microarray data analysis. *BMC Bioinformatics*. 2009 10:11.

Harington, J. The desirability function. *Industrial Quality Control*. 1965 21:494-498.

Irwin JR, McClelland GH. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research*. 2003 40:366-371.

Kuss O. The danger of dichotomizing continuous variables: A visualization. *Teaching Statistics*. 2013 35(2):78-79.

McClintick JN, Edenberg HJ. Effects of filtering by Present call on analysis of microarray experiments. *BMC Bioinformatics*. 2006 7:49.

Naggara O, Raymond J, Guilbert F, Roy D, Weill A, Altman DG. Analysis by categorizing or dichotomizing continuous variables is inadvisable: an example from the natural history of unruptured aneurysms. *AJNR Am J Neuroradiol*. 2001 32(3):437-440.

Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*. 2006 23(1):127-141.

Smyth, GK. Limma: linear models for microarray data. In: 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York. 2005, 397-420.

Venet D, Dumont JE, Detours V. Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*. 2011 7(10):e1002240.