The basic procedure for generating local dependency networks is straightforward and consists of two steps. First, we use GeneRaVE to select genes that 'best' discriminate between subtypes present in the data. In the example described below, discrimination is between two groups of individuals—smokers and never smokers—and we identify a small subset of genes that discriminates between these two groups by using the sparse logistic regression. Second, we use these genes as 'seed' genes for building the local dependency networks.

To do so, the individual seed genes become the response variables in separate sparse linear regressions in which all other genes are potential explanatory variables. Again we use GeneRaVE, so that the resulting regression model is sparse and contains only a few other explanatory genes. After this first round of regressions, the network is defined by edges between the response (seed) genes and their selected regressor (explanatory or predictor) genes.

To expand the network beyond the first round, we simply use the explanatory genes from the first round as response variables in regressions using all other genes as potential regressors. Though it is possible to repeat this procedure several times, the resulting large network will be difficult to visualize and interpret. In practice, it is more useful to construct only local networks, which are typically obtained after two or three rounds. Note that dependency networks can be constructed around any genes of interest within a dataset. Furthermore, the R function in NetRaVE that is used for constructing networks is written so that the user can easily write his/her own method for generating conditional probability distributions using alternative regression methods. Further details about NetRaVE and dependency networks appear in the Supplementary Material, along with a larger version of Figure 1 and the R script used to generate it.
There are a few options for generating and visualizing dependency that information. The dependency network in Figure 1 shows other provides the researcher with a means of organizing and visualizing the network imposes an additional structure on those lists that CYP1B1 annotated with the cross-validated (light blue nodes) in the first round of regressions (depth 1) are the directed edges pointing towards them from their regressors MUC5AC at later depths [e.g. the directed edge between genes that appear at early depths can also be explanatory of genes and TFF3 R cross-validated example, analyzing this and other datasets, we have found many of the genes NetRaVE and in the Supplementary Material. In our experience in 12 of the 32 genes in Figure 1 appear in the top-97 list of Spira expressed genes obtained by, for example, modified DISCUSSION Fig. 1. Dependency network generated from the data of Spira et al. (2004) to a depth of three around the genes CEACAMS, ALDH3A1 and CYP1B1. See text for explanation of colours and annotations. Using NetRaVE and sparse linear regression, we can generate a dependency network as illustrated in Figure 1. The seed genes (CEACAMS, ALDH3A1 and CYP1B1) are the light red nodes, and the directed edges pointing towards them from their regressors (light blue nodes) in the first round of regressions (depth 1) are annotated with the cross-validated $R^2$ of the sparse regression. For example, ALDH3A1 has two regressors (PIR and AKR1B10), with a cross-validated $R^2$ of 0.78, whereas CEACAMS has five, including MUC5AC, UPK1B, SCGB1A1, CLDN10 and KRT13 ($R^2 = 0.73$). The regressors, or neighbours, of genes in depth 1 are coloured green, and, in turn, their neighbours are coloured purple. Note that genes that appear at early depths can also be explanatory of genes at later depths [e.g. the directed edge between ALDH3A1 (depth 0) and TFF3 (depth 2)].

4 DISCUSSION There are a few options for generating and visualizing dependency networks that are discussed in the documentation accompanying NetRaVE and in the Supplementary Material. In our experience in analyzing this and other datasets, we have found many of the genes in a local network are also found in top n lists of differentially expressed genes obtained by, for example, modified t-tests. Here, 12 of the 32 genes in Figure 1 appear in the top-97 list of Spira et al. (2004). For example, they found that ALDH3A1 (depth 0) and CYP1B1 (depth 0) were highly up-regulated in smokers. However, the network imposes an additional structure on those lists that provides the researcher with a means of organizing and visualizing that information. The dependency network in Figure 1 shows other genes that were not comparatively highly up- or down-regulated, yet that are related to the effects of cigarette smoking. For example, AKR1B10 is a potential diagnostic marker of non-small cell lung carcinoma (Penning, 2005), and together with other aldo–keto reductase genes (AKR1C1, AKR1C2 and AKR1C3) participates in the metabolism of xenobiotics; the up-regulation of PIR, which appears at depth 1, represents one mechanism by which cigarette smoke induces apoptosis in the airway epithelium (Gelbman et al., 2007); and many other such examples. In molecular toxicology research, it is well documented that the transfection factor aryl hydrocarbon receptor (AhR, also called the dioxin receptor) enhances gene transcription upon activation by polycyclic hydrocarbons such as those found in cigarette smoke (reviewed in Kitamura and Kasai, 2007). It is tempting to speculate that some gene linkages within the network in Figure 1 are due to coordinated gene expression in response to activation of AhR by cigarette smoke toxins, but the truth is inevitably more complex. Consider the effect of smoking on epithelial inflammation. Linkages between some genes such as MUC5AC (mucin super family member) and the CEACAM adhesion molecules could also be due to effects of neutrophils (Fischer and Vojnov, 2007). Whether these mechanisms are behind the expression patterns observed in the dataset by NetRaVE or not, the ability to identify linkages between groups of genes provides an alternative to unstructured data exploration. We emphasize here that, like all influence networks, local dependency networks should not be viewed as a rigorous means of uncovering true causal relationships among genes; instead, they should be viewed as a very useful tool for organizing complex information in a manner that is easy to visualize. The relationships contained in them can then be used, along with related information and background knowledge about the biological system being studied, to guide further experimentation.

Conflict of Interest: none declared.