Missing 2009.zip

Note: These tables include births by county of residence and ZIP code reported on the birth certificate. ZIP code data are not edited. If ZIP codes listed in a particular county table are not from that county, the births should be considered unknown or missing for ZIP code information. The DPS Missing and Unidentified Persons Online Bulletin is supported by the Texas DPS Missing Persons Clearinghouse and the Unidentified Persons and DNA unit. This site features individual searchable bulletins regarding missing persons, abductors, runaways, traveling companions, and unidentified persons discovered in Texas. Each bulletin includes photographs (if available), descriptors, and other information regarding the circumstances of the person's disappearance or discovery.

Click here for the current criteria that must be met in order to add a missing person or an unidentified person to this online bulletin. Call 312-666-0500 to speak to Deputy Chief Investigator Earl Briggs about matching one of these unidentified bodies to the identity of a missing person. Descriptions of cases can also be found at NAMUS.gov Epistatic miniarray profiling (E-MAPs) is a high-throughput approach capable of quantifying aggravating or alleviating genetic interactions between gene pairs. The datasets resulting from E-MAP experiments typically take the form of a symmetric pairwise matrix of interaction scores. These datasets have a significant number of missing values - up to 35% - that can reduce the effectiveness of some data analysis techniques and prevent the use of others. An effective method for imputing interactions would therefore increase the types of possible analysis, as well as increase the potential to identify novel functional interactions between gene pairs. Several methods have been developed to handle missing values in microarray data, but it is unclear how applicable these methods are to E-MAP data because of their pairwise nature and the significantly larger number of missing values. Here we evaluate four alternative imputation strategies, three local (Nearest neighbor-based) and one global (PCA-based), that have been modified to work with symmetric pairwise data. We identify different categories for the missing data based on their underlying cause, and show that values from the largest category can be imputed effectively. We compare local and global imputation approaches across a variety of distinct E-MAP datasets, showing that both are competitive and preferable to filling in with zeros. In addition we show that these methods are effective in an E-MAP from a different species, suggesting that pairwise imputation techniques will be increasingly useful as analogous epistasis mapping techniques are developed in different species. We show that strongly alleviating interactions are significantly more difficult to predict than strongly aggravating interactions. Finally we show that imputed interactions, generated using nearest neighbor methods, are enriched for annotations in the same manner as measured interactions. Therefore our method potentially expands the number of mapped epistatic interactions. In addition we make implementations of our algorithms available for use by other researchers. We address the problem of missing value imputation for E-MAPs, and suggest the use of symmetric nearest neighbor based approaches as they offer consistently accurate imputations across multiple datasets in a tractable manner. One common characteristic of E-MAPs is the high proportion of missing entries that they contain. Missing entries correspond to pairs of genes for whom interaction strengths could not be measured during the high-throughput process or those that were subsequently filtered due to unreliability. These missing values can reduce the effectiveness of some techniques, *e.g*. introducing instability in clustering [6], and prevent the use of others, *e.g*. matrix factorization techniques such as SVD and PCA. As each epistatic interaction implies a functional relationship between gene pairs, individual epistatic interactions themselves may provide valuable biological insight. Consequently there is an urgent need for an effective imputation technique. Although the problem of predicting genetic interactions is not new, to our knowledge the problem of imputing quantitative epistasis values in E-MAPs has not previously been evaluated. For E-MAP imputation the goal is to achieve a complete dataset by predicting quantitative scores for all interactions between gene pairs in a given set - including those that display no significant interaction. An illustrative example of an incomplete E-MAP (with missing values) and a corresponding completed E-MAP (with imputed values) is shown in Figure 2. **E-MAP before and after imputation**. A visual representation of a pairwise symmetric E-MAP interaction matrix. On the left-hand side is shown an original E-MAP (Chromosome Biology), where gray points indicate missing values. On the right-hand side is the corresponding complete matrix, with all missing entries replaced by imputed values. Järvinen *et al*[7] have applied a matrix approximation technique to a small scale (26 genes) E-MAP-like dataset, and have shown that gene pairs whose growth diverges significantly from the expectation can be identified without the need for measurements of single mutant growth rates. While similar matrix approximation techniques could perhaps be used to address the missing value problem, this was not addressed in their work. The problem of missing value imputation has been well studied for gene expression data. For instance, Troyanskaya and co-workers [12] compared two methods *K*-Nearest Neighbors (KNNImpute) and singular value decomposition (SVD). They recommended KNNImpute as the more robust and accurate method. Since then a number of techniques have been developed, generally falling into two broad categories: local methods, such as nearest neighbor-based techniques, and global methods, generally based on matrix decomposition such as SVD and PCA. In 2008 Brock *et al*[13] provided a comprehensive analysis of different techniques across a number of datasets. Notably, they found that the optimal imputation methods were all competitive with each other, and that the effectiveness of different techniques depended on the "complexity" of the dataset (where the complexity was taken to signify the difficulty with which the data can be reliably transformed to a lower-dimensional subspace). These authors demonstrated that local methods generally

performed better on datasets with higher complexity. E-MAP datasets are pairwise and symmetric - each missing value represents the interaction between two genes measured under a specific experimental condition, rather than the expression of a given gene in a given sample or at a given time point. We observe that there are three types of missing data in E-MAP experiments which may need to be considered separately for the purpose of imputation. Missing values in gene expression datasets are effectively treated as missing at random. This is not the case with E-MAPs where we observe three categories of missing value: *Other Interactions:* This category can be divided into two sub-categories. Firstly, those that correspond to a double mutant measuring the interaction between one essential and one non-essential gene. Secondly, those that correspond to a measurement of the interaction between two non-essential genes. These cases make up the majority of the missing values in an E-MAP and can be considered in the same way for imputation purposes. They are not missing systematically, as is the case with the other categories, and can be treated as missing at random. They occur due to problems in growing the necessary mutants, inconsistencies in the results of multiple experiments, or other problems with the experimental technique. In this paper we consider four general strategies for imputing missing values in real-valued data - three local methods (nearest neighbor-based) and one global method (BPCA) - and adapt these strategies to work with symmetric data such as E-MAPs. In our evaluations we consider five E-MAPs that have been recently published. These datasets differ in their size, the subset of genes that are studied, and the proportion of missing values that they contain. Four are from the budding yeast *Saccharomyces cerevisiae*, and one is from the fission yeast *Schizosaccharomyces pombe*. As noted previously, E-MAP interaction datasets are typically normalized so that a data value close to zero indicates the absence of any interaction between a pair of genes. Therefore a simple solution to the problem of missing values is to replace those entries with zeros. While this may appear to be a naïve approach, it has some justification: the expectation is that most genes do not interact, and therefore their interaction score is likely to be close to zero. We also observe that the mean of the non-missing entries in the five E-MAP datasets described previously is approximately zero. This approach serves as a baseline for our experimental evaluations in the next section. Alternative baseline approaches are discussed in the 'Additional file 2 - alternate methods.pdf ' *K*-Nearest Neighbors neighbors (KNN) imputation is a local strategy that uses genes with similar interaction profiles to impute missing values. Standard imputation algorithms based on KNN involve imputing values in feature-based asymmetric datasets. Our proposed approach is designed to handle symmetric data. For each missing interaction (i, j), we find the K nearest neighbor(s) for both gene i and gene j. We then find the values for the interaction of i with j's neighbors, and j with i's neighbors. These values are averaged to provide an imputed value for the missing entry (i, j). An illustration of this approach is shown in Figure 3. For E-MAP data we suggest the use of Pearson's correlation measure to calculate the similarity of gene profiles, as initial experiments indicated that Euclidean distance offered significantly worse performance (data not shown). Note that the effectiveness of this method is heavily dependent on the choice of value for the parameter K. Therefore in our experiments we assess the results for a variety of values of K.

Missing 2009.zip

21f597057a