EXPLORATION OF REACTANT-PRODUCT LIPID PAIRS IN MUTANT-WILD TYPE LIPIDOMICS EXPERIMENTS

Lianqing Zheng

Gary L. Gadbury

Jyoti Shah

Ruth Welti

See next page for additional authors

Recommended Citation

Zheng, Lianqing; Gadbury, Gary L.; Shah, Jyoti; and Welti, Ruth (2012). "EXPLORATION OF REACTANT-PRODUCT LIPID PAIRS IN MUTANT-WILD TYPE LIPIDOMICS EXPERIMENTS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1035

Author Information

Lianqing Zheng, Gary L. Gadbury, Jyoti Shah, and Ruth Welti

EXPLORATION OF REACTANT-PRODUCT LIPID PAIRS IN MUTANT-WILD TYPE LIPIDOMICS EXPERIMENTS

Lianqing Zheng[1], Gary L. Gadbury[1*], Jyoti Shah[2], Ruth Welti[3]
[1]Department of Statistics, Kansas State University, Manhattan, KS 66506,
[2]Department of Biological Sciences, University of North Texas, Denton, TX 76203
[3]Division of Biology, Kansas State University, Manhattan, KS 66506

*Corresponding Author:
Gary L. Gadbury, Department of Statistics
Kansas State University
Manhattan, KS 66506
phone: (785)532-0526, email: gadbury@ksu.edu

Abstract: High-throughput metabolite analysis is very important for biologists to identify the functions of genes. A mutation in a gene encoding an enzyme is expected to alter the level of the metabolites which serve as the enzyme's reactant(s) (also known as substrate) and product(s). To find the function of a mutated gene, metabolite data from a wild-type organism and a mutant are compared and candidate reactants and products are identified. The screening principle is that the concentration of reactants will be higher and the concentration of products will be lower in the mutant than in wild type. This is because the mutation reduces the reaction between the reactant and the product in the mutant organism. Based upon this principle, we suggest a method to screen metabolite pairs for candidate reactant-product pairs. Metrics are defined that quantify the effect of a mutation on each potential reaction, represented by a metabolite pair. For reactions catalyzed by well-characterized enzymes, one or more biologically functioning reactant-product pairs are known. Knowledge of the functional reactant-product pairs informs the development of the metrics. The goal is for ranking of the metrics for all possible pairs to reflect the likelihood that a particular metabolite pair is a functional reactant-product pair.

Key words: Lipid experiment; Pathway analysis; Reactant-product lipid pairs; Metabolome; Statistic distribution;

1. Introduction

The metabolome is the total collection of the set of small molecule metabolites (Oliver et al. 1998; Oliver 2002; Griffin and Vidal-Puig 2008; Dunn et al. 2005). The metabolome includes metabolic intermediates, hormones, and other products and intermediates of metabolism. Unlike the genome and the proteome whose elements are composed of similar building blocks, the metabolome is a group of dynamic molecules with varied structures. Biologists use metabolic profiling to get a "snapshot" of the composition of metabolites to understand biomolecular functions within organisms. Since metabolites are products of gene and protein function, it can be argued that they provide the most complete description of cellular function (Wu et al. 2005; Raamsdonk et al. 2001). Metabolic studies can be used to address the question of how a gene's

mutation affects phenotypes of the organism. Many biologists advocate metabolic profiling in a functional genomics study (Dixon et al. 2006).

One subset of the metabolome, the lipidome, plays an important role in the biochemical processes in the cell. Lipids are compounds of biological origin that are poorly soluble in water but are soluble in nonpolar solvents (Blei and Odian 2006). They include well-known compounds, such as triglycerides, phospholipids, sterols, fat-soluble vitamins, fatty acids, and many others. Many lipids are structural components of cell membranes. The concentration of lipid metabolites in the cell may change due to both internal and external factors (Welti and Wang 2004). Concentrations of lipids reflect enzymatic activities which make and degrade them. The action of enzymes involved in lipid formation and break-down is dependent on the presence of genes encoding the enzymes. If a lipid-metabolizing gene is mutated and its enzyme is no longer made, the levels of the gene product's reactant(s) and product(s) will be altered. Biological reactions may be part of a long chain of reaction paths or reaction networks.

In this paper, we conduct an exploratory analysis of mutation effects on reactant-product pathways in the plant *Arabidopsis thaliana*, a model plant with many available mutants. Using knowledge of certain known pairs whose reaction is modified by the mutation, we define metrics that quantify the effect of the mutation on the reaction. An optimal metric will allow one to rank all possible metabolite pairs in order of the likelihood that the mutation modified the pathway between them. We use experimental data derived from analysis of wild-type plants and those defective in an enzyme involved in the addition of double bonds to fatty acid groups in membrane lipids. The defective enzymes are known as a "desaturases." Table 1 lists abbreviations that are used. For example, DGDG34:6 represents a lipid that has 34 acyl carbons and 6 carbon-carbon double bonds, with a head group DGDG (digalactosyldiacylglycerol). To develop the scheme used to identify a reactant-product pair whose reaction is reduced by a mutation, among all lipid pairs, the notation in table 2 is used, where WT = wild type and MT = mutant. Figure 1 illustrates the scheme used to find a reactant and product lipid pair in a metabolic pathway.

In Figure 1, (a) A → B is a general notation for an arbitrary reactant and product pair if A is a reactant and B is its product in the pathway. (b) Aw → Bw is a notation to show that Aw can generate Bw. In the wild type condition, this reaction leads to decreased concentration in Aw and increased concentration in Bw. In step (b), Am $\not\rightarrow$ Bm is the notation that indicates that the generation of Bm from Am is reduced if there is a mutation that affects the pathway between reactant and product in the mutant. A decrease in the reaction occurs because the mutation lowers the level of the enzyme that is used to catalyze the reaction. As a result, the concentration of the reactant Am increases, and the level of Bm decreases. In general, if Aw→ Bw and Am $\not\rightarrow$ Bm in Figure 1 (b), the reactant A should have higher concentration in the MT group than in the WT group, and the product B should have lower concentration in the MT group than in the WT group. This leads to the two relations shown in (c), i.e., Aw < Am and Bw > Bm. A reactant product pair adhering to the scheme in Figure 1 will be denoted an A-B pair in text that follows. The scheme illustrated in Figure 1 may seem overly simplistic. However, the usefulness of the scheme is enhanced by employing mutant and wildtype samples in the experiment. In a chemical reaction, no matter what the network, reactants (substrates) of a blocked reaction will be increased and products decreased. Other compounds may be affected also, but the substrate and

products should be among the affected compound group, provided they are measured. Here we are only interested in those points in the network that are altered by the mutation.

Table 1: Abbreviations used in this paper

DGDG	digalactosyldiacylglycerol
fad	fatty acid desaturase (deficiency)
LysoPC	lysophosphatidylcholine
LysoPG	lysophosphatidylglycerol
MGDG	monogalactosyldiacylglycerol
PA	phosphatidic acid
PC	phosphatidylcholine
PE	phosphatidylethanolamine
PG	phosphatidylglycerol
PI	phosphatidylinositol
PS	phosphatidylglycerol

Table 2: The reactant-product notation in the wild type and mutant groups

A	Reactant in the pathway
B	Product in the pathway
A_w	Reactant concentration in WT
A_m	Reactant concentration in MT
B_w	Product concentration in WT
B_m	Product concentration in MT

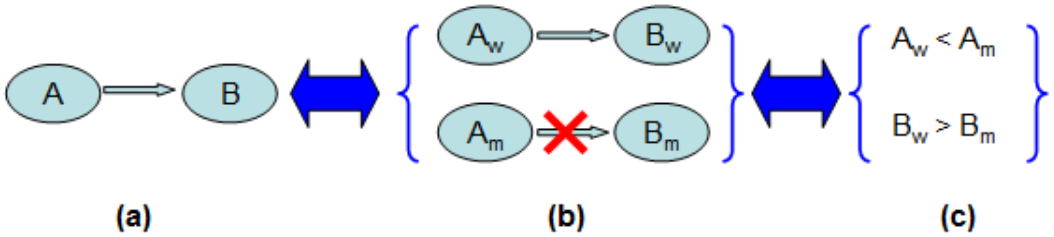

Figure 1: The principle used to find reactant and product A-B lipid pairs
If A is the reactant and B is the product in the pathway in **(a)**, then reactant A can generate B in the WT, i.e., Aw→ Bw, but A cannot generate B in the MT group, i.e., Am ↛ Bm in **(b)**. As a result, Aw < Am and Bw > Bm as shown in **(c)** (Fan 2010).

Data from six lipidomic experiments (see the detailed experimental information in Fan 2010) were collected on mutant plants with mutations in genes with known functions. These mutations were *fad2* (Okuley et al. 1994), *fad3* (Arondel et al. 1992), *fad4* (Gao et al. 2009), *fad5* (Mekhedov et al. 2000), *fad6* (Falcone et al. 1994), and *fad7* (Iba et al. 1993 and Gibson et al. 1994). A total of $2\binom{141}{2} = 19740$ lipid pairs from the 141 lipids were considered in each of the six

different lipidomic experiments. There were 5 samples in the WT group and 5 samples in the MT group. To discriminate the possible reactant-product pairs in the 19740 arbitrary lipid pairs, a list of known reactant-product pairs were used as a criterion for developing a method to identify A-B pairs whose reaction is blocked by the mutation. These biologically functional lipid pairs have attributes that were explained in Fan (2010). These criteria, combined with knowledge of the particular mutation, assist the biologist in providing candidate biologically functional pairs that can be used to establish statistical metrics that quantify characteristics of these pairs. The goal herein is to use patterns that are apparent in the data for known biologically functional pairs to propose an exploratory method and metrics to identify candidate pairs in future experiments where the function of the mutation in the lipid pathway is unknown.

Other approaches have been proposed for exploring and identifying metabolite networks. The main principle in many methods seeking to detect one metabolite in the pathway of another metabolite is to measure and analyze the change of concentration of the metabolites. Raamsdonk et al. (2001) introduced a technique to find the function of "silent" genes using metabolite level changes in a single-celled organism, *Saccharomyces cerevisiae*, a species of yeast. The researchers expected to reveal the role of unknown genes by comparing the metabolite profile of yeast with mutations in those genes to those of mutants in genes of known function using a co-response coefficient in an approach they called FANCY (Functional Analysis by Co-responses in Yeast). The method in Raamsdonk et al. (2001) is closest to the approach proposed herein; however, their method considered concentration changes in six metabolites with respect to a single reference metabolite and used a subset of the information that is used here when defining metrics.

Another method that has been used in metabolic pathway analysis is correlation analysis (Weckwerth et al. 2004; Fukushima et. al. 2011; Steuer 2006). Correlation analysis emphasizes that the metabolic fluctuation might have a linear association between the metabolite concentrations of a metabolite pair in the WT and in the MT groups. Fukushima et al. (2011) used Spearman's correlation to find correlations between pairs of metabolites that were significantly different from zero in two parts of a plant, the aerial and roots. They also tested for correlations between the pairs that were significantly different between aerial and root parts of the plant. Local False Discovery Rate (*lfdr*) was used for multiple testing control.

We used Spearman's correlation analysis as reported in Fukushima et al. (2011) to determine if the technique was effective in identifying the biologically functional pairs in our lipid data sets. The tests of correlations did not detect any biologically functional pairs that were statistically different between the WT and the MT groups, so the use of correlation analysis does not appear useful for the problem considered here. New metrics are needed for quantifying A-B pairs whose reaction is blocked by the mutation.

2. Data Exploration and Definition of Metrics

Here, we refine the supporting evidence for a mutation effect (as shown in Figure 1) into a statistic(s). The method is exploratory and does not rely on distributional assumptions and accommodates potential nonlinear relations and zero values that are present in the data sets. Another limitation for developing statistical methods with these data is small sample sizes. Small sample sizes are not uncommon in metabolomics, and they present difficulties for using

assumptions of normality or application to central limit theorem and for use of correlation for quantifying relationships. Raamsdonk et al. (2001) analyzed their metabolomic data with 3 samples in each treatment. In this experiment, 5 samples are taken for each treatment. Another challenge with metabolite data is a likely high-dimensional dependence structure among lipids and their concentrations. If there is a long chain of reactant and product pathways, one lipid's concentration change may be associated with all other lipids on the pathway. Therefore, one change in concentration of a lipid in the network might cause a sequence of changes in the pathway or the pathway networks (Steuer et al. 2003). However, as noted earlier substrates of a blocked enzyme will be increased and products decreased and this principle is used here in defining metrics to rank candidate lipid pairs whose reaction is blocked by the mutation. We do assume that the samples themselves are independent of each other.

In the following part of this analysis, data from the *fad2* experiment are used as an illustration. The other five data sets have similar properties. The unit of the data is nmol per mg dry weight. The first 5 samples are from the WT group and the last 5 samples are from the MT group. Table 3 lists all the notations for the samples before scaling.

Table 3: Notations used for one reactant A and product B in a lipid pair

n: The sample size in each group.

i : Subscript $i = 1, 2$ to denote the "treatment," 1 = WT and 2 = MT.

j : Subscript $j = 1, 2, \ldots, n$ denotes sample within treatment.

Before scaling:

x_{ij} : The concentration for the j^{th} sample in the i^{th} treatment for one lipid.

$\bar{x}_{i\bullet}$: The group mean in the i^{th} treatment for one lipid.

$\bar{x}_{\bullet\bullet}$: The overall mean across two treatment groups for one lipid, $\bar{x}_{\bullet\bullet} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} x_{ij}$.

s : Standard deviation for one lipid across two treatments, $s = \sqrt{\frac{\sum_{i}^{2}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2}{2n-1}}$.

x_{Aij} : The concentration of j^{th} sample for lipid A in the i^{th} treatment.

x_{Bij} : The concentration of j^{th} sample for B in the i^{th} treatment.

$\bar{x}_{A\bullet\bullet}$: The mean concentration of A across two treatments.

$\bar{x}_{B\bullet\bullet}$: The mean concentration of B across two treatments.

$\bar{x}_{Ai\bullet}$: The mean concentration of A in the i^{th} treatment.

$\bar{x}_{Bi\bullet}$: The mean concentration of B in the i^{th} treatment.

Different lipids are found at varying concentrations in biological samples, with some having substantially greater abundance than others. This presents challenges to evaluate reactant-product pairs in a pathway. Thus a first step is to scale lipid concentrations so that different pairs are comparable using a single metric. This scaling should not alter the relative positioning of lipids with respect to one another and, thus, alter the nature of the mutation's effect on the

reaction. Lipid cconcentrations are centered and scaled by using the standardization formula, z_{ij}, given below.

After scaling:

Let $z_{ij} = \frac{x_{ij} - \bar{x}_{\bullet\bullet}}{s}$. Then all quantities above that are defined before scaling have corresponding quantities after scaling, and are denoted by the variable *z* instead of *x*.

- **Proposition 1:** Consider a single lipid and denote the concentration by x_{ij} for the j^{th} sample in the i^{th} treatment, where $i = 1, 2$, and $j = 1, 2, \ldots, n$. Then, $|\bar{z}_{i\bullet}| \le \sqrt{1 - \frac{1}{2n}}$ for $i = 1, 2$ and $|\bar{z}_{1\bullet} - \bar{z}_{2\bullet}| \le 2\sqrt{1 - \frac{1}{2n}}$.

Proof: It is clear that $\bar{z}_{\bullet\bullet} = \frac{1}{2n}\sum_{i=1}^{2}\sum_{j=1}^{n} z_{ij} = 0$ and equal samples in each group implies $\bar{z}_{1\bullet} = -\bar{z}_{2\bullet}$. So if we focus on $\bar{z}_{1\bullet}$, we have

$$\bar{z}_{1\bullet} = \frac{1}{n}\sum_{j=1}^{n} z_{1j} = \frac{1}{n}\sum_{j=1}^{n} \frac{x_{1j} - \bar{x}_{\bullet\bullet}}{s} = \frac{1}{n}\frac{\sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet})}{s}. \tag{1}$$

The numerator of (1) can be calculated as

$$\sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet}) = \frac{1}{2}(n\bar{x}_{1\bullet} - n\bar{x}_{2\bullet}) = \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet}).$$

Then,

$$\begin{aligned}
(2n-1)s^2 &= \sum_{i=1}^{2}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{\bullet\bullet})^2 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{\bullet\bullet}) \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet} + \bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet} + \bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + n(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})\sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet}) \\
&\quad + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 + 2(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})\sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet}) \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + n(\bar{x}_{1\bullet} - \bar{x}_{\bullet\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n(\bar{x}_{2\bullet} - \bar{x}_{\bullet\bullet})^2 \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + n\left(\bar{x}_{1\bullet} - \frac{\bar{x}_{1\bullet} + \bar{x}_{2\bullet}}{2}\right)^2 + n\left(\bar{x}_{2\bullet} - \frac{\bar{x}_{1\bullet} + \bar{x}_{2\bullet}}{2}\right)^2 \\
&= \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2 \qquad (2)
\end{aligned}$$

Using (2) and (1) results in,

$$\bar{z}_{1\bullet}^2 = \left(\frac{\frac{1}{n}\sum_{j=1}^{n}(x_{1j} - \bar{x}_{\bullet\bullet})}{s} \right)^2 = \frac{(2n-1)}{n^2} \cdot \frac{\left(\frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})\right)^2}{\sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2 + \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2 + \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}.$$

For one lipid, let $SSW_1 = \sum_{j=1}^{n}(x_{1j} - \bar{x}_{1\bullet})^2$, and $SSW_2 = \sum_{j=1}^{n}(x_{2j} - \bar{x}_{2\bullet})^2$, then

$$\bar{z}_{1\bullet}^2 = \frac{(2n-1)}{4} \frac{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{SSW_1 + SSW_2 + \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}.$$

For the reciprocal of $\bar{z}_{1\bullet}^2$, $\frac{1}{|\bar{z}_{1\bullet}^2|} = \frac{4}{2n-1} \frac{SSW_1 + SSW_2 + \frac{n}{2}(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}$,

i.e., $$\frac{1}{\bar{z}_{1\bullet}^2} = \frac{2n}{2n-1} + \frac{4}{2n-1} \frac{SSW_1 + SSW_2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}. \quad (3)$$

Let $\Delta = \frac{4}{2n-1} \frac{SSW_1 + SSW_2}{(\bar{x}_{1\bullet} - \bar{x}_{2\bullet})^2}$, then $\frac{1}{\bar{z}_{1\bullet}^2} = \frac{2n}{(2n-1)} + \Delta$. So $\bar{z}_{1\bullet}^2 = \frac{1}{\frac{2n}{(2n-1)} + \Delta}$.

The max ($\bar{z}_{1\bullet}^2$) should occur when $\Delta \to 0$ which begins to happen when $\bar{x}_{1\bullet} - \bar{x}_{2\bullet} > 0$ and the two within sums of squares, SSW_1 and SSW_2, are close to zero. So $\bar{z}_{1\bullet}^2 \le \frac{2n-1}{2n}$ or $|\bar{z}_{1\bullet}| \le \sqrt{1 - \frac{1}{2n}}$. Similarly, $|\bar{z}_{2\bullet}| \le \sqrt{1 - \frac{1}{2n}}$. Therefore, $|\bar{z}_{1\bullet} - \bar{z}_{2\bullet}| \le 2\sqrt{1 - \frac{1}{2n}}$ holds. When *n* becomes large, this upper bound is close to 2. ■

After data are centered and scaled, Proposition 1 gives the following results.

1. For the data described here with *n* = 5, $|\bar{z}_{i\bullet}| \le \sqrt{1 - \frac{1}{10}} = 0.949$. As *n* gets large, $\max|\bar{z}_{i\bullet}|$ goes to 1.
2. For a lipid pair, A and B, with two dimensional means given by ($\bar{z}_{A1\bullet}, \bar{z}_{B1\bullet}$) and ($\bar{z}_{A2\bullet}, \bar{z}_{B2\bullet}$) for the wild type and mutant groups, respectively, the maximum Euclidian distance between them is 2.684 for *n* = 5 and approaches $\sqrt{8} = 2.828$ as sample size increases.

3. Defining

$$SS_{between} = SSD = (\bar{z}_{A1\bullet} - \bar{z}_{A2\bullet})^2 + (\bar{z}_{B2\bullet} - \bar{z}_{B1\bullet})^2 = SS_{between,A} + SS_{between,B} \text{ and}$$

$$SS_{within} = \sum_{i=1}^{2}\sum_{j=1}^{5}\left[(z_{Aij} - \bar{z}_{Ai\bullet})^2 + (z_{Bij} - \bar{z}_{Bi\bullet})^2\right] = SS_{within,A} + SS_{within,B},$$

   a. $SS_{between} = 0$ implies that the lipid means $|\bar{z}_{l\bullet}| \approx 0$ for each lipid.
   b. If $SS_{within} >> SS_{between}$, both group centers are close to the origin (0, 0).

Figure 2 shows the relative position in the WT and MT groups for the same lipid pair before and after scaling. The relative positions of the WT and MT groups remain the same in the two plots. Another result worth noting is that Pearson's sample correlation coefficient between a pair of lipids is unchanged after scaling the data as described above.

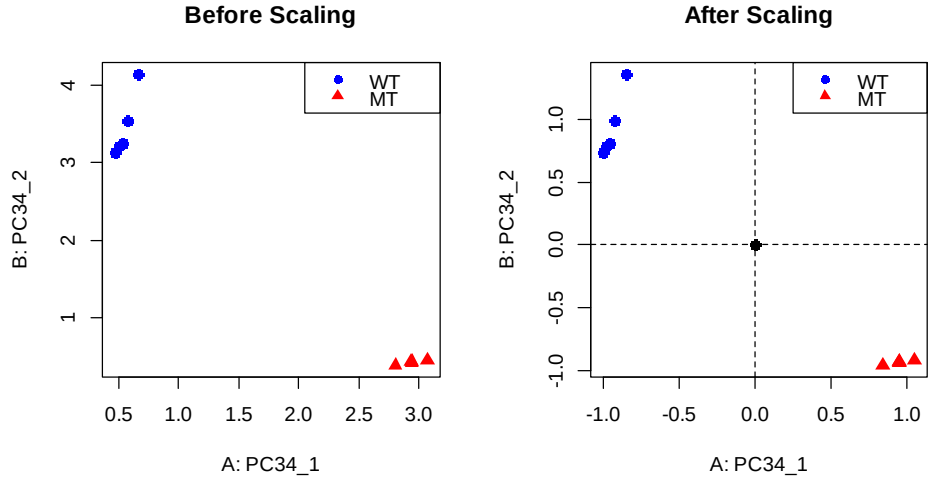

Figure 2: Example scatter plots of one lipid pair before and after scaling
Lipid PC34_1 (A, PC34:1, putative reactant) and lipid PC34_2 (B, PC34:2, putative product), which form a lipid pair, are plotted before (left panel) and after scaling (right panel).

After centering and scaling, a total 19740 lipids are paired to determine whether or not the concentration change in the pair follows the scheme shown in Figure 1 and given here by $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$. Note that each lipid is allowed to be a candidate product or reactant prior to the below screening procedure. For convenience and to quantify the scheme in Figure 1, define a variable *y*, where

$$y = I_{\{\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}\}} + I_{\{\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}\}}. \quad (4)$$

When both $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ hold, A and B are a lipid pair that satisfy the screening procedure for a reactant-product pair, and $y = 2$. The arbitrary lipid pairs that satisfy the conditions $y = 2$ will be used to prescreen the sample of all lipid pairs when defining metrics in the section that follows. Note that y = 0 reflects the same pair but with product and reactant roles reversed, and that $y = 1$ implies that both lipids are either reactants or products but the two together are not a reactant-product pair.

3. Example

The biologically functional reactant-product pairs are used as a standard to compare with all other arbitrary lipid pairs. Figure 3 shows the scatter plot characteristics of nine biologically functional reactant-product pairs in the data set *fad2*. There are in total 18 biologically functional pairs in the *fad2* data set. The remaining scatter plots from the biologically functional pairs are all similar to those in Figure 3. All have similar patterns: WT is in the upper left corner, MT is in the lower right corner. Their concentration relationships satisfy the screening scheme in Figure 1 which is $\bar{z}_{A1\bullet} < \bar{z}_{A2\bullet}$ and $\bar{z}_{B1\bullet} > \bar{z}_{B2\bullet}$ with mean differences between WT and MT near the maximum derived in the Proposition. In fact, in all other mutant data sets that we have evaluated, the same patterns as shown in Figure 3 are apparent for any known biologically functional substrate-product pairs.

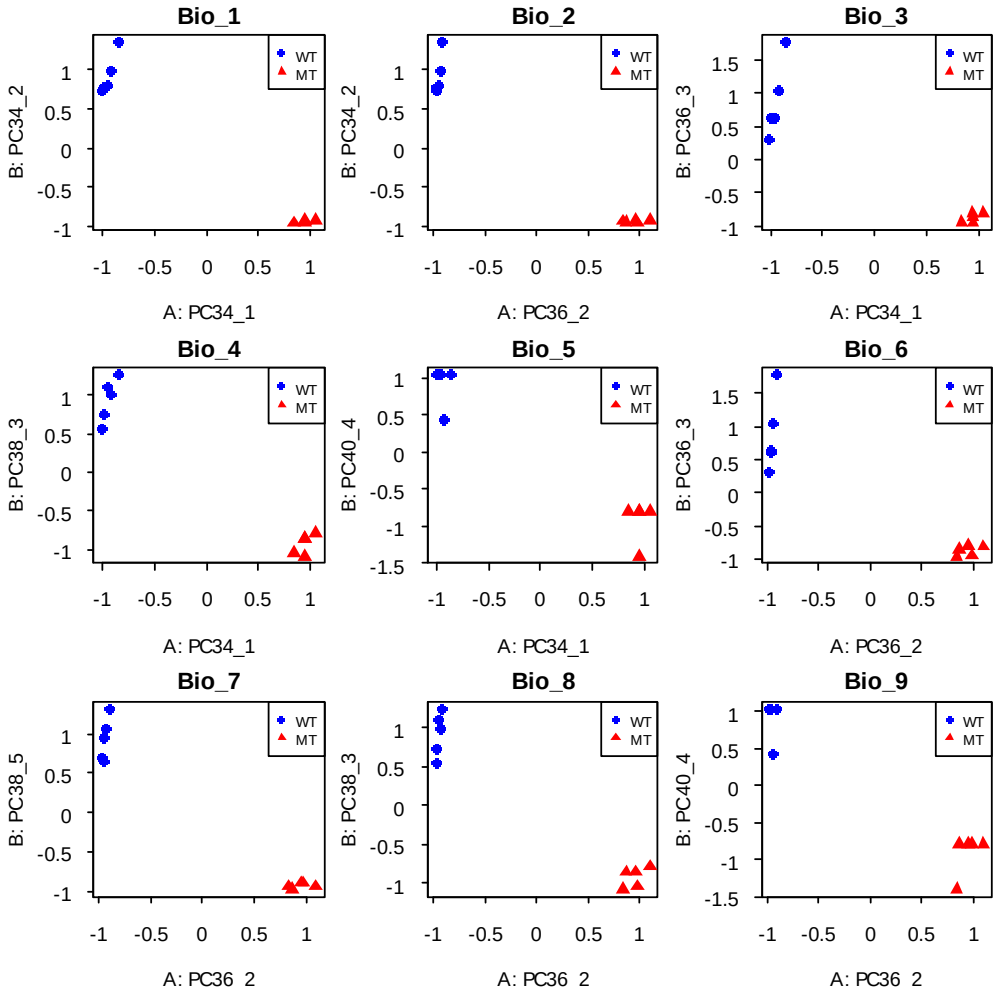

Figure 3: Scatter plots of nine biologically functional pairs in *Fad2*

In each panel, the 5 blue circles represent the WT group with coordinates (z_{A1j}, z_{B1j}) and the 5 red triangles stand for the MT group with coordinates (z_{A2j}, z_{B2j}). The x-axis is the concentration of the reactant A and the y-axis is the concentration of the product B.

Three Summary Statistics: Three summary test statistics are developed according to patterns seen in exploratory data analysis. These three are denoted tg, SSD, and –log(R) statistics. The distributions of these statistics from the *fad2* data are shown in Figure 4. The statistics are computed from each candidate lipid reactant-product pair in the scaled data (i.e., those pairs satisfying the $y = 2$ screening criteria).

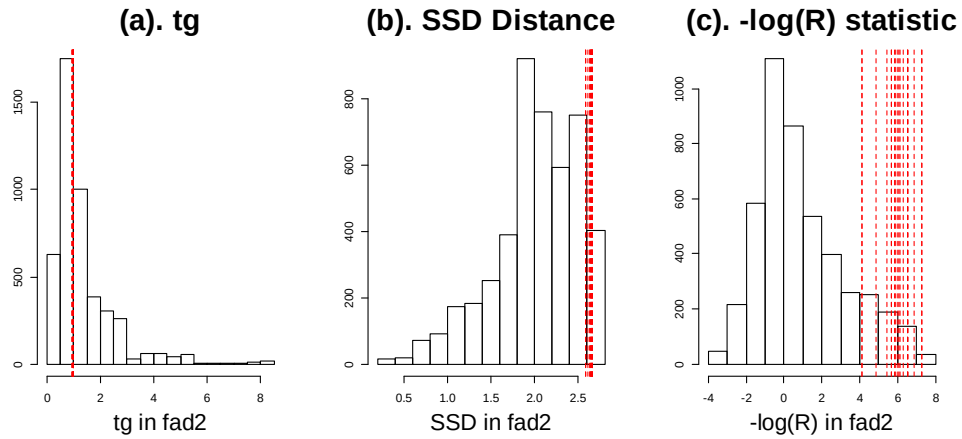

Figure 4: The distributions of the three test statistics.
The red dashed vertical lines show the statistics for the biologically functional lipid pairs in data set *fad2*.

- **Statistic 1: tg**

This tg is the ratio of a lipid product B group mean difference to a reactant A group mean difference, and is defined as

$$\text{tg} = \frac{\bar{z}_{B1\bullet} - \bar{z}_{B2\bullet}}{\bar{z}_{A2\bullet} - \bar{z}_{A1\bullet}}, \qquad (5)$$

where the means $\bar{z}_{Ai\bullet}$ and $\bar{z}_{Bi\bullet}$ were defined earlier . According to exploratory data analyses, the positions for the two dimensional groups WT and MT most representative of biologically functional pairs is a 135 degree angle with the x-axis, which leads to tg = 1. The red lines in the tg distribution in Figure 4(a) show the biologically functional pairs with tg values that are all close to 1. Raamsdonk et al. (2001) used a measurement based on tg (actually an arctangent transformation of it) in defining the co-response coefficient Ω which was a ratio of the log concentration change in their FANCY approach.

- **Statistic 2: SSD**

SSD is a squared distance between the two group centers and given by

$$\text{SSD} = SS_{between} = (\bar{z}_{A1\bullet} - \bar{z}_{A2\bullet})^2 + (\bar{z}_{B1\bullet} - \bar{z}_{B2\bullet})^2. \qquad (6)$$

Large SSD, or inter-group distance, corresponds to biologically functional pairs. These results are shown in Figure 4(b).

- **Statistic 3: R**

When used alone, the tg statistic does not capture all characteristics of potential reactant-product pairs. Similarly, the SSD statistic also has a disadvantage. If the inter-group distance is very large, but the angle between the groups is very different from 135º, then the result using just SSD may not select the true candidate reactant-product pairs. That is, using SSD alone may lead to false discoveries.

In Figure 5, the data points (bio.tg, bio.SSD) from the 18 different biologically functional pairs show them to be close to the top peak with coordinates (1, 2.684). Therefore, a statistic combining both tg and SSD at the same time is proposed. This combined statistic, called R, can eliminate the respective limitations of tg and SSD while keeping their advantages. The statistic R is defined as

$$R = (tg - 1)^2 + (SSD - \max(SSD))^2 . \quad (7)$$

The value $\max(SSD)$ is set to 2.684 from the theoretical maximum SSD from the proposition for our samples of size 5 in each group. The value of R should not ever be exactly zero. The lipid pairs that are of interest will have values of R near zero.

For improved interpretability and separation of small values of R, the R statistics are transformed by –log(R) so that large values of –log(R) are those of interest. Figure 4(c) shows the distribution of the transformed R statistic, –log(R). The biologically functional pair's –log(R) statistics (red lines) show that the larger values of –log(R) reflect results that are of interest.

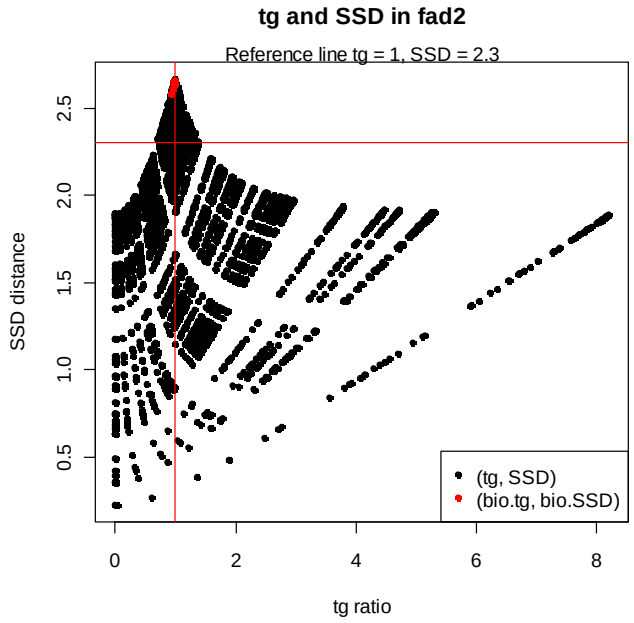

Figure 5: Illustration of the information for the R statistic, using the scatter plot of tg versus SSD in the data set involving the WT and the *fad2* mutant
The vertical red line shows tg = 1. The horizontal red line shows an arbitrary cutoff point at SSD = 2.3. The black points are the (tg, SSD) coordinates for each lipid pair. The red points at the peak are the biologically functional pairs. The peak area contains the most interesting lipid pairs with large SSD values and tg close to 1.

Note that in Figure 5, the scatter plot of tg and SSD shows curved patterns because tg and SSD are both functions of the means, $\bar{z}_{Ai\bullet}$ and $\bar{z}_{Bi\bullet}$ as shown previously.

4. Discussion and Future Work

In conclusion, the three statistics derived in section 3 were based on the exploration of data according to the screening principle illustrated in Figure 1. From the above analysis, we can see that the three statistics reflect the data characteristics for lipid pairs that are biologically functional reactant-product pairs whose reaction is modified by the mutation in the organism. They can be employed separately or combined as a whole. So as metrics themselves, they are useful quantities for ranking reactant-product pairs as potentially affected by a mutation in cases where the role of the mutation is unknown. Analysis of other mutant data sets in which biologically functional pairs were known revealed the same pattern for the defined metrics as seen above in Figures 4 and 5. As such, we propose these metrics for identifying reactions that are modified by mutations of unknown function. Work is in currently in progress and planned to do this with new data sets.

There are still improvements that can be made to a statistical method for detecting such pairs. After the three statistics are found from the above example, the empirical distributions of the three statistics can be presented as shown in Figure 4. To assess statistical significance, this empirical distribution could be compared with a distribution of the statistics under some null hypothesis. Null distributions of statistics can often be generated by bootstrap resampling methods, under a condition for which the null hypothesis is true. A challenge that arises is how to specify an appropriate null hypothesis. One null hypothesis is $H_0: F = G$, where F is a multivariate distribution of lipid concentrations for the WT group and G is the distribution for the MT group. This null hypothesis is easy to accommodate in a bootstrap procedure and some initial work has been done. However, this null hypothesis may be too restrictive in situations where a mutation substantially modifies concentrations in the entire lipidome. We have seen that in some data sets, the bootstrap null distributions of the test statistics, SSD and –log(R), deviate quite far from the empirical distributions. The results suggest strong mutation effects in the data sets. A less restrictive null hypothesis would be the following intersection-union hypotheses,

$$H_0: \mu_{Aw} \geq \mu_{Am} \text{ or } \mu_{Bw} \leq \mu_{Bm}$$
$$H_A: \mu_{Aw} < \mu_{Am} \text{ and } \mu_{Bw} > \mu_{Bm}.$$

Bootstrap sampling under these hypotheses may be more reasonable for the application considered here.

When conducting the exploratory analysis, the prescreening of candidate pairs was done with the *y* statistic as defined in equation (4). If attempting to derive a probabilistic certainty to a list of findings, the sampling variability of this prescreening step may also need to be incorporated. Also, the metrics that were defined were largely based on differences in means. One might also wonder if differences in sample variances might also be used to evaluate candidate pairs. It is unclear at this point to what extent or how the sample standard deviations may be altered by the mutation. One exception would be if the mutation completely blocked the

reaction and the formation of the lipid product. In such a case, one might expect the candidate product would be positive in the wildtype organism and exactly zero in all samples in the mutant. This has been rarely seen and it is not always clear whether zero concentrations are real zeros or simply below the limit of detection for the instrument.

As development of methods for the analysis of gene expression data progressed over the past years, attention turned to methods for simulating realistic high-dimensional data. Previously, data were often simulated (and still are) from multivariate normal distributions with restrictive dependence structures. Simulating more realistic gene expression data was considered in Gadbury et al. (2008). Simulating realistic lipidomic data is likely to be challenging. Still it will be necessary in order to evaluate the performance of new statistical methods for analyzing lipidomic data. This is another area of research to be explored.

Acknowledgements
The authors acknowledge Lixia Fan, George Milliken, Haiyang Wang, Lili Cheng, Richard Jeannotte, and Ashis Nandi for earlier work leading to that reported herein.

References

Arondel, V., Lemieux, B., Hwang, I., Gibson, S., Goodman, H.M. and Somerville, C.R. (1992). Map-based cloning of a gene controlling omega-3 fatty acid desaturation in *Arabidopsis*. *Science*, 258, 1353-1355.

Blei, I., Odian G. (2006). General, organic, and biochemistry. Second edition, New York: W.H. Freeman and Company.

Dixon, R. A., Gang, D. R., Charlton, A. J., Fiehn, O., Kuiper, H. A., Reynolds, T. L., Tjeerdema, R. S., Jeffery, E. H., German, J. B., Ridley, W. P. and Seiber, J. N. (2006). Applications of Metabolomics in Agriculture. *Journal of Agricultural and Food Chemistry*, 54, 8984-8994.

Dunn, W. B., Ellis, D. I. (2005). Metabolomics: Current analytical platforms and methodologies. *Trends in analytical chemistry*, 24, 285-294.

Falcone, D.L., Gibson, S., Lemieux, B. and Somerville, C. (1994). Identification of a gene that complements an Arabidopsis mutant deficient in chloroplast omega 6 desaturase activity. *Plant Physiology*, 106, 1453-1459.

Fan, L. (2010). An exploratory method for identifying reactant-product lipid pairs from lipidomic profiles of wild-type and mutant leaves of *Arabidopsis thaliana*. Master report. Kansas State University.

Fukushima, A., Kusano, M., Redestig H., Arita, M., Saito, K. (2011). Metabolomic correlation-network modules in Arabidopsis based on a graph-clustering approach. *BMC Systems Biology*, 5, 1-12.

Gadbury, G. L., Xiang, Q., Yang, L., Barnes, S., Page, G. P., Allison, D. B. (2008). Evaluating statistical methods using plasmode data sets in the age of massive public databases: An illustration using False Discovery Rates. *PLos Genetics*, 4(6).

Gao, J., Ajjawi, I., Manoli, A., Sawin, A., Xu, C., Froehlich, J. E., Last, R. L. Benning, C. (2009). FATTY ACID DESATURASE4 of Arabidopsis encodes a protein distinct from characterized fatty acid desaturases. *The Plant Journal*, 60, 832–839.

Gibson, S., Arondel, V., Iba, K. and Somerville, C. (1994). Cloning of a temperature-regulated gene encoding a chloroplast omega-3 desaturase from Arabidopsis thaliana. *Plant Physiol*, 106, 1615-1621.

Griffin, J. L., Vidal-Puig, A. (2008). Current challenges in metabolomics for diabetes research: a vital functional genomic tool or just a ploy for gaining funding? *Physiological Genomics*, 34, 1–5.

Iba, K., Gibson, S., Nishiuchi, T., Fuse, T., Nishimura, M., Arondel, V., Hugly, S. and Somerville, C. (1993). A gene encoding a chloroplast omega-3 fatty acid desaturase complements alterations in fatty acid desaturation and chloroplast copy number of the fad7 mutant of *Arabidopsis thaliana*. *The Journal of Biological Chemistry*, 268, 24099-24105.

Mekhedov, S., de Ilarduya, O.M. and Ohlrogge, J. (2000). Toward a functional catalog of the plant genome. A survey of genes for lipid biosynthesis. *Plant Physiology*, 122, 389-401.

Okuley, J., Lightner, J., Feldmann, K., Yadav, N., Lark, E. and Browsea, J. (1994). Arabidopsis fad2 gene encodes the enzyme that is essential for polyunsaturated lipid synthesis. *The Plant Cell*, 6, 147-158.

Oliver, S. G. (2002). Functional genomics: lessons from yeast. *Philosophical Transactions of the royal society B*, 357, 17-23.

Oliver, S.G., Winson, M.K., Kell, D.B. and Baganz, F. (1998). Systematic functional analysis of the yeast genome. *Trends in Biotechnology*, 16, 373-378.

Raamsdonk, L.M., Teusink, B., Broadhurst, D., Zhang, N., Hayes, A., Walsh, M. C., Berden, J. A., Brindle, K. M., Kell, D. B., Rowland, J. J., Westerhoff, H. V., Dam, K. V. and Oliver, S. G. (2001). A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nature Biotechnology*, 19, 45 – 50.

Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Briefings in Bioinformatics*, 7, 151-158.

Steuer, R., Kurths, J., Fiehn, O., Weckwerth, W. (2003). Observing and interpreting database for Medicago truncatula. *Bioinformatics*, 23, 1418–1423.

Weckwerth, W., Loureiro, M-E, Wenzel, K., and Fiehn, O. (2004). Differential metabolic networks unravel the effects of silent plant phenotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 101, 7809-7814.

Welti, R. and Wang, X. (2004). Lipid species profiling: a high-throughput approach to identify lipid compositional changes and determine the function of genes involved in lipid metabolism and signaling. *Current Opinion in Plant Biology*, 7, 337–344.

Wu, L., Winden, W. A. V., Gulik, W. M. V. and Heijnen, J. J. (2005). Application of metabolome data in functional genomics: A conceptual strategy. *Metabolic Engineering*, 7, 302–310.