8-2010

An Analysis of Global Gene Expression Resulting from Exposure to Energetic Materials

Vernon L. McIntosh Jr.
University of Tennessee - Knoxville, vmcintos@gmail.com

To the Graduate Council:

I am submitting herewith a dissertation written by Vernon L. McIntosh Jr. entitled "An Analysis of Global Gene Expression Resulting from Exposure to Energetic Materials." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Microbiology.

Gary S. Sayler, Major Professor

We have read this dissertation and recommend its acceptance:

Gladys Alexandre, Todd Reynolds, John Sanseverino, Erik Zinser

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Vernon L. McIntosh Jr. entitled “An Analysis of Global Gene Expression Resulting from Exposure to Energetic Materials.” I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy with a major in Microbiology.

Gary Sayler, Major Professor

We have read this dissertation
and recommend its acceptance:

Gladys Alexandre

Todd Reynolds

John Sanseverino

Erik Zinser

Accepted for the Council:

Carolyn R. Hodges
Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

AN ANALYSIS OF GLOBAL GENE EXPRESSION RESULTING FROM EXPOSURE TO ENERGETIC MATERIALS

A Dissertation

Presented for the Doctor of Philosophy Degree

University of Tennessee, Knoxville

VERNON LASHAWN MCINTOSH JR.

August 2010

Dedication

This dissertation is dedicated to my family. My mother and father Debra and Vernon McIntosh instilled in me the respect for academic excellence and the drive maximize my potential. Early on, my younger brother Kyle started showing signs of a shared interest in biology thus my desire to be a positive role model for him kept me motivated. Last but certainly not least, my loving wife and best friend Nichole has been there to offer love and support throughout my entire undergraduate and graduate degrees. It's difficult to imagine making it this far without her (and that's not just because she paid the bills).

Abstract

Characteristic transcriptional biomarkers have been identified for microbial cultures exposed to 2, 4, 6-trinitrotoluene (TNT), 2, 6-dinitrotoluene (DNT), or triacetone-triperoxide (TATP). This study describes the generation of expression profiles for exposure to each compound, the functional significance of each response, and the identification of the characteristic alterations in gene expression associated with exposure to each compound.

Expression profiles were generated from a total of three different candidate organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, and *Pseudomonas putida*. Common to all three organisms, TNT exposure resulted in increased expression of genes involved in toxin resistance and drug efflux systems. The *S.cerevisiae* and *E.coli* expression profiles were both characterized by increased expression of genes involved in iron-sulfur cluster assembly, sulfur containing amino acids, sulfate transport and assimilation and the metabolism of nitrogen compounds.

Only *E.coli* and Saccharomyces were used to generate DNT induced expression profiles; both profiles exhibited high degrees of similarity with each organism's respective TNT profiles. This was especially true of the *E.coli* profile where 25 of the 30 alterations were also observed after exposure to TNT.

A computational discriminant functional analysis was performed to identify characteristic biomarkers for each exposure. For each compound a set of

transcriptional biomarkers (10 or less) was developed. An additional set of biomarkers was developed encompassing both TNT and DNT exposure. These sets of genes serve as a transcriptional fingerprint for exposure to each respective compound. The sensitivity and specificity of each transcriptional fingerprint is sufficient to correctly identify exposure to energetic materials against a background of non-energetic compound exposures.

This study makes several novel contributions to the greater body of scientific knowledge:

- This is the first documented study of the interactions of TATP in any biological system.
- This is the first comprehensive gene expression study of the TNT response by *P. putida*, *E.coli* or *E.coli*.
- This is the first application of computational class prediction in the development of biomarkers for exposure to energetic materials

Table of Contents

List of Figures

List of Tables

Chapter 1 : Literature Review

Motivation

Arguably, the most influential moments in history have not been milestones of human accomplishment or ingenuity; neither have they been the result of human error or acts of nature. They have been deliberate acts of violence. The most broad-reaching of these events have been political or religious wars waged between bona-fide, legitimate entities. In many cases, the civilian casualties from these wars have been ongoing due to residual unexploded ordinance left in their wake. Landmines represent an indiscriminate threat, and as such nearly 85% of casualties and injuries are suffered by non-military personnel (International Campaign to Ban Landmines promotional materials).

Most recently, prevailing social psychology has been shaped most by the actions of a minority of dedicated, radical individuals with a goal to terrorize and destroy human life:

- Unabomber from 1978-1995,
- the 1995 Oklahoma City bombing,
- the 1998 US embassy bombings in Tanzania and Kenya,
- the 1999 massacre at Columbine High School,
- the and most notably the attacks of September 11, 2001

The motivation for this course of research was, and still is, to address the growing threat of domestic terrorism in the United States and abroad.

Introduction

Explosive agents in the environment are of significant concern due their toxicity and also their threat of traumatic damage upon detonation. They are present as a result of manufacture waste, unexploded ordinance (UXO), construction loss, illegal synthesis and use in improvised explosive devices (IED). The threat posed by landmines alone is considerable. The United Nations estimates that total cost of location and removal of currently installed landmines would incur a cost in excess of $30 billion and over 1100 years. To compound this issue , the installation of new landmines outpaces the removal of old mines at a rate of 30:1(Sylvia, Janni et al. 2000).

The single greatest problem facing detection of these compounds is sampling. The compounds of interest often have very low vapor pressures making their concentrations in ambient conditions very dilute. This in turn makes detection of these compounds at any reasonable distance extremely difficult. Successful detection of these compounds almost always requires direct sampling at the source. An efficient detection method will require real time, highly sensitive, sampling of ambient air or water.

There are opportunities to use genetic and transcriptional profiling methods to ascertain whether biological populations have been exposed to explosives or energetic agents and to further develop approaches for sensor detection and possible sentinel monitoring. One of these approaches is proposed as “transcriptional fingerprint biomarker analysis of energetic chemical exposure”, the objective of this investigation.

This research seeks to determine if environmental exposure to energetic agents can be recorded in the profile of gene expression from pure cultures and, potentially, the metagenome.

Extensive production and use of 2,4,6-trinitrotoluene has lead to excessive land contamination (Esteve-Nunez, Caballero et al. 2001; Jenkins, Hewitt et al. 2006; Gong, Guan et al. 2007). Environmental contamination of both TNT and its decomposition products is a cause for concern. TNT toxicity has been observed in many organisms, with symptoms including reproductive inhibition, anemia, liver damage , skin irritation, cataracts, and oxidative stress (Johnson, Ferguson et al. 2000; Reddy, Chandra et al. 2000; Cenas, Nemeikaite-Ceniene et al. 2001; Nemeikaie-Ceniene, Sarlauskas et al. 2004; Gong, Guan et al. 2007; Gong, Guan et al. 2007).

TNT contamination can occur through numerous routes including deposition of detonation residue, improper disposal, storage and faulty ordinance housing. The usual fate is soil contamination (Esteve-Nunez, Caballero et al. 2001; Jenkins, Hewitt et al. 2006). The extent of groundwater contamination is ultimately determined by the mobility and speciation of TNT in soils. Mobility and speciation are influenced by a number of factors including chemical transformation, covalent bonding with organic matter, and absorption by soil particles(Pennington and Brannon 2002). Biological processes play a role in each of these factors.

Microbial transformation of nitrotoluenes

Although generally considered to be recalcitrant, several studies have demonstrated reductive transformation of nitrotoluenes by microbial species (Wittich, Ramos et al. 2009). Reductive transformation of TNT and DNT by fungal and bacterial species typically occurs either through the formation of hydride Meisenheimer complexes followed by the subsequent rearomatization and release of nitrogen as nitrite, or through the reduction of the nitro groups to hydroxylamine or amino groups followed by the subsequent release of nitrogen as ammonia (Roldan, Perez-Reinado et al. 2008). While there are no previously published studies describing reductive transformation of nitrotoluenes by the model yeast organism Saccharomyces cerevisiae, nitrogen release from TNT through the reductive hydroxylamine pathways have been described in other Saccharomyces strains (Zarlpov, Naumov et al. 2002).

Previous studies have established that *E. coli* has the ability to both reduce TNT to its ADNT and HADNT metabolites and ultimately release nitrogen from the aromatic ring to be used for growth (Yin, Wood et al. 2005; Gonzalez-Perez, van Dillewijn et al. 2007) . This process is catalyzed by FMN dependent-NAD (P) H nitroreductases NfsA and NfsB as well as a xenobiotic reductase NemA. N-ethylmalemimide reductase (NemA) is a member of the old yellow enzyme (OYE) family of proteins. Bacterial OYE enzymes are characterized by the ability to reduce TNT nitro groups in vivo (Williams, Rathbone et al. 2004; Gonzalez-Perez, van Dillewijn et al. 2007; Roldan, Perez-Reinado et al. 2008).

Reductive transformation of TNT has also been observed in cultures of *Pseudomonas putida* (Caballero, Esteve-Nunez et al. 2005; Caballero and Ramos 2006). These studies have established that, like *E. coli*, *P. putida* has the ability to metabolize nitrotoluenes to release nitrogen for growth. While the precise mechanism of nitrogen release has yet to be resolved, the involvement of glutamine syntethase-glutamate synthase (GS-GOGAT) has been suggested (Caballero, Esteve-Nunez et al. 2005). Mutants deficient in any component of this protein complex show impaired reduction of TNT. Further work by Caballero and colleagues suggested that nitrite reductase nasB and nitroaromatic reductase pnrA play essential roles in TNT reductive transformation. Double knockout mutants resulted in complete growth inhibition on media containing TNT as the sole nitrogen source.

Toxicogenomics and the use of Transcriptional Profiling

The term *toxicogenomics* was first coined by Emile Nuwaysir and colleagues in 1999 (Nuwaysir, Bittner et al. 1999). It was in that paper that the concept of combining the newly emerging DNA microarray technology with the growing wealth of publicly accessible sequence data to address some of the most pressing challenges in toxicology was born. The goal of toxicogenomics was thus defined as "… to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant." With the refinement of microarray technology and the growing capacity for transcriptional analysis, the scope and definition of toxicogenomics has broadened. Toxicogenomics has since been defined

simply as “the study of toxicological processes at the transcriptome level of a target organ or cell” (de Longueville, Bertholet et al. 2004). Thus toxicogenomics is the marriage of genomics and classical toxicology. Most previously developed *in vivo* and *in vitro* toxicology assays rely on the presentation of physiological effects or symptoms of the toxicant exposure. Implicit was the idea that all of these effects are preceded by altered gene expression. This was a concept recognized early on (Nuwaysir, Bittner et al. 1999).

The National Center for Toxicogenomics (NCT) has adopted, as its goal, the challenge of providing a repository of expression data as a knowledge base for toxicogenomic analysis (Tennant 2002). This goal has been realized in the form of the Chemical Effects in Biological Systems database (CEBS).

Predictive Toxicogenomics

In the context of this study, predictive toxicogenomics describes the use of transcriptional profiling to develop a characteristic list of responsive genes for compound exposure. Expression data resulting from known compound exposure conditions are designated functional classifications. Expression profiles from unknown exposure conditions are then identified based on their similarity to the database of known gene expression patterns. This serves two purposes. First it becomes possible to identify, or at least functionally categorize, an unknown compound based on the changes it induces in the transcriptome of a model organism. In other words,

information regarding the biological effect of an unknown compound is applied to classify that compound in relation to other known compounds. The second purpose served by predictive toxicogenomics is to develop biomarkers for compound exposure. Ideally these biomarkers would be specific to a particular compound. However, in the case of xenobiotic compounds, it is likely that biomarkers from such a study would be specific to a class of compounds or a specific type of molecular interaction. These biomarkers can now be used to either indicate exposure to a specific compound or to predict possible toxicological effects of the compound in question (de Longueville, Bertholet et al. 2004; Gatzidou, Zira et al. 2007).

Mechanistic Toxicogenomics

Mechanistic toxicogenomics deals with providing information about the specific pathways and cellular activities associated with a given exposure (Gatzidou, Zira et al. 2007). This can provide insight into the mode of action, or potential physiological effects of the compound. To make an analogy, if toxicogenomics is the marriage of genomics and toxicology, then mechanistic toxicogenomics is specifically the marriage of *functional* genomics and toxicology. The question posed by such studies is not simply "What genes have altered expression?" but rather "What is the *function* of the genes with altered expression?" The core assumption is that the global mRNA pool is indicative of the current state of the cell, and will be altered during any response to a toxicant (Gatzidou, Zira et al. 2007).

DNA Microarray Technology

DNA microarrays allow the expression thousands of gene targets to be measured from a single biological sample in tandem. Their broad use has resulted in a paradigm shift from single gene expression studies to whole-genome expression profiling (Kuhn, Baker et al. 2004). Most methodologies exploit basic nucleic acid hybridization chemistry for detection of individual transcripts (Schena, Shalon et al. 1995). Briefly, nucleic acid sequences, or probes, are affixed to a solid substrate. Probes can either be long cDNAs (up to 2000bp) derived from PCR amplification of the target transcript, or short oligonucleotide probes (usually 25-50bp)(Brown and Botstein 1999; Ragoussis 2009) Target nucleic acid samples are then fluorescently labeled and hybridized to the immobilized probes. The abundance of each target sequence is measured as a function of the fluorescence of each probe after hybridization(Lemieux, Aharoni et al. 1998). Recently numerous commercial microarray platforms have emerged, most utilizing some variation of this hybridization schema. Affymetrix GeneChip (Affymetrix 2005) arrays were chosen for the present study due to the commercial availability of arrays for each of the studied organisms.

The Affymetrix Platform

Oligonucleotide Probe Synthesis

Affymetrix GeneChips are ultra high density short oligonucleotide microarrays. Each feature, or gene probe set, is composed of multiple (usually 16) 25mer complimentary oligonucleotide sequences. As a quality control measure, Affymetrix also includes mismatch (MM) probes that are identical in sequence save for a single nucleotide mismatch. Probes are synthesized directly onto a silicone matrix in parallel via photolithographic protection chemistry (Deyholos and Galbraith 2001). Briefly, the silicone matrix is coated with a photosensitive linker molecule (Fig. 1-1 c). This linker molecule “protects” the array matrix from nucleotide addition. Ultraviolet radiation is used to de-protect specific loci on the array (Fig 1-1 d). Nucleotides are added to each de-protected spot in a coupling reaction (Fig 1-1 e). This sequence is repeated for each nucleotide until all probes have been synthesized.

Sample Preparation and Hybridization

The Affymetrix platform utilizes a single color detection format with absolute transcript quantification. This is in contrast with many other platforms such as cDNA microarrays that use a two color format with subsequent competitive hybridization and relative quantification (Brown and Botstein 1999). RNA targets are first reverse-transcribed into cDNA. First-strand cDNA serves as the template for transcription of the cRNA that ultimately serves as the target used during hybridization. Biotinylated nucleotides are

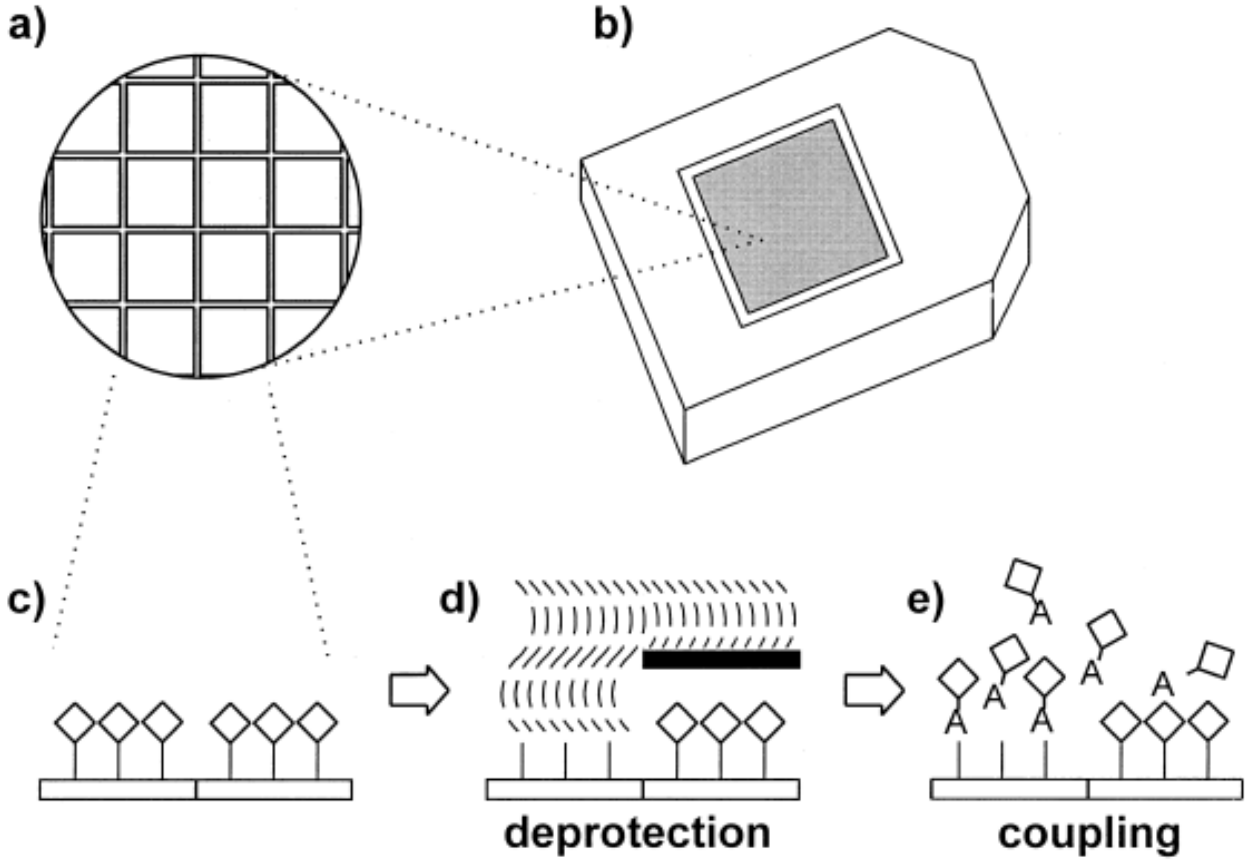


Figure 1-1: Photolithographic Oligonucleotide Probe Synthesis

incorporated into each cRNA molecule during the transcription reaction. After hybridization, the samples are then labeled with a streptavadin-fluorochrome conjugate dye (Deyholos and Galbraith 2001). Fluorescent signals are visualized using a laser excitement followed by detection with a charge-coupled device (CCD).

.CEL file generation

The final product of the physical experimentation is a high resolution digital bitmap image. As such, the “data” referred to in microarray studies are some function of the color saturation of the image locus corresponding to a given probe on the original chip. The signal intensity are calculated from this these saturation levels. A compressed version of the scanned image along with the calculated signal intensities are used to generate a .CEL file. All subsequent statistical calculations and analyses are performed on data obtained from the .CEL file.

Expression measure

Affymetrix GeneChips employ multiple short oligonucleotide probes to assay each gene target. Each set of 11-20 probes is roughly 20 base pairs in length. Each probe assays a different segment of the target gene. After background normalization and signal intensity calculations are performed on each probe, data from the probe set is condensed into signal intensity for that target gene.

Microarray Normalization

An important step in the analysis of microarray data is accounting for sources of variation. Generally, all sources of variation can be categorized as either obscuring or interesting variation. *Interesting variation* is the variation that is caused by the intended biological phenomenon being observed. In the case the present study, interesting variation would refer to the differences resulting from the actual changes in transcript abundance. All other sources of variation would be considered obscuring variation. These sources would include variation introduced by experimenter error, optical distortion by the scanning equipment, and inconsistencies in the array printing. The goal of normalization is to eliminate the effect of the obscuring sources of variation so that differences between samples are attributed only to the sources of interesting variation.

Scaling Normalization

In standard Affymetrix normalization, each array is scaled such that they all have the same average signal intensity. This occurs after probe set expression measures have been calculated. Calculations are performed sequentially on each array independently, thus it is less computationally intensive than the multi-array method employed by RMA and GCRMA. Additionally, this method of scaling normalization is useful when building databases of expression data for later comparison. Since the normalization is independent of all other microarrays experiments can be added piecewise over time without the need to re-normalize the entire database.(Affymetrix 2002; Lu 2004)

Background correction

An intrinsic limitation with microarrays is that they are a high-throughput hybridization-based technology. With any hybridization method accuracy, precision and specificity are all highest when stringency is optimized for an individual nucleic acid sequence with a discrete melting temperature and G-C content. When tens of thousands of targets are hybridized in tandem, the ability to optimize hybridization conditions is limited. Therefore, the possibility of non-specific hybridization is of major concern. Affymetrix addresses the cross-hybridization issue with measures physically implemented on the chip. For each probe targeting a specific transcript, or perfect match probe (PM), there is a corresponding mismatch probe (MM) which is identical in sequence with the exception of a single nucleotide substitution in the center. The default background correction for Affymetrix GeneChips uses these MM probes to adjust the PM intensity values. The details of this procedure will be discussed along with the section labeled *statistical algorithms.*

Statistical Algorithms

Normalization, background correction and signal intensity probe set signal calculations are three distinct facets of data acquisition but have been streamlined to such an extent that they are most appropriately discussed as a single process. Several mathematical algorithms have been developed to achieve these tasks (Affymetrix 2002; Irizarry, Bolstad et al. 2003; Harr and Schlotterer 2006) but only three will be described in any

detail in this text. Each of the algorithms discussed here, MAS5, RMA, and GCRMA, incorporate these three tasks. The characteristic differences among these algorithms

deals with their respective approaches to background correction, so this will be the primary focus of contrasts made between them here. Additionally, the scope of this project is limited to the application of these algorithms thus the discussion will be limited to the strengths and weaknesses of each method and the rational for choosing GCRMA over the other two.

The basic assumption governing each method of background correction is that the observed PM signal is inflated by sources of obscuring variation. In its simplest form this assumption is represented by the equation:

$$P=T+B \quad (1)$$

Where P is the signal intensity of the PM probe, T is the contribution made by the target mRNA, and B is background "noise" resulting from obscuring variation sources such as nonspecific hybridization and optical noise during image acquisition.

MAS5

Microarray Suite 5.0 or MAS5 is the current iteration of Affymetrix's default background correction and normalization package(Affymetrix 2002). The distinctive feature of MAS5 is the direct application of MM probe hybridization data to correct for non-specific target binding. Referencing equation 1, MAS5 treats MM signal as such:

$$M=f(B) \quad (2)$$

That is, that MM probe intensities (M) are a function of the background noise. Thus background correction is performed based solely on the information derived from each PM's corresponding MM signal intensity. For most probes this involves simply subtracting the MM signal intensity from the PM value. Irizarry et al 2003 reported that about 1/3 of the MM probes on any given chip had a higher overall binding efficiency than the corresponding PM probe. Also, MM signal generally includes some proportion of target signal as well. As such MM $\geq$ PM for 1/3 of probe pairs resulting in a null or negative corrected value. Affymetrix has responded to these findings by adjusting the background correction for probes in which this is the case. For cases in which MM> PM an idealized PM value is estimated based on the behavior of the entire probe set. To further ensure that no negative or null values result from background correction, physical "zones" are designated on the chip and minimum or "floor" values are assigned to each zone based on the performance of the probes in that zone.

With the introduction of GCRMA, it would seem that MAS5 background correction underperforms in every regard. However, it is still supported by Affymetrix and in the literature due to its simplicity, minimal computational requirements and flexibility in database applications (Affymetrix 2002; Pepper, Saunders et al. 2007).

RMA

Robust Multi-array Analysis (RMA) was introduced by Irizarry ET all 2003 as an alternative to MAS5. Several interesting observations motivated the development of this algorithm

1. Spike-in experiments suggest that MM probes detect target sequence as well as background noise. This causes PM-MM based correction methods to produce artificially truncated values.
2. The average difference between PM and MM values increased with target mRNA concentration.
3. PM-MM correction has a greater distortion effect on small PM values than larger ones, meaning that while MM values include target sequence, they are not sequence dependent.

Simply, they make two assumptions based on these observations:

1. On average, PM probe intensities are adequate representations of the target gene expression.
2. Target signal intensity follows an exponential distribution while background signal intensity follows a normal distribution.

Consequently, the RMA background correction ignores MM probes completely. Instead, the expected distributions of target and background signals are derived from the observed data and used to adjust the PM intensity values.

GCRMA

RMA achieves much better precision than MAS5 (Irizarry, Bolstad et al. 2003; Irizarry, Hobbs et al. 2003) , but with lower accuracy. Since RMA background subtraction is performed globally, using estimated distributions among all probes, there is no consideration given specifically to non-specific binding or probe-level effects. GCRMA addresses this issue by implementing probe-level background correction using the on-chip MM probes. The major feature that distinguishes GCRMA background correction from that of MAS5 is the additional consideration given to sequence specific binding affinities. GCRMA was motivated by 2 general observations:

1. MM probe intensities vary widely from sequence to sequence. Theoretically, MM probes do not match any target sequence; this suggests that there is an intrinsic difference in the binding affinity amongst the MM probes. This logic could also be extended to the PM probes.
2. There is a clear relationship between signal intensity and GC content of the probe sequence.(Zhang, Miles et al. 2003; Wu and Irizarry 2004)

From these observations, an assumption is made: The intrinsic variance attributed to non-specific hybridization is sequence specific, with the governing feature being GC

content of the probe. Accordingly, the GCRMA considers GC content when adjusting PM values using MM data. This is achieved by first using sequence data to group all probes according to their GC content and predicted hybridization properties. These groups are referred to as *pseudo*-MM. Next, PM values are adjusted according to the data collected from the MM and *pseudo*-MM probe sets. This method provides additional statistical power over MAS5 due to the larger set of probes used for correction, thus increasing precision. Additionally, the consideration given to non-specific binding provides increased accuracy over RMA (Wu, Irizarry et al. 2004). However; the adjusted values are still biased because both MM and *pseudo*-MM probes detect target signal as well as cross-hybridization.

Statistical Tests for Identifying Differentially Expressed Genes

The statistical significance of a given expression ratio, as measured by a pair wise analysis of variance (ANOVA), is expressed as that ratio's "p-value". This number represents the probability that the observed expression ratio would have occurred by chance alone. It is appropriately interpreted as the strength of evidence for rejecting the null hypothesis. That is, a p-value of 0 indicates that the null hypothesis is false and that the observed changes are due to a measurable source of variation. The problem with a standard ANOVA is that these are performed on a per-gene basis. When thousands of these tests are performed, the type 1 error rate (the occurrence of falsely rejecting a null hypothesis), even based on small pvalues, increases. For example, if a statistical significance threshold of p=0.05 is used, 500 false positive results would be expected by

chance alone from a microarray study involving 10,000 probes. Multiple test correction can limit the occurrence of false positive results in large datasets.

Bonferroni Multiple Test Correction

The Bonferroni method of multiple test correction controls the probability of making a single type 1 error among all observations. It is considered to be the simplest yet most conservative method of multiple test correction. It assumes that all tests are independent (which is often not the case for microarray data) and calculates adjusted significance scores as a product of a given p-value and the total number of observations being made. Values above 1 are rounded down to 1. To use the example given above, a pvalue of 0.05 would result in an adjusted value of 1 meaning that the null hypothesis would not be rejected. The highest acceptable p-value in this example would be $p=10^{-4}$ The Bonferroni method is often the least powerful and thus inappropriate for many analyses. It may not be necessary to control false positives among *all* tests. It may be sufficient to simply limit the occurrence of false discoveries among a selected subset of the tests.

Controlling the False Discovery Rate (FDR)

FDR correction is a compromise between the ANOVA significance test and the Bonferroni method. Rather than to prevent any type 1 errors, FDR is only concerned with defining the statistical significance threshold such that it controls the occurrence of

false positive results among tests falling below that threshold. Two different FDR corrections were utilized in this study.

The Hochberg or “step-up” correction was used for the bacterial expression data. This is achieved by first ranking some number (n) of pvalues in ascending order. The correction is calculated as $FDR \geq P_{(m)}*(n/m)$ where P is the value to be adjusted, n is the number of pvalues being adjusted and m is the rank of the pvalue being adjusted. A Holm or “step-down” correction was applied to the yeast expression data. After ranking all p-values in ascending order, a correction is performed according to the equation $FDR \geq P_{(m)} *(n-m+1)$. In either case, the adjusted threshold now reflects the probability of falsely rejecting the null hypothesis within the sample size chosen for further analysis. The difference between the two methods is that the Hochberg method is more powerful since it increases (steps up) in stringency as each p-value is tested. The Holm method results in fewer genes passing the statistical threshold since it starts at maximum stringency and decreases (steps down) its stringency as each successive p-value is tested.

Computational Class Prediction

Microarray studies are often criticized for bearing the “curse of dimensionality”. This describes a situation in which the expression values of thousands of genes are used to compare just a few biological samples (Radmacher, McShane et al. 2002). This is particularly a problem when expression data is applied to diagnostics, sample identification and risk assessment. In these cases, expression profiles resulting from a

particular condition (class) of interest are compared to expression profiles of a larger class made up of many different conditions. These groups make up a *training set* from which patterns of gene expression characteristic of the condition of interest are identified. The large numbers of genes involved often result in patterns that are characteristic to the differences in that particular training set rather than to the broader biological significance of the observed condition. This leads to *overfitting* in which the resulting classification schema fits the training set perfectly but may not be useful for classifying new samples (Breiman 2001). If not controlled, spurious conclusions may pass even the most stringent statistical criteria. Limiting the number of genes assayed may mitigate much of the risk of overfitting the data. Ideally the number of samples should be much greater than the number of genes used to characterize them (Lee 2008).

Selection of Genes for Class Prediction

There are several methods of gene selection with the simplest and most common being based on univariate significance criteria. For each gene, an expression ratio is calculated between the classes in the training set. Genes are selected for inclusion in the classifier if they have a t-statistic or α value that satisfies the desired significance level. This straightforward approach ensures that all genes in the classifier have statistically significant fold changes between classes. The number of genes selected cannot be directly controlled using univariate significance; rather the target significance threshold acts as a tuning parameter for indirectly adjusting the number of genes. It is important to note that there is not often a linear relationship between this tuning

parameter, the number of genes, and the performance of the final classifier. This is because the class predictions for all methods described here are based on multivariate comparisons not the univariate statistics of individual genes.

Multivariate approaches to gene selection result in lists of genes that perform well collectively. Typically, the number of genes to be included in the classifier is pre-defined and the algorithm will select the best set of genes, in terms of predictive capacity, to populate the classifier. The method used in the present study is recursive feature elimination (RFE) (Guyon, Weston et al. 2002). The classifier is initially composed of all genes in all profiles. For each class, RFE first ranks all genes according to their weight in determining the class distinctions. The lowest ranked gene is removed from the classifier and a new set of prediction rules along with new weight and ranks are assigned to the remaining genes. This is repeated until the pre-determined number of genes is reached

Mathematical Models for Class Prediction Rules

Compound Covariate Prediction (CCP)

Compound covariate prediction is based on weighted log expression values. First, t-statistics are calculated each log expression ratio. A weight value is calculated for each gene as the t-statistic given a sign dependent on the correlation of that genes expression to the experimental conditions. In other words, if high expression of a gene is indicative of condition A, its weight will have a positive sign. Genes in which high expression is indicative of condition B will have a negative sign. Genes are selected for

inclusion in the predictor and a compound covariate value is calculated for each condition as the inner sum of the weighted expression ratios. The prediction threshold is defined as the mean compound covariate value of the two conditions being compared. To identify an unknown sample, its compound covariate value is calculated using the selected genes and is identified as the condition to which it is closest in value. (Radmacher, McShane et al. 2002)

Support Vector Machines (SVM)

Support vector machines prediction is based on a linear discriminant function. First, each expression profile in the training set is plotted as a vector of log expression values then a linear function is established to separate the two classes. This function is defined such that it maximizes the distance between the two closest expression profiles from each class in the training set. That is, the two worst classified samples in each class. These samples are known as the support vectors. This serves as the threshold by which new samples will be identified. Each gene in the classifier is then assigned a weight score based on its overall contribution to that linear function. The prediction rule is made up of the list of genes, their corresponding weight scores and the linear threshold.

To identify an unknown sample, the signal intensity for each gene is multiplied by that genes weight score in the classifier. The sum of these products, known as the inner sum, is calculated for all genes in the classifier. The unknown sample is then compared to the linear threshold. A class is identified depending on which side of that linear

function it falls on the coordinate hyper plane; that is, samples which meet or exceed the threshold value are identified as a “positive” sample for that classifier.(Chih-Wei Hsu 2009)

Nearest Neighbors

Nearest neighbors prediction involves plotting each expression profile as a vector of the log expression values for each gene in the classifier. Unknown samples are plotted and the Euclidean distance is calculated from all other profiles in the training set. The new sample is identified based on the profile to which it is closest. A modification of this method, termed k-nearest neighbors, uses a weighted vote among some number (k) of the nearest profiles. The identity of profiles which lie closest to the unknown sample will weigh more in the identification than those further away.

Nearest Centriod

Nearest centroid prediction is similar to nearest neighbors. Each class is plotted as a vector of average expression values for each gene among all profiles in the training set belonging to that class. This average vector is known as the centroid. The unknown sample is classified based on the centroid to which it is closest.

A nearest centroid prediction can be applied in such a way that it also results in gene selection. For each gene, the class average is adjusted toward the average for that gene among all classes. This results in “shrinkage” of the Euclidean distance between class centroids. As this distance decreases, the effect of genes that had

original values close to the average becomes negligible. In this application, the shrinkage factor becomes a tuning parameter for determining the number of genes to be included in the final classifier (Tibshirani, Hastie et al. 2002). New samples are then predicted based on these new centroid distances. This method is known as nearest shrunken centroid and is the prediction method employed by the popular Prediction Analysis of Microarrays (PAM).

Objectives of the study

This investigation focuses on the problem of energetic materials produced by IED and terrorist, principally, Triacetone triperoxide (TATP), 2, 4, 6-trinitrotoluene (TNT) and a related compound 2, 6-dinitrotoluene (DNT). It is of direct interest of this research to test the hypotheses that microbial exposure to energetic materials will result in distinct patterns of gene expression, and that these profile patterns of gene expression are reproducible biomarkers of such exposure in microbial populations. A second goal of this investigation is to produce a knowledge base and fundamental understanding of the gene expression response profiles in exposed populations. Such a mechanistic understanding is important not only in developing more effective strategies for real-time biosensor technology but also from a toxicogenomic perspective.

Hypotheses

I. Incubation of microbial cultures with each of the energetic materials TNT, DNT, and TATP will result in distinct, characteristic patterns of gene expression.

II. Profile patterns of gene expression are reproducible biomarkers exposure to energetic materials in microbial populations.

References

Affymetrix (2002). "Statistical Algorithms Description Document."
Affymetrix (2005). GeneChip Expression Analysis Technical Manual. Santa Clara, CA.
Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5-32.
Brown, P. O. and D. Botstein (1999). "Exploring the new world of the genome with DNA microarrays." Nature Genetics **21**: 33-37.
Caballero, A., A. Esteve-Nunez, et al. (2005). "Assimilation of Nitrogen from Nitrite and Trinitrotoluene in Pseudomonas putida JLR11." J. Bacteriol. **187**(1): 396-399.
Caballero, A. and J. L. Ramos (2006). "A double mutant of Pseudomonas putida JLR11 deficient in the synthesis of the nitroreductase PnrA and assimilatory nitrite reductase NasB is impaired for growth on 2,4,6-trinitrotoluene (TNT)." Environmental Microbiology **8**(7): 1306-1310.
Cenas, N., A. Nemeikaite-Ceniene, et al. (2001). "Quantitative structure-activity relationships in enzymatic single-electron reduction of nitroaromatic explosives: implications for their cytotoxicity." Biochimica Et Biophysica Acta-General Subjects **1528**(1): 31-38.
Chih-Wei Hsu, C.-C. C., and Chih-Jen Lin (2009). A Practical Guide to Support Vector Classification Technical Report. Taipei, National Taiwan University.
de Longueville, F., V. Bertholet, et al. (2004). "DNA microarrays as a tool in toxicogenomics." Combinatorial Chemistry & High Throughput Screening **7**(3): 207-211.
Deyholos, M. K. and D. W. Galbraith (2001). "High-density microarrays for gene expression analysis." Cytometry **43**(4): 229-238.
Esteve-Nunez, A., A. Caballero, et al. (2001). "Biological degradation of 2,4,6-trinitrotoluene." Microbiology and Molecular Biology Reviews **65**(3): 335-+.
Gatzidou, E. T., A. N. Zira, et al. (2007). "Toxicogenomics: a pivotal piece in the puzzle of toxicological research." Journal of Applied Toxicology **27**(4): 302-309.
Gong, P., X. Guan, et al. (2007). "Transcriptomic analysis of RDX and TNT interactive sublethal effects in the earthworm Eisenia fetida." Bmc Genomics **9**.
Gong, P., X. Guan, et al. (2007). "Toxicogenomic analysis provides new insights into molecular mechanisms of the sublethal toxicity of 2,4,6-trinitrotoluene in Eisenia fetida." Environ Sci Technol **41**(23): 8195-8202.
Gonzalez-Perez, M. M., P. van Dillewijn, et al. (2007). "Escherichia coli has multiple enzymes that attack TNT and release nitrogen for growth." Environmental Microbiology **9**(6): 1535-1540.
Guyon, I., J. Weston, et al. (2002). "Gene selection for cancer classification using support vector machines." Machine Learning **46**(1-3): 389-422.
Harr, B. and C. Schlotterer (2006). "Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons." Nucl. Acids Res. **34**(2): e8-.
Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of affymetrix GeneChip probe level data." Nucleic Acids Research **31**(4).
Irizarry, R. A., B. Hobbs, et al. (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-264.
Jenkins, T. F., A. D. Hewitt, et al. (2006). "Identity and distribution of residues of energetic compounds at army live-fire training ranges." Chemosphere **63**(8): 1280-1290.
Johnson, M. S., J. W. Ferguson, et al. (2000). "Immune effects of oral 2,4,6-trinitrotoluene (TNT) exposure to the white-footed mouse, Peromyscus leucopus." International Journal of Toxicology **19**(1): 5-11.
Kuhn, K., S. C. Baker, et al. (2004). "A novel, high-performance random array platform for quantitative gene expression profiling." Genome Research **14**(11): 2347-2356.

Lee, S. (2008). "Mistakes in validating the accuracy of a prediction classifier in high-dimensional but small-sample microarray data." Statistical Methods in Medical Research **17**(6): 635-642.
Lemieux, B., A. Aharoni, et al. (1998). "Overview of DNA chip technology." Molecular Breeding **4**(4): 277-289.
Lu, C. (2004). "Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays." BMC Bioinformatics **5**.
Nemeikaie-Ceniene, A., J. Sarlauskas, et al. (2004). "Enzymatic redox reactions of the explosive 4,6-dinitrobenzofuroxan (DNBF): implications for its toxic action." Acta Biochimica Polonica **51**(4): 1081-1086.
Nuwaysir, E. F., M. Bittner, et al. (1999). "Microarrays and toxicology: The advent of toxicogenomics." Molecular Carcinogenesis **24**(3): 153-159.
Pennington, J. C. and J. M. Brannon (2002). "Environmental fate of explosives." Thermochimica Acta **384**(1-2): 163-172.
Pepper, S., E. Saunders, et al. (2007). "The utility of MAS5 expression summary and detection call algorithms." BMC Bioinformatics **8**(1): 273.
Radmacher, M. D., L. M. McShane, et al. (2002). "A paradigm for class prediction using gene expression profiles." Journal of Computational Biology **9**(3): 505-511.
Ragoussis, J. (2009). "Genotyping Technologies for Genetic Research." Annual Review of Genomics and Human Genetics **10**: 117-133.
Reddy, G., S. A. M. Chandra, et al. (2000). Toxicity of 2,4,6-trinitrotoluene (TNT) in hispid cotton rats (Sigmodon hispidus): Hematological, biochemical, and pathological effects, Taylor & Francis Ltd.
Roldan, M., E. Perez-Reinado, et al. (2008). "Reduction of polynitroaromatic compounds: the bacterial nitroreductases." Fems Microbiology Reviews **32**(3): 474-500.
Schena, M., D. Shalon, et al. (1995). "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray." Science **270**(5235): 467-470.
Sylvia, J. M., J. A. Janni, et al. (2000). "Surface-enhanced Raman detection of 1,4-dinitrotoluene impurity vapor as a marker to locate landmines." Analytical Chemistry **72**(23): 5834-5840.
Tennant, R. W. (2002). "The national center for toxicogenomics: Using new technologies to inform mechanistic toxicology." Environmental Health Perspectives **110**(1): A8-A10.
Tibshirani, R., T. Hastie, et al. (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression." Proceedings of the National Academy of Sciences of the United States of America **99**(10): 6567-6572.
Williams, R. E., D. A. Rathbone, et al. (2004). "Biotransformation of explosives by the old yellow enzyme family of flavoproteins." Applied and Environmental Microbiology **70**(6): 3566-3574.
Wittich, R.-M., J. L. Ramos, et al. (2009). "Microorganisms and Explosives: Mechanisms of Nitrogen Release from TNT for Use as an N-Source for Growth." Environmental Science & Technology **43**(8): 2773-2776.
Wu, Z. and R. A. Irizarry (2004). "Preprocessing of oligonucleotide array data." Nat Biotech **22**(6): 656-658.
Wu, Z. J., R. A. Irizarry, et al. (2004). "A model-based background adjustment for oligonucleotide expression arrays." Journal of the American Statistical Association **99**(468): 909-917.
Yin, H., T. K. Wood, et al. (2005). "Reductive transformation of TNT by Escherichia coli resting cells: kinetic analysis." Applied Microbiology and Biotechnology **69**(3): 326-334.
Zarlpov, S. A., A. V. Naumov, et al. (2002). "Models of 2,4,6-trinitrotoluene (TNT) initial conversion by yeasts." Fems Microbiology Letters **217**(2): 213-217.

Zhang, L., M. F. Miles, et al. (2003). "A model of molecular interactions on short oligonucleotide microarrays." Nat Biotech **21**(7): 818-821.

Chapter 2 : Gene Expression Profiling

Introduction

Oligonucleotide microarrays provide a platform for measuring the abundance of specific nucleic acid sequences in a population of cells. This is especially useful for assessing the composition of the mRNA pool for a given cell population and comparing the abundance of specific transcripts among different populations. In practice, changes in gene expression can be associated with different physiological conditions or alterations in culture conditions.

Oligonucleotide array technology was applied to quantify changes in gene expression resulting from exposure to 5 different chemicals in relation to a solvent control. The 5 compounds chosen were 2,4,6- trinitrotoluene (TNT) , 2,6-dinitrotoluene (DNT) , triacetone-triperoxide (TATP), hydrogen peroxide (H2O2), and a 50% v/v mixture of H2O2 and acetone (MIX). Chemical structures of each compound are presented in Figure 2-1.The choice of chemicals provides a range of chemical properties and potential physiological and metabolic interactions. There are, for example, compounds that are chemically, and potentially biologically similar to TNT (DNT) as well as chemicals that are chemically, potentially biologically distinct from TNT (TATP, H2O2, MIX). TNT and DNT represent nitrogen rich aromatic hydrocarbons, TATP is insoluble, chemically unstable, cyclical organic peroxide, and Hydrogen peroxide is a source of reactive oxygen leading to oxidative stress.

The major hypothesis being tested is as follows:

I. Incubation of microbial cultures with each of the energetic materials TNT, DNT, and TATP will result in distinct, characteristic patterns of gene expression.

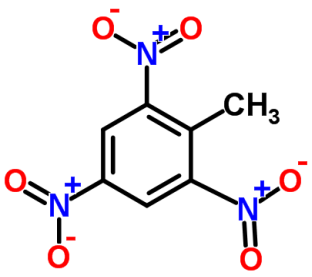
Related to the biochemical rationale for the selected compounds, each compound serves as controls for several null sub-hypotheses:

TNT

Null Hypothesis 1: *The observed responses to TNT will not be indicative of TNT exposure but rather a characteristic response to exposure to nitroaromatic compounds.*

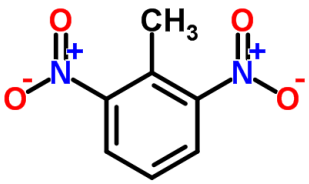
Null Hypothesis 2: *The observed response to TNT will not be specific to TNT but rather a characteristic response to exposure to energetic materials.*

Null Hypothesis 3: *The observed responses to TNT will not be specific to TNT but rather a generic response to chemical stimulus.*



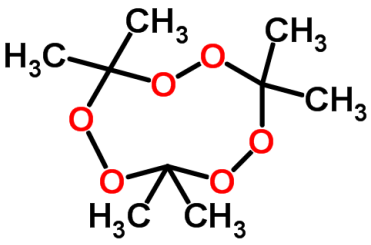
2,4,6 Trinitrotoluene

(TNT)



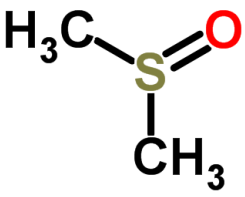
2,6-Dinitrotoluene

(DNT)



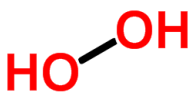
Triacetone-triperoxide

(TATP)



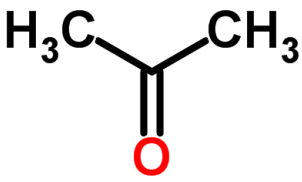
Dimethylsulfoxide

(DMSO)



Hydrogen Peroxide

(H_2O_2)



Acetone

Figure 2-1: Chemicals Used to Generate Gene Expression Profiles

Null hypothesis 1 will be confirmed if the TNT induced changes in expression are identical to those observed upon DNT exposure. The TATP induced changes will serve as a measure of the response to energetic materials thereby testing null hypothesis 2. Finally, an observation that the TNT induced changes in transcription are identical to two or more of the other chemicals would be cause for further investigation of null hypothesis 3.

TATP

Currently there is no published literature describing the biological relevance of TATP. Without any prior knowledge or expected results, it will be difficult to interpret the results of the TATP exposure. Therefore, in addition to the biochemical, and null hypothesis rationales stated for the TNT exposures, the two exposures H2O2 and MIX will aid interpretation of the results from TATP exposure. The appropriateness of these chemicals for that purpose has yet to be empirically confirmed but in such cases logical reasoning prevails. Preliminary efforts to elucidate the chemical nature of TATP revealed that it is virtually insoluble in both water and DMSO. It was also observed that, at biologically relevant temperatures, it rapidly dissipates, typically losing nearly half of its mass within 24 hours. These observations are consistent with the literature (Bellamy 1999; Matyas, Pachman et al. 2009). It was hypothesized three likely outcomes of TATP exposure would be observed:

1) The exposed populations will interact with intact TATP molecules.

2) The exposed populations will interact with chemicals resulting from chemical transformation of TATP.

3) The exposed populations will have no interaction with TATP.

H2O2 and MIX represent two possible components of chemical transformation of TATP. No controls for the interaction with the dimer and tetramer products of TATP transformation were considered because a) the synthesis of TATP inevitably results in formation of these compounds as byproducts, thus all TATP samples will contain some amount of them, and b) in practice, there would be little motivation to distinguish TATP from its byproducts as presence of any one would have no legitimate purpose.

The primary focus of the present study is elucidating the characteristic alterations in gene expression observed in *Escherichia coli* cultures exposed to TNT and TATP. *E.coli* is *the* model bacterial species, it represents a potential tractable model for gene regulation, metabolic perturbation, and physiological stress associated energetic compound exposure. In addition to the work with *E. coli,* gene expression profiles were developed for 3 other organisms, each presenting a unique opportunity for modeling energetic compound exposure. *Pseudomonas putida,* a soil bacterium, has been characterized by its versatile metabolic potential. Several studies have employed its use in biodegradation and remediation of toluene and polystyrene (Nelson, Weinel et al. 2002; Ward, Goff et al. 2006). Further, *E. coli* and pseudomonades including *P. putida* have been shown to partially degrade TNT and DNT as well as utilize its nitro groups as a nitrogen source (Spanggord, Spain et al. 1991; Duque, Haidour et al. 1993; Esteve-Nunez, Lucchesi et al. 2000; Esteve-Nunez, Caballero et al. 2001; Caballero, Esteve-

Nunez et al. 2005; Stenuit, Eyers et al. 2006; Gonzalez-Perez, van Dillewijn et al. 2007). *Saccharomyces cerevisiae* was profiled as well as a representative as lower eukaryotes. *S. cerevisiae* serves much the same role as *E.coli;* it is a tractable model that is well characterized. Numerous sources of annotation, gene regulation and metabolic data are available and it is a fast growing, non-pathogenic eukaryote. There is interest in deciphering the signature alterations in gene expression upon exposure to these compounds. These interests are related to clinical diagnostics, toxicology as well as biomarker development for security and military applications.

All expression data for microarray experiments presented in this document are presently deposited in the University of Tennessee Microarray Database (UTMD: genome.ws.utk.edu). Additional deposits will be made to the NCBI's Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) upon publication of manuscripts prepared from this study.

Escherichia coli

Materials and Methods

Growth and exposure media

To limit possible effects of the growth media on gene expression, a minimal salts media was prepared to serve as both the growth and exposure media.

Minimal Salts Media (MSM) 500µM KH2PO4, 10mM K2HPO4, 831µM MgSO4 ·7H2O, 125µM NH4NO3, 55.5µM glucose, 0.01% trace metals solution

Trace metals solution: 10g/L MgO, 2g/L CaCO3, 5.4g/L FeCl3 ·6H20, 1.44g/L ZnSO4 · 7H20, 250mg/L CuSO4, 62mg/L H3BO4, 490mg/L NaMoO4 · 2H20

Bacterial cultures and treatment

E. coli K12 (ATCC #29425) was grown from freezer stock in Luria-Bertini broth overnight. Revived cultures were used to inoculate MSM and grown overnight at 37℃ with 200rpm orbital shaking. Overnight cultures of *E. coli* were diluted in 37℃ MSM and grown to an $O.D._{600}$ of 0.5 at 37℃ with shaking. The cultures were then sp lit into 6mL aliquots and spiked with either TNT; 2,6-DNT; Peroxide; Triacetone-Triperoxide (TATP); a solution (MIX) containing 50% acetone, 15% hydrogen peroxide and 35% H20; or a vehicle control (DMSO). An exposure concentration of 220µM TNT, 275µM DNT, 244µM peroxide, 750µM TATP, and 0.1% MIX was achieved. All cultures contained 0.008% DMSO by volume. TNT and DNT concentrations were determined, through growth curves, as effective, sub-lethal concentrations. TATP proved to be

completely insoluble in both DMSO and the growth media, so 1mg of freshly synthesized TATP crystals was measured and added directly to the growth media resulting in a final concentration of 732 µM. There was no data available that would predict how the TATP may act in solution or once exposed to the bacterial culture. As such, the MIX exposure served as a control for the abiotic decomposition of the TATP crystals. The MIX concentration was determined based on percent volume of each of the TATP synthesis compounds diluted into the DMSO vehicle. Total cellular RNA was extracted from each culture 1 hour after spiking.

Growth Curves

E. coli K12 (ATCC #29425) was grown from freezer stock in Luria-Bertani broth overnight. Revived cultures were used to inoculate MSM and grown overnight at 37℃ with 200rpm orbital shaking. Overnight cultures of *E. coli* were diluted in 37℃ MSM and grown to an $O.D._{600}$ of 0.1 at 37℃ with shaking. The cultures were then split into 1mL aliquots and distributed, in triplicate, into a 24 well microtiter plate. Each well was then spiked with TNT over a range of concentrations from 0-70mg/L. The cultures were incubated at 37 ℃ with continuous orbital shaking. Growth was measured as an increase in absorbance at 600nm. Measurements were taken every 15 minutes over a 24 hour period. This assay was repeated for DNT and TATP. Preliminary absorbance measurements indicated that TNT causes a measurable difference in sterile media. To control to effect of TNT concentration, the absorbance of sterile controls containing TNT were subtracted from the growth curve measurements.

Determination of Appropriate Exposure time

It was determined from the preliminary growth curve experiments that there is a positive correlation between TNT concentration and absorbance at 600nm over time. The results from the growth curves described here were also used to determine the appropriate exposure duration. The correlation between TNT concentration and absorbance was calculated for each time point. The selected exposure duration was approximately half of the earliest time point in which a positive correlation above 90% was observed.

RNA Isolation

The TNT, DNT, and DMSO RNA samples were comprised of total cellular RNA collected from a total of 6 biological replicates on 2 separate dates. The TATP, peroxide and MIX RNA samples included 3 biological replicates all collected on the same date. Isolation was performed using the RNeasy Protect Bacteria Mini Kit (Qiagen, Valencia, CA) according to manufacturer's instructions. The recommended RNAse-free DNAse treatment was performed on the samples prior to final elution.

cDNA synthesis, labeling, hybridization, staining and array scanning

This experiment involved hybridization of 6 single color arrays for each of the TNT, DNT , and DMSO samples and 3 single color arrays for each of the peroxide, TATP and MIX samples. All processing of isolated RNA including cDNA synthesis, labeling,

hybridization, staining and array scanning were performed by the UT Affymetrix Core facility according to Affymetrix standardized procedures for *E.coli* 2.0 GeneChip arrays. Labeled and fragmented cRNA was then hybridized to Affymetrix *E. coli* 2.0 arrays. A detailed protocol is available (Affymetrix 2005).

Gene Expression Measurements

Raw array image data was collected in CEL format using Affymetrix gene chip operating software (ver. 3.4.2152.32776). The raw CEL files were imported to Partek Genomics Suite (Partek 2008). Guanine-cytosine robust multi-array analysis (GC-RMA) was applied to all arrays to achieve background subtraction and normalization across all arrays. The data were then adjusted to remove any variation due to batch effects then the averages were calculated among each treatment type. All statistical analysis performed in Partek Genomics Suite are detailed in the Partek Genomic Suite online manual (Partek 2008).

Functional Analysis of Significant gene changes

Gene ontology data for each of the significant genes were retrieved from EcoCyc (Keseler, Bonavides-Martinez et al. 2009). A gene ontology enrichment analysis was performed using the retrieved data using the Gene Ontology Enrichment Analysis Toolkit (omicslab.genetics.ac.cn) using AmiGo source version OBO v.1.2 (Ashburner, Ball et al. 2000)

Quantitative Reverse Transcriptase PCR

Quantitative Reverse Transcription Polymerase Chain Reaction (Q-RTPCR) was performed to validate the results of the microarray experiment. Transcript abundance was quantified using the non-specific DNA binding dye SYBR Green. The reactions were performed in a thermocycler equipped with a Chromo4 fluorescence detection unit (MJ Research Inc., Waltham, MA) Three genes were chosen from the list of differentially expressed genes resulting from TNT exposure. These genes were azoR, soxS, and nhoA.

Primer Design

Each primer pair was designed to target a 150-250 bp region of the indented transcript sequence. Sequences for each transcript were obtained from the National Center for Biotechnology Information (NCBI). Primers were designed using Primer 3 software. Primer sequences are provided in **Error! Reference source not found.**.

Verification of Primer Specificity

The Primer-BLAST feature of the NCBI primer design tool was used to verify the specificity of each primer against the *Escherichia coli* K12 non-redundant nucleotide database (taxid: 83333). To further verify the specificity of the q-rtpcr primers, standard polymerase chain reaction was performed using genomic DNA as a template.

Table 2-1 Q-RTPCR detection primers and sequences

Target Gene	Primer Sequences	Amplicon Size
azoR	F: 5' CTTTCCGATGAGTTGATTGCC 3'	173bp
	R: 5' TTACCCGTTACCAGACCTTCC 3'	
soxS	F: 5' TCAGACGCTTGGCGATTACA 3'	150bp
	R: 5' TCAAACTGCCGACGGAAAA 3'	
nhoA	F: 5' TGCGAGCAGCAACAAAGC 3'	236bp
	R: 5' TCCACGCCCAGACCAAA 3'	

Table 2-2 Full length gene Amplification primers and sequences

Target Gene	Primer Sequences	Amplicon Size
azoR	F:5' AACAAGCAACGGGGCATC 3'	731bp
	R:5' GCTGAGATTATGGGAAAACAGG 3'	
soxS	F:5' TGCGTTTCGCCACTTCG 3'	576bp
	R:5' GCCAGGGATGGTTCTTTGC 3'	
nhoA	F:5' GAGAAAACCACTAAGGGAAACG 3'	940bp
	R:5' CAGGTCTACAACCGGGCTAA 3'	

Preparation of cDNA plasmid standards

cDNA plasmid standards were prepared to facilitate optimization of the q-rtpcr protocol as well as to provide the standard curve for quantification. Full length coding regions of each gene were amplified using primers complimentary to genomic sequences flanking the 5' and 3' ends of the gene sequence. (**Error! Reference source not found.**) The amplification products were visualized using gel electrophoresis through 1% agarose gel. Resulting band size was compared to expected fragment length as verification of product amplification. The amplification products were cloned into the PCR2.0 vector using a TOPO TA cloning kit (Invitrogen, Carlsbad, CA) according to the manufacturer's instructions. Plasmid DNA was extracted from the resulting clones using Wizard Plus MiniPreps kit (Promega, Madison, WI) according to the manufacturer's instructions. Purified plasmid DNA was subjected to restriction digest using EcoRI restriction endonucleases. The products of the restriction digest were visualized using gel electrophoresis through 1% agarose gel. Plasmids containing inserts of the appropriate size were delivered to the Molecular Biology Resource Facility at the University of Tennessee for sequencing via the Sanger method. Standard M13 primers were used for the sequencing reactions. Sequencing results were uploaded to the NCBI Basic Local Alignment Search Tool (BLAST) for identification.

Clones bearing plasmids with sequences matching the intended sequences were grown overnight and plasmids were purified using Wizard Plus MiniPreps kit (Promega, Madison, WI) according to the manufacturer's instructions. An additional RNAse

treatment insured that no contaminating RNA was present in the final elution. Plasmids were eluted in 30 µL of nuclease free water.

Total DNA was quantified for each plasmid preparation using a DyNa Quant 200 flourometer and Hoechst dye. Plasmid concentration was calculated as total DNA concentration divided by the combined mass of the vector and insert. Plasmid standards were prepared via serial dilutions of an initial preparation with a plasmid concentration of 10^9 plasmids per µL.

Optimization of Q-RTPCR Assay

To ensure accurate quantification during the Q-RTPCR analysis, the reaction conditions were optimized for each gene. The optimization included adjustments to the primer concentrations, annealing temperature, extension time, and the point at which measurements were taken. To determine optimum conditions, cDNA plasmids standards of known concentrations were quantified. Assays were performed over a range of primer concentrations as well as a range of annealing temperatures and extension times. It was determined that for all three assays, the optimum primer concentration is 0.8µM. The optimum annealing temperature is 56℃, 58℃, and 60℃ for azoR, soxS, and nhoA respectively. The optimum extension time for all assays is 0.5 minutes. Typically, fluorescence is measured after the extension step; however it was observed that the R^2 value of the standard curve could be improved if an additional 80℃ incubation step was added before the fluorescence m easurements were taken. The dissociation curve revealed that the amplification products of each assay had a

dissociation peak near 80℃. It was reasoned that, at tha t temperature, most products of non-specific binding will denature and not skew the fluorescence measurement.

Q-RTPCR quantification

Q-RTPCR experiments were performed in triplicate. The abundance of each mRNA transcript was quantified by plotting the threshold cycle (C_T) along the standard curve. Transcript abundance was then approximated as copies per µL of RNA solution. Calculated transcript abundance was normalized to total RNA for each reaction to correct for differences in RNA extraction efficiency between samples, thus all values are reported as copies per ng of total RNA.

Results

Growth Curves

None of the compounds assayed, TNT, DNT, or TATP, had adverse effects on growth of *E.coli* K12 as measured by optical density at 600nm. However; it appeared initially that there that TNT had a positive effect on growth. Upon re-evaluation, the observed differences in optical density were attributed to discoloration of the media associated with increasing TNT concentrations Figure 2-2. These changes in absorbance were not observed in the sterile controls that had been spiked with TNT. Therefore, although the data should not be considered a measure of growth effects, they do serve as a measurement a biological response. From the growth curve experiments, it was determined that 1 hour is the appropriate exposure duration and that a concentration of 50mg/L (50ppm or 220 µM) is sufficient to cause a measurable biological effect.

Quantitative Reverse Transcriptase PCR

For TNT exposed cultures, the results of the QRTPCR results for all three genes were in agreement with those observed in the microarray experiments. The assays indicated increased transcription of the gene nhoA DNT exposure although this was not observed in the microarray experiments. It is likely that the RTQPCR assay is more sensitive than the microarray measurements because each assay was optimized specifically for each target. Another possible explanation, which would account for the increase in observed fold changes, involves the GCRMA algorithm used to normalize

and background-adjust the microarray experiments. It has been documented that the incorporation of mismatch probe values for non-specific signal correction can lead to truncated values (Irizarry, Bolstad et al. 2003).

Gene Expression Profiles

In order to determine the gene expression changes resulting from exposure to each compound, total RNA was harvested from *E. coli* cultures following 60 minutes of exposure. The harvested RNA was fluorescently labeled and hybridized to an Affymetrix *E.coli* 2.0 GeneChip. The *E.coli* 2.0 GeneChip contains 10208 sets of sequence specific features (probe sets) consisting of 15-20 different sequences (probes) matching a target mRNA sequence.

The microarray experiments produced lists of GeneChip features and their corresponding signal intensities (Appendix). All values presented in this section are calculated fold changes based on averages of those signal intensities. For any given compound exposure, the average value of each GeneChip feature was divided by the corresponding average value in the control experiment. In the case of genes with decreased expression, negative inverse values are reported.

Statistical significance was determined for each feature. A one-way analysis of variance test (ANOVA) was first applied to each experiment against the DMSO control. The results of the ANOVA are a measure of the probability that the observed expression ratio is not due to chance, but rather actual changes in expression. The resulting p-values were then adjusted to reflect the false discovery rate (FDR) for each comparison.

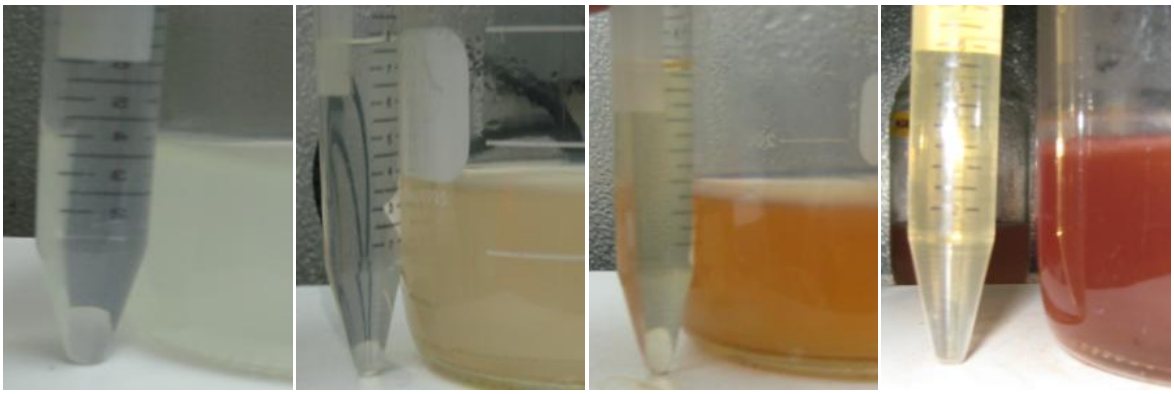


Figure 2-2: Growth in TNT results in a dose-dependent discoloration of the media

***E.coli* K12 was grown in MSM growth media at 37℃ and aera ted with a magnetic stir bar. In each picture is the bacterial culture on the right and the sterile control containing the same concentration of TNT on the left.**

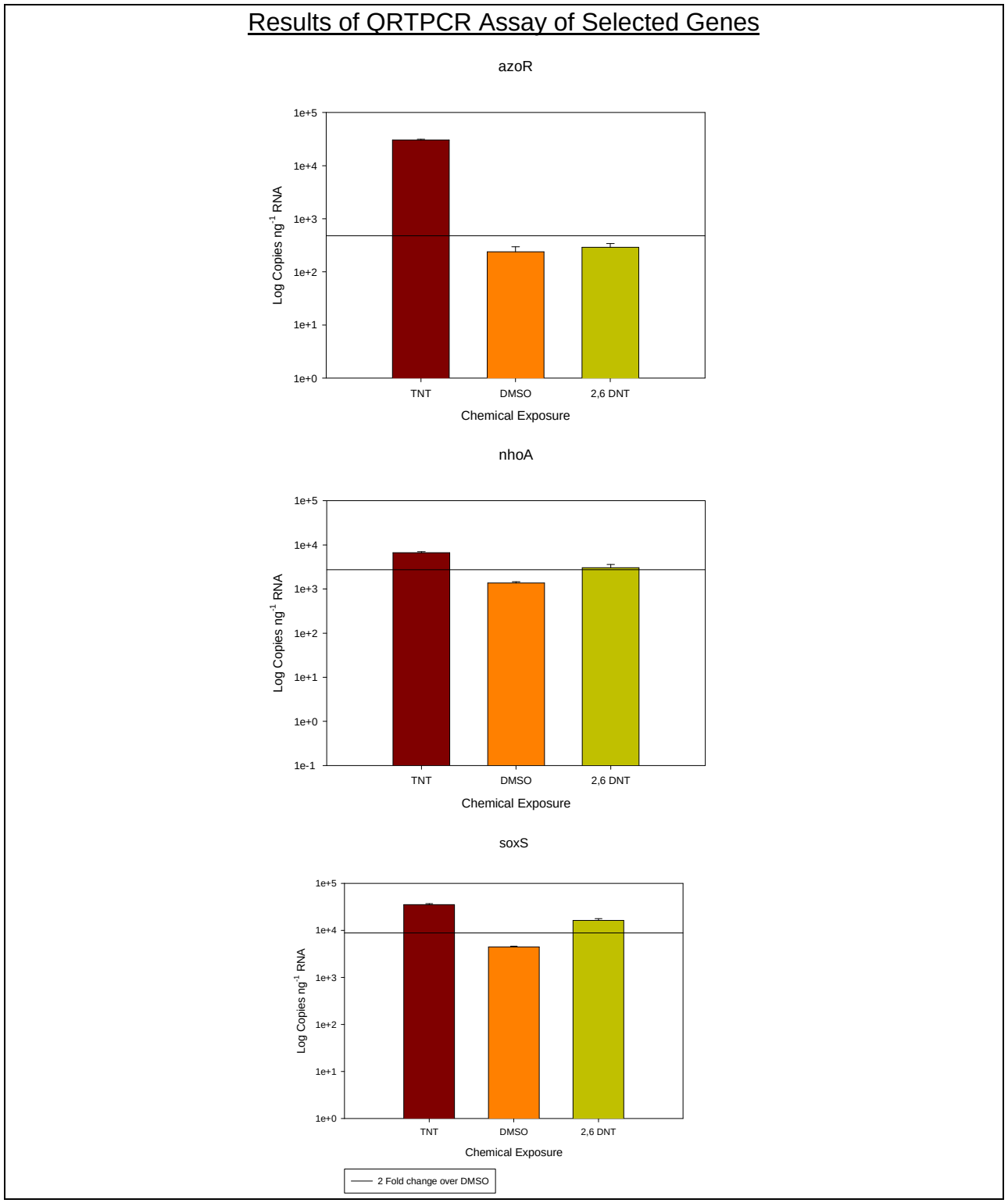


Figure 2-3 Quantitative PCR Analysis Confirms Microarray Results

Table 2-3 : Results of the *E.coli* expression profile statistical filtering

Experiment	P-value Threshold	# of Genes
DNT	1.47E-04	30
H2O2	8.23E-04	168
MIX	6.86E-05	14
TATP	4.02E-04	82
TNT	2.45E-03	501

The FDR adjustment allows the incidence of false positives to be controlled. The expression ratio of a gene was determined to be statistically significant if its FDR adjusted p-value (q-score) was 0.05 or less. The results of the FDR significance thresholds are presented in

Table 2-3.

Changes induced by TNT exposure

Filtering the data to leave a list of genes with an overall false discovery rate of 5% indicates that 501 genes (179 up, 322 down) are differentially expressed after 1 hour of TNT exposure. 117 of these genes have an absolute fold change of 2 fold or greater. Approximately 30% of the genes discovered to have significant differences in expression are genes of unknown function or completely unannotated transcripts. Of the 117 genes with absolute fold changes of 2 or more, 71 have increased expression and 46 have decreased expression. The most highly up regulated gene is azoR, a FMN-dependent NADH-azoreductase with approximately 63 fold increase in

transcription. It has previously been described as a component in tryptophan metabolism (Khodursky, Peter et al. 2000), and thiol-specific stress resistance (Liu, Zhou et al. 2009). The most down regulated gene, with greater than 7 fold decrease in transcription was csgB, a gene encoding the minor subunit of the curli complex which is involved in biofilm formation (Nenninger, Robinson et al. 2009). The functional composition of the TNT expression profile is discussed in greater detail in chapter 4.

Changes induced by DNT

Filtering the data to leave a list of genes with an overall false discovery rate of 5% indicates that 30 genes (17 up, 13 down) are differentially expressed after 1 hour of TNT exposure. 7 of these genes have an absolute fold change of 2 fold or greater. Again, 30% of the genes discovered to have significant differences in expression are genes of unknown function or completely unannotated transcripts. Of the 7 genes with absolute fold changes of 2 or more, 5 have increased expression and 2 have decreased expression. 25 genes are also found to have similar changes in expression upon TNT exposure. The remaining 5 genes (ycdL, yjbJ, chaC, dam, and a hypothetical protein of unknown function) have no similar change in expression in any other profile. No functional congruence among these 5 genes is evident.

Fumarase C (fumC) has the greatest increase in transcription with an approximate fold change of 6.70. It is involved in the conversion of fumarate to malate during the citric acid cycle. It is also a member of the superoxide stress regulon SoxRS, indicating a functional role in the response to oxidative stress. micF, which has an

approximate fold change of 2.88, is involved in global stress response as well as oxidative stress. It is an antisense RNA that serves as a negative regulator of ompF; ompF has a 2.45 fold decrease in expression. micF is also a member of the SoxRS regulon which indirectly links the decrease in ompF to oxidative stress as well. inaA which has a 2.85 fold increase in expression is also associated with oxidative stress through the SoxRS regulon.

Changes induced by TATP

Filtering the data to leave a list of genes with an overall false discovery rate of 5% indicates that 82 genes (34 up, 48 down) are differentially expressed after 1 hour of TATP exposure. 25 of these genes have an absolute fold change of 2 fold or greater. Approximately 26% of the genes discovered to have significant differences in expression are genes of unknown function or completely unannotated transcripts. Of the 25 genes with absolute fold changes of 2 or more, 17 have increased expression and 8 have decreased expression. The “gene” with the highest increase in transcription is an unannotated RNA sequence (probe 1760446_s_at). Unannotated probes target transcriptionally active regions of DNA that have not had mature gene products or transcripts identified in the literature.

The gene ontology enrichment analysis indicates that TATP exposure is characterized by increased expression of genes involved in copper ion binding and transport, DNA repair and modification, amino acid biosynthesis, galactitol and glucose metabolism, responses to stress, and energy metabolism (Appendix 2). The most

significantly overrepresented functional class was made up of genes involved in galactitol metabolic processes (GO: 0019402). These four genes (gatACYZ) all have increased expression. The most significantly overrepresented functional classes with decreased expression are those involved in copper ion transport (GO: 0006825). These genes, cusABCFX, encode the proteins involved in the copper/silver efflux system.

There were 26 genes that had significantly altered expression (>1.1 fold and FDR< 0.05) upon TATP exposure but not hydrogen peroxide or the synthesis mixture (Appendix 3). The presence of these genes in the TATP induced expression profile suggests that the observed changes are due to an interaction with intact TATP molecules and not simply a response to residual synthesis reagents or the products of its decomposition. Among this subset of genes there were few functional classes with significant overrepresentation. The copper/silver efflux system (genes cusABCF) was the most significantly enriched functional class.

Pseudomonas putida

Materials and Methods

Growth and exposure media

To limit possible effects of the growth media on gene expression, a minimal salts media was prepared to serve as both the growth and exposure media.

Minimal Salts Media (MSM) 500µM KH_2PO_4, 10mM K_2HPO_4, 831µM $MgSO_4 \cdot 7H_2O$, 125µM NH_4NO_3, 55.5µM glucose, 0.01% trace metals solution

Trace metals solution: 10g/L MgO, 2g/L $CaCO_3$, 5.4g/L $FeCl_3 \cdot 6H_2O$, 1.44g/L $ZnSO_4 \cdot 7H_2O$, 250mg/L $CuSO_4$, 62mg/L H_3BO_4, 490mg/L $NaMoO_4 \cdot 2H_2O$

Bacterial cultures and treatment

P. putida KT2440 (ATCC #47054) was grown from freezer stock in Luria-Bertini broth overnight. Revived cultures were used to inoculate MSM and grown overnight at 37℃ with 200rpm orbital shaking. Overnight cultures of *P. putida* were diluted in 37℃ MSM and grown to an $O.D._{600}$ of 0.7 at 37℃ with shaking. The cultures were then split into 6mL aliquots and spiked with either TNT, Triacetone-Triperoxide (TATP), or a vehicle control (DMSO). An exposure concentration of 220µM TNT, and 800µM TATP was achieved. All cultures contained 0.001% DMSO and 0.0002% tetrahydrofuran (THF) by volume. Growth assays indicated no statistically significant effect of TNT over all concentrations, Therefore 220µM TNT was chosen to match the concentrations used in the *E.coli* assay. TATP proved to be completely insoluble in both DMSO and the growth media, so freshly synthesized TATP crystals were dissolved in THF to achieve a

4M solution. This solution was used as the TATP spike. Total cellular RNA was extracted from each culture 1 hour after spiking.

Growth Curves

P. putida KT2440 (ATCC #47054) was grown from freezer stock in Luria-Bertini broth overnight. Revived cultures were used to inoculate MSM and grown overnight at 37℃ with 200rpm orbital shaking. Overnight cultures of *P. putida* were diluted in 37℃ MSM and grown to an $O.D._{600}$ of 0.1 at 37℃ with shaking. The cultures were then split into 1mL aliquots and distributed, in triplicate, into a 24 well microtiter plate. Each well was then spiked with TNT over a range of concentrations from 0-100mg/L. The cultures were incubated at 37 ℃ with continuous orbital shaking. Growth was measured as absorbance at 600nm. Measurements were taken every 15 minutes over a 24 hour period. Preliminary absorbance measurements indicated that TNT causes a measurable difference in sterile media. To control to effect of TNT concentration, the absorbance of sterile controls containing TNT were subtracted from the growth curve measurements.

Determination of Appropriate Exposure time

1 hour exposure time was chosen to match the *E.coli* assay.

RNA Isolation

RNA was isolated as described for the *E.coli* assay.

cDNA synthesis, labeling, hybridization, staining and array scanning

This experiment involved hybridization of 3 single color arrays for each of the TNT, TATP and DMSO samples. All processing of isolated RNA including cDNA synthesis, labeling, hybridization, staining and array scanning were performed by the UT Affymetrix Core facility according to Affymetrix standardized procedures for prokaryotic GeneChip arrays. Labeled and fragmented cRNA was then hybridized to custom made *P. putida* arrays. The arrays were designed and manufactured by Affymetrix upon request. Probes were designed using sequence data from the J. Craig Venter Institute Comprehensive Microbial Resource (CMR). Probes were designed for each locus in the database. Affymetrix accession number for the Affymetrix *P. putida* array is P_putida530130. A detailed protocol is available (Affymetrix 2005).

Gene Expression Measurements

Raw array image data was collected in CEL format using Affymetrix gene chip operating software (ver. 3.4.2152.32776). The raw CEL files were imported to Partek Genomics Suite (Partek 2008). Robust Multi-Array Analysis (RMA) was applied to all arrays to achieve background subtraction and normalization. GC-RMA normalization was not an option for these custom arrays. All *P. putida* microarray experiments were

performed on the same day, using the same lot of reagents and GeneChips, therefore no batch effect removal was necessary. All statistical analysis performed in Partek Genomics Suite are detailed in the Partek Genomic Suite online manual (Partek 2008).

Results

Gene Expression Profiles

A two-step FDR multiple test correction was used to control the number of type 1 errors. Relatively few genes pass the initial FDR adjusted significance threshold. In fact, the FDR control may be overly conservative for the TNT sample. With an FDR of 5%, the probable number of type 1 errors is far less than 1. The threshold was re-adjusted to reflect the probability of making a maximum of 0.99 type 1 errors Table 2-4. The significance threshold readjustment did little to provide heightened analytical power. The number of significant TATP induced changes actually decreased by 1.

There are only 37 genes differentially expressed among all treatment groups meaning that, comparatively, there is little effect of any of the compounds. More significant results were expected, especially in the wake of the *E.coli* expression profile results. The decrease in magnitude of the response may be due the difference in the exposure protocol. The *E.coli* profiles were generated using DMSO as the solvent control; the *P. putida* exposures contained both DMSO and THF. A recent study conducted by Dr. Theodore Henry and colleagues investigated the effects of THF on gene expression of the model organism *Danio rerio* (zebra fish) (Henry, Menn et al. 2007).

Table 2-4: Results of the *P. putida* expression profile statistical Filtering

Experiment	P-value Threshold	# of Genes
TATP	1.80×10^{-4}	28
TNT	1.46×10^{-4}	12

Through gene expression profiling, they found that the gene expression profiles that had previously been attributed to exposure to C-60 aggregates were actually the result of oxidative damage cause by THF which was used as the vehicle control. This may also explain the decreased number of genes found in the present study. THF may have a masking effect on the gene expression profiles. That is, THF may present a stronger stimulus than TATP or TNT. Thus the greatest contributor to the expression profiles of each culture was the THF resulting in few significant changes between them.

TNT induced changes

Twelve genes pass the statistical filtering (Appendix 4). The "gene" with the greatest increase in expression was a 207 bp intergenic region spanning from base 2295676-2295883. Annotation as an intergenic region would indicate that this locus is a non-coding region of the chromosome. Because of the high fold change and statistical significance ($p=4.3x10^{-6}$), it is unlikely that this is a spurious finding. This result would suggest a novel transcript or small regulatory RNA molecule. Otherwise, the expression profile is dominated by membrane transport proteins and efflux pumps. The limited power resulting from the statistical analysis makes further characterization of this response challenging. A gene ontology enrichment analysis revealed little more than what is apparent from a casual review of the list of genes. The most significantly overrepresented functional classes are drug resistance and response to chemical stimulus.

TATP induced changes

Twenty-eight genes pass the statistical filtering (Appendix 5). Only 1 gene is common to the TNT and TATP exposures. The most highly up-regulated gene encodes a cytochrome c-type protein. Altered expression of several supposed intergenic regions is observed.

Saccharomyces cerevisiae

Materials and Methods

Growth and exposure media

To limit possible effects of the growth media on gene expression, a yeast minimal media (YMM) was prepared to serve as both the growth and exposure media. (Routledge and Sumpter 1996).

The toxicity assay using the bioluminescent yeast strain CEB585 utilized a modified YMM lacking uracil and leucine.

Bacterial cultures and treatment

Saccharomyces cerevisiae W303 (ATCC # 200060) was grown from freezer stock in YPD media (1% yeast extract, 2% peptone, 1% dextrose) overnight. Revived cultures were used to inoculate fresh YPD media and grown overnight at 30℃ with 200rpm orbital shaking. Overnight cultures were washed 3 times in phosphate buffered saline (PBS) then diluted in 30℃ YMM and grown to a n $O.D._{600}$ of 0.65 at 30℃ with shaking. The cultures were then split into 10mL aliquots and spiked with either TNT; 2,6-DNT; Peroxide; Triacetone-Triperoxide (TATP); a solution (MIX) containing 50% acetone, 15% hydrogen peroxide and 35% H_20; or a vehicle control (DMSO). Exposure concentrations of 22 µM TNT, 220µM TNT, 82.5 µM DNT, 275µM DNT, 244µM peroxide, 750µM TATP, and 0.1% MIX were achieved. The lower TNT and DNT concentrations were derived from toxicity assays described below. All other

concentrations and exposure times were chosen to match those used for the *E.coli* study. Total cellular RNA was extracted from each culture 1 hour after spiking.

TNT and DNT effect on the bioluminescence of strain CEB585

S. cerevisiae bioluminescent strain CEB585 (Gupta, Patterson et al. 2003) was grown in modified YMM lacking uracil and leucine. Upon reaching an optical density of OD_{600}=0.15, cultures were spiked with concentrations of TNT or DNT ranging from 0-100 mg/L. Controls consisted of 1% DMSO. Optical density (OD_{590}) and bioluminescence (centilations per second; CPS) were measured every 30 minutes. Bioluminescence was normalized to optical density.

RNA Isolation

Total cellular RNA from the 82.5 µM DNT, 22 µM TNT, and 3 of the DMSO exposed cultures was isolated using the yeast protocol for RNeasy mini kits (Qiagen, Valencia, CA). These extractions started with the generation of spheroplasts using a 30 minute zymolase treatment of the yeast cell pellet as per the manufacturer's instructions.

A second set of microarray experiments were performed using a higher concentration of TNT and DNT (220µM and 275 µM respectively) and a modified RNA extraction protocol. This set also includes the TATP, peroxide and MIX exposures. RNA samples included 3 biological replicates all collected on the same date. Isolation was performed using a hot phenol RNA purification (Kohrer and Domdey 1991). Briefly, yeast cultures were grown to an OD_{600} of 1. Cells were pelleted and flash frozen using

dry ice and 100% ethanol. The cells were lysed at 65℃ by incubating in a pre-heated mixture (50% v/v) of TES buffer and water equilibrated acid phenol. During the incubation, the cells were forcefully agitated every 15 minutes. The lysate was incubated on ice for 5 minutes followed by centrifugation with a phase separation gel (Qiagen MaXtract cat#:129056, Qiagen, Valencia, CA). The aqueous layer was removed and a phenol wash was repeated followed by a second wash using phenol/chloroform/isoamyl alcohol. RNA was precipitated using 3M sodium citrate and 100% cold ethanol. The RNA pellet was washed in 70% ethanol, allowed to dry and eluted in HPLC grade nuclease free water. Total RNA concentration and purity were assessed based on OD 260/280 measured using a Nanodrop-1000 (Nanodrop Technologies Inc, Wilmington, DE).

cDNA synthesis, labeling, hybridization, staining and array scanning

This experiment involved hybridization of 3 biological replicates for each of the exposure conditions. All processing of isolated RNA including cDNA synthesis, labeling, hybridization, staining and array scanning were performed by the UT Affymetrix Core facility according to Affymetrix standardized procedures for Yeast Genome 2.0 GeneChip arrays. Labeled and fragmented cRNA was then hybridized to Affymetrix Yeast Genome 2.0 GeneChips. A detailed protocol is available (Affymetrix 2005).

Gene Expression Measurements

Raw array image data was collected in CEL format using Affymetrix gene chip operating software (ver. 3.4.2152.32776). The raw CEL files were imported to Partek Genomics

Suite (Partek 2008). Guanine-cytosine robust multi-array analysis (GC-RMA) was applied to all arrays to achieve background subtraction and normalization across all arrays. No batch effect removal was necessary because all samples were collected and processed on the same day, using the same batch of reagents. All statistical analysis performed in Partek Genomics Suite are detailed in the Partek Genomic Suite online manual (Partek 2008).

Functional Analysis of Significant gene changes

Gene ontology data for each of the significant genes were retrieved from the Saccharomyces Genome Database (SGD accessible at www.yeastgenome.org). A gene ontology enrichment analysis was performed using the retrieved data using the Gene Ontology Enrichment Analysis Toolkit (omicslab.genetics.ac.cn) using AmiGo source version OBO v.1.2 (Ashburner, Ball et al. 2000)

Results

Bioluminescence and growth assays

S. cerevisiae strain CEB585 is a constitutive bioluminescent strain derived from strain W303. The bioluminescence is driven by *lux*C, D, A, B, and E gene expression under the control and GPD and ADH1 promoters. Exposure to TNT and DNT caused a dose-dependent decrease in bioluminescence at 1 hour (Figure 2-4). It was from these dose-response experiments that the concentration and exposure duration was determined for the expression profiling study.

Gene Expression Profiling

The gene expression analysis from the low concentration TNT and DNT exposures yielded few significant alterations in gene expression (Table 2-5); and those that were statistically significant had low to moderate fold changes. While it is possible that neither TNT nor DNT alter the expression of a large number of genes. It was hypothesized that the stimulus presented by the 30 minute zymolase incubation was sufficient to alter the gene expression patterns induced by the TNT and DNT treatments. The result being that the dominant determinant for all of the resulting expression profiles was the zymolase treatment, which would be similar for all samples. Therefore, a second set of microarray experiments was performed using a higher concentration of TNT and DNT as well as a modified extraction protocol omitting the zymolase treatment.

Table 2-5: Results of the Statistical Significance Filter

Treatment	Pvalue Threshold	# of Genes	Cell Lysis Method
5mg/L TNT	4.30×10^{-6}	28	Zymolase
15mg/L DNT	3.72×10^{-6}	17	
50mg/L TNT	5.01×10^{-6}	537	Hot Phenol
50mg/L DNT	5.12×10^{-6}	665	
TATP	5.07×10^{-6}	557	
H2O2	5.16×10^{-6}	644	
MIX	1.11×10^{-6}	2	

There was a marked increase in the observed number of genes with significantly altered expression between the two lysis methods. The increased response could be attributed to either the higher concentration of each respective compound or the change in lysis protocol. The bioluminescence assay demonstrated a dose-dependent effect on light production. This response is independent of any differences in cell density; at 1 hour, there was little difference in optical density and any observed difference would be accounted for by the normalization procedure. This would suggest that the differences in bioluminescence are due to some change in the physiological state of the cell. These differences are either related to the energy charge of the cell, as the production of light via the *lux* operon is ATP and NADPH dependent (Deweger, Dunbar et al. 1991; Hill, Rees et al. 1993; Neilson, Pierce et al. 1999) , or due to decreased activation of the promoters governing lux expression (GAD or ADH1).

From the expression profiles of the two control groups, it is clear that the RNA extraction protocols had a significant effect on global gene expression. Both groups of control cultures received the exact same DMSO treatment, and cells were harvested at

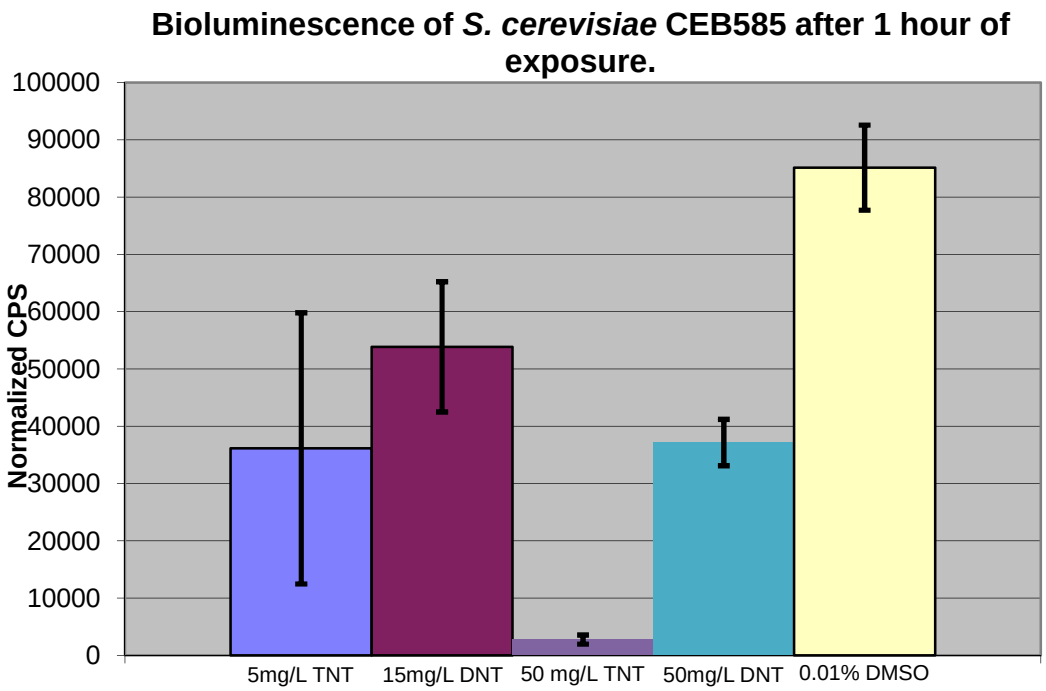


Figure 2-4: Bioluminescence of *S. cerevisiae* strain CEB585 is diminished by exposure to TNT or DNT

the same time point and cell density. There was no difference in the exposure parameters of the DMSO cultures yet there are 2950 genes that are differentially expressed between the two groups.

The differences between the zymolase (set 1) and hot phenol (Set 2) control groups limits the appropriateness of a direct comparison of the two sets of microarray experiments. That is not so say that such a comparison is futile, the few differences observed between the TNT, DNT and DMSO exposures in the zymolase treated cells do represent *real* changes that must be attributed to the compound exposure as this was the only difference between samples.

The statistical criteria used to determine significance is more conservative than the step-down multiple test correction described for the *E.coli* experiments. This was necessary due to the large number of genes found to be significant in the hot phenol treated cultures. To better control the occurrence of false positive results a Holm multiple test correction was employed instead. This is very similar to the step-up procedure used with the bacterial data sets with the exception being that p-values are listed in descending order and a much simpler adjustment is made. If (n) is the number of p-values (p) in the dataset, the highest ranking (largest) p is adjusted according to pa=p*n. The next highest is adjusted according to p_a=p*(n-1), the next p_a=p*(n-2) and so on. This method is more conservative than the step-up method but much less conservative than the Bonferroni method.

TNT induced changes

From the set 2 exposures, 537 genes pass the FDR adjusted statistical significance threshold; 242 of these are up-regulated, and 297 are down regulated. 333 have an absolute fold change of 2 or greater. The gene ontology enrichment analysis revealed that the most significantly up-regulated biological processes are those involved in sulfur metabolism, amino acid synthesis, general response to drugs and toxins, and cell wall synthesis and repair. In terms of molecular function, the most significantly up-regulated functional groups were related to nitrogen metabolism, sulfur amino acid synthesis, oxidoreductase activity acting on sulfur donors and carbon-nitrogen bonds, and transport of sulfur compounds and amino acids. None of these functional groups were significantly enriched among the down-regulated genes. The genes with decreased expression were predominantly involved in heat shock, metabolism of carbohydrates, alcohols and cellular energy reserves, glutamate and glutamine catabolic processes, copper / iron transport and binding, and cell division. The molecular functions with decreased expression were generally related to respiration, and oxidoreductase activity acting on metals and carbohydrates.

DNT induced changes

From the set 2 exposures, 665 genes pass the FDR adjusted statistical significance threshold; 333 of these are up-regulated, and 332 are down regulated. There are a total of 281 genes which were found to have similar alterations in gene expression upon TNT exposure. These included increased expression of genes involved in methionine, cysteine and serine family amino acid biosynthesis, sulfate

assimilation, de novo IMP biosynthesis, and positive regulation of the cell cycle. Genes common to the TNT exposure with decrease expression include genes involved in the response to reactive oxygen, cellular respiration, and response to heat.

There were also 384 genes which represent distinct differences from the TNT response. It was surprising to find that although this list of genes was larger than that of those common to the TNT exposure, functional enrichment of ontology groupings associated with these genes was less pronounced. Enriched functional groupings composed of up-regulated genes include responses to pheromone leading to conjugation, arginine biosynthesis, and the response to osmotic stress. Functional ontology groups with decreased expression include telomere maintenance via telomerase, proteosome assembly, and DNA replication.

TATP induced changes

From the set 2 exposures, 557 genes pass the FDR adjusted statistical significance threshold; 242 of these are up-regulated, and 297 are down regulated. 333 have an absolute fold change of 2 or greater. The gene ontology enrichment analysis revealed that the most significantly up-regulated biological processes are those involved in cell division, cell wall biosynthesis, sulfur metabolism, and amino acid metabolism. None of these functional groups were significantly enriched among the down-regulated genes. The genes with decreased expression were predominantly involved in heat shock, structural/ ribosomal RNAs, DNA damage/ repair, pyrimidine nucleoside metabolism, and transmembrane sugar transport.

There is a great deal of similarity between the TATP and hydrogen peroxide induced expression profiles. 452 of the 557 genes have similar expression upon peroxide exposure. Of the remaining 105, only 48 are unique to the TATP exposures. Among these were increased copper ion homeostasis, regulation of cell cycle, and a generic response to organic cyclic compounds. There was also significantly decreased expression of rRNA maturation networks, ribosomal subunit assembly, and, response to xenobiotic substances.

References

Affymetrix (2005). GeneChip Expression Analysis Technical Manual. Santa Clara, CA.
Ashburner, M., C. A. Ball, et al. (2000). "Gene Ontology: tool for the unification of biology." Nature Genetics **25**(1): 25-29.
Bellamy, A. J. (1999). "Triacetone triperoxide: Its chemical destruction." Journal of Forensic Sciences **44**(3): 603-608.
Caballero, A., A. Esteve-Nunez, et al. (2005). "Assimilation of nitrogen from nitrite and trinitrotoluene in Pseudomonas putida JLR11." Journal of Bacteriology **187**(1): 396-399.
Deweger, L. A., P. Dunbar, et al. (1991). "USE OF BIOLUMINESCENCE MARKERS TO DETECT PSEUDOMONAS SPP IN THE RHIZOSPHERE." Applied and Environmental Microbiology **57**(12): 3641-3644.
Duque, E., A. Haidour, et al. (1993). "CONSTRUCTION OF A PSEUDOMONAS HYBRID STRAIN THAT MINERALIZES 2,4,6-TRINITROTOLUENE." Journal of Bacteriology **175**(8): 2278-2283.
Esteve-Nunez, A., A. Caballero, et al. (2001). "Biological degradation of 2,4,6-trinitrotoluene." Microbiology and Molecular Biology Reviews **65**(3): 335-+.
Esteve-Nunez, A., G. Lucchesi, et al. (2000). "Respiration of 2,4,6-trinitrotoluene by Pseudomonas sp strain JLR11." Journal of Bacteriology **182**(5): 1352-1355.
Gonzalez-Perez, M. M., P. van Dillewijn, et al. (2007). "Escherichia coli has multiple enzymes that attack TNT and release nitrogen for growth." Environmental Microbiology **9**(6): 1535-1540.
Gupta, R. K., S. S. Patterson, et al. (2003). "Expression of the Photorhabdus luminescens lux genes (luxA, B, C, D, and E) in Saccharomyces cerevisiae." Fems Yeast Research **4**(3): 305-313.
Henry, T. B., F. M. Menn, et al. (2007). "Attributing effects of aqueous C-60 nano-aggregates to tetrahydrofuran decomposition products in larval zebrafish by assessment of gene expression." Environmental Health Perspectives **115**(7): 1059-1065.
Hill, P. J., C. E. D. Rees, et al. (1993). "THE APPLICATION OF LUX GENES." Biotechnology and Applied Biochemistry **17**: 3-14.
Irizarry, R. A., B. M. Bolstad, et al. (2003). "Summaries of affymetrix GeneChip probe level data." Nucleic Acids Research **31**(4).
Keseler, I. M., C. Bonavides-Martinez, et al. (2009). "EcoCyc: A comprehensive view of Escherichia coli biology." Nucleic Acids Research **37**: D464-D470.
Khodursky, A. B., B. J. Peter, et al. (2000). "DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli." Proceedings of the National Academy of Sciences of the United States of America **97**(22): 12170-12175.
Kohrer, K. and H. Domdey (1991). "PREPARATION OF HIGH-MOLECULAR-WEIGHT RNA." Methods in Enzymology **194**: 398-405.
Liu, G., J. Zhou, et al. (2009). "The Escherichia coli Azoreductase AzoR Is Involved in Resistance to Thiol-Specific Stress Caused by Electrophilic Quinones." J. Bacteriol. **191**(20): 6394-6400.
Matyas, R., J. Pachman, et al. (2009). "Study of TATP: Spontaneous Transformation of TATP to DADP - Full Paper." Propellants Explosives Pyrotechnics **34**(6): 484-488.
Neilson, J. W., S. A. Pierce, et al. (1999). "Factors influencing expression of luxCDABE and nah genes in Pseudomonas putida RB1353(NAH7, pUTK9) in dynamic systems." Applied and Environmental Microbiology **65**(8): 3473-3482.
Nelson, K. E., C. Weinel, et al. (2002). "Complete genome sequence and comparative analysis of the metabolically versatile Pseudomonas putida KT2440." Environmental Microbiology **4**(12): 799-808.
Nenninger, A. A., L. S. Robinson, et al. (2009). "Localized and efficient curli nucleation requires the chaperone-like amyloid assembly protein CsgF." Proceedings of the National Academy of Sciences of the United States of America **106**(3): 900-905.

Partek (2008). Partek® Genomics SuiteTM. St. Louise, Partek Inc.
Partek (2008). Partek® software On-line Help, version 6.3, build 6.08.0110. St. Louise, Partek Inc.
Routledge, E. J. and J. P. Sumpter (1996). "Estrogenic activity of surfactants and some of their degradation products assessed using a recombinant yeast screen." Environmental Toxicology and Chemistry **15**(3): 241-248.
Spanggord, R. J., J. C. Spain, et al. (1991). "BIODEGRADATION OF 2,4-DINITROTOLUENE BY A PSEUDOMONAS SP." Applied and Environmental Microbiology **57**(11): 3200-3205.
Stenuit, B., L. Eyers, et al. (2006). "Aerobic growth of Escherichia coli with 2,4,6-trinitrotoluene (TNT) as the sole nitrogen source and evidence of TNT denitration by whole cells and cell-free extracts." Applied and Environmental Microbiology **72**(12): 7945-7948.
Ward, P. G., M. Goff, et al. (2006). "A two step chemo-biotechnological conversion of polystyrene to a biodegradable thermoplastic." Environmental Science & Technology **40**(7): 2433-2437.

Chapter 3 : Exploratory Interpretation of the *E. coli* TNT response

Introduction

From the expression profiling and prediction analysis a list of characteristic gene expression modulations have emerged. *E.coli*'s response to TNT was the most robust in terms of the number genes unique to that response. It was for this reason that a significant portion of this study was dedicated to understanding that response. The goal of this section is to provide a descriptive analysis of the changes in gene expression resulting from TNT exposure in *E.coli* cultures.

Methods

Meta-Analysis of Significant gene changes

Functional analysis: Gene ontology data for each of the significant genes were retrieved from EcoCyc (Keseler, Bonavides-Martinez et al. 2009). A gene ontology enrichment analysis was performed using the retrieved data using the Gene Ontology Enrichment Analysis Toolkit (omicslab.genetics.ac.cn)

Expression Data Mining: Several sources of published expression data were mined for results relevant to the present study. These databases include MetaCyc, RegulonDb, GenExpDb, and the Gene Expression Omnibus (GEO). Data collected from these sources was used as a catalyst for further literature review.

Discussion

Meta-analysis of Gene expression data Reveals Signs of Oxidative Stress

The first indication of oxidative stress came from the initial review of the gene list; the expression of at least 31 genes characteristic of oxidative stress have significantly increased expression. The bulk of these genes are members of the SoxRS regulatory network of which SoxS is the activator. Expression of soxS is dependent on activation by its transcriptional factor SoxR; however there was no significant change in expression of soxR in this study. SoxS regulation occurs through activation of soxR through post-translational modification of the SoxR protein (Nunoshiba 1996; Touati 2000; Pomposiello, Koutsolioutsou et al. 2003). Activation occurs through either oxidation of the Fe-S center by superoxide or direct nitrosylation by a reactive nitrogen species. As such, there may not be any change in expression of SoxR.

SoxR was not differentially expressed in any of the other exposures; neither was soxS. A literature review provided little insight. The cited references were all enzymatic activity or protein modification analyses; none involved gene expression measurements. A meta-analysis of publicly available expression data was performed to see if other researchers have observed similar results (Figure 3-1). This analysis was limited to those experiments publicly accessible from the Gene Expression Omnibus (GEO http://www.ncbi.nlm.nih.gov/geo/). There are currently 142 published gene expression experiments deposited in which soxS has an expression ratio of +/- 2 or greater. Of those, less than 20 also show soxR to have an expression ratio of +/-2 or greater in the same direction. Furthermore, in all of the 792 experiments in which soxR expression

data is available, only 57 show any significant change in expression. The details of these experiments reveal that several of the studies in which soxR has significantly altered expression have been conducted under conditions which would cause global transcription to be increased in general.

Increased expression of azoR suggests cytosolic interactions

The idea of TNT induced oxidative stress is supported in the literature (Kumagai, Kikushima et al. 2004; Gong, Guan et al. 2007; Miliukiene and Cenas 2008). What it is unclear from the literature is the source of this oxidative stress. The intuitive answer would involve an interaction with the nitro groups positioned around the aromatic ring of TNT. TNT, like many nitro-aromatics tends to be recalcitrant to biological metabolism (Kulkarni and Chaudhari 2007). It is a large bulky compound that does not readily permeate the cell wall. Because of this, it seemed unlikely that the biological interaction with TNT would extend beyond the cell surface. However, the increased expression of azoR suggests that this interaction is not just limited to surface proteins and the cell wall.

AzoR is the most highly up-regulated gene with a 62 fold increase in expression. Further, these changes were observed in the TNT exposed cultures, but not the DNT, peroxide, MIX or TATP cultures. AzoR encodes an FMN-dependent azoreductase which has yet to be associated with TNT biodegradation, mineralization, or detoxification in any capacity.

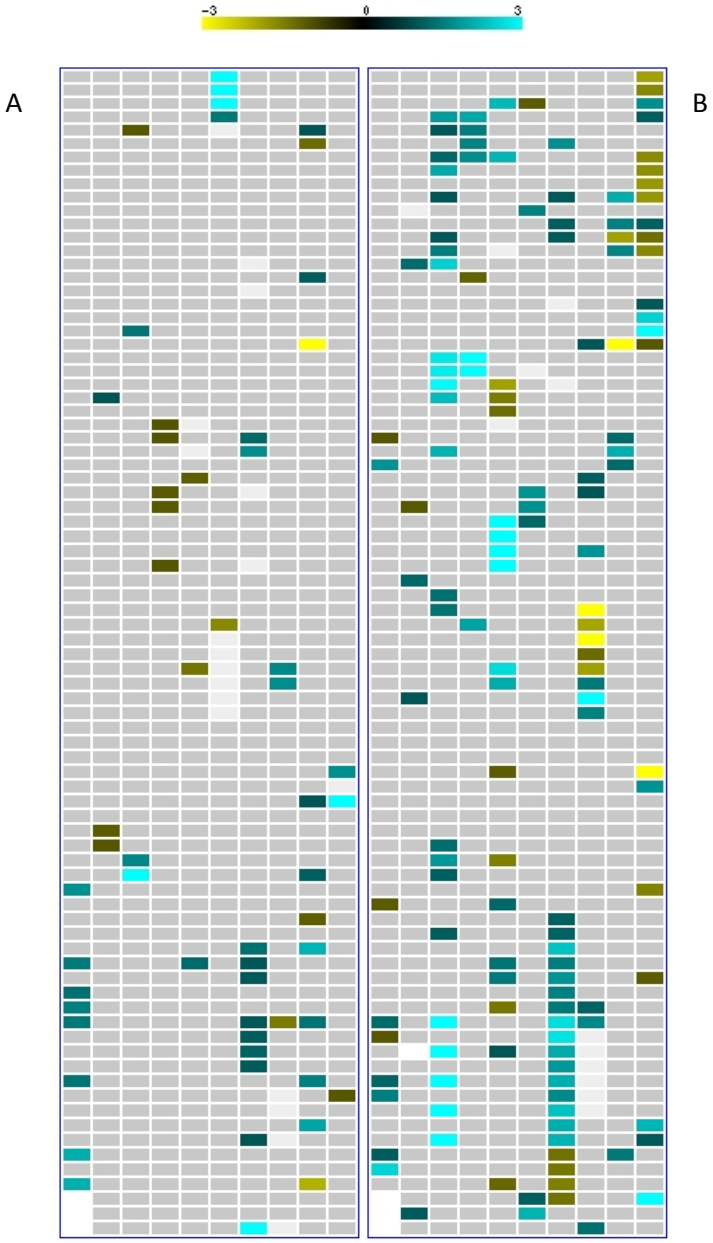


Figure 3-1: Meta-analysis reveals little correlation between expression of SoxS and SoxR

This figure represents the collective body of published expression data for SoxR (A) and SoxS (B). Each heat map (blue outline) is essentially an array of arrays for these two genes. Each of the small rectangles represents and individual microarray experiment that has been deposited in NCBI's Gene Expression Omnibus. The coordinates of the microarray experiments have been preserved between heat maps for comparison.

The literature provides no direct indication as to why such a marked increase in azoR expression would be induced by TNT exposure. It has been associated with the reductive transformation of Azo dyes (Hawari, Halasz et al. 1999; Nakanishi, Yatome et al. 2001). An early study (Kaplan and Kaplan 1982), observed the formation of azoxytoluene (AZTs) (

Figure 3-2) products from TNT metabolism in mixed culture composting experiments. Their formation was attributed to abiotic condensation of hydroxyamino-dinitrotoluene (HADNT) intermediates (Figure 3-3). These findings have since been supported by other researchers (Bumpus and Tatarko 1994; Hawari, Halasz et al. 1998; Hawari, Halasz et al. 1999). This could explain the biologically mediated discoloration of the growth media observed during the growth and toxicity assays (Figure 2-2) although the formation of hydride-Meisenheimer complexes (discussed later) is a more likely explanation. However, there is no evidence in the literature that AZTs undergo any further biological degradation so it is unlikely that azoR is acting on the N=N bond joining the two aromatic rings. It has been reported that *Rhodobacter sphaeroides* is able to reduce TNT through the activity of a related FMN dependent azoreductase, AZR, however similar assays involving azoR have not reported any reductase activity toward TNT (Liu, Zhou et al. 2009).

Although AzoR has been characterized as an azoreductase based on its ability to reduce azo dyes *in vitro*, the physiological role of AzoR has been recently called into question (Rau and Stolz 2003; Liu, Zhou et al. 2009). The findings of all previous studies suggest that the protein, AzoR, is exclusively cytosolic under physiological

conditions and it is unlikely that large polar molecules such as an azo dye would diffuse across the cell membrane. Also, the products of azoreductase activity on these

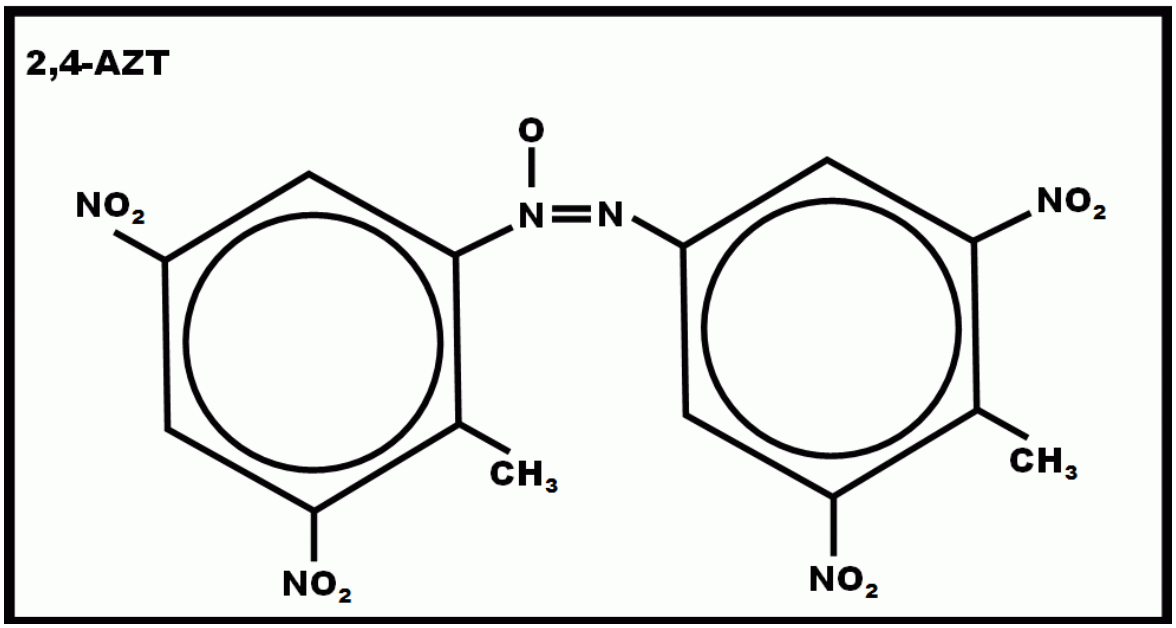


Figure 3-2 : 2, 4-Azoxytoluene

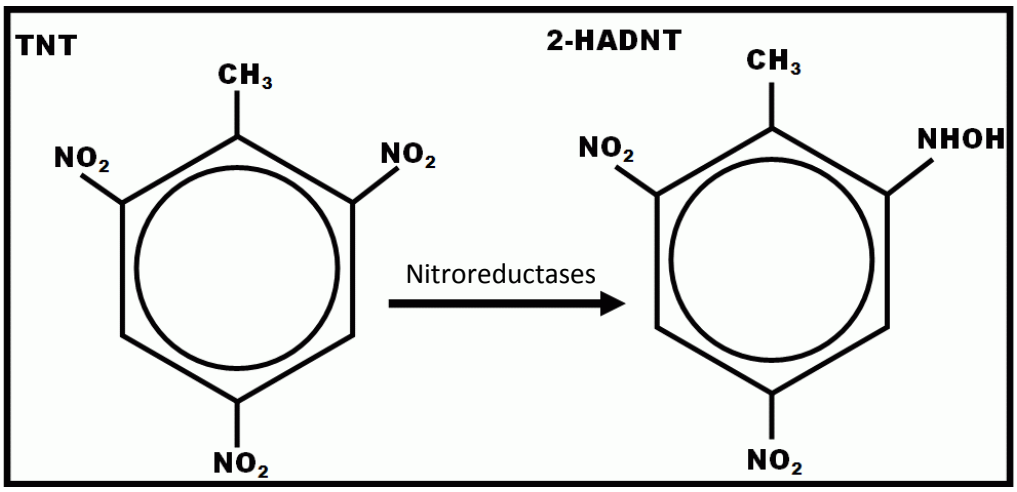


Figure 3-3 Reductive Transformation of TNT to 2-HADNT

compounds are often toxic. Specific to this scenario, HADNT and ADNT intermediates are more toxic than the AZT end products so there would be little benefit to regenerating them through azoR activity. *In vivo* experiments comparing azoreductase activity of azoR mutants to wild type have demonstrated no difference in their reduction of azo dyes.

Liu et al proposed an alternate role for AzoR as a cytosolic defense mechanism against thiol specific stress caused by electrophilic quinones. Toxicity of these quinones has been attributed to their ability to covalently modify thiol groups of proteins and their tendency disrupt redox cycling thereby generating reactive oxygen species within the cell (Rodriguez, Fukuto et al. 2005). These two modes of action are virtually indistinguishable under physiological conditions. It was only when Rodriguez and colleagues observed that these quinones caused thiol-related growth inhibition under anaerobic conditions that direct covalent modification mechanism was apparent. Liu et al found azoR expression to be induced by 2-methylhydroquinone, catechol, menadione and to a much lesser extent 10mM hydrogen peroxide. They concluded that AzoR did offer protection against oxidative stress, but specifically thiol-related stress resulting from this covalent modification.

The meta-analysis of azoR expression revealed that, like SoxR, few studies have found significant increases in expression of azoR. This includes at least one set of experiments in which the effect of the superoxide generating agent paraquat was studied (Blanchard JL 2007). This suggests that the observed increased expression of azoR, while not unique to TNT exposure is distinct from that induced by reactive oxygen. The lack of any significant alteration of azoR expression in the peroxide and

MIX exposed cultures is supportive of these findings.

A possible explanation for the marked increase of azoR is that the nitro groups are presenting a source of oxidative stress; specifically nitrosative stress. Nitrosative stress describes the toxic effect of reactive nitrogen species (RNS). The effect RNS on microbes has been well documented due to the role they play in host immune function (Svensson, Marklund et al. 2006) where they serve as antimicrobial agents by macrophages. The most well studied RNSs are nitric oxide (NO) and NO generating compounds such as S-nitrosylglutathione (GSNO). These compounds have been shown to damage thiol containing proteins which would explain the high induction of azoR (Spiro 2006; Brandes, Rinck et al. 2007).

E.coli, like many microbes, has mechanisms to combat nitrosative stress, but identifying this type of stress through expression profiling is challenging. This is because gene networks involved in nitrosative stress resistance are often identical to those involved in oxidative stress caused by reactive oxygen species (Inoue, Nishikawa et al. 1999). One such example is the aforementioned SoxRS response. Reactive nitrogen damages Fe-S cluster containing proteins by displacing the sulfur group; the RNS mediated activation of SoxR exploits this (Figure 3-4) (Brunelli, Crow et al. 1995; Ding and Demple 2000; Vasil'eva, Stupakova et al. 2001; Lo, Chen et al. 2008) . The relationship between sulfur and RNS is not limited to Fe-S containing proteins. They will readily complex with any available sulfur containing compounds (Andrews, Hassanzadeh et al. 1996). This includes the sulfur containing amino acids cystine, cysteine and homocysteine (Spiro 2006).

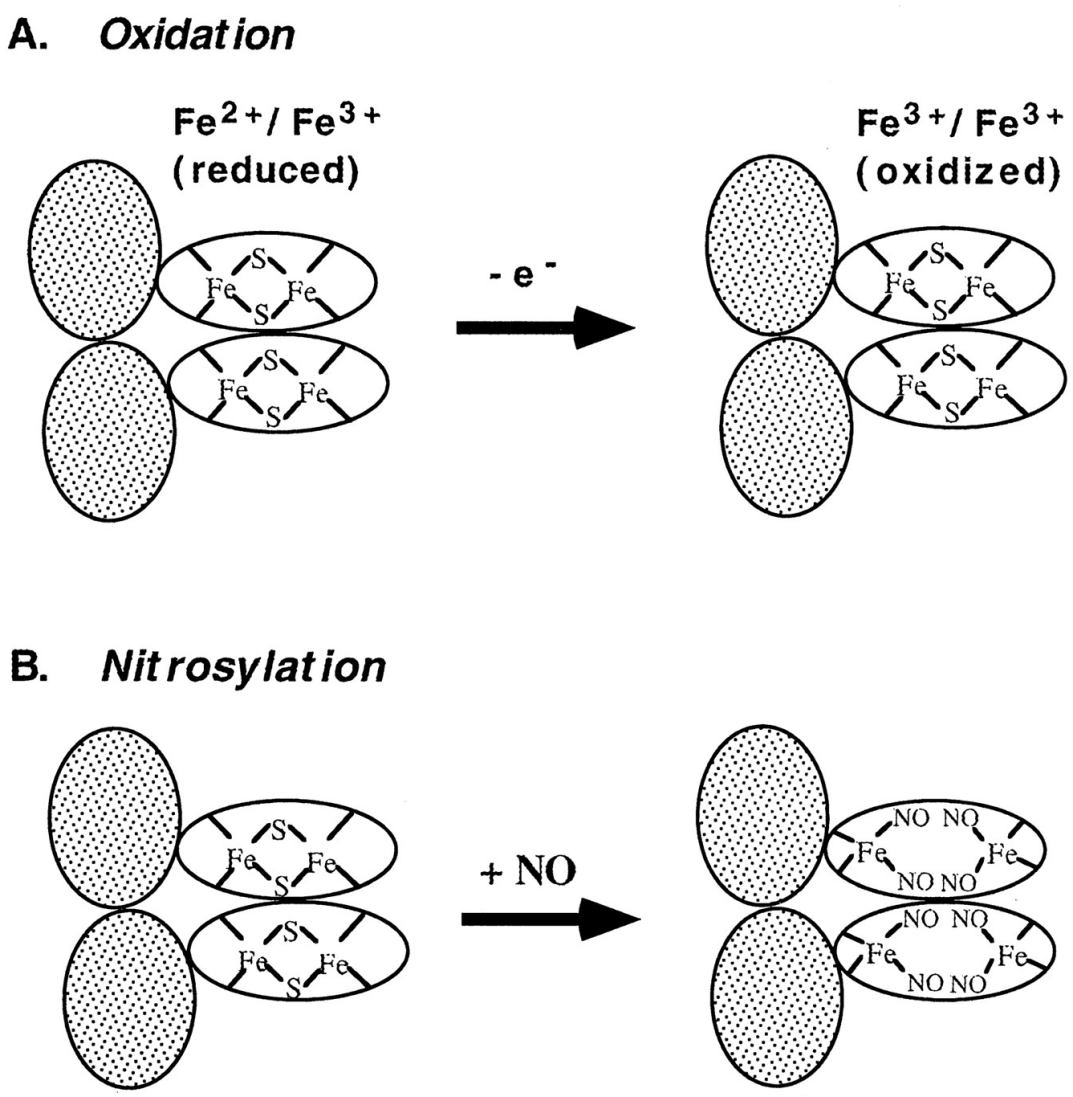


Figure 3-4: Activation of the SoxR transcriptional factor occurs via two distinct mechanisms (Ding and Demple 2000).

To ascertain whether the results of the microarray study support the hypothesis that TNT is causing RNS-like perturbations, a gene ontology enrichment analysis was performed. The genes with decreased expression produced the largest number of statistically significant enrichments. These include decreases in biosynthesis of most essential and non-essential amino acids (the exceptions being cysteine, serine, glutamine and argentine). Also among the biological processes with decreased expression are biosynthesis of DNA, lipids, and cell wall components. There is also a significant decrease in energy metabolism. Taken together, the genes involved in most biological processes have decreased expression. Fewer functional groups were significantly overrepresented by the genes with increased expression. Among these are cystine, cysteine, serine, glutamine and argentine biosynthesis, fatty acid catabolic processes, sulfate transport, reduction and assimilation, Fe-S cluster assembly, electron transport chain assembly, metabolism of xenobiotic compounds and responses to stress (oxidative, xenobiotic, drug, antibiotic). In addition to the effects mentioned previously, RNS are also characterized by their ability to inhibit energy metabolism and ATP synthesis through disruption of the electron transport machinery (Minamiyama, Takemura et al. 1997; Inoue, Nishikawa et al. 1999). The significant enrichment of genes involved in electron transport chain assembly is suggestive of this effect.

Reductive Transformation of TNT

As mentioned, it is unlikely that a large extracellular compound would cause the alterations in gene expression observed here. The most pressing evidence is the increased expression of azoR which has been shown to not respond to extracellular

stimuli (Rau and Stolz 2003; Liu, Zhou et al. 2009). A plausible explanation is that TNT is undergoing reductive transformation and it is a metabolite of TNT that is the source of these alterations.

Reductive transformation of polynitroaromatic compounds by bacteria can occur through the reduction of the nitro groups to hydroxylamine or amino groups (Roldan, Perez-Reinado et al. 2008). Previous studies have established that *E. coli* has the ability to both reduce TNT to its ADNT and HADNT metabolites and ultimately release nitrogen from the aromatic ring to be used for growth (Yin, Wood et al. 2005; Gonzalez-Perez, van Dillewijn et al. 2007) . This process is catalyzed by FMN dependent-NAD (P) H nitroreductases nfsA and nfsB. There is a moderate (1.8 fold) increase in nfsA expression upon TNT exposure; however no significant increase in nfsB expression was observed. While increased expression of nfsB was expected, two important points must be considered. It is important to first recognize that this study was conducted under different conditions than the referenced studies. In previously published studies, *E.coli* was exposed to TNT either in PBS or minimal media with no other nitrogen source. The experiments of the present study were conducted in minimal growth media containing nitrogen. Additionally, the referenced studies have focused on the enzymatic activity of these gene products; none measured gene expression in response to the TNT amendments. This study represents the first published microarray experiments involving *E. coli* exposure to TNT, TATP, DNT or any other munitions compound.

E. coli N-ethylmalemimide reductase (NemA) is a xenobiotic reductase that is a member of the old yellow enzyme (OYE) family of proteins. Bacterial OYE enzymes are characterized by the ability to reduce TNT nitro groups in vivo (Williams, Rathbone et al.

2004; Gonzalez-Perez, van Dillewijn et al. 2007; Roldan, Perez-Reinado et al. 2008). A study by Gonzolez-Perez and colleagues established that nemA, nfsA and nfsB each play a role in the sequential reduction of TNT nitro groups (Gonzalez-Perez, van Dillewijn et al. 2007). Single knockout mutants for each gene showed no decrease in TNT reduction, double mutants of nfsA and nfsB showed marked decrease in TNT reduction but were still able to perform at 30% of wild type capacity. In the present study, a nearly 3 fold increase in nemA expression was observed upon TNT exposure. To this extent, two of the three genes previously associated with TNT metabolism by wild type *E.coli* are significantly expressed here.

Increased expression of N-hydroxyarylamine O-acetyltransferase suggests a possible role in TNT transformation.

NhoA encodes an n-hydroxyarylamine-o-acetyltransferase. A 5.08 fold increase in expression of nhoA was observed in the TNT induced expression profile. No statistically significant change in expression was observed in any other expression profile. While no previous study has described activity of this enzyme on TNT specifically, these enzymes are involved in the acetylation of hydroxylamine derivatives of N-aryl compounds such as nitroaromatics (Figure 3-6). Their role in toxicity and mutagenesis by nitro compounds has been documented (McCoy, Rosenkranz et al. 1981; Hein 2000).

The proposed mechanisms by which *E.coli* liberates a nitro group from TNT involves a Bamberger rearrangement via activity of yet unidentified enzymes

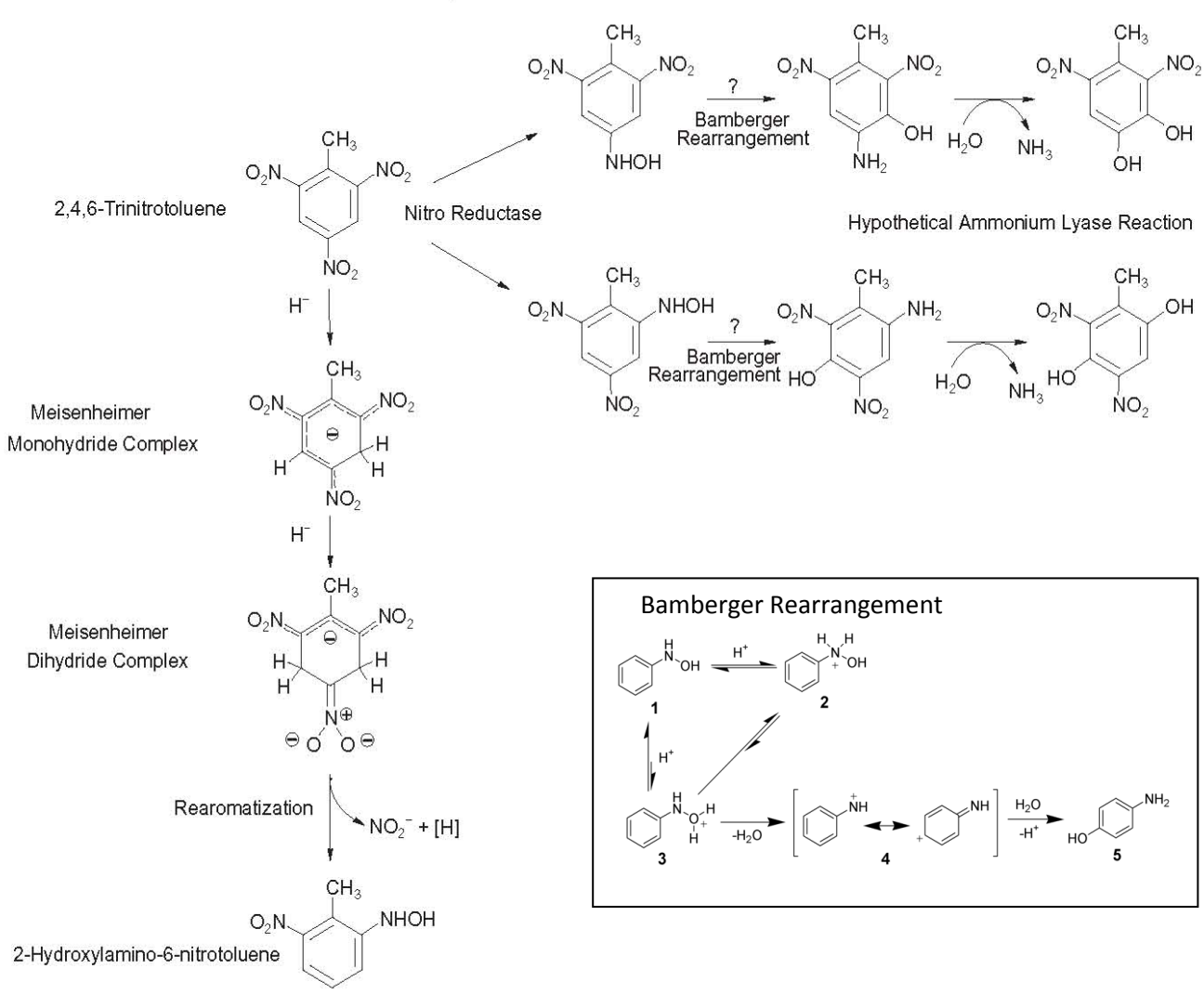


Figure 3-5: Microbial Reductive Transformation of TNT

(Figure 3-5). Acetyltransferase activity on aryl amines results in the spontaneous production of nitrenium ions which are the transition state leading to the final aminophenol product of the Bamberger rearrangement. N-acetylation of HADNT intermediates as well as the accumulation of acetyl-amino dinitrotoluenes has been observed during TNT transformation by *Pseudomonas fluorescense* (Gilcrease and Murphy 1995). As applied to the proposed TNT transformation mechanism, nhoA may be acting on the HADNTs produced by nitrogenase activity (nfsA, nfsB, and nemA) to facilitate the Bamberger rearrangement.

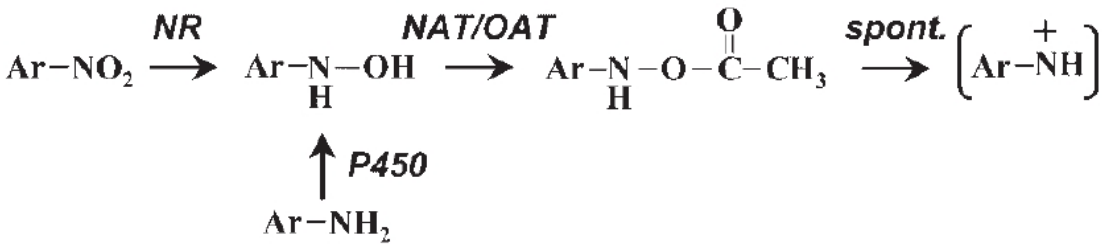


Figure 3-6: NhoA catalyzes the acetlyation of aromatic hydroxylamines (Josephy, Summerscales et al. 2002)

NR: nitroreductase, NAT: N-acetyltransferase, OAT: O-acetyltransferase. In mammalian cytochrome P450 shuttles aryl amines into this pathway.

References

Andrews, L., P. Hassanzadeh, et al. (1996). "Reactions of Nitric Oxide with Sulfur Species. Infrared Spectra and Density Functional Theory Calculations for SNO, SNO+, SSNO, and SNNO in Solid Argon." The Journal of Physical Chemistry **100**(20): 8273-8279.

Blanchard JL, C. W., Conlon EM, Pomposiello PJ (2007). Expression data from a paraquat time course experiment in wild type and SoxR deficient strains.

Brandes, N., A. Rinck, et al. (2007). "Nitrosative stress treatment of E-coli targets distinct set of thiol-containing proteins." Molecular Microbiology **66**(4): 901-914.

Brunelli, L., J. P. Crow, et al. (1995). "THE COMPARATIVE TOXICITY OF NITRIC-OXIDE AND PEROXYNITRITE TO ESCHERICHIA-COLI." Archives of Biochemistry and Biophysics **316**(1): 327-334.

Bumpus, J. A. and M. Tatarko (1994). "BIODEGRADATION OF 2,4,6-TRINITROTOLUENE BY PHANEROCHAETE-CHRYSOSPORIUM - IDENTIFICATION OF INITIAL DEGRADATION PRODUCTS AND THE DISCOVERY OF A TNT METABOLITE THAT INHIBITS LIGNIN PEROXIDASES." Current Microbiology **28**(3): 185-190.

Ding, H. and B. Demple (2000). "Direct nitric oxide signal transduction via nitrosylation of iron-sulfur centers in the SoxR transcription activator." Proceedings of the National Academy of Sciences of the United States of America **97**(10): 5146-5150.

Gilcrease, P. C. and V. G. Murphy (1995). "BIOCONVERSION OF 2,4-DIAMINO-6-NITROTOLUENE TO A NOVEL METABOLITE UNDER ANOXIC AND AEROBIC CONDITIONS." Applied and Environmental Microbiology **61**(12): 4209-4214.

Gong, P., X. Guan, et al. (2007). "Toxicogenomic analysis provides new insights into molecular mechanisms of the sublethal toxicity of 2,4,6-trinitrotoluene in Eisenia fetida." Environmental Science & Technology **41**(23): 8195-8202.

Gonzalez-Perez, M. M., P. van Dillewijn, et al. (2007). "Escherichia coli has multiple enzymes that attack TNT and release nitrogen for growth." Environmental Microbiology **9**(6): 1535-1540.

Hawari, J., A. Halasz, et al. (1999). "Biotransformation of 2,4,6-trinitrotoluene with Phanerochaete chrysosporium in agitated cultures at pH 4.5." Applied and Environmental Microbiology **65**(7): 2977-2986.

Hawari, J., A. Halasz, et al. (1998). "Characterization of metabolites in the biotransformation of 2,4,6-trinitrotoluene with anaerobic sludge: Role of triaminotoluene." Applied and Environmental Microbiology **64**(6): 2200-2206.

Hein, D. W. (2000). "N-acetyltransferase genetics and their role in predisposition to aromatic and heterocyclic amine-induced carcinogenesis." Toxicology Letters **112**: 349-356.

Inoue, M., M. Nishikawa, et al. (1999). "Cross-talk of NO, superoxide and molecular oxygen, a majesty of aerobic life." Free Radical Research **31**(4): 251-260.

Josephy, P. D., J. Summerscales, et al. (2002). "N-hydroxyarylamine O-acetyltransferase-deficient Escherichia coli strains are resistant to the mutagenicity of nitro compounds." Biological Chemistry **383**(6): 977-982.

Kaplan, D. L. and A. M. Kaplan (1982). "Thermophilic biotransformations of 2,4,6-trinitrotoluene under simulated composting conditions." Appl. Environ. Microbiol. **44**(3): 757-760.

Keseler, I. M., C. Bonavides-Martinez, et al. (2009). "EcoCyc: A comprehensive view of Escherichia coli biology." Nucleic Acids Research **37**: D464-D470.

Kulkarni, M. and A. Chaudhari (2007). "Microbial remediation of nitro-aromatic compounds: An overview." Journal of Environmental Management **85**(2): 496-512.

Kumagai, Y., M. Kikushima, et al. (2004). "Neuronal nitric oxide synthase (nNOS) catalyzes one-electron reduction of 2,4,6-trinitrotoluene, resulting in decreased nitric oxide production and increased nNOS gene expression: Implication for oxidative stress." Free Radical Biology and Medicine **37**(3): 350-357.

Liu, G., J. Zhou, et al. (2009). "The Escherichia coli Azoreductase AzoR Is Involved in Resistance to Thiol-Specific Stress Caused by Electrophilic Quinones." J. Bacteriol. **191**(20): 6394-6400.

Lo, F. C., C. L. Chen, et al. (2008). "A study of NO trafficking from dinitrosyl-iron complexes to the recombinant E-coli transcriptional factor SoxR." Journal of Biological Inorganic Chemistry **13**(6): 961-972.

McCoy, E. C., H. S. Rosenkranz, et al. (1981). "EVIDENCE FOR THE EXISTENCE OF A FAMILY OF BACTERIAL NITROREDUCTASES CAPABLE OF ACTIVATING NITRATED POLYCYCLICS TO MUTAGENS." Environmental Mutagenesis **3**(4): 421-427.

Miliukiene, V. and N. Cenas (2008). "Cytotoxicity of nitroaromatic explosives and their biodegradation products in mice splenocytes: Implications for their immunotoxicity." Zeitschrift Fur Naturforschung Section C-a Journal of Biosciences **63**(7-8): 519-525.

Minamiyama, Y., S. Takemura, et al. (1997). "Irreversible inhibition of cytochrome P450 by nitric oxide." Journal of Pharmacology and Experimental Therapeutics **283**(3): 1479-1485.

Nakanishi, M., C. Yatome, et al. (2001). "Putative ACP Phosphodiesterase Gene (acpD) Encodes an Azoreductase." Journal of Biological Chemistry **276**(49): 46394-46399.

Nunoshiba, T. (1996). "Two-stage gene regulation of the superoxide stress response soxRS system in Escherichia coli." Critical Reviews in Eukaryotic Gene Expression **6**(4): 377-389.

Pomposiello, P. J., A. Koutsolioutsou, et al. (2003). "SoxRS-Regulated Expression and Genetic Analysis of the yggX Gene of Escherichia coli." J. Bacteriol. **185**(22): 6624-6632.

Rau, J. and A. Stolz (2003). "Oxygen-Insensitive Nitroreductases NfsA and NfsB of Escherichia coli Function under Anaerobic Conditions as Lawsone-Dependent Azo Reductases." Appl. Environ. Microbiol. **69**(6): 3448-3455.

Rodriguez, C. E., J. M. Fukuto, et al. (2005). "The interactions of 9,10-phenanthrenequinone with glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a potential site for toxic actions." Chemico-Biological Interactions **155**(1-2): 97-110.

Roldan, M., E. Perez-Reinado, et al. (2008). "Reduction of polynitroaromatic compounds: the bacterial nitroreductases." Fems Microbiology Reviews **32**(3): 474-500.

Spiro, S. (2006). "Nitric oxide-sensing mechanisms in Escherichia coli." Biochemical Society Transactions **34**: 200-202.

Svensson, L., B. I. Marklund, et al. (2006). "Uropathogenic Escherichia coli and tolerance to nitric oxide: The role of flavohemoglobin." Journal of Urology **175**(2): 749-753.

Touati, D. (2000). "Sensing and protecting against superoxide stress in Escherichia coli - how many ways are there to trigger soxRS response?" Redox Report **5**(5): 287-293.

Vasil'eva, S. V., M. V. Stupakova, et al. (2001). "Activation of the Escherichia coli SoxRS-regulon by nitric oxide and its physiological donors." Biochemistry-Moscow **66**(9): 984-988.

Williams, R. E., D. A. Rathbone, et al. (2004). "Biotransformation of explosives by the old yellow enzyme family of flavoproteins." Applied and Environmental Microbiology **70**(6): 3566-3574.

Yin, H., T. K. Wood, et al. (2005). "Reductive transformation of TNT by Escherichia coli resting cells: kinetic analysis." Applied Microbiology and Biotechnology **69**(3): 326-334.

Chapter 4 : Identification of Transcriptional Biomarkers

Introduction

Microarray studies have earned the reputation of overwhelming researchers with data. This has become known as the “analytical bottleneck”. The high throughput nature results in hundreds, thousands or even tens of thousands of statistically significant results. Directly testing a specific hypothesis becomes the proverbial “needle in a haystack”. Additionally, when conclusions are drawn, they are often based on the altered expression of hundreds of genes making it difficult to decipher biologically meaningful information or use that information for downstream application.

The present study aims to identify a small number of mRNA transcripts that function as indicative biomarkers for exposure to energetic materials. These biomarkers can be thought of as a molecular fingerprint for each compound. Ideally, such a suite of biomarkers is concise, specific to the condition that they are intended to be indicative of, and are amenable to analysis through traditional molecular techniques.

Finding the justification for declaring any subset of the data as the *most* significant is a challenge. This problem can be addressed by using more conservative criteria for identifying statistically significant expression ratios. However; relying on the statistical significance threshold to achieve a dimensional reduction of data has two important limitations. First consider that the statistical significance measurements are intended to be an indication of the probability that a given observation is *real*. They are derived from a comparison of the within class variability and the between class variability; it is a measure of how well a particular observation disproves the null hypothesis that the observed results are due to chance alone.

Ideally, every experiment will be well controlled, the methods will be accurate and precise, and the analytical equipment will have been calibrated to such an extent that every observation is statistically significant at any significance threshold. Therefore, the genes eliminated by more conservative thresholds have less *analytically* significance; this is not necessarily a reflection of their actual biological significance. The second limitation is a reduction of analytical power. Genes passing the more conservative thresholds represent the best candidates with regard to statistical significance, but further analysis may prove difficult with smaller datasets.

The predictive approach taken here is purely computational in nature, thus it alleviates the need for a guided selection of the *most* significant genes. It does this without relying solely on univariate statistical thresholds. All data from each of the experiments was used during the initial discriminant analysis, providing an adequate level of analytical power. After the initial training and classification genes were selected based on their ability to identify TNT, DNT or TATP exposure within a designated confidence threshold. The method employed here results in small lists of genes with expression patterns that are characteristic of exposure. The goal here is not to describe the physiological impact of each exposure rather it is to define a set of characteristic transcriptional biomarkers.

Methods

Normalization of Microarray Data

CEL files were imported to Partek Genomics Analysis software for normalization and background correction using the GCRMA algorithm. The resulting signal intensity values were then normalized specifically to remove batch effects. The batch effect removal was necessary due to intrinsic differences in signal intensities obtained from each batch of experiments. This batch effect is attributed to variations in sample preparation efficiency and differences in production lots of materials.

Optimization of the binary Class Prediction Assays

A binary class prediction assay was optimized for each compound. This section details the optimization procedure for the TNT class prediction. An identical procedure was performed for each of the other assays.

Selection of the prediction algorithm

The normalized signal intensities were exported to a text document as tab separated values. The exported file was opened in Microsoft Excel 2007 equipped with the BRB Array Tools add-in. Normalized expression profiles from 27 MSM grown cultures exposed to either TNT, DNT, TATP, DMSO, H2O2, or MIX were categorically labeled with TNT induced expression data being labeled “TNT” and all others labeled “x”. Class prediction rules were defined to discriminate between the two groups using 6 different mathematical models: compound covariate predictor (CCP), diagonal linear

discriminant analysis, k-nearest neighbors, support vector machine (SVM), nearest centroid and Bayesian compound covariate (BCC) methods. For each method, features were selected for inclusion into the classifier that were significantly different between classes at the α=0.001 significance threshold. Each grouping scheme was ranked according to the misclassification rate achieved for that method. The BCC method consistently had the highest rate of misclassification for any given set of genes. All other methods performed similarly; support vector machines was chosen because of its intrinsic compatibility with the recursive feature elimination method of selecting genes for inclusion in the classifier.

Provision for the external validation set

An additional 9 arrays were available from preliminary experiments. These expression profiles were obtained from *E.coli* K12 cultures that were grown and exposed in Luria-Bertani media (LB). These cultures were exposed to TNT, DNT, or TATP. Other than the growth/exposure media, these experiments were carried out exactly as described for the 27 MSM cultures. These arrays were included in all subsequent analyses providing a dataset consisting of 36 microarray experiments in total.

The minimum number of samples needed to perform an accurate prediction rule was calculated using the method developed by Simon and colleagues (Dobbin and Simon 2007; Dobbin, Zhao et al. 2008). They have developed and made available a web application for calculating the optimum sample size for class prediction (http://linus.nci.nih.gov/brb/samplesize/samplesize4GE.html). First, the highest

standardized fold change was calculated between each group. Standardized fold change is defined by the following formula:

$$\text{Standardized fold change} = [(i_1/i_2)/S_a] \times 0.8$$

where i_1 and i_2 are the average signal intensities (on a log base 2 scale) for the gene with the highest fold change between groups group 1 and group 2 respectively, and S_a is the average within-class standard deviation for that gene. In addition to the standardized fold change of the most differentially expressed gene, the number of genes on the array and the proportion of total arrays that are in the largest group are considered when calculating the minimum sample size. The result of these calculations indicates the minimum number of arrays to be included in each group to achieve accuracy within 5% of the optimum. For this group of arrays a training set of 19 arrays (14 in group “not TNT” and 5 in group “TNT”) is sufficient to produce a prediction rule that is accurate within 5% of the mathematical maximum.

In order to provide profiles for an external validation of the classifier, 1 array from each set of biological replicate experiments was withheld from the dataset used to develop the prediction rules (the training set). This resulted in a final training set consisting of 24 arrays (18 from the group “not TNT” and 6 in the group “TNT”)

Optimization of the classifier size

A key benefit of the RFE method of selection genes for inclusion in the classifiers is that the size of the classifier can be preselected independently of the statistical analysis. This is as much a virtue as it is a liability. There is an inherent risk of either

selecting so few genes that there is little power to make an accurate prediction or selecting so many genes that overfitting becomes an issue. For each compound, classifiers were developed over a range of sizes. Performance was assessed based on the misclassification rate of the training set cross validation and the withheld samples.

Class Prediction and Gene Selection

All class prediction calculations were performed prior to any further statistical analysis or filtering; all 10208 GeneChip features were considered during the initial training and classification steps. Therefore, the resulting classifiers have been generated from a pool initially consisting of all *E.coli* genes with no preference based on their statistical significance.

Fluorescence intensity measurements were median-centered on a gene-by-gene basis to minimize the dominating effect of genes with relatively high baseline expression. Gene expression data were analyzed from 24 different compound exposures: 4 TNT in minimal media (MSM) , 2 TNT in Luria-Bertani media (LB) , 4 DNT in MSM, 2 DNT in LB, 2 TATP in MSM, 2 TATP in LB, 4 DMSO in MSM, 2 MIX in MSM, 2 H2O2 in MSM. There were a total of four training sets formed from different groupings of these 24 arrays, one each for TNT, DNT, TATP and a fourth class including both TNT and DNT arrays. Class prediction rules were established based on a linear support vector machine to discriminate between each class and all others.

Genes were selected for inclusion in the final classifier using Recursive Feature Elimination (Guyon, Weston et al. 2002; Chih-Wei Hsu 2009). Briefly, from the initial classification, all 10208 array features are scored based on their overall contribution to

the classification determinant. Support vector machines; for example, utilizes as its determinant a distance measurement between the worst classified arrays in each class. The feature that least contributes to that determinant is eliminated from each array. A new classifier is then built using the remaining data. This process is repeated, recursively, until a pre-determined number of genes are remaining.

Class Prediction and Gene Selection for an optimized TATP classifier

An additional SVM classifier for TATP expression profiles was developed as described previously with one exception. The 12 arrays withheld from the initial training set were re-introduced forming a larger training set composed of all 36 expression profiles. There was no external validation test performed on the larger classifier because all available arrays were included in the training set.

Random Forest Class Prediction for All Exposure Classes

A random forest can be thought of as a group of hierarchical decision trees (Figure 4-1). The split at each node of each tree represents a decision that is determined by some characteristic feature or variable distinguishing the two groups created by the split. Arrays are classified based on which terminal nodes (Figure 4-1 red nodes) they fall in on the decision tree. Random forests generate multiple permutations of this decision tree and each resulting tree *votes* on the classification of a given array.

A random forest was generated using the algorithm developed by Leo Breiman (Breiman 2001). Briefly, 20 arrays are drawn randomly with replacement from the training set to serve as the training set for each decision tree; 500 decision trees are generated. At each node, 100 genes are selected at random and used as the basis of

that split. This continued until all classes have been segregated. In generating each tree, arrays are randomly drawn with replacement for inclusion in the training set. This typically results in a third of the arrays being left out of the training set for any given tree. These arrays are termed “out-of-the-bag” (OOB) (Breiman 2001). After as each tree is generated and terminal nodes are assigned classifications, the oob arrays are segregated down the tree using the prediction rules established by the training set. The error rate is recorded and the expression value of each gene is randomly reassigned. The arrays with the permuted gene values are segregated down the tree again. The average difference between the error rate of the original array and its permuted version among all trees is the raw *importance* score for that gene. This score is divided by the standard error to produce a z-score. A significance p-value is assigned based on this z score. Genes with differences in expression at the p=0.001 significance level were chosen for inclusion in the classifier.

Due to the limitations imposed by the cross validation steps, only the TNT, DNT, TATP and DMSO expression profiles were used for the fandom forest class prediction assay. Inherent to the method, each class must be composed of at least 3 arrays to achieve proper cross validation. The data formatting and preparation for the random forest prediction assay was similar to that described for the binary class prediction; the one difference being the class labels them. Each of the 20 arrays was labeled according to their compound exposure (6 TNT, 6 DNT, 4 TATP, 4 DMSO).

Assessing the Performance of the Classifiers

There are four characteristics of mathematical class prediction performance: sensitivity, specificity, positive predictive value (PPV), and negative predictive value. The value of

each variable is calculated for each class. Sensitivity represents the probability that an expression profile of class N will be identified as such.

Specificity represents the probability that an expression profile of not of class N will identified as a non class N profile. PPV represents the probability that a profile that has been identified as class N is truly a class N profile. NPV represents the probability that a profile that has been identified as non class N is indeed not a class N profile.

These variables are calculated as follows:

Let, for some class N,

n11 = number of class N samples predicted as N

n12 = number of class N samples predicted as non-N

n21 = number of non-N samples predicted as N

n22 = number of non-N samples predicted as non-N

Then the following parameters can characterize performance of classifiers:

Sensitivity = n11/(n11+n12)

Specificity = n22/(n21+n22)

Positive Predictive Value (PPV) = n11/(n11+n21)

Negative Predictive Value (NPV) = n22/(n12+n22)

Binary Class Prediction

The gene expression profiles are inherently high-dimensional due to the number of genes included in the initial classification. The use of RFE to select genes that make up the classifier may mitigate some of the risk of over-fitting the data because the classifier is rebuilt at every step using each subset of the data. However, because the

“not TNT” class is made up of otherwise unrelated samples, the within-class variability of any given gene’s expression can be high. Usually, the goal of classification methods such as those presented here is to be able to identify a single unknown sample that was not a part of the initial training set. Therefore the risk of misclassification of individual samples must be addressed (Radmacher, McShane et al. 2002).

To address the risk of spurious classification of individual samples, the misclassification rate of the classifier was assessed through leave-one out cross validation (Molinaro, Simon et al. 2005). Briefly, an array was removed from the training set and the prediction rule, using the existing classifier, is built based on the remaining arrays in the training set. The removed array is then identified based on the new prediction rule. This is repeated for each array in the training set. The misclassification rate of the cross-validation serves as a measure of the appropriateness of the chosen genes.

Cross validation alone may not be sufficient to accurately demonstrate the reliability of the developed classifier. Several researchers have reported low cross validated misclassification rates when assessing classifiers built from experiment groupings that were made from technical replicate experiments (Ambroise and McLachlan 2002; Radmacher, McShane et al. 2002; Molinaro, Simon et al. 2005).

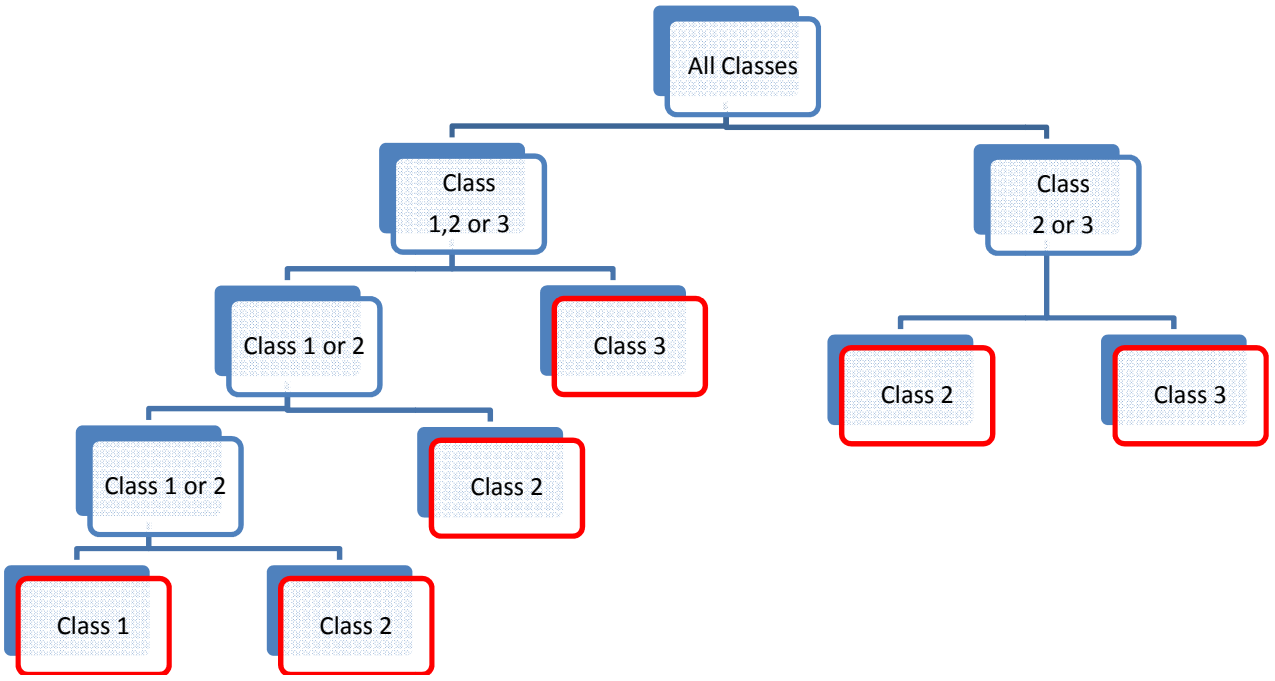


Figure 4-1 Example of a Class Prediction Decision Tree

One would expect no reliable discrimination to be had from a training set composed entirely of replicate experiments as there no real difference between classes. Therefore any discrimination is actually due to chance alone and the cross validated misclassification rates should be near 50%. Radmacher and collogues found misclassification rates as low as 0.1% although the expression profiles making up each of the opposing classes were generated from replicate data sets. Therefore, in some situations, the results of the cross validation tests do not conclusively indicate the accuracy of the predictor.

A better assessment of the classifier performance is achieved through parametric permutations of the cross validation assay. The purpose of the permutation testing is to determine the statistical significance of the misclassification rate determined by the cross validation. This is achieved by randomly reassigning the class labels of each array in the training set. A new cross validation is performed on the re-labeled training set. This process is repeated n times with n being designated number of permutations to test. The probability that the differences identified between the two classes are due to chance alone is a function of the misclassification rate for all cross validated permutations of the array classes. This serves as a statistical significance measurement of the developed classifier.

Due to the reiterative nature of the recursive feature elimination, a minimum of 10,199 classifiers are built during the process because removal of each feature requires the rebuilding of the classifier. As a consequence an error made at any step could negatively affect all consecutive steps. To control this, the cross validation and

permutation assays were performed at each iteration of the recursive feature elimination.

Random Forest Class Prediction

Performance assessment for the random tree analysis is performed on the OOB arrays. For each tree, the OOB arrays are segregated according to the prediction rules established during the previous classification steps. The percentage of misclassifications averaged among all generated trees for any given class is interpreted as a measure of performance. There is no need for permutation testing because the performance for each tree is validated by a data set that is external to that trees training set.

Results

The TNT classifier

Recursive feature elimination based on prediction rules established using support vector machines was used to select 10 genes that are sufficient to distinguish the gene expression profiles obtained from TNT exposed cultures from those obtained from cultures exposed to DNT, TATP, H2O2 or MIX Table 4-1 A. The cross-validation permutation testing resulted in no misclassifications in any of the 2000 permutations. Therefore statistical significance of the classifiers predictive ability is $p=0.0005$ ($5x10^{-4}$), the mathematical maximum significance for an assay consisting of 2000 permutations (Table 4-1). Likewise, the external validation using the 12 withheld samples produced no misclassifications.

A preliminary assay using another method, Prediction Analysis of Microarrays, suggested that a 2-gene classifier yielded the optimum misclassification rate for that method. While, in terms of misclassification rate, no improvements can be made on the 10-gene classifier, a classifier consisting of only 2 genes may be desirable for future applications. The SVM was performed again, this time selecting only 2 genes. As expected, these genes were the two highest ranked genes in the larger classifier. This classifier performed as well as the larger classifier. The compositions and performance of the resulting classifiers are presented in Table 4-1.

Table 4-1: Transcriptional Fingerprint for TNT exposure

Probe	Gene	Fold Change	10 Weight	2 Weight
1766518_s_at	azoR	131.40	0.0958	0.3227
1762628_s_at	Unannotated	22.39	0.0972	0.2740
1762061_s_at	ybiJ	17.78	0.0703	N/A
1761862_s_at	c1838	12.46	0.0688	N/A
1759339_at	oxyS	7.31	0.0643	N/A
1763543_s_at	marR	9.56	0.0637	N/A
1761528_s_at	soxS	7.04	0.0506	N/A
1768514_s_at	nhoA	2.47	0.0543	N/A
1760938_s_at	marA	3.51	0.0610	N/A
1760593_s_at	fldA	1.64	0.0558	N/A

The threshold for the 10 Gene predictor is 7.009
The threshold for the 2 Gene predictor is 6.238

Performance of the TNT Support Vector Machine Classifiers

$P=5 \times 10^{-4}$

B

Class	Sensitivity	Specificity	PPV	NPV
TNT	1	1	1	1
x	1	1	1	1

Recursive feature elimination (RFE) was applied to select 10 genes that distinguish the TNT generated expression profiles from all other exposures. These 10 genes serve as a transcriptional fingerprint that is indicative of TNT exposure. All 10 genes are statistically significant at the FDR adjusted $p \leq 0.05$ level in Chapter 2. The fold change reported is the TNT vs. DMSO fold change as described in Chapter 2. A Prediction Analysis of Microarrays (PAM) identified two genes as the minimum number of genes required to achieve the lowest misclassification rate using that method. These same two genes were identified using the SVM classifier with RFE gene selection. The gene weights of the 10-gene classifier and 2-gene classifier are listed in the columns labeled "10 Weight" and "2 Weight" respectively.

The DNT Classifier

Recursive feature elimination based on prediction rules established using support vector machines was used to select 10 genes that best distinguish the gene expression profiles obtained from DNT exposed cultures from those obtained from cultures exposed to TNT, TATP, H2O2 or MIX. The statistical significance of the classifiers predictive ability is $p=2x10^{-3}$ based on 2000 permutations of the cross validation test. The specific nature of the classifier's significance reflected in the performance variables. The composition and performance of the resulting classifier is presented in Table 4-2.

The TATP Classifier

Recursive feature elimination based on prediction rules established using support vector machines was used first to select 10 genes that best distinguish the gene expression profiles obtained from TATP exposed cultures from those obtained from cultures exposed to TNT, DNT, H2O2 or MIX. The composition of the resulting classifier is presented in Table 4-4.

A second classifier was developed in an attempt to improve the low specificity observed in the cross-validation of the first classifier. The rational governing the second classifier's development was simple: The additional analytical power offered by the inclusion of more training arrays should improve the performance of the resulting classifier. The composition of the optimized classifier is presented in Table 4-5.

The statistical significance of the initial classifier's predictive ability is p=0.07 based on 2000 permutations of the cross validation test. The optimized classifier has a

Table 4-2: Transcriptional Fingerprint for DNT exposure.

Probe	Gene	Fold Change	Weight
1762699_at	micF	3.10	0.3469
1764812_s_at	nhoA **	1.71	-0.2402
1762628_s_at	Unannotated Probe	1.05	-0.2326
1769097_at	rydB *	2.59	0.2442
1760063_s_at	ybdK *	-1.53	-0.1792
1761847_s_at	lysA *	1.42	0.1734
1761146_s_at	livG	-2.62	-0.1609
1768185_at	Ybfd *	-2.41	-0.1676
1768540_at	c4419	3.19	0.1657
1768359_s_at	fumC	6.70	0.1584

The threshold for the Support Vector Machine predictor is 4.85

An asterisk (*) indicates that the gene was not found to be significantly different in chapter 2.
** nhoA in this list is not the same probe set as the nhoA identified as a TNT biomarker.

Performance of the DNT Support Vector Machine Classifier

P=2 x 10^{-3}

Class	Sensitivity	Specificity	PPV	NPV
DNT	1	0.944	0.857	1
x	0.944	1	1	0.857

Recursive feature elimination was applied to select 10 genes that best distinguish the DNT generated expression profiles from all other exposures. These 10 genes serve as a transcriptional fingerprint that is indicative of DNT exposure. Only 5 genes are statistically significant at the FDR adjusted p ≤ 0.05 level described in chapter 2.

Table 4-3: Identification of samples during the external validation.

	Predicted Exposure				
True Exposure	**TNT**	**DNT**	**NT**	**TATP**	**RF**
DMSO	x	x	x	x	DMSO
DNT	x	DNT	NT	x	DNT
TNT	TNT	**DNT**	NT	x	TNT
TNT	TNT	x	NT	x	TNT
DNT	x	DNT	NT	x	DNT
TATP	x	x	x	TATP	**DMSO**
H2O2	x	x	x	x	excluded
DMSO	x	x	x	x	DMSO
MIX	x	x	x	x	excluded
TNT_LB	TNT	x	NT	x	TNT
DNT_LB	x	DNT	NT	x	DNT
TATP_LB	x	x	x	TATP	TATP

One array from each set of biological replicates was withheld from the training set and subjected to each classifier for identification. The H2O2 and MIX were excluded when testing the random forest classifier because those arrays were not included in the training set.

Table 4-4 : Transcriptional Fingerprint for TATP exposure.

Probe	Gene	Fold Change	Weight
1760518_s_at	**torA***	-1.56	-0.3898
1759632_s_at	**codB***	2.45	0.2875
1769254_s_at	**ydjY***	1.31	0.2227
1762007_s_at	**cusB**	-2.64	-0.1657
1762177_s_at	**metA***	2.55	0.1746
1768650_s_at	**gntK**	3.84	0.1931
1762115_s_at	**cusF**	-3.65	-0.1510
1768442_s_at	**zntA***	1.90	0.1376
1761682_s_at	**cpxP***	-1.36	-0.2272
1764983_s_at	**cpxP***	-1.60	-0.1859

The threshold for the Support Vector Machine predictor is 4.85

An asterisk (*) indicates that the gene was not found to be significantly different in chapter 2.

Performance of the TATP Support Vector Machine Classifier

P=0.07

Class	Sensitivity	Specificity	PPV	NPV
TATP	0.5	0.95	0.667	0.905
x	0.95	0.5	0.905	0.667

Recursive feature elimination was applied to select 10 genes that best distinguish the TATP generated expression profiles from all other exposures. These 10 genes serve as a transcriptional fingerprint that is indicative of TATP exposure. Only 3 genes are statistically significant at the FDR adjusted p ≤ 0.05 level described in chapter 2.

Table 4-5 : Optimized Transcriptional Fingerprint for TATP exposure

Probe	Gene	Fold Change	Weight
1768650_s_at	gntK	3.85	0.3236
1764983_s_at	spy*	-1.61	-0.2601
1761682_s_at	cpxP*	-1.36	-0.2276
1760518_s_at	torA*	-1.56	-0.3567
1761379_s_at	tdcG*	-1.40	-0.2288
1767449_s_at	yqfA*	-2.13	-0.2144
1766131_s_at	yagU*	1.67	0.2198
1764483_s_at	spy*	-1.61	-0.1837
1762115_s_at	cusF	-3.65	-0.1999
1762007_s_at	cusB*	-2.64	-0.1597

The threshold for the Support Vector Machine predictor is -10.086

An asterisk (*) indicates that the gene was not found to be significantly different in chapter 2.

Performance of the Optimized TATP Support Vector Machine Classifier

$P=5 \times 10^{-4}$

Class	Sensitivity	Specificity	PPV	NPV
TATP	0.833	1	1	0.968
x	1	0.833	0.968	1

Recursive feature elimination was applied to select 10 genes that best distinguish the TATP generated expression profiles from all other exposures. This optimized classifier was generated from a training set consisting of all 36 arrays. No arrays were withheld to perform an external validation test. Only 2 genes are statistically significant at the FDR adjusted p ≤ 0.05 level described in chapter 2.

significance of p=0.0005. The specific nature of the classifier's significance reflected in the performance variables (Table 4-4 and Table 4-5).

A combined Nitrotoluene Classifier

Recursive feature elimination based on prediction rules established using support vector machines was used to select 5 genes that best distinguish the gene expression profiles obtained from both TNT and DNT exposed cultures (NT) or from those obtained from cultures exposed to TATP, H2O2 or MIX (Table 4-6). The statistical significance of the classifiers predictive ability is p=0.0005 based on 2000 permutations of the cross validation test. The specific nature of the classifier's significance reflected in the performance variables. The composition and performance of the resulting classifier is presented inTable 4-6.

Random Forest

The random forest prediction resulted in 37 genes that are significant at the p=0.001 level. The composition and performance of the random forest classifier are presented in Table 4-7.

Table 4-6 Transcriptional Fingerprint for Nitrotoluene Exposure

Probe	Gene	Fold Change	Weight
1762699_at	**micF**	**3.75**	0.4677
1767981_s_at	**ompF**	**-2.80**	-0.2441
1762061_s_at	**ybiJ**	**5.61**	0.2925
1763931_s_at	**inaA**	**2.19**	0.3447
1761146_s_at	**livG**	**-3.95**	-0.1953

The threshold for the Support Vector Machine predictor is 7.291

Performance of the NT Support Vector Machine Classifier

$P=5 \times 10^{-4}$

Class	Sensitivity	Specificity	PPV	NPV
DNT	1	1	1	1
x	1	1	1	1

Recursive feature elimination was applied to select 5 genes that best distinguish the two nitrotoluene (TNT and DNT) generated expression profiles from all other exposures. These 5 genes serve as a transcriptional fingerprint that is indicative of nitrotoluene exposure.

Table 4-7: Composition of the Random Forest Classifier

Probe	Gene	Probe	Gene
1762061_s_at	ybiJ	**1764272_s_at**	guaC
1762628_s_at	---	**1762674_s_at**	torC
1764316_s_at	yhhW	**1760223_s_at**	yhaK
1766518_s_at	azoR	**1760768_s_at**	marB
1759339_at	oxyS	**1763981_s_at**	aceA
1760797_s_at	---	**1768822_s_at**	ybiM
1761485_s_at	tatA	**1762825_s_at**	iscA
1765694_s_at	ypfH	**1759182_s_at**	yhcN
1768980_s_at	c3865	**1764745_at**	yqjG
1761964_s_at	trxC	**1763422_s_at**	smpA
1761862_s_at	c1838	**1762699_at**	micF
1759494_s_at	bhsA	**1764483_s_at**	spy
1768064_s_at	sgrS	**1766286_s_at**	yhaV
1769137_s_at	qorB	**1764937_s_at**	torD
1760911_s_at	yceI	**1759429_s_at**	feoB
1759254_at	c4973	**1762672_s_at**	yqjF
1765578_s_at	ygiD	**1760938_s_at**	marA
1763835_s_at	---	**1766996_s_at**	c2142
1761528_s_at	soxS		

Performance of the Random Forest Classifier

Class	Sensitivity	Specificity	PPV	NPV
DMSO	0.25	0.812	0.25	0.812
DNT	0.667	0.786	0.571	0.846
TATP	0.75	1	1	0.941
TNT	1	1	1	1

Discussion

The SVM Classifiers

With only one exception, each of the SVM classifiers is able to correctly identify all of the expression profiles used in the external validation tests (Table 4-3); the one exception being the DNT classifier's incorrect classification of a TNT expression profile as DNT. This is, arguably, an inconsequential misclassification; any practical application of this assay would weigh DNT and TNT exposure similarly. It was for this reason that the nitrotoluene classifier was developed.

The DNT SVM classifier is sensitive, specific, and will, with 100% certainty, rule out the possibility of DNT exposure. The PPV is 0.875 indicating that 85.7% of profiles identified as DNT, in fact, result from DNT exposure. Inversely, this suggests a false positive rate of 14.3% which corresponds to a single misclassification during the cross validation. This was once again a situation in which a TNT profiles was incorrectly classified as DNT. The nitrotoluene classifier provides improved performance at the cost of a loss of the ability to resolve the difference between the TNT and DNT profiles.

The initial, 10-gene TATP classifier is specific but lacks sensitivity. Initially, it would seem that the classifier is of relatively limited use. There is a 50% chance of incorrectly identifying TATP profiles as *something* other than TATP. There are three considerations that possibly mitigate the effect of spuriousness sensitivity. (i) While the sensitivity is low, the specificity is high. With 91% certainty, the classifier will rule out the possibility of TATP exposure. This translates to a low false negative rate. (ii) The PPV

indicates a false positive rate of only 33%. (iii) With a pvalue of 0.07, the classifier is much more accurate than what would be expected by chance alone.

The most likely explanation for the weak performance of the TATP classifier is the fact that the composition of the training set was less than the optimum identified by the sample size calculation (Dobbin and Simon 2007; Dobbin, Zhao et al. 2008). The ideal training set includes at least 60 samples with 10 belonging to the TATP class. The option of retrieving microarray experiments from public databases is feasible, and was attempted. The problem with this approach is that there are no TATP exposure experiments in any of the public databases, nor has expression data been published in the literature. Without adding additional arrays to the TATP training class, the addition of non-TATP training arrays serves only to marginally strengthen the specificity and NPV. However, it did little to improve the specificity or PPV of the classifier. It was possible to develop an improved TATP classifier by reintroducing the withheld array experiments for inclusion in the training set. This has the disadvantage of eliminating the pool of arrays to be used as an external validation tests. This is an acceptable compromise. The initial TATP classifier correctly identified all of the external arrays; there is no reason to assume that an optimized classifier with a higher significance value and specificity would perform poorly.

An inherent limitation of the SVM method, as applied in the present study, is the binary nature of the prediction. Profiles are identified, for example, as being the result of TNT exposure, or *any* other compound. Collectively, the four SVM classifiers provide a level of flexibility for identifying each exposure individually; such an application would necessitate a hierarchically structured analysis. Each compound would be eliminated

sequentially as a possibility. The benefit to such an analysis is that it is adaptable. Additional sets of biomarkers can be identified and added as needed. If identification of additional compounds is desired it may not be necessary to alter the existing classifiers. Additionally, these classifiers were developed under the mantra of “less is more”, with the goal being to identify the smallest subset of genes that produce an accurate classifier. They were developed from the top down, eliminating genes sequentially based on how well the collective set of genes identified the members of the training set. Strengthening the classifier is simply a matter of adding more profiles to the training set. This was made evident by the improved performance of the optimized TATP classifier (Table 4-5).

The Random Forest Classifier

The random forest classifier was developed as a larger suite of transcriptional biomarkers for specific identification of each expression profile in tandem. The result is a list of 37 genes (Table 4-7). The external validation test was successful with one exception. One TATP expression profile was incorrectly identified as a DMSO expression profile. The performance variables suggest that, in terms of its predictive accuracy, TNT>TATP>DNT>>DMSO. This was no surprise considering that TNT>TATP>DNT in terms of the number of genes with significant differences in expression resulting from exposure to each compound. Additionally, all 25 of the 30 genes in which a statistically significant change in expression was observed upon DNT exposure were also observed to have similar changes in expression upon TNT exposure. Thus the performance of the classifier is a reflection of the overall

differences in the expression profiles generated from each exposure to each compound. DMSO is the control condition, all exposures contained equal amounts of DMSO. This is likely the reason for the poor performance when identifying the DMSO arrays. The control state, from a training standpoint, is poorly defined due to the experimental design.

Intuitively, since all exposures include DMSO as the solvent, the resulting differences in expression profile between the experimental and control arrays represents the effect of each compound beyond the effects observed in the control. In those cases, the training set consists of classes in which the exposure conditions used to generate each profile are distinctly different from those found in all other classes. For example, the only class that contains expression profiles which were the result of TNT exposure is the TNT class. This is not the case for the DMSO class. All classes contain expression profiles which were generated from cultures exposed to DMSO. Thus the control arrays are likely a poor choice for a class intended to represent the default state of “no exposure”.

Using the random forest classifier eliminates the need for a binary analysis. A profile can be identified as resulting from exposure to TNT, DNT, or TATP from a single suite of transcriptional biomarkers. Though this does not necessarily result from a smaller list of genes, in fact, there are 37 genes in the random forest predictor as opposed to the 30 included in the 3 SVM classifiers combined. However, the random forest predictor is much faster. The machine used for the computational analyses is equipped with a dual core 2.13 GHz processor and 2Gb of memory. Each of the SVM predictors required a minimum of 5 hours (15+ total) of processing time to develop the

classifier and perform the permutation tests. Calculations for the random forest, which included all 3 compounds, were completed in less than an hour. In the identification of future samples, this time discrepancy is mitigated once the classifier has been developed because predictions using the classifier only involve the selected genes.

References

Ambroise, C. and G. J. McLachlan (2002). "Selection bias in gene extraction on the basis of microarray gene-expression data." Proceedings of the National Academy of Sciences of the United States of America **99**(10): 6562-6566.

Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5-32.

Chih-Wei Hsu, C.-C. C., and Chih-Jen Lin (2009). A Practical Guide to Support Vector Classification Technical Report. Taipei, National Taiwan University.

Dobbin, K. K. and R. M. Simon (2007). "Sample size planning for developing classifiers using high-dimensional DNA microarray data." Biostatistics **8**(1): 101-117.

Dobbin, K. K., Y. Zhao, et al. (2008). "How large a training set is needed to develop a classifier for microarray data?" Clinical Cancer Research **14**(1): 108-114.

Guyon, I., J. Weston, et al. (2002). "Gene selection for cancer classification using support vector machines." Machine Learning **46**(1-3): 389-422.

Molinaro, A. M., R. Simon, et al. (2005). "Prediction error estimation: a comparison of resampling methods." Bioinformatics **21**(15): 3301-3307.

Radmacher, M. D., L. M. McShane, et al. (2002). "A paradigm for class prediction using gene expression profiles." Journal of Computational Biology **9**(3): 505-511.

Chapter 5 : Conclusions

Exposure to energetic materials results in distinct patterns of gene expression

The microarray experiments and subsequent gene expression profiling were conducted in the testing of the first major hypothesis:

I. *Microbial exposure to energetic materials will result in distinct, characteristic patterns of gene expression.*

This hypothesis was directly addressed through contrasts of the expression ratios resulting from each compound's exposure. Consistently, TNT, DNT, and TATP exposure resulted in patterns of gene expression that are not found in any other exposure including the two alternate control exposures (H_2O_2, or H_2O_2 + Acetone). From the composition of the gene expression profiles alone, it is reasonable to conclude that there are clearly differences in the transcriptional profile obtained from cultures exposed to each compound. This is definitive confirmation of the stated hypothesis.

TATP is not sensed as a xenobiotic toxin by *E.coli, S.cerevisiae* or *P. putida*

P. putida and *E.coli* expression profiles resulting from TATP exposure were composed of genes facilitating metabolism of cyclic organic compounds. The most dominant functional class in the *E.coli* profile was galactitol metabolism, also present were genes involved in metabolism of various hexose sugars. Furthermore there were significant decreases increase in expression of genes characteristic of xenobiotic compounds and drug resistance. The data does seem to suggest a toxic interaction however. There were significant similarities between the TATP and peroxide exposures

for both the *E.coli* and *S.cerevisiae* cultures. Many of these include increases in DNA repair mechanisms and copper/iron homeostasis.

Exposure to energetic materials can be identified from the resulting gene expression profiles alone

Indicative transcriptional biomarkers were developed for exposure to each of the energetic materials. The *E. coli* training set used to develop the SMV classifiers consisted of expression profiles obtained from both MSM and LB cultures. The result is a collection of classifiers that is sufficient to indicate exposure of TNT, DNT, or TATP in each of the externally tested *E.coli* cultures. Further, the LB media represents a complex chemical environment. It has a vast array of constituents that are of higher concentration than the TNT, are biologically available and have well documented effects on the global gene expression and metabolism of the cultures. The ability of these transcriptional biomarkers to identify each compound under such potentially masking conditions is evidence that these transcriptional fingerprints represent the core transcriptional changes that are characteristic of these materials.

These genes represent ideal candidates for future development of PCR based assays as well as bioluminescent bioreporters for these compounds. Further, the *ability* to develop these classifiers serves as definitive evidence that the stated hypothesis is true; these energetic materials indeed induce distinct, characteristic changes in transcription, and these alterations are indicative of exposure to the respective

compound. These changes have been shown to be distinct from those induced by DMSO, hydrogen peroxide, or a mixture of acetone and hydrogen peroxide.

TNT induces expression of the superoxide stress response regulon in *E.coli*.

TNT exposure resulted in the increased expression of 31 genes associated with the SoxRS response regulon. Among these genes is the SoxS transcriptional activator.

TNT induced alterations in gene expression are suggestive of nitrosative stress.

The physiological effects of reactive nitrogen stress are often similar to those caused by reactive oxygen. This is because, in many cases, the cellular components involved in responding to both classes of compounds are the same. Typically, the mechanism of reactive nitrogen stress involves displacement of sulfur atoms from iron sulfur centers, sulfur containing amino acids and thiol containing proteins.

- Both *E.coli* and *S. cerevisiae* have significantly increased expression of genes involved in iron-sulfur center assembly, sulfate transport, and sulfur containing amino acid synthesis.
- The *E.coli* expression profiles indicates increased expression cysteine biosynthetic networks yet cellular abundance of cysteine is decreased (Metabolite profile: Appendix 1)
- AzoR which has a 62 fold increase in expression responds specifically to thiol damage resulting from covalent modification.

- Transcription of SoxS is regulated by the transcription factor SoxR. SoxR activation can occurs through displacement of its iron-sulfur center.

E.coli NhoA may play a role in the reductive transformation of TNT

A 5 fold increase in nhoA was observed upon exposure to TNT. While increased expression of this gene was not observed in any other gene expression profile, nhoA was selected by the support vector machine as a potential transcriptional biomarker for DNT as well. The inclusion of this gene as biomarkers for TNT and DNT exposure but none of the other chemical exposures confirms that this response is specific to nitroaromatics. The reported activity of this enzyme is capable of initiating the HADNT Bamberger rearrangement proposed by Gonzolez and others. It is possible that this protein is directly involved in the liberation of nitrogen from TNT via reductive transformation.

Suggested Future Work

Time-course Gene Expression Profile

The mechanistic insight into the interactions with each compound was less productive than what was originally envisioned. This was realized very early in the project; as early as the first preliminary set of *E.coli* microarrays. By taking an expression "snapshot", that is measuring expression values at a single time point and at a single compound concentration, there is simply not enough information to identify all of the biologically relevant changes in gene expression. What is missing is a direct link between each compound exposure and the specific changes in gene expression observed. This raises several pertinent questions:

- Are these changes primary or secondary responses?
- Are they due to a specific interaction with each compound or are they the result of a combination of less specific primary responses?
- Are the differences observed among the expression profiles truly unique or are those differences limited to that specific time point? In other words, might DNT or TATP induce an expression profile similar to TNT at an earlier or later time point, or perhaps at a different concentration?

These questions are best answered through a comprehensive pathway analysis. This would require a time-course expression study. This would provide more information about nature of the response with regards to the regulatory networks involved. The ontology enrichment analysis was based on the statistical significance of a given biological process with respect to the distribution of genes in the data set. Networks and pathways that lacked significant representation in this study may emerge as highly significant in the context of the entire exposure period. Additionally, the time course

experiment may yield a more complete list of responsive genes, since the most significant changes may occur at an earlier or later time point. Although such an approach has become standard practice in the literature, it was not included in the original design of this study because these questions were outside of the scope of this project which was to simply identify biomarkers for compound exposure. That specific aim required expression data from a diverse array of compounds for the comparisons. Had the study focused on a single compound, it may have resulted in more biologically meaningful results but there would have been no way if identifying the most characteristic results.

The Effect of Growth Media

E. coli gene expression data were generated from cultures grown in both Luria-Bertani (LB) media and a minimal salts media. As expected, there were considerable differences between the two sets of expression data. A proper comparison could not be made because data resulting from DMSO control exposures were not collected for the LB grown cultures.

If TNT is inducing nitrosative stress as suggested, a comparison of the LB and MSM grown cultures may yield informative results. It has been established that the effects of nitrosative stress are more pronounced in cells exposed in LB media than defined minimal media (Flatley, Barrett et al. 2005; Spiro 2006). This is due to the fact that, although LB is regarded as a rich media, it is relatively iron poor when compared to defined media which is supplemented with iron. Direct comparisons of the TNT-LB and

TNT-MSM expression profiles support this idea but without the DMSO-LB control, any such comparison is inconclusive. However, the role of iron should be addressed by performing an expression study in iron limited MSM. It is a more direct method of assessing this relationship.

Metabolomics

The metabolite profiling study was understated to a great extent in this text. In fact, it was briefly mentioned in the conclusions and only presented in full in Appendix 1. This facet of the project is incomplete; there not enough technical/biological replicates performed to provide the statistical power to draw conclusions. This is especially true in the context of this study in which much emphasis was placed on the statistical analysis of each experiment.

The next logical step in completing the metabolomics study would be to first repeat the experiments using more replicates. The data presented in Appendix 1 resulted from a single technical replicate for each of the 2 biological replicates for each condition. Given the variability observed, there should be at least 3 technical replicates for each biological replicate. Also, specific attention should be given to TNT and DNT transformation products. From the expression data, expected metabolites would include hydroxylamino-dinitrotoluenes (HADNTs), acetylated HADNTs, and azoxy-nitrotoluenes. The evidence of oxidative and nitrosative stress would also suggest an increased abundance of s-nitrosylthiols and nitrosylated proteins, heme groups and amino acids.

The inherent limitation of the current experimental design is also worth consideration. As designed, these experiments will only yield information about the abundance of a given metabolite. What may be more meaningful is to gain information about the flux of metabolites through the cell. Are these metabolites accumulating because of increased biosynthesis or a decrease in the activities that deplete them? Are these changes in metabolite abundance originating from the metabolism of the energetic compounds or are they due to overall changes in cellular metabolism caused by exposure to the compounds? These questions could be answered using isotope-labeled substrates. Spiking with labeled carbon and/or nitrogen at the time of the exposure should help resolve these questions. Labeled TNT, DNT and TATP would allow for the identification of compounds originating from metabolism of each compound.

References

Flatley, J., J. Barrett, et al. (2005). "Transcriptional Responses of Escherichia coli to S-Nitrosoglutathione under Defined Chemostat Conditions Reveal Major Changes in Methionine Biosynthesis." Journal of Biological Chemistry **280**(11): 10065-10072.

Spiro, S. (2006). "Nitric oxide-sensing mechanisms in Escherichia coli." Biochemical Society Transactions **34**: 200-202.

Appendix

Appendix 1

Metabolite Profile of TNT exposed *E.coli* Cultures

Metabolomic analysis

The gene expression data was supplemented with an analysis of the changes in intracellular metabolite abundance associated with TNT exposure. *E.coli* cultures were grown and exposed according to the method described for the microarray experiments.

Extraction and ESI-MS/MS

Cells were lysed and metabolites were collected following one hour of exposure. 10mL of each culture was passed through a 0.45µm nylon filter to collect cells. The filters were immediately transferred , cell side down, to a petri dish containing the -20℃extraction solvent (40% MeOH, 40% ACN, and 20% 0.1 M formic acid). Filters were incubated at -20℃ for 15 minutes.

After the 15 minute incubation, filters were forcibly rinsed in the extraction solvent using a pasture pipette to remove cellular debris and metabolites from the filter. Extraction solvent was collected and centrifuged for 5 minutes at 4°C at 1300 rpm in pre-cooled microcentrifuge tubes.

For each extraction 2 auto sampler vials were prepared containing 36µL 15% ammonium bicarbonate and 5µL internal standard. The positive polarity standard was Tris (+) and the negative polarity standard was benzoic acid (-).

420µL of the cell extract was added to each vial and thoroughly mixed by vortexing.

Metabolites were analyzed via ESI-MS. Metabolite abundance was normalized to each of the internal standards.

Table of Metabolite Abundances from TNT exposed *E.coli* Cultures

KEGG ID	Metabolite	Fold Change
C00498	ADP-D-glucose	7.15
C00300	Creatine	6.08
C00024	(-)acetyl-CoA	5.49
C00100	Propionyl-CoA	5.19
C00015	UDP	5.03
C00460	dUTP	4.69
C00440	5-methyl-tetrahydrofolate	4.17
C00091	Succinyl-CoA	4.14
C02557	Methylmalonyl-CoA	4.14
C00363	TDP	3.24
C00061	FMN	3.18
C00214	Thymidine	2.55
C00286	dGTP	2.54
C01179	Hydroxyphenylpyruvate	2.45
C00767	Glucarate	2.43
C00257	Gluconate	2.36
C00417	Aconitate(cis+trans)	2.04
C02341	trans aconitate	2.04
C04256	N-acetyl-glucosamine-1-phosphate	2.02
C00169	Carbamyl phosphate	1.94
C00015	UDP	1.89
C00864	Pantothenic acid	1.80
C00526	Deoxyuridine	1.79
C00882	Dephospho-CoA	1.69
C06400	Trehalose	1.65
C00074	Phosphoenolpyruvate	1.63
C00254	Prephenic acid	1.51
C06893	6-Phospho-D-gluconic acid	1.51
C00043	UDP-N-acetyl-glucosamine	1.44
C00365	dUMP	1.42
C03539	S-Ribosyl-Homocysteine(SRH)	1.42
C00158	Citrate+IsoCitrate	1.38
C00311	isocitrate	1.38
C00437	N-acetyl-Ornithine	1.34
C00354	Fructose-1,6-bisphosphate	1.32
C00236	glycerate-1,3-diphopshate	1.32
	Acadesine (AICAR without phosphate)	1.22
C00103	Glucose-1P	1.22
C00711	Malate	1.22

KEGG ID	Metabolite	Fold Change
C02672	D-hexose-phosphate	-1.21
C00062	Arginine	-1.23
C00385	Xanthine	-1.24
C00542	Cystathionine	-1.24
C00408	DL-Pipecolic acid	-1.24
C00438	N-Carbamoyl-L-aspartate	-1.25
C00806	Tryptophan	-1.26
C02057	Phenylalanine	-1.26
C01103	Orotidine-phosphate	-1.27
C16434	isoleucine	-1.27
C16439	(Iso)Leucine	-1.27
C00197	3-Phospho-glycerate	-1.28
C00303	Glutamine	-1.30
C00689	Trehalose-6-Phosphate	-1.31
C00979	O-acetyl-L-serine	-1.37
C00295	Orotic acid	-1.41
C00559	Deoxyadenosine	-1.43
C16436	Valine	-1.45
C00499	Allantoic acid	-1.54
C00624	N-acetyl-glutamate	-1.54
C00119	5-Phospho-D-ribose-1-diphosphate(PRPP)	-1.57
C00114	Choline	-1.66
C00736	Cysteine	-1.68
C05512	Deoxyinosine	-1.70
C00118	D-glyceraldehdye-3-phosphate	-1.72
C01026	Dimethylglycine	-1.75
C00025	Glutamate	-1.78
C00003	NAD	-1.79
C00353	Geranyl-pryophosphate	-1.80
C00035	GDP	-1.83
C00458	dCTP	-1.84
C00448	trans, trans-farnesyl diphosphate	-1.86
C00166	Phenylpyruvate	-1.92
C00294	Inosine	-1.93
C00008	ADP	-1.97
C00002	(-)ATP	-2.02
C01762	Xanthosine	-2.08
C00020	AMP	-2.08
C00198	Glucono-?-lactone	-2.11
C00063	(-)CTP	-2.16
C00258	glycerate	-2.26

KEGG ID	Metabolite	Fold Change
C00008	ADP	-2.28
C00131	dATP	-2.46
C00575	cyclic-AMP	-2.46
C00147	Adenine	-2.60
C05330	Homocysteine	-2.65
C00104	IDP	-2.92
C00493	Shikimate	-3.18
C00299	Uridine	-3.66
C00504	Folate	-4.46
C00362	dGMP	-7.67
C00224	Adenosine 5'-phosphosulfate (APS)	-8.66
C00108	Anthranilic acid(oABA)	-10.09
C01081	Thiamine-phosphate	-18.50

Appenix 2

Functional Characterization of the *E.coli* TATP response

Galactitol Metabolism		
gatC	galactitol-specific PTS system component IIC	2.84
gatA	galactitol-specific PTS system component IIA	1.65
gatY	tagatose-bisphosphate aldolase	1.62
gatZ	putative tagatose 6-phosphate kinase 1	1.42

Glucose Metabolism		
gntK	gluconate kinase 1 & 2	3.85
pckA	phosphoenolpyruvate carboxykinase	1.65
gatY	D-tagatose 1,6-bisphosphate aldolase 2	1.62

DNA Repair/Modification		
hsdS	specificity determinant for hsdM and hsdR	1.50
pyrD	dihydroorotate dehydrogenase 2	2.30
yebG	DNA damage-inducible protein , regulated by LexA	1.86

Amino Acid Biosynthesis			
ilvC	ketol-acid reductoisomerase, NAD(P)-binding	1.73	Valine & Isoleucine
asnA	asparagine synthetase	2.56	Asparagine
cysZ	putative inner membrane sulfate transport protein	1.58	Cysteine
hisG	ATP phosphoribosyltransferase	3.25	Histidine
hisD	histidinol dehydrogenase	1.16	
thiP	thiamine transporter membrane protein subunit	2.60	Thiamine
thiL	thiamine monophosphate kinase	2.01	
thrB	homoserine kinase	-1.74	Threonine
thrC	threonine synthase	-1.92	

Copper ion Binding and Transport		
cusA	copper/silver efflux system	-2.06
cusB	copper/silver efflux system	-2.64
cusF	periplasmic copper-binding protein	-2.99
cusX	periplasmic copper-binding protein	-3.65

Cellular Response to Stress		
rpoS	RNA polymerase sigma factor	-1.28
osmC	osmotically inducible protein	-1.45
rseA	anti-RNA polymerase sigma factor	-1.20
psiE	phosphate-starvation-inducible protein	-1.95
cusB	copper/silver efflux system	-2.64
dps	Fe-binding and storage protein , DNA starvation/stationary phase protection	-1.69
yebG	DNA damage-inducible protein , regulated by LexA	1.86

RedOx and Respiration		
c2467	putative 3-hydroxyacyl-CoA dehydrogenase	-1.01
gapA	glyceraldehyde-3-phosphate dehydrogenase	-1.19
osmC	osmotically inducible protein	-1.45
dps	Fe-binding and storage protein , DNA starvation/stationary phase protection	-1.69
torC	trimethylamine N-oxide reductase cytochrome c-type subunit	-1.87
gltA	type II citrate synthase	-1.52

Appendix 3

E.coli Genes responding to TATP but not H2O2 or MIX

Gene		Fold Change
edd	phosphogluconate dehydratase	3.50
c2215	hypothetical protein	2.55
borD	DLP12 prophage; predicted lipoprotein	2.44
fecE	KpLE2 phage-like element; iron-dicitrate transporter subunit	2.02
thiL	thiamine monophosphate kinase	2.01
yraQ	predicted permease	1.89
yedE	putative inner membrane protein	1.88
rbsC	ribose ABC transporter permease protein	1.81
ilvC	ketol-acid reductoisomerase, NAD(P)-binding	1.73
gatA	galactitol-specific PTS system component IIA	1.65
cysZ	putative sulfate transport protein	1.58
rbsB	D-ribose transporter subunit	1.56
hsdS	specificity determinant for hsdM and hsdR	1.50
rtn	hypothetical protein	1.32
hisD	histidinol dehydrogenase	1.16
yabI	conserved inner membrane protein	-1.51
gltA	type II citrate synthase	-1.52
thrB	homoserine kinase	-1.74
mglA	galactose/methyl galaxtoside transporter ATP-binding protein	-1.89
cusA	putative cation efflux system protein cusA	-2.06
1764701_s_at	unannotated probe-set	-2.18
ilvG	acetolactate synthase 2 catalytic subunit (pseudogene)	-2.21
1768644_s_at	unannotated probe-set	-2.53
cusB	copper/silver efflux system membrane fusion protein CusB	-2.64
cusC	copper/silver efflux system, outer membrane component	-2.67
cusF	periplasmic copper-binding protein	-3.65

Appendix 4

Gene expression profile of *Pseudomonas putida* resulting from TNT exposure

JCVI Locus	CMR annotation	Fold Change
---	intergenic region 2295676-2295883	32.27
PP_3426	multidrug efflux RND transporter MexF	20.89
PP_3425	multidrug efflux RND membrane fusion protein MexE	14.70
PP_3427	multidrug efflux RND outer membrane protein OprN	6.12
PP_2022	hypothetical protein	4.06
PP_1684	transporter, putative	2.62
PP_2944	sensor histidine kinase	-1.34
PP_2260	sugar ABC transporter, ATP-binding protein	-1.29
---	intergenic region 4102513-4102659	-1.23
PP_3541	transporter, MgtC family	-1.18
PP_1761	sensory box protein GGDEF family protein	-1.14
PP_5096	YGGT family protein	1.13

Appendix 5

Gene expression profile of *Pseudomonas putida* resulting from TATP exposure

JCVI Locus	CMR Annotation	Fold Change
PP2675	cytochrome c-type protein	17.04
PP2334	carboxyvinyl-carboxyphosphonate phosphorylmutase, putative	14.07
PP2335	methylcitrate synthase, putative	13.03
PP2336	aconitate hydratase, putative	11.31
PP2676	periplasmic binding protein, putative	9.92
PP2680	aldehyde dehydrogenase family protein	8.99
PP2674	quinoprotein ethanol dehydrogenase	7.72
PP2669	outer membrane protein, putative	6.81
PP2681	coenzyme PQQ synthesis protein D, putative	6.65
PP2677	hypothetical protein	5.10
PP2672	DNA-binding response regulator, LuxR family	4.18
PP0378	coenzyme PQQ synthesis protein C	4.01
---	intergenic region 2663973-2664134	3.69
PP2337	hypothetical protein	3.66
PP0056	oxidoreductase, GMC family	3.60
---	intergenic region 66596-66692	3.35
PP2663	hypothetical protein	3.18
PP0376	coenzyme PQQ synthesis protein E	2.78
PP0379	coenzyme PQQ synthesis protein B	2.74
PP2668	ABC efflux transporter, ATP-binding protein	2.70
PP0377	coenzyme PQQ synthesis protein D	2.54
PP2664	sensory box histidine kinase response regulator	2.04
PP2678	hydrolase, putative	1.79
---	intergenic region 3070430-3070782	1.78
PP2422	carboxymuconolactone decarboxylase family protein	1.49
---	intergenic region 458779-458941	1.40
PP0375	prolyl oligopeptidase family protein	1.30
PP3541	transporter, MgtC family	-1.17

Vita

Vernon Lashawn McIntosh Jr. was born to Vernon and Debra McIntosh in Memphis, TN on February 18, 1983. He graduated from the University of Tennessee in Knoxville, TN with Bachelor of Science degree in Microbiology. During his undergraduate career, he participated in research internships at the Tennessee Health Science Center in Memphis, TN and the Center for Environmental Microbiology at the University of Tennessee in Knoxville, TN. It was this research experience that motivated the pursuit of his PhD. Vernon entered the Department of Microbiology at the University of Tennessee in Knoxville Tennessee as a graduate student in August of 2005. There he worked under the guidance of Dr. Gary Sayler at the Center for Environmental Biotechnology until he completed the requirements for his PhD in May 2010.