molecular systems biology

Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates

Jeffrey M. Skerker, Dacia Leon, Morgan N. Price, Jordan S. Mar, Daniel R. Tarjan, Kelly M. Wetmore, Adam M. Deutschbauer, Jason K. Baumohl, Stefan Bauer, Ana Ibáñez, Valerie Mitchell, Cindy H. Wu, Ping Hu, Terry Hazen, Adam P. Arkin

Corresponding author: Adam P. Arkin, University of California, Berkeley

Review timeline:

Submission date:	30 November 2012
Editorial Decision:	09 January 2013
Revision received:	09 April 2013
Editorial Decision:	03 May 2013
Revision received:	09 May 2013
Accepted:	12 May 2013

Editor: Thomas Lemberger

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision 09 January 2013

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the three referees who agreed to evaluate your manuscript. As you will see from the reports below, the referees find the topic of your study of potential interest. They raise however a series of concerns and recommendations on your work, which should be convincingly addressed in a revision of this work. The suggestion provided by the reviewers are very clear in this regard.

On a more editorial level, we would also kindly ask you to deposit your data in public repositories when appropriate or as 'datasets' in supplementary information. For long term preservation, we would prefer not to rely exclusively on authors' websites. Please add a "Data availability" sub-section to Materials & Methods to list the datasets and their respective accession numbers.

We would agree with referee 3 that the title of the study should be shortened, perhaps something along the line: "Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates" or "Chemogenomic profiling of plant hydrolysate stress"

Referee reports:

Reviewer #1 (Remarks to the Author):

Project Summary
The authors generate barcoded pools of random gene knockouts in Zymomonas mobilis using TN5

transposon-bombing, testing these libraries along with a previously-generated knockout library in Saccharomyces cerevisiae for their relative fitness in the presence of miscanthus and switchgrass hydrolysates, synthetic mixtures of inhibitory hydrolysate compounds, individual hydrolysate inhibitors, and various other stress conditions. Competitive growth assays were performed using the knockout pools grown in the various conditions and fitness values for a given knockout in a specific condition were determined based on log2 normalized ratios of barcode abundance [log2(mutant-specific barcode after competition / mutant-specific barcode prior to competition)]. Barcode abundance was determined via hybridization to Affymetrix microarrays and reported fitness values averaged from multiple experiments.
Of the 1578 knockout Z. mobilis strains for which data was obtained, 1184 (63% of the 1892 annotated protein coding genes) grew sufficiently well in rich media to reach appropriate relative starting abundance for condition-specific comparisons. Of these 1184 knockouts, fitness was significantly impaired in 44 strains when grown in hydrolysate versus rich media (these 44 gene knockouts were henceforth termed putative "tolerance genes"). Of these 44 strains, 13 were confirmed to diminish growth in lignocellulosic hydrolysate, and in three instances the fitness defect was complemented via plasmid-borne expression of the wild-type gene. Analogously, the authors identified 28 putative "tolerance genes" in S. cerevisiae though did not validate these knockouts for decreased fitness nor complement these candidate tolerance genes.
Subsequently, the authors compared knockout-fitness in synthetic mixtures of hydrolysate-derived inhibitory compounds, largely recapitulating the hydrolysate fitness landscape with mixtures of 37 and 10 chemicals. Despite acceptable overall correlation, hydrolysate toxicity could not be explained for eight of the 44 putative tolerance genes using the synthetic mixtures of known inhibitors, suggesting the presence of additional hydrolysate-associated inhibitory compounds. To investigate, the authors modeled composite hydrolysate toxicity as the linear combination of individual inhibitor toxicities, identifying 16 of 37 compounds as significantly contributing to the predictive power of their model. Ultimately, the authors identify methylglyoxal, present in miscanthus hydrolysate, as the compound for which inclusion in the model most improves its ability to predict the fitness landscape of the knockouts grown in lignocellulosic hydrolysate. The authors also show that the overexpression of one identified tolerance gene improves the growth rate and ethanol productivity of the parent Z. mobilis strain in the presence of miscanthus hydrolysate.

Criticisms/Clarifications
1) Although the inclusion of methylglyoxal improved the authors' linear model, when added to their synthetic mixture SYN-10 the correlation with fitness in actual hydrolysate was worse (R2 of 0.60 versus 0.67; Figure 5). Perhaps inclusion in a more representative mixture (e.g. SYN-37, or a "SYN-16" of the 16 compounds used in their linear model), or inclusion at a different concentration (table S1 indicates methylglyoxal is by far the most prevalent compound detected in miscanthus "Batch 2", leading this referee to wonder a) why had it been previously unidentified and/or b) is the concentration actually accurate?) would improve the correlation of synthetic mixture with the actual hydrolysate. Yes, the fitness defect of six outliers is better explained in the SYN-10 + methylglyoxal mixture than SYN-10 alone, but an even greater number of new outliers are observed and the overall correlation is worse. With no acknowledgement or explanation of this discrepancy, this referee is concerned regarding the accuracy of methylglyoxal as an important hydrolysate inhibitor. Moreover, why was SYN-10 + methylglyoxal highlighted in the discussion as "a reasonable proxy for dilute-acid miscanthus hydrolysate" when SYN-10 mixture alone was better correlated with actual miscanthus hydrolysate and SYN-37 (Fig 3; R2 = 0.81) significantly better than both SYN-10 variants?

2) Additionally, this referee has no doubt methylglyoxal is toxic to the six outlying mutants, but is not convinced it is toxic to wild-type Z. mobilis (and therefore is a relevant lignocellulosic-associated inhibitory compound). If the right mutant(s) are considered, many compounds can appear toxic. For instance, galE-null Bacillus subtilis mutants rapidly lyse in the presence of exogenous galactose due to the buildup of a toxic intermediate [J Bacteriol. 1998 Apr;180(8):2265-70; MBio. 2012 Aug 14;3(4):e00184-12]. Nonetheless, wild-type B. subtilis readily metabolizes galactose and its growth is unaffected by the sugar. Figure S18 attempts to address the referee's concern regarding methylglyoxal toxicity, but no toxicity experiments are performed with wild-type Z. mobilis (they should be). Moreover, extremely minimal (in some cases imperceptible) toxicity is observed in the mutant pools, with the discrepancy in growth curves between pools larger than the discrepancy between the curves depicting presence and absence of methylglyoxal. As only one trial was performed, this referee has no confidence the toxicity is reproducible (and therefore real) and

encourages multiple trials with standard error bars before claiming methylglyoxal "contributes to overall [miscanthus hydolysate] toxicity" in Z. mobilis.

3) The final linear model ("Model-16") includes coefficients for each of the 16 compounds, reflecting their contribution to overall hydrolysate toxicity (supplementary table 10). In six instances, these coefficients are negative. If this referee understands the model correctly, this would imply the fitness of a mutant in these six compounds is inversely correlated with its fitness in bulk hydrolysate. What, if anything, distinguishes these six compounds from the remaining ten? Additionally, do mutants with increased susceptibility to these compounds show experimentally increased tolerance to bulk hydrolysate?

4) When calculating the log2 normalized fitness scores, what exactly was used as the "start" library (against which the post-selection libraries were normalized)? This referee assumes (hopes) the up-pool and dn-pool were resuscitated as described (grown for 5 hours shaking at 30C in rich media), and then split with a portion being used to seed competitive growth assays and a portion used as the "start-condition" control. If, instead, the frozen up- and dn-pools were used directly as the baseline control, then each growth experiment involves 5 hours of competitive growth in rich media not accounted for in the control, which could drastically swing relative abundances in the library (assuming a doubling time of 1.4 hrs for Z. mobilis [Appl Biochem Biotechnol. 1991 Spring;28-29:221-36], this would equate to 3.6 generations of unaccounted competition when the actual experiments were conducted with as little as 5 generations). Regardless of the true scenario, this referee requests simple clarification in the methods. If, however, the appropriate control was not used, this referee has serious reservations about the microarray data used.

5) Rather than referring to mutants as "sick", this referee would prefer a less anthropomorphized term (e.g. "growth-impaired").

6) Some justification of why the parameters for what constituted a "tolerance gene" are different between Z. mobilis and S. cerevisiae should be included

7) On page 10, where it states "...versus average fitness in batch 1 hydrolysate uncovered 11 outliers (labeled on plot in Figure 4C)...", this referee encourages a modification to the sentence that reflects the fact many more than 11 outliers are evident in Figure 4C, with a subset of 11 of these outliers overlapping with the set of 28 previously-discussed S. cerevisiae "tolerance genes".

8) On page 12, the first paragraph references figures 5C and 5D where it appears the authors intend to reference 5D and 5E, respectively.

9) Was it tested whether or not ZMO1875 improves EtOH production anaerobically? If not, this is a relatively straightforward experiment that could improve the industrial relevance of the ORF to reactor-conditions.

10) On page 23, in the methods, Supplementary Table 3 is referenced where it appears the authors intend to reference Supplementary Table 2

11) On figures 1 and 6, OD600 is plotted but no optical density values are given on either of the y-axes. Thus, although the reader can compare data-series against one another, there is no absolute reference for cell-growth.

12) On figure 3, the green colors used to differentiate between 'cytochrome c' and 'regulator' are too similar. Thus, it is impossible to distinguish between the two categories in figure 3B.

Recommendation
The majority of the criticisms this referee has are small, and readily addressed. However, as the manuscript is currently presented, this referee has major reservations regarding the claim of

methylglyoxal's role as a novel hydrolysate-associated inhibitor, as well as its contribution to improving the predictive power of the synthetic compound mixtures (points 1 and 2 above). Additionally, the referee would like to see further clarification regarding the linear model and experimental scheme (as discussed in points 3 and 4 above) before considering the manuscript for publication. However, the project addresses an important concern in the field, using innovative techniques, with thorough genetic follow-up experimentation, and reports novel findings of high import. In its current form, this manuscript's fatal flaw is its emphasis on the discovery/importance of methylglyoxal to miscanthus hydolysate toxicity. The prose, unfortunately, seems to overstate the data. Accordingly, this referee suggests authors re-submit a revised draft, with the aforementioned concerns satisfactorily addressed.

Reviewer #2 (Remarks to the Author):

In this work, Skerker et al. use chemical-genomic profiling and fitness modeling to identify genes important for hydrolysate tolerance in Zymomonas and to identify a missing toxin in dilute-acid pretreated miscanthus or switchgrass hydrolysates. Toxic byproducts of pretreatment methods present a major bottleneck for cellulosic biofuel production, since the inflicted stress limits microbial fermentation and metabolism - yet the full suite of toxins in different hydrolysates, as well as their effect on cells, remains incompletely understood. This is a major problem in renewable energy production.

Skerker et al. generate a transposon library in Zymomonas mobilis, allowing fitness profiling of strains mutated in 83% of protein-coding Z. mobilis genes. They screened several hydrolysates, known inhibitors applied in isolation, and combined inhibitors in several synthetic hydrolysates to identify genes required for optimal fitness under the different conditions. Conditions were also used to screen the S. cerevisiae deletion library, for a comparative analysis across biofuel organisms. The results were used to model the interactions of individual inhibitors - contrary to some other studies, the results suggest that most of the inhibitors (at least as studied here) act additively rather than synergistically. Based on the modeling, which implicated a missing component in synthetic hydrolysate, and the functional annotations of important genes, the group identified methylglyoxal as a missing component in the synthetic mix.

The work is very nicely done, well controlled, and well presented. This is an interesting paper on an important topic. I have mainly minor points to be addressed.

1. Throughout the paper (in text, Methods, Modeling description, and figures), the term "genes" is used when in fact the fitness data represent strains mutated at those genes. For example, rows of hierarchically clustered data are listed as genes when in fact each row represents a strain. The terminology should be clarified throughout the manuscript.

2. The term 'fitness' is used seemingly in multiple ways in the paper. In the early text, relative fitness is described as strain abundance before and after selection, but in the Methods it seems to be reflecting normalized strain abundance (signal intensity) on a single array; in other parts, fitness is used to compare growth in hydrolysate to growth in inhibitor-free medium. It would be useful to differentiate the terminology in these cases, as I became confused which 'fitness' measure was being used in different parts of the manuscript.

3. At least for the S. cerevisiae growth experiments, there is no mention of the generation times used - but this is really critical in calculating fitness. Normally, control cells are grown for exactly the same (population) generations as stressed cells, to optimally distinguish slow-growing strains from strains removed by selection. Minimally, the Methods should better describe how many generations each library went so that users of the data will have all the information.

4. The heterozygous essential knockouts was interesting. The authors implied that the species is polyploidy, but I wondered if instead they simply selected for polyploidy mutants when essential genes were being knocked out?

5. Several datasets are made available on the authors' website - these should be made available by the journal and/or deposited into a public array repository, since over time it's harder to maintain lab

websites.

6. I was not clear how the doses of stresses were chosen for the synthetic hydrolysate - it is not clear from the text if the doses were chosen to match hydrolysate? From the table, some of the doses match and others seemingly don't. Does the dose of stresses used in this analysis affect whether the stresses might interact? In other words, might their be dose-specific interactions?

7. Along the lines of interactions terms, it is often harder to find enough statistical power to identify interactions (although that probably does mean the effect size is small, as the authors suggest). I wasn't sure how the p-value for interaction was chosen, and it looked like the test correction was the stringent Bonferorni correction - I wondered if a less stringent FDR correction (Benjamini and Hochberg or q-value) might identify other statistically significant interactions?

Reviewer #3 (Remarks to the Author):

Review of "Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates and 37 components" MSB-12-4223

This study employs chemogenomic profiling to identify genes in Z. mobilis and S. cerevisiae that confer resistance to dilute acid-treated plant biomass. I find this paper is to be interesting and well-written with convincing evidence to support the major claims. I recommend publication pending the changes/clarifications listed below.

Summary of Study:

-The chemical composition of miscanthus and switchgrass hydrolysates were analyzed using GC/MS and HPLC to quantify 4 sugars and 37 potential inhibitors. The 10 and 37 most abundant, identified compounds in miscanthus batch 1 hydrolysate were combined to make synthetic hydrolysate mixtures, SYN-10 and SYN-37, that were used to test the mutant pools.

-A library of 1.4e4 Z. mobilis transposon mutants was created, mapped, and pooled. Mutants with transposon insertions in essential genes were viable and PCR of the genes showed two bands, one with and one with a transposon insertion. The authors conclude that Z. mobilis is polyploid with multiple copies of the main 2Mb chromosome.

-DNA-barcoded libraries of mutants were used to identify Z. mobilis and S. cerevisiae mutants with reduced growth in the presence of plant hydrolysates. Z. mobilis fitness was tested in miscanthus and switchgrass hydrolysates, two types of synthetic hydrolysate mixtures, 37 individual hydrolysate components, and 11 other stress conditions. S. cerevisiae fitness was tested in miscanthus batch 1 and batch 2 hydrolysates and the two synthetic mixtures. These experiments revealed 44 Z. mobilis genes and 28 S. cerevisiae genes involved in resistance to plant hydrolysates.

-The average fitness of each mutant was modeled as a linear combination of its fitness values in each individual inhibitor. Comparison of fitness in the hydrolysates and the modeled values revealed mutants with lower than expected fitness in the hydrolysates. Several of these mutants had defects in genes for detoxification of methylglyoxal. Adding fitness for growth in medium containing methylglyoxal helped explain differences between synthetic and natural hydrolysates, supporting that methylglyoxal is an inhibitor present is plant hydrolysate.

-21 Z. mobilis tolerance genes were overexpressed to identify targets to improve growth in presence of plant hydrolysates. Overexpression of one gene, ZMO1875, resulted in increased hydrolysate resistance and 2.4-fold improved specific ethanol productivity in the presence of miscanthus hydrolysate. This gene of unknown function contains a DUF1476 domain, which has been shown to be an inhibitory subunit of the F0F1 ATP synthase. Based on its genomic proximity to another tolerance gene, BolA, the authors propose that ZMO1875 is instead involved in Fe-S cluster assembly, which enhances growth because Fe-S clusters are damaged by ROS during growth in the presence of hydrolysates.

-While the fitness experiments in this study were done in aerobic conditions, industrial fermentations are generally anaerobic. The fitness of the Z. mobilis mutant pools in rich medium

was thus compared to batch 2 hydrolysate under anaerobic conditions to see if the aerobic fitness results translate to anerobic conditions. Four of the 44 tolerance mutants were sick in anaerobic hydrolysate but not rich medium, supporting the aerobic fitness results are largely specific to aerobic conditions.

Points to address:

It would be helpful to provide more details in the supplementary info about the rationale and methods for the TagModules barcodes used to identify/quantify transposon insertions. This manuscript references Deutschbauer et al, 2011, which, in turn, references Oh et al, 2010, making it cumbersome to learn about the method.

Could the authors clarify which experiments were used to identify the 44 Z. mobilis resistance genes? The text and Fig 3 legend suggests these genes were identified based on the averages of 37 hydrolysate experiments. At first, I thought this meant the 37 individual inhibitors, but Fig S7 shows there were also 37 experiments on natural hydrolysates. Is this just a confusing coincidence?

Fitness values were averaged for all insertions in the same gene. How large are the standard deviations for insertions in the same gene? Do some genes have little apparent effect on fitness even though one insertion had a large effect because other insertions did not?

Are any of the hydrolysate resistance enzymes putatively secreted? I wonder if this method using mutant pools misses secreted enzymes because the lack of an enzyme in one strain could be compensated by another strain.

The only enzyme whose inactivation improved growth in hydrolysate was PEP carboxylase (pg 8), but Fig 3A suggests that other mutants also had higher fitness in hydrolysate. Does this statement on pg 8 mean that mutants with apparent fitness advantages were tested individually and no difference was found? Or was the PEP carboxylase mutant the only one in the pool that passed some statistic threshold?

PEP carboxylase has been shown in Z. mobilis to irreversibly convert PEP into oxaloacetate, which is then used for anabolic reactions (Bringer-Meyer and Sahm 1989). What do the authors propose is the role of PEP carboxylase in hydrolysate tolerance? Also, please show growth data for the fitness advantage of the PEP carboxylase mutant over wild-type, similar to what is shown for overexpression of ZMO1875 (Fig 6).

It is stated on pg 11 that "adding methylglyoxal to the regression significantly improved the fit (adjusted R2=0.903)". It appears in Fig 5A that the correlation without methylglyoxal was R2=0.88. Are the authors suggesting that an improvement from 0.88 to 0.90 was statistically significant? If so, how was this calculated?

In the aerobic experiments, tolerance genes were identified as mutants having a fitness >-1 in the normal medium and <-1 in hydrolysate with a fitness difference of >1. Were the same criteria used to identify genes in anerobic conditions (Fig S17)? If so, please state in the text.

Most tolerance gene mutants identified in the aerobic experiments were sick in rich medium without added inhibitors (pg 13). How was this 'sickness' defined? Does this mean the mutants grow slower in anerobic conditions than anaerobic conditions?

Fig 1. Both plots show glucose concentration on the left vertical axis and ethanol on the right. However, the legend includes OD600. How is OD to be read on these plots?

Fig 1. Z. mobilis ferments efficently to ethanol. What measures were taken to assure that accumulation of ethanol in the cultures was not itself an inhibitor?

Fig 2 is a heat map showing the fitness of mutants for all 1586 genes across 58 conditions. It appears that there is a yellow column of mutants with fitness advantages in many inhibitors. Is this the case? If so, why are these mutants not discussed in the text?

Fig 3A compares fitness in rich medium and hydrolysate and highlights the 44 Z. mobilis tolerance genes. According to the text, tolerance genes were identified as those mutants having a fitness >-1 in rich medium and <-1 in hydrolysate with a fitness difference of >1 (pg 8). However, it appears in Fig 3A that several of the mutants have a fitness <-1 in rich medium. Why were these mutants included?

Fig 4A compares fitness in SYN37 versus hydrolysate. A subset of 'outlier' mutants that are sick in hydrolysate, but not SYN37, are enclosed in an ellipse. What formula was used to define this ellipse?

Fig 4B. It appears that many of the 44 mutants have higher relative fitness in DMSO than ZRMG. Does DMSO rescue defects of the tolerance mutants?

Fig 6. Overexpression of ZMO1875 appears to have resulted in much larger increases in ethanol production rates (2.4x) and concentrations (nearly 2x) with relatively small increases in growth and cell yield. Does expression of this gene simply enabled the cells to grow in the presence of hydrolysate or did it also shift metabolism to favor ethanol production?

Fig S11. It is stated in the text that both synthetic inhibitor cocktails were less potent than the batch hydrolysates. Fig S11 compares growth in the batch hydrolysates and synthetic cocktails, but curves in both the batch and cocktail culltures are cut off as soon as growth begins. Please modify this figure to include the entire growth curves.

Table SI: The M1 and M2 batch samples have glucose:xylose ratios of ~1.5:1 whereas the pure miscanthus samples are 1:4. Why do the batch miscanthus samples have such different sugar ratios than any of the pure samples? The batch samples were used to identify inhibitors for the synthetic cocktails, which were then compared to the pure samples. Wouldn't the differences in sugar composition of the plant matter confound these comparisons?

The authors state that addition of methionine and cysteine to the medium should improve hydrolysate tolerance (pg 15). Was this hypothesis tested? Also, do the authors propose that adding cysteine increases tolerance by filling an increased metabolic demand or by acting as a reductant of ROS?

Do the authors propose that the ROS that affect growth in the hydrolysates came directly from the plant hydrolysates or are the ROS produced by the microbes as a result of metabolic changes during growth in the presence of hydrolysates?

Writing/syntax

When I read the title describing "plant hydrolysates and 37 components", I asked myself 37 components of what? I would shorten the title to "Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates
".

At 6.3e4 characters, I found the manuscript to be quite long. I would recommend shortening both the introduction and the discussion.

The authors state that "We first modeled the average fitness of each gene" (pg 10). The mutant strains have fitness, but I do not think that genes themselves have fitness. I would suggest changing this to "the average fitness of each mutant".

The authors state that "most of the tolerance genes we identified in our aerobic Z. mobilis studies were sick under anaerobic growth conditions" (pg 13). A mutant strain can be sick, but gene itself cannot be sick. I suggest changing this to " most of mutant strains we identified in our aerobic Z. mobilis studies were sick under anaerobic growth conditions".

1st Revision - authors' response 09 April 2013

We thank the reviewers for their useful comments and criticisms. We believe we have addressed all of their concerns and have submitted a revised manuscript for your consideration. All of our fitness data has now been submitted to MSB as Supplementary datasets. In addition, we have shortened the title to: "Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates". A short summary of our revision is below, followed by a point-by-point response to the reviewers questions. All of the revisions to manuscript are in red text to make it easy to identify changes. Reviewers #2 and #3 had only "minor points to be addressed" and "changes/clarifications" to the manuscript. Reviewer #1 was most concerned about the importance of methylglyoxal as a hydrolysate inhibitor, and whether we have overstated this conclusion.

As requested, we have performed additional growth inhibition experiments using wild type *Z. mobilis* (20 biological replicates) in the presence of 0.56 mM methylglyoxal (MG) and demonstrate that MG has a small, but significant effect on growth. Thus, MG is a bona fide inhibitor present in plant hydrolysates. We include two new references in our manuscript that indicate that MG can be formed from glucose, both in a model buffer system, and after the supercritical water treatment of cellulose derived from red cedar (Thornalley *et al*, 1999; Nakata *et al*, 2006). We believe we have not overstated our conclusion that "MG is a previously unidentified hydrolysate inhibitor" -- addition of methylglyoxal fitness data to our model improves the fit, addition of MG to our SYN-10 mixture removes outlier genes, MG is present in our batch 2 hydrolysate, and MG inhibits the growth of wild type *Z. mobilis*.

All three reviewers wanted additional clarification of our pooled fitness methods. We now explain in detail that "gene fitness", and not the fitness of an individual strain, is what is shown in most of our plots, and we give more details about how the fitness experiments were performed in the Results and Methods. We have changed our thresholds for identifying yeast tolerance genes to match those for *Z. mobilis*, and the description of *Z. mobilis* thresholds was corrected (because it was incorrect in the original submission). We also provide more detail about the meaning of the coefficients in our linear model, and explain that the model is statistical and not mechanistic. Finally, we corrected an error in how interaction terms were considered for inclusion in the model (for looking at "synergy"). One of the reviewers suggested an alternative analysis but as described below, it does not affect the results much.

Sincerely,

Adam P. Arkin

Point-by-point response to Reviewers (our responses are shown in red text):

Reviewer #1

1) Although the inclusion of methylglyoxal improved the authors' linear model, when added to their synthetic mixture SYN-10 the correlation with fitness in actual hydrolysate was worse (R2 of 0.60 versus 0.67; Figure 5). Perhaps inclusion in a more representative mixture (e.g. SYN-37, or a "SYN-16" of the 16 compounds used in their linear model), or inclusion at a different concentration (table S1 indicates methylglyoxal is by far the most prevalent compound detected in miscanthus "Batch 2", leading this referee to wonder a) why had it been previously unidentified and/or b) is the concentration actually accurate?) would improve the correlation of synthetic mixture with the actual hydrolysate. Yes, the fitness defect of six outliers is better explained in the SYN-10 + methylglyoxal mixture than SYN-10 alone, but an even greater number of new outliers are observed and the overall correlation is worse. With no acknowledgement or explanation of this discrepancy, this referee is concerned regarding the accuracy of methylglyoxal as an important hydrolysate inhibitor. Moreover, why was SYN-10 + methylglyoxal highlighted in the discussion as "a reasonable proxy for dilute-acid miscanthus hydrolysate" when SYN-10 mixture alone was better

correlated with actual miscanthus hydrolysate and SYN-37 (Fig 3; R2 = 0.81) significantly better than both SYN-10 variants?

RESPONSE:
We do not know why the correlation of SYN-10 + MG is worse than SYN-10; our text now mentions this discrepancy. We have also removed the discussion of six outlier genes in our results and instead focus on only two genes (*ZMO0759* and *ZMO0760*), which encode the GloAB detoxification system. In the revised manuscript we now emphasize that addition of MG to SYN-10 recapitulates the fitness defects of *ZMO0759* and *ZMO0760* in genuine hydrolysate (compare arrows in revised Figure 5D & E). This suggests that *ZMO0759* and *ZMO0760* are important for growth in real hydrolysate because they are directly involved in methylglyoxal detoxification.

The concentration of MG in batch 2 hydrolysate is 41.4 mg/mL (the reviewer may have confused the units and thought it was mg/mL). In fact, methylglyoxal is not the most prevalent inhibitor. For example, acetate and furfural are present at much higher levels (3.9 mg/mL and 5.5 mg/mL, respectively).

As requested by the reviewer, we have now tested the growth inhibition of wild type *Z. mobilis* in the presence of 0.56 mM methylglyoxal (new Figure S18), and show that it does in fact have a small, but significant growth defect (n = 20, unpaired t-test, $P < 10^{-5}$). We have removed the statement that SYN-10 is a reasonable proxy for hydrolysate, and edited our Discussion to indicate that SYN-37 is the better proxy, since its correlation is better ($R^2 = 0.810$ versus 0.669).

We have removed the SYN-10 + MG growth experiments from Figure S18. It was difficult to show the small amount of growth inhibition in these original experiments because they were done at only 45% (v/v) SYN-10, which corresponds to only 0.25 mM methylglyoxal. Instead, our wild type growth data at 0.56 mM (the amount present in batch 2 hydrolysate) demonstrates that methylglyoxal is a bona fide plant hydrolysate inhibitor.

2) Additionally, this referee has no doubt methylglyoxal is toxic to the six outlying mutants, but is not convinced it is toxic to wild-type Z. mobilis (and therefore is a relevant lignocellulosic-associated inhibitory compound). If the right mutant(s) are considered, many compounds can appear toxic. For instance, galE-null Bacillus subtilis mutants rapidly lyse in the presence of exogenous galactose due to the buildup of a toxic intermediate [J Bacteriol. 1998 Apr;180(8):2265-70; MBio. 2012 Aug 14;3(4):e00184-12]. Nonetheless, wild-type B. subtilis readily metabolizes galactose and its growth is unaffected by the sugar. Figure S18 attempts to address the referee's concern regarding methylglyoxal toxicity, but no toxicity experiments are performed with wild-type Z. mobilis (they should be). Moreover, extremely minimal (in some cases imperceptible) toxicity is observed in the mutant pools, with the discrepancy in growth curves between pools larger than the discrepancy between the
curves depicting presence and absence of methylglyoxal. As only one trial was performed, this referee has no confidence the toxicity is reproducible (and therefore real) and encourages multiple trials with standard error bars before claiming methylglyoxal "contributes to overall [miscanthus hydolysate] toxicity" in Z. mobilis.

RESPONSE:
See above. Supplemental Figure 18 now shows that methylglyoxal inhibits the growth of wild type *Z. mobilis* (n = 20, unpaired t-test, $P < 10^{-5}$), at the concentration known to be present in batch 2 miscanthus hydrolysate (0.56 mM).

3) The final linear model ("Model-16") includes coefficients for each of the 16 compounds, reflecting their contribution to overall hydrolysate toxicity (supplementary table 10). In six instances, these coefficients are negative. If this referee understands the model correctly, this would imply the fitness of a mutant in these six compounds is inversely correlated with its fitness in bulk hydrolysate. What, if anything, distinguishes these six compounds from the remaining ten? Additionally, do mutants with increased susceptibility to these compounds show experimentally increased tolerance to bulk hydrolysate?

RESPONSE:

The meaning of the negative coefficients is somewhat confusing. We have added text to the legend for Supplemental Table 7 (was Table 10 in the original submission). Even though the coefficients of the model do not have biological meaning, we still have evidence that our regression modeling was useful:

1). Using our model, without overfit, we can predict growth in hydrolysate from growth in the individual inhibitors without postulating synergies amongst the inhibitors.
2). We also show that lack of fit of key strains to the model can identify new components of the mixture

Revised legend for Table S7:

Supplementary Table 7. Linear regression model results. List of significant components identified, list of coefficients from the R function lm, and results of the ANOVA F test (P value, and F value) for Model-16 (rich + 16 components), Model-17 (rich + 17 components), and Model-24 (rich + 24 components). Note: Because the variables in the regression are all highly correlated with each other, the regression coefficients do not have any biological meaning. For example, consider the problem of fitting gene fitness in hydrolysate using gene fitness data from two similar components X and Y. Since X and Y are correlated, we can get a good fit using either X, Y, or a mixture of the two. Multiple regression will choose coefficients that give the best fit, but these coefficients tell us nothing about the relative importance of X versus Y. Next, consider a case where X and Y are similar stresses in which the same pathways are important for fitness, but to differing extents. If the pathways that are slightly more important on Y than on X tend to be slightly less important in hydrolysate, then the best fit will be something like be X - Y/10. This illustrates why the coefficients can be negative. For example, 5 of the components have negative coefficients in the regression, but for all of these components, fitness is positively correlated with fitness in hydrolysate (all $R > 0.6$). The R value for each of the components in Model-17 is listed in the last column of the table.

4) When calculating the log2 normalized fitness scores, what exactly was used as the "start" library (against which the post-selection libraries were normalized)? This referee assumes (hopes) the up-pool and dn-pool were resuscitated as described (grown for 5 hours shaking at 30C in rich media), and then split with a portion being used to seed competitive growth assays and a portion used as the "start-condition" control. If, instead, the frozen up- and dn-pools were used directly as the baseline control, then each growth experiment involves 5 hours of competitive growth in rich media not accounted for in the control, which could drastically swing relative abundances in the library (assuming a doubling time of 1.4 hrs for Z. mobilis [Appl Biochem Biotechnol. 1991 Spring;28-29:221-36], this would equate to 3.6 generations of unaccounted competition when the actual experiments were conducted with as little as 5 generations). Regardless of the true scenario, this referee requests
simple clarification in the methods. If, however, the appropriate control was not used, this referee has serious reservations about the microarray data used.

RESPONSE:
The referee is correct; the correct control (which we did in fact use, called "start") is after pool recovery, not the frozen pool. After competitive growth in the condition of interest, we collect another sample, called "END". We have added text to our Results and Methods section to clarify our methods. In both our *Z. mobilis* and *S. cerevisiae* fitness experiments, we calculate strain fitness = $\log_2$ (END/START). We also more clearly define "strain fitness" and "gene fitness" in our Results and Methods.

5) Rather than referring to mutants as "sick", this referee would prefer a less anthropomorphized term (e.g. "growth-impaired").

RESPONSE:
All uses of the word "sick" has been removed from our paper and replaced with "important for growth" or "detrimental for growth". In addition, we have added this text to the colorbar of our revised Figure 2, which helps clarify the meaning of negative and positive gene fitness values for the reader.

6) Some justification of why the parameters for what constituted a "tolerance gene" are different between Z. mobilis and S. cerevisiae should be included

RESPONSE:
We thank the reviewer for pointing out this discrepancy. The parameters for defining a tolerance gene have been corrected and now clearly stated in the text ($fitness_{hydrolysate} < -1$ and $fitness_{hydrolysate} < fitness_{rich} -1$). Using this criterion (it is now the same for *Z. mobilis* and *S. cerevisiae*) we have expanded our list of *S. cerevisiae* genes from 28 to 99. We have updated the text and relevant figures to reflect this correction and slightly modified our tolerance gene categories (see new Table II and Supplemental Dataset 3). We believe the new list of tolerance genes is biologically consistent, and captures information that we missed in our original set of 28 genes. For example, our expanded list has 2 new genes in the ERG pathway (ERG4, ERG5), and two targets of the YAP1 regulator (GSH1, CYS3). Nine out of 99 of the tolerance genes we identified have previously been implicated in single inhibitor or hydrolysate tolerance studies, which further supports the validity of our expanded list.

7) On page 10, where it states "...versus average fitness in batch 1 hydrolysate uncovered 11 outliers (labeled on plot in Figure 4C)...", this referee encourages a modification to the sentence that reflects the fact many more than 11 outliers are evident in Figure 4C, with a subset of 11 of these outliers overlapping with the set of 28 previously-discussed S. cerevisiae "tolerance genes".

RESPONSE:
We have modified the text to explain that we are only referring to outliers relative to the set of previously discussed “tolerance genes”. We now have 20 outliers (out of the 99 tolerance genes), which has been updated on the plot in Figure 4C.

8) On page 12, the first paragraph references figures 5C and 5D where it appears the authors intend to reference 5D and 5E, respectively.

RESPONSE:
The text has been corrected to reference 5D and 5E.

9) Was it tested whether or not ZMO1875 improves EtOH production anaerobically? If not, this is a relatively straightforward experiment that could improve the industrial relevance of the ORF to reactor-conditions.

RESPONSE:
Yes, we agree, it would be interesting to test whether *ZMO1875* improves ethanol production anaerobically; however, our proposed role for *ZMO1875* is the repair of Fe-S clusters after oxidative damage; thus, it is unlikely that *ZMO1875* overexpression would have a positive effect on anaerobic ethanol productivity. In addition, *ZMO1875* does not have a fitness defect under anaerobic conditions, consistent with its proposed aerobic function. We have chosen to not do any additional fermentations for this study. Instead, we think that future studies could focus on the eleven anaerobic hydrolysate tolerance genes that we identified (Supplemental Figure 17) and make this point in our Discussion.

10) On page 23, in the methods, Supplementary Table 3 is referenced where it appears the authors intend to reference Supplementary Table 2

RESPONSE:
This has been corrected.

11) On figures 1 and 6, OD600 is plotted but no optical density values are given on either of the y-axes. Thus, although the reader can compare data-series against one another, there is no absolute reference for cell-growth.

RESPONSE:
We have corrected this mistake. The Y-axis should have included both labels: “OD600” and “Ethanol (g/L)”. This has been corrected in Figure 1, Figure 6, and Supplemental Figure 16.

12) On figure 3, the green colors used to differentiate between 'cytochrome c' and 'regulator' are too similar. Thus, it is impossible to distinguish between the two categories in figure 3B.

RESPONSE:
We have changed the color of 'regulator' to brown to make is easier to differentiate. We have revised all of the relevant plots with the new color scheme (Figures 3A and 5, Supplemental Figures 12A, 12B, and 17).

Detailed responses for Reviewer #2:

1. Throughout the paper (in text, Methods, Modeling description, and figures), the term "genes" is used when in fact the fitness data represent strains mutated at those genes. For example, rows of hierarchically clustered data are listed as genes when in fact each row represents a strain. The terminology should be clarified throughout the manuscript.

RESPONSE:
We now clearly define "strain fitness" and "gene fitness" in our Results and Methods section. Yes, our pool assay does measure the fitness of a mutated strain, not of a gene. However, in this paper we have multiple transposon insertions per gene and each gene has multiple "strain fitness" values (on average we made 3.5 strain fitness measurements per gene). "Gene fitness" is the average of all the "strain fitness" values for a particular gene. In most of our *Z. mobilis* fitness plots we are showing "gene fitness", not the fitness of individual mutant strains. In addition, "gene fitness" values are further averaged for replicate conditions (e.g. 24 rich media experiments), or for identical concentrations of an inhibitor (e.g. 10 mM 4-hydroxybenzoic acid, n = 2), or across all of our plant hydrolysate (37 different experiments). We have updated the Figures and Figure legends to indicate this. Thus, most plots in this paper are average "gene fitness" values (experiments that were averaged are indicated in Supplemental Table 5). For *S. cerevisiae*, only one mutant per gene exists in the deletion pool, so "strain fitness" and "gene fitness" are the same thing.

2. The term 'fitness' is used seemingly in multiple ways in the paper. In the early text, relative fitness is described as strain abundance before and after selection, but in the Methods it seems to be reflecting normalized strain abundance (signal intensity) on a single array; in other parts, fitness is used to compare growth in hydrolysate to growth in inhibitor-free medium. It would be useful to differentiate the terminology in these cases, as I became confused which 'fitness' measure was being used in different parts of the manuscript.

RESPONSE:
See above. We now use the terms "gene fitness" and "strain fitness" in our revised manuscript. Fitness only refers to fitness values derived from pool experiments. We no longer use the term "fitness" to refer to follow-up growth studies with single mutants. However, we believe that fitness in this context is in fact, the same fitness that we measure in our pool format. However, to avoid confusion, we have revised this throughout the text.

3. At least for the S. cerevisiae growth experiments, there is no mention of the generation times used - but this is really critical in calculating fitness. Normally, control cells are grown for exactly the same (population) generations as stressed cells, to optimally distinguish slow-growing strains from strains removed by selection. Minimally, the Methods should better describe how many generations each library went so that users of the data will have all the information.

RESPONSE:
Although the reviewer suggested that it is important for control cells to be grown for the same number of generations as unstressed cells, this is not necessary for our fitness assays, in which we compare the abundance of each strain after growth in either media to the abundance before inoculation. For *Z. mobilis*, growth in inhibitors often led to a lower final OD, implying fewer generations. Indeed, if you compare the general trend for the comparison of gene fitness in hydrolysate and gene fitness in rich media (and ignore the tolerance genes and other outliers) in Figure 3A to the x=y line, you can see that the slope is slightly less than one. This corresponds to the (slightly) fewer number of doublings, but it has little impact on our analyses.

For yeast, growth in hydrolysate (either genuine or synthetic) led to an increased lag, without much effect on the final OD or the total number of generations. Indeed, if you examine the scatterplot of gene fitness in hydrolysate versus gene fitness in YPD (Figure S13, again ignoring tolerance genes and other outliers), it fits the x=y line.

In the revised Methods, we mention the number of generations for the yeast fitness assays (around 7).

4. The heterozygous essential knockouts was interesting. The authors implied that the species is polyploidy, but I wondered if instead they simply selected for polyploidy mutants when essential genes were being knocked out?

RESPONSE:
The fact that the rate of insertion in essential versus non-essential genes was about the same supports our hypothesis that *Z. mobilis* is normally polyploid. If polyploidy were somehow selected for only when essential genes were being knocked out, we would expect the rate of insertions in essential versus non-essential genes to be different. We have added text to the Results section to address this question.

5. Several datasets are made available on the authors' website - these should be made available by the journal and/or deposited into a public array repository, since over time it's harder to maintain lab websites.

RESPONSE:
We have submitted our *Z. mobilis* and *S. cerevisiae* fitness data to MSB as "Supplemental Datasets".

6. I was not clear how the doses of stresses were chosen for the synthetic hydrolysate - it is not clear from the text if the doses were chosen to match hydrolysate? From the table, some of the doses match and others seemingly don't. Does the dose of stresses used in this analysis affect whether the stresses might interact? In other words, might their be dose-specific interactions?

RESPONSE:
The SYN-10 and SYN-37 mixtures were made based on the composition of batch 1 miscanthus hydrolysate. The concentrations of four compounds in SYN-10 were in error: (94.9 µg/mL versus 30.3 µg/mL for vanillin, 53.2 µg/mL versus 19.5 µg/mL for syringaldehyde, 75.4 µg/mL versus 22.2 µg/mL for vanillic acid, 0.22 mg/mL versus 0.05 mg/mL for furoic acid. For SYN-37, two compounds were in error: 4.6 µg/mL versus 0.8 µg/mL for benzoic acid, and 1.04 mg/mL versus 0.56 mg/mL for cellobiose. The fitness profiles for SYN-37 and SYN-10 in both *Z. mobilis* and *S. cerevisiae* are highly correlated ($R^2 = 0.807$ for *Z. mobilis*, $R^2 = 0.860$ for *S. cerevisiae*); thus, it appears that these errors in concentration don't have a significant biological impact. We can't rule out subtle effects on the fitness profiles of SYN-10 and SYN-37, however, this does not affect the conclusions of our paper. We have included text in the Methods and Supplementary Table 1 to explain these discrepancies in the expected concentrations.

7. Along the lines of interactions terms, it is often harder to find enough statistical power to identify interactions (although that probably does mean the effect size is small, as the authors suggest). I wasn't sure how the p-value for interaction was chosen, and it looked like the test correction was the stringent Bonferorni correction - I wondered if a less stringent FDR correction (Benjamini and Hochberg or q-value) might identify other statistically significant interactions?

RESPONSE:
Based on the reviewer's question, we redid the analysis based on the false discovery rate, as described in more detail below. In this process we found an error in our original submission: we had considered only interaction terms that involved furfural. After considering all possible interactions, we still find that the linear model fits reasonably well and that adding interaction terms does not materially improve the fit.

If we add every possible interaction term of the form x*y to Model-16, then there are 20 terms (of the form x*y) that significantly improve the fit ($P < 0.0001$, ANOVA), not 3 as stated in the original

submission. If we use a Bonferroni correction with $P < 0.05$ then that rises to 27 significant terms, and if we require a false discovery rate of under 0.05 then that rises to 53 terms (out of 1536).

If we take the terms with $P < 0.0001$, sort them by their p-value, and use ANOVA to remove terms that are insignificant ($P > 0.0001$) in combination, then three significant interactions remain: formic acid * levulinic acid ($P < 10^{-13}$), furfural * 4-hydroxyphenylacetic acid ($P < 10^{-15}$), and furfural * vanillin ($P < 10^{-5}$). Adding these terms makes relatively little difference overall (adjusted R^2 rises from 0.880 to 0.893). However, the predictions for four genes do change by 0.5 or more; these genes are *ZMO0975*, *ZMO1430*, *ZMO1431*, and *ZMO1432*. Also *ZMO1429* is near the threshold. The operon *ZMO1429-ZMO1432* encodes a putative efflux system. *ZMO0975* (*hpnL*) is probably related to the synthesis of hopanoids, a sterol-like membrane component. Oddly, all five of these genes were identified as having non-stable insertions. So we are not sure if these genes are really evidence of biologically significant interactions between the components.

As the reviewer asked about the FDR, we did a similar analysis starting with the 53 significant terms, and used ANOVA to remove terms that were not significant in combination with a more lax threshold ($P > 0.01$). This left eight terms. This model gave very similar results as the model with 3 interaction terms ($R^2 = 0.994$) and roughly the same outliers. For example, see: http://morgannprice.org/tmp/zm4/model16_with_interaction_terms.pdf

In our revised submission, we have corrected the summary of the interaction terms in the Results (using the $P < 0.0001$ threshold) and we have added more detail about this analysis to the Methods section.

Detailed responses for Reviewer #3:

It would be helpful to provide more details in the supplementary info about the rationale and methods for the TagModules barcodes used to identify/quantify transposon insertions. This manuscript references Deutschbauer et al, 2011, which, in turn, references Oh et al, 2010, making it cumbersome to learn about the method.

RESPONSE:
We have added additional text in the Results and Methods to better explain our pooled fitness assay (e.g. definition of "strain fitness" versus "gene fitness", how we collected "START" and "END" samples and calculated fitness values). This should help the reader follow our methods in sufficient detail without having to refer to our earlier papers (Deutschbauer *et al*, 2011 and Oh *et al*, 2010).

Could the authors clarify which experiments were used to identify the 44 Z. mobilis resistance genes? The text and Fig 3 legend suggests these genes were identified based on the averages of 37 hydrolysate experiments. At first, I thought this meant the 37 individual inhibitors, but Fig S7 shows there were also 37 experiments on natural hydrolysates. Is this just a confusing coincidence?

RESPONSE:
Tolerance genes were identified from analysis of the data shown in Figure 3A. This is average gene fitness in rich media (n = 24, replicate experiments) and average gene fitness in hydrolysate (average of 37 hydrolysate experiments, see Supplemental Figure 7). The number 37, for the number of inhibitors examined, and the number of hydrolysate experiments is just a coincidence.

Fitness values were averaged for all insertions in the same gene. How large are the standard deviations for insertions in the same gene? Do some genes have little apparent effect on fitness even though one insertion had a large effect because other insertions did not?

RESPONSE:
The strain fitness values for multiple insertions in the same gene are quite consistent. For example, after averaging all the hydrolysate experiments the strain correlation (different insertions in the same gene, all within central 5-80%) was r = 0.85. The mean absolute difference was 0.27. We have added this statistic to our Methods section.

Are any of the hydrolysate resistance enzymes putatively secreted? I wonder if this method using mutant pools misses secreted enzymes because the lack of an enzyme in one strain could be compensated by another strain.

RESPONSE:
Yes, we agree that a pooled fitness assay can miss secreted enzymes. To get an idea of how many genes might have been missed, we used the PSORTb algorithm to predict secreted proteins, and found only 29 genes in *Z. mobilis* that are predicted to be "extracellular". None of our 44 tolerance genes are predicted to be extracellular. Overall, we believe the pooled fitness assay is still a valid approach to identify tolerance genes, given the above caveat.

The only enzyme whose inactivation improved growth in hydrolysate was PEP carboxylase (pg 8), but Fig 3A suggests that other mutants also had higher fitness in hydrolysate. Does this statement on pg 8 mean that mutants with apparent fitness advantages were tested individually and no difference was found? Or was the PEP carboxylase mutant the only one in the pool that passed some statistic threshold?

RESPONSE:
We identified only one gene, PEP carboxylase that met our cutoff for genes that are detrimental for growth in hydrolysate. Our selection criterion was: $fitness_{hydrolysate} > 0.5$ and $fitness_{hydrolysate} > fitness_{rich} + 0.75$, and it is now labeled on Figure 3A, as dashed black lines.
The value reported for PEP carboxylase is average gene fitness, based on 5 insertions in that gene. We have high confidence in the positive gene fitness value reported for *ZMO1496*, but we did not perform any single-mutant follow-up studies.

PEP carboxylase has been shown in Z. mobilis to irreversibly convert PEP into oxaloacetate, which is then used for anabolic reactions (Bringer-Meyer and Sahm 1989). What do the authors propose is the role of PEP carboxylase in hydrolysate tolerance? Also, please show growth data for the fitness advantage of the PEP carboxylase mutant over wild-type, similar to what is shown for overexpression of ZMO1875 (Fig 6).

RESPONSE:
Although it's interesting to understand why the PEP carboxylase gene is detrimental for growth in hydrolysate, we chose to focus on the 44 genes that were important for growth in plant hydrolysate. We did not do any follow-up studies on this mutant, although of potential interest to other researchers working on *Z. mobilis* tolerance; nor did we speculate on its proposed role in hydrolysate tolerance.

It is stated on pg 11 that "adding methylglyoxal to the regression significantly improved the fit (adjusted R2=0.903)". It appears in Fig 5A that the correlation without methylglyoxal was R2=0.88. Are the authors suggesting that an improvement from 0.88 to 0.90 was statistically significant? If so, how was this calculated?

RESPONSE:
As described in the revised Results, we used an ANOVA test to determine that this increase in R^2 is significant ($P < 10^{-15}$).

In the aerobic experiments, tolerance genes were identified as mutants having a fitness >-1 in the normal medium and <-1 in hydrolysate with a fitness difference of >1. Were the same criteria used to identify genes in anerobic conditions (Fig S17)? If so, please state in the text.

RESPONSE:
The rule was stated incorrectly in our original submission. We have corrected this in our revised submission, and now use the same criterion for aerobic and anaerobic tolerance genes in *Z. mobilis* and aerobic tolerance genes in *S. cerevisiae* ($fitness_{hydrolysate} < -1$ and $fitness_{hydrolysate} < fitness_{rich} - 1$). As a result, the number of yeast tolerance genes is now 99 instead of 28, and the number of anaerobic hydrolysate tolerance genes in *Z. mobilis* is now 11.

Most tolerance gene mutants identified in the aerobic experiments were sick in rich medium without

added inhibitors (pg 13). How was this 'sickness' defined? Does this mean the mutants grow slower in anerobic conditions than anaerobic conditions?

RESPONSE:
This paragraph has been rewritten in our discussion. We no longer refer to "sick" genes, and use the cutoff described above. We identified 11 anaerobic hydrolysate tolerance genes: 4 overlap with our aerobic studies, and 7 new tolerance genes were identified (see revised Supplemental Figure 17).

Fig 1. Both plots show glucose concentration on the left vertical axis and ethanol on the right. However, the legend includes OD600. How is OD to be read on these plots?

RESPONSE:
The secondary Y-axis had been corrected to include OD600. Both OD600 and ethanol are read from the same secondary Y-axis.

Fig 1. Z. mobilis ferments efficently to ethanol. What measures were taken to assure that accumulation of ethanol in the cultures was not itself an inhibitor?

RESPONSE:
In our fermentation experiments we produce a maximum of 10 g/L (~1.26% v/v) ethanol. *Z. mobilis* is known to be ethanol tolerant up to 16% v/v (Seo *et al*, 2005). Therefore, it is unlikely that ethanol accumulation has much of an inhibitory effect in our experiments. In contrast, a recent paper (He *et al*, *Biotechnol Biofuels* 2012 5:75) has reported ethanol stress at 5% (v/v); however, this concentration is still well above the maximum in our studies.

Fig 2 is a heat map showing the fitness of mutants for all 1586 genes across 58 conditions. It appears that there is a yellow column of mutants with fitness advantages in many inhibitors. Is this the case? If so, why are these mutants not discussed in the text?

RESPONSE:
We do discuss the one gene (*ZMO1496*) that seems to be detrimental in hydrolysate but not in rich media (see above). As the reviewer noted, some genes are frequently detrimental to fitness (vertical yellow stripe in Figure 2, a little to the right of center). However, according to our fitness data, these genes are not specifically detrimental to fitness in hydrolysate and so we do not expect that they are important for understanding hydrolysate stress. Also, although we believe that the detrimental fitness effects of many of these genes are genuine, 18 of these genes are in a putative prophage region *ZMO1920-ZMO1952*. In our comparative genome hybridization data, this region appeared to have variable copy number (Supplementary Figure 4H). So the positive fitness of these genes could be an artifact -- if the prophage increases its copy number when the cell is stressed, then the barcodes in these strains will be amplified and their fitness values will be positive, even though those cells have not increased in abundance. We have added an explanation of this putative prophage region in our Methods section.

Fig 3A compares fitness in rich medium and hydrolysate and highlights the 44 Z. mobilis tolerance genes. According to the text, tolerance genes were identified as those mutants having a fitness >-1 in rich medium and <-1 in hydrolysate with a fitness difference of >1 (pg 8). However, it appears in Fig 3A that several of the mutants have a fitness <-1 in rich medium. Why were these mutants included?

RESPONSE:
In the original submission, we made an error in how we stated the criterion for selecting tolerance genes. The criterion has been corrected in the revised text ($fitness_{hydrolysate} < -1$ and $fitness_{hydrolysate} < fitness_{rich} - 1$). All 44 tolerance genes lie within this cutoff (see revised Figure 3A).

Fig 4A compares fitness in SYN37 versus hydrolysate. A subset of 'outlier' mutants that are sick in hydrolysate, but not SYN37, are enclosed in an ellipse. What formula was used to define this ellipse?

RESPONSE:

Instead of using a hand-drawn ellipse, we now define the outlier genes using ($fitness_{hydrolysate} < -1$ and $fitness_{SYN-37} > -1/3$), and this is now indicated on the revised graph in Figure 4A with dashed black lines. The same cutoff is used to define outliers in Figure 4C.

Fig 4B. It appears that many of the 44 mutants have higher relative fitness in DMSO than ZRMG. Does DMSO rescue defects of the tolerance mutants?

RESPONSE:
It is true that some of the tolerance genes (*ZMO0200*/*ZMO0201*/*ZMO0468* = *trpDGE*, *ZMO1429-ZMO1432*, and *ZMO1221*) have higher average fitness values in DMSO than in ZMRG. But these genes are not consistently higher fitness in rich media -- only in a subset of the replicates (see Supplemental Figure 7). So we doubt that this difference is real.

Fig 6. Overexpression of ZMO1875 appears to have resulted in much larger increases in ethanol production rates (2.4x) and concentrations (nearly 2x) with relatively small increases in growth and cell yield. Does expression of this gene simply enabled the cells to grow in the presence of hydrolysate or did it also shift metabolism to favor ethanol production?

RESPONSE:
We have added this text to our Results: "Glucose is fully consumed in both the wild type and overexpression strains, yet the wild type strain makes both less biomass and less ethanol. This suggests that the improvements in ethanol productivity in the *ZMO1875* overexpression strain are due to a metabolic shift resulting in the production of less byproducts (Amin *et al*, 1983; Yang *et al*, 2009b)."

Fig S11. It is stated in the text that both synthetic inhibitor cocktails were less potent than the batch hydrolysates. Fig S11 compares growth in the batch hydrolysates and synthetic cocktails, but curves in both the batch and cocktail culltures are cut off as soon as growth begins. Please modify this figure to include the entire growth curves.

RESPONSE:
We have replaced Figure S11 with a more detailed inhibition analysis. By measuring the growth rate over a range of concentrations, we determined that batch hydrolysates are more inhibitory than synthetic cocktails (for both *Z. mobilis* and *S. cerevisiae*). We have revised the Results section to reflect this new data. Using an ANOVA test, we demonstrate that these differences are significant (for *Z. mobilis*, $P < 10^{-5}$, for *S. cerevisiae* $P < 10^{-15}$).

Table SI: The M1 and M2 batch samples have glucose:xylose ratios of ~1.5:1 whereas the pure miscanthus samples are 1:4. Why do the batch miscanthus samples have such different sugar ratios than any of the pure samples? The batch samples were used to identify inhibitors for the synthetic cocktails, which were then compared to the pure samples. Wouldn't the differences in sugar composition of the plant matter confound these comparisons?

RESPONSE:
As explained in the text, the batch 1 and batch 2 samples were prepared under more harsh pretreatment conditions, so more of the xylose was converted into inhibitory products; therefore the ratio of glucose:xylose was closer to 1:1, as pointed out by the Reviewer. All of the sugars present in our hydrolysate are at rather low concentrations, and we don't expect them to have any adverse osmotic effects. In addition, the fitness profiles of SYN-37 and SYN-10 are highly correlated; yet, SYN-10 has no added sugars. This implies that the addition of sugars to SYN-37 has little effect on the fitness profile and would not confound our results or alter our conclusions.

The authors state that addition of methionine and cysteine to the medium should improve hydrolysate tolerance (pg 15). Was this hypothesis tested? Also, do the authors propose that adding cysteine increases tolerance by filling an increased metabolic demand or by acting as a reductant of ROS?

RESPONSE:
As stated in our original Discussion, we propose that increased demand for cysteine is due to an increased demand for glutathione. We suggest that addition of cysteine to the medium might

improve tolerance; however we did not test this hypothesis, but reference papers in support of this idea.

Do the authors propose that the ROS that affect growth in the hydrolysates came directly from the plant hydrolysates or are the ROS produced by the microbes as a result of metabolic changes during growth in the presence of hydrolysates?

RESPONSE:
In our revised Discussion, we clarify this point and propose that intracellular ROS results from metabolic changes during growth in the presence of hydrolysate. ROS are likely not stable enough to be present in our plant hydrolysates.

When I read the title describing "plant hydrolysates and 37 components", I asked myself 37 components of what? I would shorten the title to "Dissecting a complex chemical stress: chemogenomic profiling of plant hydrolysates".

RESPONSE:
We changed the title to the shorter version, as suggested.

At 6.3e4 characters, I found the manuscript to be quite long. I would recommend shortening both the introduction and the discussion.

RESPONSE:
We recognize that this paper is quite long, but feel that the length is appropriate for the amount of data we are presenting.

The authors state that "We first modeled the average fitness of each gene" (pg 10). The mutant strains have fitness, but I do not think that genes themselves have fitness. I would suggest changing this to "the average fitness of each mutant".

RESPONSE:
Yes, we agree, that genes themselves don't have fitness. As discussed earlier, we have now clearly defined "strain fitness" and "gene fitness" in our revised manuscript. Based on our definition, we did, in fact, model "average gene fitness", and not the "average fitness of each mutant". The text has been edited, and all of our text references to fitness have been changed to "gene fitness", as necessary. In addition, the axes of our scatterplots have been changed to "gene fitness" in our revised Figures. By defining these fitness terms early in the revised Results section, we now believe our terminology is correct and consistent throughout the manuscript.

The authors state that "most of the tolerance genes we identified in our aerobic Z. mobilis studies were sick under anaerobic growth conditions" (pg 13). A mutant strain can be sick, but gene itself cannot be sick. I suggest changing this to " most of mutant strains we identified in our aerobic Z. mobilis studies were sick under anaerobic growth conditions".

RESPONSE:
In the revised submission, we remove all references to genes being "sick". Instead, we say that a "gene is important for growth" or a "gene is detrimental for growth". In addition, we have now clearly defined "strain fitness" versus "gene fitness" in our Results and Methods. The fitness of a "gene" is calculated based on the fitness of mutant strains (it is the average of all the strain fitness values for that particular gene). We don't refer to "mutant strain fitness" in this paper, because all of our fitness values (for *Z. mobilis*) are average "strain fitness" values based on multiple insertions per gene. On average, we made 3.5 fitness measurements per gene. For *S. cerevisiae*, "strain fitness" and "gene fitness" have the same meaning, because there is only one deletion mutant per gene present in the yeast pool. We have revised the text in the Discussion relating to anaerobic hydrolysate tolerance genes using the new cutoff we described earlier. In the revised submission, we use the same cutoff in Figure 3A and S17, and identified 11 genes that were important for growth in anaerobic hydrolysate (4 were from our previous set of 44 aerobic tolerance genes).

2nd Editorial Decision 03 May 2013

Thank you again for submitting your revised work to Molecular Systems Biology. We have now heard back from the two referees who accepted to evaluate the revision. As you will see, the referees are now fully supportive and I am pleased to inform you that we will be able to accept your manuscipt for publication pending the following minor amendments:

- reviewer #2 feels that the expression "gene fitness" should be amended. Please edit the text as you see fit.

Referee reports:

Reviewer #1 (Remarks to the Author):

The authors have satisfactorily addressed major reviewer concerns. The authors now demonstrate methylglyoxal is toxic to wild-type Z. mobilis, a necessary step to claiming the compound "contributes to overall [miscanthus hydrolysate] toxicity", addressing the major criticism of this reviewer. It is worth noting, that while acknowledging reviewer suggestions in their rebuttal, the authors have opted to omit multiple experiments that could have improved the rigor of this work and facilitated direct comparisons between in-silico models and synthetic inhibitor-mixtures (#1: testing the addition of methylglyoxal to the SYN-37 mixture; #2: recapitulate the biological effects of Models-16, -17, and/or -24 with synthethic mixtures of 16, 17, and 24 compounds). However, in my opinion, these experiments are not critical requirements for publication, but would certainly have improved the support for the authors' conclusions. The project as described addresses a major concern in the field, uses innovative techniques, and is generally well-supported. Therefore, I am in support of publication of this revised manuscript describing this largely high-quality work.

Reviewer #2 (Remarks to the Author):

The authors have addressed all my original concerns (and made a number of key corrections in the process, which is good). I think the manuscript will be broadly interesting to people in the biofuels area and is a nice piece of work overall.

One remaining nit-picky point, which may seem like semantics but is important I think: a gene does not have a fitness, but rather a fitness contribution. I would strongly suggest that the authors change the language from "gene fitness" to "gene fitness score" or "fitness contribution". To a geneticist, "gene fitness" is grating every time I read it.

2nd Revision - authors' response 09 May 2013

We are excited to publish our work in *Molecular Systems Biology* and thank the reviewers for their comments on the revised manuscript. We have made the following changes to the main text:

Page 1: Added middle initial to Jason K. Baumohl
Page 1: Dan R. Tarjan was changed to Daniel R. Tarjan
Page 15: We corrected the sentence from four to five OCA genes: "...we identified five OCA genes (OCA1, OCA2, OCA4, OCA5, OCA6)..."

We have decided to keep the term "gene fitness" in our text without amendment. We clearly define the term in our revised paper, and have used the same terminology in two previous Arkin lab publications (one of which was recently accepted in MSB).

