Identification of genes conferring tolerance to lignocellulose-derived inhibitors by functional selections in soil metagenomes

Kevin J. Forsberg
Washington University School of Medicine in St. Louis

Sanket Patel
Washington University School of Medicine in St. Louis

Evan Witt
Washington University School of Medicine in St. Louis

Bin Wang
Washington University School of Medicine in St. Louis

Tyler D. Ellison
Washington University School of Medicine in St. Louis

See next page for additional authors

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
Forsberg, Kevin J.; Patel, Sanket; Witt, Evan; Wang, Bin; Ellison, Tyler D.; and Dantas, Gautam, "Identification of genes conferring tolerance to lignocellulose-derived inhibitors by functional selections in soil metagenomes." Applied and Environmental Microbiology, 82, 2, 528-537. (2016).
http://digitalcommons.wustl.edu/open_access_pubs/4489

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Identification of Genes Conferring Tolerance to Lignocellulose-Derived Inhibitors by Functional Selections in Soil Metagenomes

Kevin J. Forsberg, Sanket Patel, Evan Witt, Bin Wang, Tyler D. Ellison, Gautam Dantas

Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA; Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri, USA; Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri, USA; Department of Molecular Microbiology, Washington University School of Medicine, St. Louis, Missouri, USA.

The production of fuels or chemicals from lignocellulose currently requires thermochemical pretreatment to release fermentable sugars. These harsh conditions also generate numerous small-molecule inhibitors of microbial growth and fermentation, limiting production. We applied small-insert functional metagenomic selections to discover genes that confer microbial tolerance to these inhibitors, identifying both individual genes and general biological processes associated with tolerance to multiple inhibitory compounds. Having screened over 248 Gb of DNA cloned from 16 diverse soil metagenomes, we describe gain-of-function tolerance against acid, alcohol, and aldehyde inhibitors derived from hemicellulose and lignin, demonstrating that uncultured soil microbial communities hold tremendous genetic potential to address the toxicity of pretreated lignocellulose. We recovered genes previously known to confer tolerance to lignocellulosic inhibitors as well as novel genes that confer tolerance via unknown functions. For instance, we implicated galactose metabolism in overcoming the toxicity of lignin monomers and identified a decarboxylase that confers tolerance to ferulic acid; this enzyme has been shown to catalyze the production of 4-vinyl guaiacol, a valuable precursor to vanillin production. These metagenomic tolerance genes can enable the flexible design of hardy microbial catalysts, customized to withstand inhibitors abundant in specific bioprocessing applications.

Many lignocellulosic feedstocks (e.g., switchgrass) are preferred to maize, sugarcane, and other traditional food crops for the production of fuels and chemicals because they are able to grow on marginal land, often require little attention or energy input, and do not compete directly with the food supply (1–6). However, lignocellulose requires harsh thermochemical pretreatment methods to liberate fermentable monosaccharides (7–9), producing an additional compendium of compounds inhibitory to microbial growth that ultimately reduce production efficiencies (10–12). These small-molecule inhibitors derived from lignocellulose pretreatment (here called lignocellulosic inhibitors) are typically aldehydes, organic acids, furans, or phenolics and can originate from the cellulosic, hemicellulosic, and lignified fractions of the feedstock (11–15).

Engineering harder microbial production hosts with elevated tolerance to these inhibitors offers potential to ameliorate the toxic effects of these compounds without incurring the high process costs associated with detoxifying the lignocellulosic hydrolysate (16, 17). Unfortunately, the modes of toxicity of many of these toxins are poorly described, and genes conferring tolerance to many of these compounds have not been identified (18, 19). An expanded catalog of tolerance-conferring genotypes may shed light on mechanisms of toxicity and enable synthetic biology approaches for the design of diverse production hosts with broad-spectrum tolerance.

Soil microorganisms, including white-rot fungi (20) and many bacteria (21), are likely a valuable reservoir of genetic elements that confer tolerance to lignocellulosic inhibitors and next-generation biofuels (22–25). Indeed, our previous work using large-insert (40- to 50-kb) functional metagenomic libraries identified three such genes conferring improved tolerance to two biomass inhibitors (19). Characterization of these large-insert libraries required numerous rounds of sequencing to assemble tolerance-conferring DNA fragments, followed by transposon mutagenesis and targeted subcloning to identify the three causal tolerance genes on the two fragments interrogated. These time- and cost-intensive methods are not feasible strategies for discovering genes conferring tolerance to many inhibitors, and accordingly, we applied a higher-throughput approach.

We functionally interrogated 16 agricultural and grassland soils for their repertoire of lignocellulosic inhibitor tolerance-conferring genotypes (26–28). Since our earlier work with large-insert libraries demonstrated that tolerance phenotypes can be conferred by individual open reading frames (ORFs) (19), we elected to construct small-insert (1- to 5-kb) metagenomic libraries, which allow screening with greater clone diversity and sampling depth compared to screening with large-insert libraries (29, 30). To facilitate the interrogation of large, diverse clone libraries with direct application to bioprocessing goals, we performed functional selections in Escherichia coli, as it exhibits high transformation efficiencies and is a leading producer of advanced biofuels and commodity chemicals (31–34).
MATERIALS AND METHODS

Construction of soil metagenomic libraries. Construction of soil metagenomic libraries was performed exactly as described in previous work (28). Briefly, metagenomic DNA was extracted from soils (see Table S1 in the supplemental material) using the PowerMax soil DNA isolation kit (MoBio Laboratories), and 10 g was used for the construction of a given library. Small-insert metagenomic libraries were created by shearing this DNA into 1- to 5-kb fragments before ligation into the pZE21 expression vector (35) and electrotransformation into E. coli MegaX cells (Invitrogen). These cells do not contain the TetR repressor, and therefore, expression from the PLtetO-1 promoter of pZE21 is constitutive. Titers of libraries were determined by plating out 0.1 l and 0.01 l of recovered cells onto Luria-Bertani (LB) agar (5 g yeast extract, 5 g NaCl, 10 g tryptone, 12 g agar in 1 liter of water) plates containing 50 mg/ml kanamycin (Kan). For each library, insert size distribution was estimated by gel electrophoresis of PCR products obtained by amplifying the insert from 12 randomly picked clones using primers flanking the HinCII site of the multiple cloning site of the pZE21 MCS1 vector (which contains a selectable marker for kanamycin resistance). The average insert size across all libraries was determined to be 2,000 bp, and library size estimates were calculated by multiplying the average PCR-based insert size by the number of CFU from titer determination after transformation recovery. The rest of the recovered cells were inoculated into 50 ml of LB medium containing 50 mg/ml kanamycin and grown overnight. The overnight culture was frozen with 15% glycerol and stored at −80°C for subsequent selection.

Determining MICs of lignocellulosic inhibitors. MICs for solid and liquid toxicity assays were determined separately, each in LB medium using E. coli MegaX cells (DH10B-derived) containing an empty pZE21 MCS vector. Initial MIC predictions were informed by previous work (13–15, 19) and tested using LB medium formulated with 50 mg/ml kanamycin and a range of concentrations for each inhibitor spanning the predicted MIC (typically, four inhibitor concentrations were chosen for agar MICs and seven for liquid MICs). For agar MIC determination, overnight cultures were plated using glass sterile beads and growth was monitored at 37°C for 2 days. For liquid MICs, a 2.5% inoculum of mid-log culture was added to 200 ml of broth medium and growth was monitored in 96-well plates using the Synergy H1 microplate reader (BioTek Instruments) for up to 5 days at 37°C. If an empirically determined MIC fell between two tested concentrations, a subsequent round of MIC determination was performed to narrow the step size between putative MICs. If the MIC fell outside the initial tested range, the initial assay was repeated to set boundaries for a MIC estimate and a third assay was performed to narrow the MIC further. Because the range of inhibitor concentrations chosen for the first round of MIC testing varied across compounds, so, too, did the resolution for the final rounds of MIC determination.

Functional selections for tolerance to lignocellulosic inhibitors. For each soil metagenomic library, selections for tolerance to each of the 20 inhibitors tested were performed on LB agar containing 50 mg/ml kanamycin and supplemented with concentrations of lignocellulosic inhibitor outlined in Table S2 in the supplemental material. For each metagenomic library, the number of plates collected on each type of selective medium represented 10 times the number of unique CFU in the library, as determined from titers during library creation. After plating (using sterile glass beads), selections for compound tolerance were incubated at 37°C for up to 5 days to allow the growth of tolerant E. coli transformants. If microbial growth was observed, cells were collected in two ways to interrogate tolerance-conferring genotypes: (i) by picking and validating individual transformants, as described below, or (ii) by collecting all microbial growth. For the latter method, after individual transformants were picked, all remaining transformants from a single plate (soil by inhibitor selection) were collected by adding 750 ml of 15% LB-glycerol to the plate and scraping with an L-shaped cell scraper (Fisher Scientific catalog no. 03-392-151) to gently remove colonies from the agar. The liquid “plate scrape culture” was then collected, and this process was performed a second time to ensure that all colonies were removed from the plate. The bacterial cells were then stored at −80°C before PCR amplification of metagenomic fragments and Illumina library creation.

Picking individual transformants and verifying tolerance phenotypes. After microbial growth was observed, but before total growth was collected, between four and 10 individual colonies were picked from a positive selection, amplified in LB broth with 50 mg/ml kanamycin (LB-Kan broth), and stored in 96-well plates with 15% glycerol at −80°C. These picked clones were subsequently subjected to liquid growth assays in Mueller-Hinton (MH) broth (2 g beef infusion solids, 1.5 g starch, 17.5 g casein hydrolysate, pH 7.4, in a final volume of 1 liter) supplemented with 50 mg/ml kanamycin and containing MICs of the compound on which the clones were originally selected, per Table S2 in the supplemental material. Before liquid growth assays, clones were passaged twice through LB-Kan broth by transfer with a 96-pin stamp transfer tool (each passage was allowed to grow overnight) to eliminate residual glycerol from freezer stocks. To test for tolerance in liquid medium, cultures were transferred into 96-well plates containing 250 ml of medium with inhibitory compound and growth was profiled by absorbance measurements at 600 nm (optical density at 600 nm [OD600]) taken every 20 min using the Synergy H1 microplate reader (BioTek Instruments) for up to 4 days at 37°C. Clones positive for growth were rearrested, stored at −80°C, and subjected to a second round of liquid growth assays, performed exactly as described above. Only clones showing growth in both liquid growth assays were considered tolerant and maintained for downstream analyses.

Amplification of metagenomic DNA fragments from selected transformants. Amplification of metagenomic fragments from picked transformants was performed by PCR, using 1 ml of cells from overnight culture. A sample PCR mixture consisted of 1 ml of template, 2.5 ml of Thermopol reaction buffer (New England Biolabs), 0.5 ml of 10 mM deoxynucleotide triphosphates (dNTPs; New England Biolabs), 0.5 ml of Taq polymerase (New England Biolabs; 5 U/ml), 3 ml of a custom primer mix (see Table S3 in the supplemental material), and 17.5 ml of nuclease-free H2O to bring the final reaction volume to 25 ml. PCR mixtures were then amplified using the following thermocycler conditions: 94°C for 10 min and 35 cycles of 94°C for 5 min, 55°C for 45 s, 72°C for 5.5 min, and 72°C for 10 min. The amplified metagenomic inserts were then purified using the QIAquick 96-well PCR purification kit (Qiagen), per the manufacturer’s recommendations. After purification, total eluate from fragments from colonies on the same selection was pooled for Illumina library preparation and sequencing, resulting in 50 pools, each representing a unique selection.

Amplification of metagenomic DNA fragments from total bacterial growth. Freezer stocks of plate scrape cultures from positive selections were thawed, and 200 to 250 ml of cells was pelleted by centrifugation at 13,000 rpm for 2 min and gently washed with 1 ml of nuclease-free H2O. Cells were then pelleted again and resuspended in 20 to 25 ml nuclease-free water (1/10 original volume), frozen at −20°C, and then thawed to promote cell lysis. Subsequently, 2.5 ml of thawed suspensions was used as the template for amplification of resistance-conferring DNA fragments by PCR with Taq DNA polymerase (New England Biolabs) exactly as described above, except that 25 cycles of PCR, rather than 35, were performed.

Illumina sample preparation and sequencing. Pooled amplicons from picked colonies and amplified metagenomic inserts from total bacterial growth were processed similarly as previously described (28). For each input, 500 ng of PCR product was sheared to 150- to 200-bp fragments using nine 10-minute cycles of 30 s on and 30 s off on the Bioruptor XL sonicator (high-power setting). The ends of sheared DNA fragments were blunt-ended with T4 DNA ligase and polymerase (New England Biolabs), and barcoded adapters were ligated to blunt fragments by T-A cloning with T4 DNA ligase (New England Biolabs). Next, 10 ml of adapter-ligated samples was combined into pools of 12 samples and concentrated by elution through a Qiagen MinElute PCR purification column. DNA sized 300 to 400 bp was purified by gel extraction, and adapter-
ligated samples were enriched by PCR using 12.5 μl 2X Phusion HF master mix, 1 μl of 10 μM Illumina PCR primer mix, and 2 μl of purified DNA as the template in a 25-μl PCR mixture. PCR proceeded as follows: 30 s at 98°C followed by 17 cycles of 98°C for 10 s, 65°C for 30 s, and 72°C for 30 s before a final 5-min extension at 72°C. Finally, a 9-PF M reaction was used for Illumina HiSeq 101-bp sequencing with the HiSeq 2500 platform at GTAC (Genome Technology Access Center, Washington University in St. Louis, MO, USA).

Assembly and annotation of metagenomic contigs. Illumina paired-end sequence reads were binned by barcode (perfect match required), such that independent selections were assembled and annotated in parallel. Metagenomic DNA fragments from each selection were assembled using PARFuMs (parallel assembly and reassessment of functional metagenomic selections), a tool developed for the high-throughput assembly and annotation of functional metagenomic selections (36). Assembly with PARFuMs consists of three iterations of variable job size with the short-read assembler Velvet (37), two iterations of assembly with PHRAP (38), and custom scripts to clean sequence reads, remove chimeric assemblies, and link contigs by coverage and common annotation, as described previously (36). In total, 116 sequencing libraries were assembled with PARFuMS, encompassing 84 unique selections (for 32 selections, picked colonies and total bacterial growth were sequenced and assembled separately). Fifty-two selections yielded metagenomic fragments greater than 500 bp, totaling 1,932 metagenomic contigs assembled. To annotate these contigs, we predicted ORFs using the gene-finding algorithm MetaGeneMark (39) and predicted gene function by searching the amino acid sequences against the TIGRFAMs (40) and Pfam (41) profile HMM databases with HMMER3 (42). MetaGeneMark was run using default gene-finding parameters while hmmscan (HMMER3) was run with the option “-cut_ga,” requiring that genes meet profile-specific gathering thresholds (rather than a global, more permissive, default log odds cutoff) before receiving annotation. The hit with the best E value was used to designate an annotation, and ORFs over 350 bp were used in downstream analyses.

Percent identity comparisons of recovered ORFs against NCBI database. All recovered ORFs greater than 350 bp were compared against the NCBI protein Non-Redundant (NR) database (retrieved 15 September 2014) using BLASTX to determine the amino acid identity between recovered ORFs and their nearest neighbor in an NCBI protein. For each ORF, the NCBI entry that generated the best local alignment was used to seed global alignments with estwite (http://denodeum.ucdavis.edu/resources/tooldocs/wise2/doc_wise2.html). The following options were used in global alignment: “-init global” and “-alg 333.” From this alignment, global percent identities were calculated as the number of matched amino acids divided by the full length of the shorter of the two sequences compared.

Subcloning putative tolerance-conferring genes. Putative tolerance-conferring genes were amplified from picked colonies twice verified for tolerance by liquid growth assay. Plasmids (pZE21 plus a soil metagenomic fragment) were purified by minipreparation using a Qiagen plasmid purification kit and used as the template for PCR of individual tolerance genes. Open reading frames were amplified using PFX polymerase (Life Technologies), with the following components in a 50-μl reaction mixture: 1.5 μl purified plasmid as the template, PFX buffer (10 μl), 10 mM dNTPs (New England BioLabs; 1.5 μl), 50 mM MgSO4 (1 μl), PFX enhancer solution (5 μl), polymerase (0.4 μl), 10 μM forward primer (0.75 μl), and 10 μM reverse primer (0.75 μl). Thermocycler conditions were as follows: 95°C for 5 min followed by 35 cycles of 95°C for 45 s, 55°C for 45 s, and 72°C for 3.5 min, finishing with a 5-min extension at 72°C. Reaction-specific primers can be found in Table S4 in the supplemental material. Amplicons were then purified using RapidTip2 PCR purification tips (Diffrinity Genomics) and quantified using the BioStack automated microplate stacker, with absorbance readings taken as described above at 37°C every 45 min for 3 days (cultures were shaken at medium intensity for 2 min before measurement).

Nucleotide sequence accession numbers. All assembled sequences have been deposited in GenBank and are available via the BioProject identifier PRJNA294310 (GenBank accession numbers KT941423 to KT943354).

RESULTS
We constructed 16 small-insert metagenomic libraries using soils from two sites in the Long Term Ecological Research network: grassland soils from Cedar Creek (CC) and agricultural soils from Kellogg Biological Station (KBS) (26–28). Metagenomic libraries ranged in size from 1.4 to 16.9 Gb (see Table S1 in the supplemental material) and were selected for tolerance to one of 20 compounds (15 common lignocellulosic inhibitors and five current and next-generation biofuels) on solid medium at the MIC of each compound against the E. coli host (see Table S2). Growth was monitored for 5 days, with tolerance identified to seven of 20 compounds (four lignocellulosic inhibitors and three fuels). Tolerance appeared to be largely compound specific (as opposed to library or soil specific), indicating that for most compounds, edaphic properties did not influence the ability to capture a tolerance phenotype (Fig. 1). For a subset of the 13 compounds where tolerant clones were not identified, organisms within the soil are known to metabolize the chemical (23, 45), and in at least one
instance, tolerance to 2-furoic acid has been transferred to *E. coli* using soil-derived large-insert metagenomic libraries (19). Thus, the absence of tolerant growth in our experiments does not signify the absence of these properties within the soil microbial community or the inability to transfer the trait to *E. coli*. Instead, it is likely that tolerance to these compounds either requires the action of multiple, distal genes or is encoded by genes at low to intermediate abundance, such that the functional gene was not incorporated into the original metagenomic libraries. Alternative strategies (e.g., *in situ* enrichment of soil communities [44, 45]) are likely required to expand both the number of compounds to which tolerance is observed and the diversity of tolerance-conferring genes recovered.

Some selections yielded unambiguous tolerance phenotypes,

<table>
<thead>
<tr>
<th>Compound / Soil</th>
<th>Grassland Soils</th>
<th>Agricultural Soils</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alcohol (inhibitor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Furfuryl alcohol</td>
<td>●●●●</td>
<td>●●●●●●●●●●●</td>
</tr>
<tr>
<td>4-Methylicatechol</td>
<td>●●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Vanillinalcohol</td>
<td>●●</td>
<td>●●●●●</td>
</tr>
<tr>
<td>Furfural</td>
<td>●</td>
<td>●●●●</td>
</tr>
<tr>
<td>4-hydroxybenzaldehyde</td>
<td>●</td>
<td>●●●</td>
</tr>
<tr>
<td>Vanillin</td>
<td>●</td>
<td>●●</td>
</tr>
<tr>
<td>Aldehyde (inhibitor)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acetic acid</td>
<td>●</td>
<td>●●</td>
</tr>
<tr>
<td>Caproic acid</td>
<td>●</td>
<td>●●</td>
</tr>
<tr>
<td>Ferulic acid</td>
<td>●</td>
<td>●●</td>
</tr>
<tr>
<td>Formic acid</td>
<td>●●</td>
<td>●●</td>
</tr>
<tr>
<td>2-Furoic acid</td>
<td>●●</td>
<td>●●</td>
</tr>
<tr>
<td>Gallic acid</td>
<td>●●</td>
<td>●●</td>
</tr>
<tr>
<td>4-hydroxybenzoic acid</td>
<td>●</td>
<td>●●</td>
</tr>
<tr>
<td>Syringic acid</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Vanillic acid</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Short Chain Alcohol (fuel)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-butanol</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Ethanol</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>2-methyl-1-butanol</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>3-methyl-1-butanol</td>
<td>●</td>
<td>●</td>
</tr>
<tr>
<td>Isobutanol</td>
<td>●</td>
<td>●</td>
</tr>
</tbody>
</table>

Tolerance observed on solid media | Contig Assembled > 500bp | At least one picked clone verified for tolerance in liquid medium

**FIG 1** Functional selections of 16 soil metagenomes for tolerance to 20 inhibitory compounds. (A) Gray cells designate selections where a tolerance phenotype was observed, hatched cells designate selections where a metagenomic contig greater than 500 bp was assembled, and dotted cells indicate selections from which individual clones were twice validated for tolerance in liquid medium. (B) Structures of the inhibitors and fuels used in functional selections.
defined as punctate colonies on experimental plates (those with metagenomic libraries) and no observable growth on matched, empty-vector controls. In other cases, small satellite colonies surrounded larger, likely bona fide tolerant clones. This phenotype may result from a larger colony that detoxifies the compound and allows for the survival of nearby nontolerant cells (e.g., see Fig. S1 in the supplemental material). Accordingly, further validation was often required to distinguish truly tolerant clones from background. As a means of validation, 483 colonies from the 99 selections that showed a positive phenotype (Fig. 1) were randomly selected and verified for tolerance by passage through liquid medium containing MICs of the appropriate compound (see Table S2). In total, 132 clones from 50 selections reproducibly exhibited a tolerance phenotype across two independent liquid growth assays and were selected for downstream analyses (Fig. 1).

For 66 of the 99 positive selections (~67%), all bacterial growth from solid selection plates was collected and putative tolerance-conferring metagenomic fragments were amplified by PCR and prepared in Illumina sequencing libraries. These 66 libraries were barcoded by selection and combined with another 50 libraries constructed from tolerance-conferring DNA fragments amplified from the 132 verified tolerant clones, also barcoded by selection. The 116 sequencing libraries included (i) 34 selections where total growth, but not verified clones, was interrogated; (ii) 18 selections where only verified clones were sequenced; and (iii) 32 selections where data of both types were generated. All 116 libraries were sequenced and assembled using PARFuMS (36), with 52 of 84 selections yielding at least one assembled contig longer than 500 bp (Fig. 1). In total, 1,932 metagenomic fragments were assembled ($N_{50}$ of >1.35 kb) containing 2,283 ORFs which generally appeared unlike sequences deposited in public repositories, consistent with past observations from soil metagenomes (28, 46–51). The average amino acid identity of all 2,283 ORFs to their closest homolog in the NCBI database was 52.7% ± 19.0% (mean ± standard deviation) (Fig. 2).

The most comprehensive sequence data were available for selections using furfuryl alcohol, ferulic acid, and furfural, and we focused downstream work on sequences identified from selections using these compounds. These chemicals encompass alcohol, acid, and aldehyde inhibitors and include structures derived from both the hemicellulosic and lignified fractions of lignocellulose.

Glyoxylase resistance proteins were the most common function identified in furfural selections (see Table S5 in the supplemental material), which are archetypal aldehyde-detoxifying enzymes (52). The heterologous expression of a glyoxalase in E. coli from the soil bacterium Pseudomonas putida has been demonstrated to improve tolerance and fuel yield in the presence of the aldehyde inhibitor methylglyoxal (53), but our data are the first to implicate a putative lactoylglutathione lyase in furfural tolerance. From selections on furfuryl alcohol, we recovered ORFs with vague or uncharacterized predicted function (e.g., HD domain and domain of unknown function [DUF] 20 [see Table S5]). Given the sequence divergence of the cloned soil metagenome from known functions (Fig. 2), vague or unassignable annotations are not surprising and may be particularly common with selections using furfuryl alcohol given the compound’s poorly characterized modes of toxicity (14).

The most frequently recovered genes from ferulic acid selections were homologous to galE, a UDP-glucose 4-epimerase which catalyzes the interconversion between UDP-glucose and UDP-galactose (54) (see Table S5 in the supplemental material). Interestingly, many substituted phenolic compounds inhibit UDP-glucose 4-epimerases (54, 55). Functional deficiencies in these enzymes result in cell wall defects (in the absence of galactose) (56) or cell death (in the presence of galactose, a component of plant hemicellulose [6]) (57). The overexpression of galE homologs may rescue the native UDP-glucose 4-epimerase of E. coli in the presence of ferulic acid, a guaiacyl phenolic compound. Given the structural diversity of phenolic compounds capable of inhibiting GalE (54, 55), its rescue may represent a generalized strategy for engineering tolerance toward products of lignin hydrolase. Indeed, the overexpression of galE homologs confers tolerance against the syringyl phenolic monomer syringaldehyde (19) and the substituted aromatic menadione (58) when expressed in E. coli.

To identify individual ORFs that confer a tolerance phenotype and to quantify an ORF’s impact on tolerance when removed from its native genetic context (as would occur in engineered production strains), we subcloned ORFs from their original metagenomic fragment and assayed for tolerance to the compound on which the gene was originally selected. We prioritized functions predicted to encompass broad mechanistic diversity, amplifying ORFs from colonies previously verified for tolerance (rather than from total bacterial growth on functional selections), as this allowed us to select transformants that displayed strong tolerance phenotypes. Across furfuryl alcohol, ferulic acid, and furfural selections, we cloned 23 predicted ORFs from 14 assembled contigs, seven of which were confirmed to provide tolerance when removed from their native context, spanning all three compounds (Fig. 3; see also Table S6 in the supplemental material). No tolerance-conferring gene impaired growth when expressed in the absence of inhibitor (see Fig. S2), suggesting that these genotypes do not generally reduce fitness and are therefore suitable for bioprocessing applications.

Two genes with unknown function conferred tolerance to ferulic acid and furfuryl alcohol, with shortened lag times and increased growth rates when expressed heterologously in E. coli compared to empty-vector controls (Fig. 3B and G). This highlights the potential of uncharacterized gene products from the soil metagenome to impact bioprocessing applications. We also identified two thymidylate synthases, homologs of ThyA and ThyX, which confer furfural tolerance (Fig. 3E and F). These enzymes catalyze the production of deoxothymidine monophosphate (dTMP) from dUMP, and their detection is consistent with pre-
A phenolic acid decarboxylase (PadC) with perfect amino acid identity to homologs from the genus *Pantoaea* (NCBI GenInfo number 497939809) was cloned and confirmed to confer tolerance to ferulic acid, a lignin monomer (Fig. 3A). Ferulic acid decarboxylases are of high industrial interest, as they catalyze the conversion of ferulic acid to 4-vinyl guaiacol, an important precursor for vanillin production with approximately 30 times the value of ferulic acid (61). Indeed, multiple groups have expressed PadC homologs in *E. coli* and have shown that these enzymes can produce 4-vinyl guaiacol from ferulic acid (61, 62). In contrast to previous work, which suggested that expressing a phenolic acid decarboxylase impaired the growth of *E. coli* and other Gram-negative bacteria (62), we demonstrate improved ferulic acid tolerance when expressing the decarboxylase (Fig. 3A). Toxicity differences between our study and previous work may result from differences in PadC expression level or the time course used in growth assays (PadC confers delayed ferulic acid tolerance at some inhibitor concentrations [see Fig. S3A in the supplemental material]). Nonetheless, PadC holds the potential to couple biomass detoxification with the production of the high-value chemical 4-vinyl guaiacol, offering a means to offset costs associated with biomass conversion.

The observed enrichment for galactose metabolism in ferulic acid selections, particularly *galE* homologs (see Table S5 in the supplemental material), suggested that the overexpression of *galE* may provide ferulic acid tolerance. Consistent with this hypothesis, a *galE* homolog conferred strong tolerance to ferulic acid when removed from its native context; a recombinant *E. coli* strain expressing this gene grew robustly at concentrations of ferulic acid entirely inhibitory to a vector-only control (Fig. 3C). Because substituted phenolics inhibit UDP-glucose 4-epimerases (54, 55), we speculate that *galE* overexpression may provide tolerance by rescuing a compromised native UDP-glucose 4-epimerase. In addition to *galE* homologs, ferulic acid selections also yielded numerous homologs of galactokinase (*galK*), another nucleotide-galactose active enzyme (see Table S5), which conferred tolerance to ferulic acid when overexpressed (Fig. 3D). The *galK* homologs identified by functional metagenomic selections all contain mutations at residues predicted to eliminate UDP-glucose 4-epimerase activity (63–65) (see Table S7), suggesting that tolerance is conferred by functionally impaired GalK variants.

Sugar is the predominant carbon source in most bioprocessing applications, whereas they constitute a minor source of carbon in LB medium compared to small peptides. To determine whether the observed ferulic acid tolerance genes improved *E. coli* growth under conditions more representative of real-world bioreactors, we tested each tolerance gene against a panel of lignin-derived phenolic monomers in M9 minimal medium supplemented with glucose as the primary carbon source. These minimal medium experiments typically reproduced the tolerance phenotypes observed using LB medium, though the growth curves occasionally differed between experiments (compare Fig. 3 to Fig. S3 in the supplemental material). Although it is tempting to attribute these growth differences to medium composition, other growth conditions also differed between experiments (namely, periods of orbital shaking, as described in Materials and Methods), and we therefore compare growth differences within, but not across, experiments.

With the exception of *galK*, each gene conferred tolerance to multiple phenolic monomers in minimal medium (Table 1; see...
also Fig. S3 in the supplemental material), despite being recovered from only ferulic acid in the original selections using LB medium. Interestingly, expression of galE conferred tolerance in both LB medium, which contains galactose (66), and M9-plus-glucose medium, which lacks galactose. Expressing our galK variant, however, provided tolerance only in the presence of galactose (Fig. 3; see also Fig. S3). This is consistent with previous observations that suggest an impaired GalE enzyme can exhibit galactose-independent toxicity (56) but indicates that our GalK variants likely ameliorate only galactose-dependent forms of toxicity.

**DISCUSSION**

Analogous to the predicted mechanism for galE-mediated ferulic acid tolerance, simple complementation of an impaired GalK enzyme could underlie the tolerance conferred by galK overexpression. Although intuitive, this hypothesis is unlikely to explain our observed ferulic acid tolerance, as it is inconsistent with previous observations; galK mutants rescue galE mutants (36, 67), likely by preventing the accumulation of GalK’s reaction product, galactose-1-phosphate (67). Therefore, increased GalK enzyme is expected to exacerbate the deleterious effects of an impaired GalE, and the overexpression of a functional galK gene is unlikely to compensate for a phenolic-inhibited GalE.

Instead, we hypothesize that the tolerance-conferring GalK variants may exhibit a dominant negative phenotype. For instance, a GalK variant may be catalytically inactive and yet able to inhibit E. coli’s native GalK by competition for substrate, relieving galactose-1-phosphate accumulation and ameliorating the galactose-dependent toxicity associated with an impaired GalE. Supporting this hypothesis, each of the five full-length galK homologs assembled from ferulic acid selections contains mutations in core motifs that are conserved across diverse bacteria, archaea, yeast, and mammals (see Table S7 and Text S1 in the supplemental material). In four of five cases (including in the enzyme expressed in Fig. 3D), mutations at these residues have been shown to eliminate galactokinase activity (63–65), with the fifth enzyme predicted to contain an alternative start codon that would truncate a key functional motif (68) (see Table S7), suggesting that loss-of-function variants of galK provide ferulic acid tolerance. Phenolic acids are also key toxins in bioproduction efforts that utilize the yeast *S. cerevisiae* (11), for which GALE deficiency similarly triggers galactose toxicity which is rescued upon GALK deletion (69). Though phenolic acids inhibit GalE in many organisms (e.g., trypanosomes and mammals [55]), whether they exert toxicity via GALE impairment in *S. cerevisiae* is unknown and warrants future work.

We also identified two thymidylate synthase variants, ThyA and ThyX, which can confer furfural tolerance on *E. coli* when expressed heterologously. Despite furfural toxicity resulting (in part) from NADPH starvation (59, 70), both ThyA and ThyX are known to cause NADPH oxidation as a consequence of dTMP synthesis, though oxidation occurs at different catalytic steps with each enzyme. ThyA can use methylenetetrahydrofolate (MTHF) as both a carbon donor and a reductant to synthesize dTMP from dUMP, producing dihydrofolate (DHF). DHF must be reduced to tetrahydrofolate (THF) before it can be recycled back into the original methyl donor, MTHF (60, 71, 72); this reduction requires an increase of DHF. In contrast, ThyX produces THF as a direct product of dTMP synthesis but oxidizes NADPH during this reaction, obviating the need for DHF reduction but not NADPH oxidation (60, 71, 72). Because both thymidylate synthases result in oxidized NADPH and yet both confer furfural tolerance when overexpressed, these enzymes likely ameliorate furfural toxicity independently of the compound’s effect on NADPH levels (59, 70, 73, 74).

Expression of ThyX, but not ThyA, reduces dependence on DHF-reducing enzymes (60, 71, 72), and yet the two thymidylate synthases confer similar levels of furfural tolerance (Fig. 3E and F). This suggests that the furfural tolerance conferred by these enzymes is also independent of their differential folate-reducing capacities and implies that furfural does not inhibit folate reduction reactions, as has been theorized elsewhere (59). Furfural is also known to cause DNA damage (75, 76), perhaps via the induction of reactive oxygen species (77). Accordingly, the increased furfural tolerance that results from ThyA and ThyX overexpression may result from an increased supply of pyrimidine deoxyribonucleotides to aid in DNA repair (59). Since both enzymes are expected to increase dTMP production, this explanation for furfural-induced toxicity most satisfactorily describes the similar tolerance profiles conferred by each thymidylate synthase.

We describe gain-of-function tolerance to acid, alcohol, and aldehyde inhibitors derived from hemicellulose and lignin, demonstrating that functional selections using soil metagenomes have tremendous potential to address the toxicity associated with thermochemical lignocellulose depolymerization. Because bioprocessing applications are necessarily diverse and are expected to utilize variable lignocellulosic feedstocks (8, 9, 78), our capacity to understand and engineer tolerance into microbial catalysts must become comparatively flexible. By performing selections against individual chemical inhibitors, rather than actual pretreated hydrolysate, we identified tolerance genes that yield insight into the mechanisms of individual inhibitor toxicity (e.g., the putative inhibition of GalE by phenolic compounds) and associated microbial tolerance (e.g., increased pyrimidine supply to provide furan tolerance). This strategy allows for flexible bioengineering efforts, where custom production hosts may be designed with inhibitor tolerance tailored particularly to their bioprocessing applications.

The toxicity associated with treated lignocellulosic hydrolysate likely results from the combination of many inhibitors at low concentrations (9). In contrast, the tolerance genes that we describe were recovered from selections on individual compounds. These genes nonetheless hold the potential to confer broad-spectrum

---

**TABLE 1 Summary of tolerance genes active against phenolic inhibitors**

<table>
<thead>
<tr>
<th>Compound</th>
<th>ORF9/</th>
<th>ORF12/unknown function</th>
<th>ORF8/</th>
<th>ORF7/</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>palDC</td>
<td></td>
<td>galE</td>
<td>galK</td>
</tr>
<tr>
<td>Ferulic acid (LB medium)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Ferulic acid (M9 medium)</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>ND</td>
</tr>
<tr>
<td>Vanillic acid</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Vanillic alcohol</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>--</td>
</tr>
<tr>
<td>4-Methylcatechol</td>
<td>--</td>
<td>+</td>
<td>--</td>
<td>+</td>
</tr>
<tr>
<td>Syringic acid</td>
<td>+</td>
<td>--</td>
<td>+</td>
<td>+/-</td>
</tr>
<tr>
<td>4-Hydroxybenzoic acid</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+/-</td>
</tr>
<tr>
<td>4-Hydroxybenzaldehyde</td>
<td>+</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

---

*All data were generated using M9 defined medium, except as indicated. +, tolerance conferred; --, no tolerance; +/-, ambiguous tolerance; ND, not determined.*
tolerance. We provide several examples of genes identified on ferulic acid selections that confer tolerance to many additional phenolic compounds (see Fig. S3 in the supplemental material). Future work should test combinations of genes for broad-spectrum tolerance to synthetic inhibitor mixtures and real-world hydrolyses, enabling the empirical construction of custom tolerance operons for diverse inhibitor profiles.

Despite the diversity of our functional metagenomic libraries, selections did not identify genes for tolerance to several lignocellulosic inhibitors known to be tolerated by soil microorganisms (19, 23, 43). This outcome is likely a result of the incredible diversity of soil metagenomes, which confounds efforts to identify low-abundance genotypes. Methods to enrich these important yet low-abundance functions should improve the capture efficiency of functional selections and enable the discovery of new gene functions (44, 45). Even without applying enrichment strategies, we discover numerous tolerance genes active against hemicellulose- and lignin-derived inhibitors, including genes of entirely unknown function (e.g., Fig. 3B and G) as well as enzymes with well-defined applications (61, 62) (e.g., PadC [Fig. 3A]). Our discoveries highlight the potential for the soil metagenome to enable process improvements in industrial biotechnology. The tremendous genetic diversity in soil (28, 46–51) (Fig. 2), however, remains mostly unexplored, and this potential is largely waiting to be realized.

ACKNOWLEDGMENTS

We thank members of the Dantas lab for their help with protocol optimization and for thoughtful discussions on the results and analyses presented here, the Genome Technology Access Center at Washington University School of Medicine for Illumina sequencing services, and N. Fierer and C. Lauber for providing soil samples.

Research reported in this publication was supported in part by awards to G.D. through the Children’s Discovery Institute (MD-II-2011-117), the International Center for Advanced Renewable Energy and Sustainability at Washington University, the National Academies Keck Futures Initiatives (Synthetic Biology, SB2), and the NIH Director’s New Innovator Award (DP2-DK-098089). K.J.F. is a National Science Foundation graduate research fellow (award number DGE-1143954) and received support from training grants through the National Human Genome Research Institute (NHGRI; https://www.genome.gov/) and the NIGMS through award numbers T32HG000045 and T32GM070767. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

FUNDING INFORMATION

Children’s Discovery Institute provided funding to Gautam Dantas under grant number MD-II-2011-117. International Center for Advanced Renewable Energy and Sustainability at Washington University provided funding to Gautam Dantas under grant number 2011-Dantas. National Academies – Keck Futures Initiative provided funding to Gautam Dantas. National Science Foundation (NSF) provided funding to Kevin J. Forsberg under grant number DGE-1143954. HHS | NIH | National Human Genome Research Institute (NHGRI) provided funding to Kevin J. Forsberg under grant number T32 HG000045. HHS | NIH | National Human Institute of General Medical Sciences (NIGMS) provided funding to Kevin J. Forsberg under grant number T32 GM07067. HHS | NIH | NIH Office of the Director (OD) provided funding to Gautam Dantas under grant number DP2 DK098089.

REFERENCES


32. Forsberg et al. 536


39. Mori T, Suenaga H, Miyazaki K. 2008. A metagenomic approach to the identification of UDP-glucose 4-epimerase as a menadione resistance pro-
plasmid-based expression of thyA.


