nature biotechnology

Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides

Joseph R Warner[1], Philippa J Reeder[1], Anis Karimpour-Fard[2], Lauren B A Woodruff[1] & Ryan T Gill[1]

A fundamental goal in biotechnology and biology is the development of approaches to better understand the genetic basis of traits. Here we report a versatile method, trackable multiplex recombineering (TRMR), whereby thousands of specific genetic modifications are created and evaluated simultaneously. To demonstrate TRMR, in a single day we modified the expression of >95% of the genes in *Escherichia coli* by inserting synthetic DNA cassettes and molecular barcodes upstream of each gene. Barcode sequences and microarrays were then used to quantify population dynamics. Within a week we mapped thousands of genes that affect *E. coli* growth in various media (rich, minimal and cellulosic hydrolysate) and in the presence of several growth inhibitors (β-glucoside, D-fucose, valine and methylglyoxal). This approach can be applied to a broad range of traits to identify targets for future genome-engineering endeavors.

© 2010 Nature America, Inc. All rights reserved.

Microbial genomes hold the potential for tremendous combinatorial diversity, comprising a sequence space of $4^{4,600,000}$. Researchers' ability to search this diversity for genetic features that affect pertinent traits remains limited by the number of individuals that can be tested, which is a small fraction of all possibilities. Thus, there is a demand for strategies for first defining relevant genetic variation and then thoroughly searching that space. This issue has been studied in great depth at the level of individual genes[1,2], where high-throughput protein engineering methods are available for introducing specific mutations and then mapping the effects of such mutations onto protein activity. Advances in genomics[3], and more recently multiplex DNA synthesis[4–8] and homologous recombination (or recombineering)[9–11], now enable the extension of such a strategy to the genome scale.

Advances in genomics have resulted in several methods for highly parallel mapping of genes to traits, such as profiling of gene-knockout and plasmid-based libraries[12–20]. In some instances, microarray technology has been used to enable parallel tracking of genetically distinct individuals throughout growth in selective environments. One such tool, molecular barcoding[12,17], involves the replacement of every gene in *Saccharomyces cerevisiae* with a specific DNA sequence that could be tracked via microarray. Although these tools are a powerful way to profile the effect of mutation, the difficulty of specifically creating new mutations limits these studies to one of two types of mutations that have previously been introduced (insertions or increases in copy number). These limitations have challenged efforts to apply these methods for dissecting phenotypes and reengineering phenotypes that rely upon the coordinated action of multiple genes and mutations.

Research over the past decade has resulted in recombination-based methods (recombineering) that make it easier to specifically modify the *E. coli* genome using synthetic DNA (synDNA)[9–11,21–23]. Recently, a recombineering-based method, called MAGE, was reported[24], whereby the expression levels of 24 genes were optimized in parallel to improve lycopene production more than all previously reported efforts, in considerably less time. This demonstration was enabled by *a priori* knowledge of what genes to modify, which is not known in many genome-engineering efforts, such as engineering growth and tolerance. Here we describe TRMR, a complementary method for simultaneously mapping genetic modifications that affect a trait of interest. The method combines parallel DNA synthesis, recombineering and molecular barcode technology to enable rapid modification of all *E. coli* genes (**Fig. 1** and **Supplementary Fig. 1**). We demonstrate this general approach through the construction of two comprehensive *E. coli* genomic libraries comprising 8,000 distinct mutations and gene-trait mapping of these cells in seven environments.

RESULTS

Synthetic DNA cassettes for promoter replacement

We designed a comprehensive library of synDNA cassettes that have predictable effects when inserted into the genome of *E. coli*. Although various genetic features could have been incorporated into the cassettes (such as point mutations or sequences affecting mRNA stability, translational efficiency and other processes), we chose to demonstrate TRMR using functional modifications that either generally increase the expression of a target gene, called 'up', or generally decrease the gene's expression, called 'down'. The up cassette contains a strong and repressible $P_{LtetO-1}$ promoter[25] and ribosome binding site (RBS)[26] sequences, which in general will increase downstream gene transcription and translation (**Fig. 2**). The down cassette was designed to replace the native RBS with an inert sequence that will generally cause a decrease in translation initiation. Both cassette designs include a blasticidin-S resistance gene[27], allowing for selection of recombinant alleles. Molecular barcodes[12] (also called 'tags') were incorporated to track the presence of each synDNA oligo and to

[1]Department of Chemical and Biological Engineering, University of Colorado, Boulder, Colorado, USA. [2]School of Medicine, University of Colorado at Health Science Center, Denver, Colorado, USA. Correspondence should be addressed to R.T.G. (rtg@colorado.edu).

Received 4 February; accepted 8 June; published online 18 July 2010; doi:10.1038/nbt.1653

track each allele (engineered cell) within the mixed population on a barcode microarray[28] (**Supplementary Notes**).

Because the length of the synDNA cassettes used here is beyond the current capabilities of commercially available oligo library synthesis, we developed a strategy for multiplex cassette construction that involves the ligation of sequences shared by all cassettes to a mixture of shorter oligos specific to each targeted gene. Construction of this library was complicated by the fact that each synDNA cassette must contain unique sequences in the flanking positions that are homologous to the chromosome where the cassette is to be inserted. This is traditionally accomplished by using PCR to amplify a DNA cassette with primers that contain the flanking homology regions[21,29]. Using such a method to construct thousands of alleles is resource- and time-intensive[3,11,19], thus limiting the number and type of allelic libraries that can be investigated.

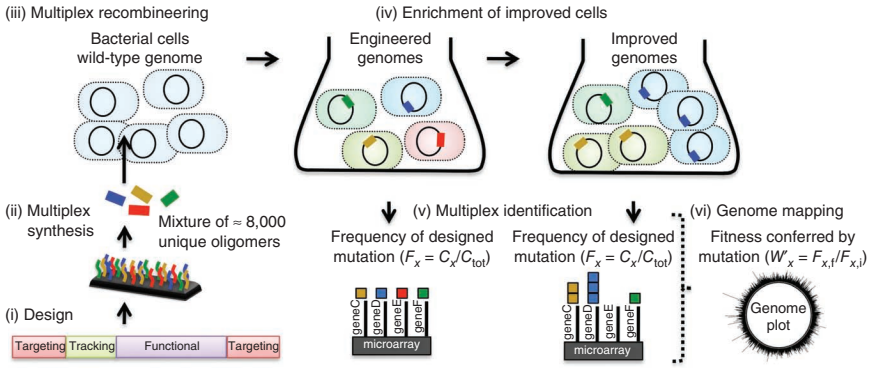

Figure 1 TRMR method. (i) Design DNA cassettes encoding the suite of mutations of interest. (ii) Synthesize those cassettes, along with associated molecular barcodes, in a single pool. (iii) Introduce cassettes into recombination-proficient *E. coli*[46] and produce thousands of variants, each with a distinct region of the chromosome that is engineered. (iv) Perform selections or screens on the mixture of variants to enrich for those possessing a desired trait. (v) Quantify changes in allele frequency using molecular barcode technology[47]. (vi) Use these frequency measurements to map specific genetic changes onto the trait of interest. C_x, concentration of allele x; C_{tot}, total concentration; $F_{x,f}$ and $F_{x,i}$, final and initial allele frequencies (see equations in Results).

To address these issues, we developed a procedure to generate thousands of synDNAs containing multiple desirable sequence features (such as homology regions and expression modulators) that can be carried out in a complex mixture. Briefly, 'targeting oligos' were first synthesized on a microarray. Then, we ligated these to the cassette that modifies gene function, amplified the resulting product with rolling-circle amplification and then cleaved the long amplified DNA molecule into the synDNAs (**Fig. 2a–c**).

Targeting oligos were designed for every protein-coding gene in the *E. coli* MG1655 genome (**Supplementary Table 1** and **Supplementary Notes**). In all, 8,154 targeting oligos were designed to create two possible expression alleles for 4,077 genes. Targeting regions were chosen such that DNA cassettes would insert upstream of genes, replace the translation start codon and account for gene overlap. Once designed, the set of targeting oligos, each 189 nucleotides long, was purchased through limited access at a cost of roughly $1 per unique oligo (Oligonucleotide Library Synthesis, Agilent).

To test cassette design and construction and to optimize the procedure for allele production, we attempted promoter replacement for the *lacZ* and *galK* genes. After optimizing design, we were able to efficiently generate these alleles using the procedures outlined in **Figure 2**. Alleles were isolated as colonies and all showed the expected change in regulation and expression of the *lacZ* gene (**Fig. 2d,e**) or the *galK* gene. Furthermore, in PCR confirmations and sequencing, 30 of 30 alleles tested showed the correct site of insertion. By counting colonies we estimated that we were able to routinely generate at least 75 alleles per microliter of cells transformed and determined that yields increased linearly with transformation volumes from 40 ml up to 400 ml tested. With increases in scale, it is conceivable that one could generate 10^5–10^7 alleles in a single day, enough to profile several modifications of every *E. coli* gene.

Efficient construction of genome-scale allele libraries

Using a library of 8,154 targeting oligos, we attempted to construct 4,077 up synDNA oligos and 4,077 down synDNA oligos in separate pools. Both oligo pools were constructed in 1 week and resulted in enough material for several rounds of multiplex recombineering. The synDNA oligos were then used in a day of recombineering experiments, separately generating thousands of up and down recombinant colonies. Colonies were scraped from plates and frozen in aliquots for subsequent experiments.

To confirm that desired mutant alleles were generated, we PCR amplified and sequenced barcode tags from 390 colonies. Sequencing of the cassette and neighboring chromosome DNA indicated that in 34 of 34 distinct alleles, the cassettes had inserted into the correct location of the genome. Sequencing also provided an estimate of the number of alleles containing an error in DNA sequence. Outside of the barcode sequences, DNA errors were observed in only three of 34 alleles, two of which had errors in regions of the cassette that should not affect allele identification or function. The barcode tag sequences provide an estimate of DNA errors present in the initial oligo libraries because barcodes are not subject to the experimental bias (bias includes selection for correct sequences during PCR amplification and during homologous recombination) that would filter out incorrect sequences. High fidelity of the molecular barcode sequences is also required to accurately detect the presence of each allele in cell mixtures. Only 5% of the 390 sequenced tags showed an error, usually substitution or loss of a single nucleotide. The high percentage of correct alleles observed here is a first indication that complex oligonucleotide mixtures may be used to engineer and identify thousands of distinct genomic loci with high fidelity.

To assess our ability to make complete and uniform libraries in multiplex, we used Affymetrix Geneflex TAG4 arrays[28] to measure the concentration of each barcode tag in the synDNA mixture (before recombineering) and in genomic DNA from cell mixtures (after recombineering). We observed microarray signals from hybridization of each of the 8,154 library tags, ten positive-control tags that we spiked into the samples to calculate tag concentrations (see **Supplementary Fig. 2**), and 1,642 negative-control tags used to provide a measure of background hybridization and noise. The barcode signals from the synDNA mixtures indicated that 8,016 of the oligos were present (detected above background). Therefore, we successfully generated nearly complete (98%) up and down oligo libraries. Microarray analysis of the cell mixtures indicated successful generation of at least 7,829 unique alleles (96% of designed alleles; **Fig. 3a** and **Supplementary Table 2**). We found that the concentration of each unique allele depended on the concentration of synDNA used

© 2010 Nature America, Inc. All rights reserved.

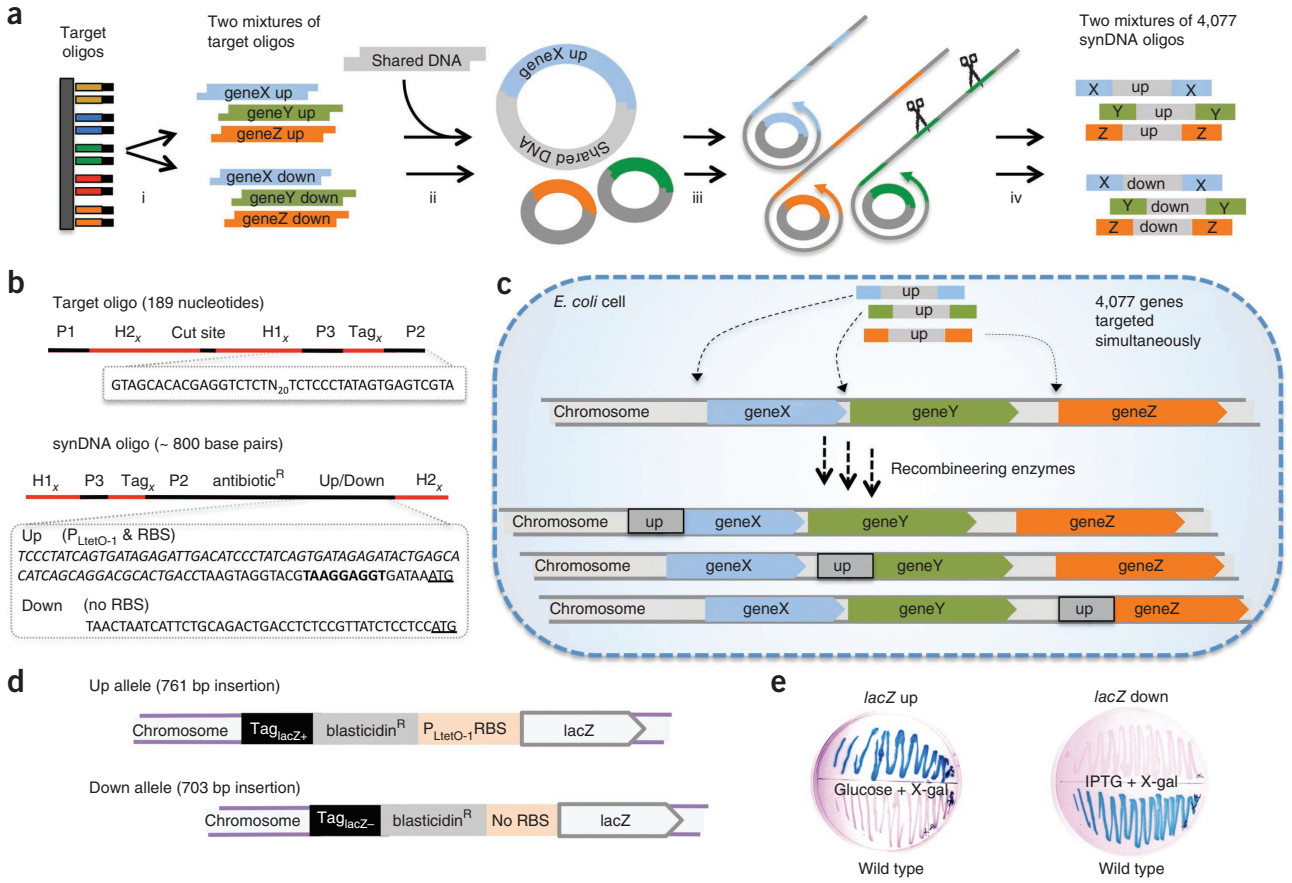

Figure 2 Multiplex strategy to rapidly generate cell mixtures with defined genetic modifications. (**a**) Construction of synDNA library. (i) 'Target' oligos that contain chromosome homology and barcodes are synthesized on a chip, cleaved from the chip, amplified by two rounds of PCR and modified with (ligation) sequences by uracil excision[48]. (ii) This pool of target oligos is ligated with oligos containing a selectable marker and promoter and RBS variants (Shared DNA), resulting in a pool of DNA circles. (iii) DNA circles are copied into a pool of linear concatemers by rolling-circle amplification[49]. (iv) Concatemers are cleaved at a repeating site linking the homology regions to provide a pool of synDNA ready for multiplex recombineering. (**b**) Schematic of target oligos and synDNA oligos for gene *x*. Red, unique regions; black, shared regions; P, PCR priming site; H, chromosome targeting region; Tag, barcode tag sequence; Up/Down, functional region. Sequence is shown for amplifying barcode tags and for functional regions (promoter sequence in italic, RBS in bold, start codon underlined). (**c**) Pool of synDNA oligos is inserted into electrocompetent *E. coli* cells. Recombineering enzymes catalyze the insertion of the synDNA oligos at thousands of unique loci in the genome. (**d**) Schematic of *lacZ* alleles used to test the method. Up allele is designed to increase gene transcription and translation. Down allele is designed to decrease translation. (**e**) *LacZ* up and down alleles yield the intended phenotypes. Up mutation of the *lacZ* gene causes cells to turn blue on the surface of agar containing glucose and X-gal. Down mutation of the *lacZ* gene causes cells to remain colorless on the surface of agar containing IPTG and X-gal.

© 2010 Nature America, Inc. All rights reserved.

to construct that allele (**Supplementary Fig. 3**). After normalization of the concentration of each allele for differences in synDNA concentrations used in recombineering, the s.d. for generating each mutant was ± 65% of the average, distributed uniformly around the genome (**Fig. 3b**). We also observed a modest dependence of recombineering frequency on the hybridization free energy[30] of the homology regions (**Supplementary Fig. 4**).

A small percentage of the alleles were not detected (4%), and in all these cases the preceding synDNA was either absent or found in low concentrations. In subsequent attempts to create allele libraries, most of these missing alleles were detected, suggesting that the alleles were initially not detected because of low concentrations of the synDNA oligos. These results indicate that the uniformity of cell mixtures in future multiplex recombineering experiments may easily be improved by supplementation with synDNA oligos that are initially present in low concentrations. Improvements in the uniformity of the initial mixture should enable the more efficient identification of cells with improved traits.

Notably, a single researcher was able to create these two genome-scale up and down allele libraries in a single day, demonstrating that multiplex recombineering is a rapid strategy for reprogramming thousands of genes.

Genome-scale mapping of alleles to selectable traits

To illustrate the potential of TRMR to rapidly generate and identify cells with new traits, we plated the cell mixtures on agar medium supplemented to create four different conditions (salicin, D-fucose, methylglyoxal and valine) in which wild-type *E. coli* typically do not grow. Colonies representing resistant mutants arose from our allele mixtures at frequencies >100-fold greater than from unmodified control cells that relied on spontaneous mutation to generate resistance (**Supplementary Table 3**). We characterized individual colonies (83 total) by sequencing the barcode tags. Additionally, we used TAG4 microarrays to characterize the populations obtained by scraping all colonies off of the surfaces of selection plates. Using microarray data, we ranked each allele in each condition according to fitness (fitness of

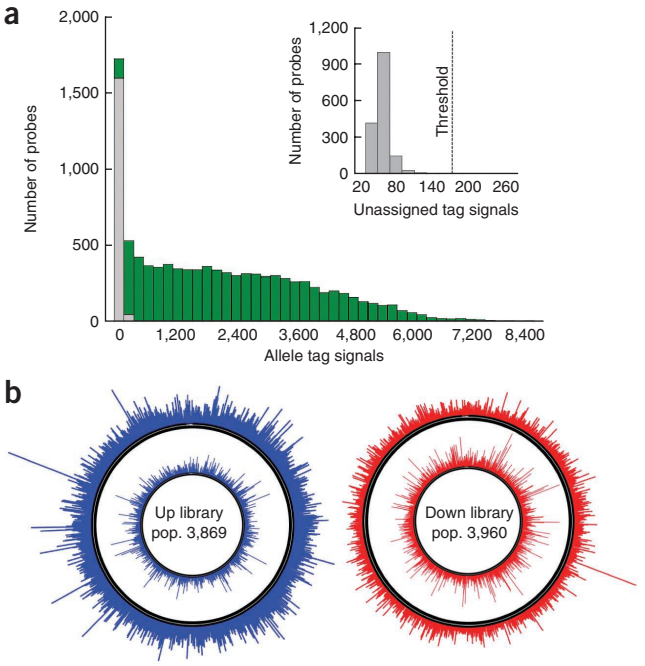

Figure 3 Analysis of synDNA and cell library. (**a**) Histogram showing the distribution of barcode signals of the up and down allele libraries detected by the TAG4 microarray. The unassigned tag signals (shown in gray) provide a measure of the background signal for each probe on the microarray. Probes that are assigned to unique alleles are shown in green. The unassigned tag signals have a low signal distribution (inset), and the threshold is shown for signals that are significantly above the background signal. The threshold for detection was such that the rate of false positives would be less than 2.2%. (**b**) TAG4 microarray results showing the distribution of synDNA oligos and alleles plotted by genomic location on the circular *E. coli* genome. Blue, up library; red, down library; inner circles, the concentration of each unique synDNA oligo before recombineering; outer circles, efficiency of generating each allele, calculated by dividing allele concentration by synDNA concentration.

allele $x = W'_x = F_{x,f} / F_{x,i}$, which is the ratio of the final allele frequency (F_x = concentration of x/total concentration) after growth to the initial allele frequency). The allele fitness determined by microarray agreed well with the results from picking and sequencing individual colonies (**Fig. 4** and **Supplementary Table 4**).

Constructing mutants with beneficial traits and identifying the genetic cause has traditionally been a slow and laborious process. Using TRMR, we were able to rapidly identify traits present in our cell mixtures that are consistent with previous studies and identify unexpected genetic modifications that could be used in future metabolic engineering. The allele(s) that conferred the highest frequency or fitness from these selections were reconstructed separately to confirm that improved growth is due to the insertion of the identified cassette. These alleles are summarized in **Figure 4** and described in detail below.

Salicin is a carbon source that *E. coli* normally cannot metabolize owing to repression of the enzymes BglF and BglB. We identified the *hns* down mutation, using both array results and sequencing, as having the greatest effect on fitness in medium supplemented with salicin. Mutations in the *hns* (histone-like nucleoid structuring protein) regulator[31] are known to confer improved growth on salicin. Its identification here confirms that the TRMR method can effectively uncover gene-trait relationships.

D-fucose is a nonmetabolizable analog of arabinose that inhibits the ability of *E. coli* to use arabinose as a carbon source by inhibiting induction of the L-arabinose operon. We identified the *xylA* up allele, which causes overexpression of *xylA* and *xylB*, as conferring the ability to grow in the presence of D-fucose. Notably, these results suggest that *E. coli* xylose isomerase (XylA) may have *in vivo* L-arabinose isomerase activity. This discovery is corroborated by the observation that overexpression of *E. coli xylAB* in *Pseudomonas putida* confers the ability to metabolize both xylose and L-arabinose[32]. Such a trait is of potential value for the efficient use of cellulosic biomass as a renewable feedstock.

Methylglyoxal is an important intracellular metabolite because it can be used as an intermediate for production of commodity chemicals and because, when metabolism is disrupted, it can accumulate,

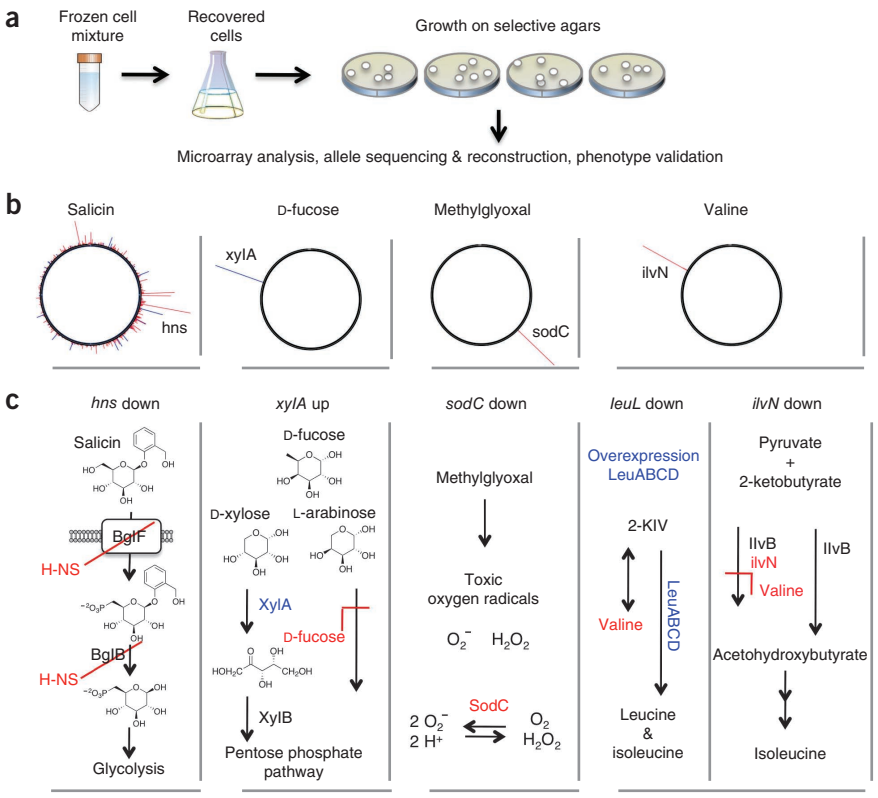

Figure 4 Trait-conferring genotypes identified in four selective environments. (**a**) Up and down alleles were recovered from frozen cultures and spread on agar medium in conditions where wild-type cells would not grow (indicated as column headings). (**b**) Fitness (W') calculated by microarray detection of barcode tags was plotted for each allele by genomic location. Blue, up allele; red, down allele. (**c**) Known or hypothesized mechanisms whereby the identified genomic modifications confer the ability to grow. High-fitness alleles were detected on microarrays, except for *leuL* down, which was identified by sequencing of barcode tags within colonies.

© 2010 Nature America, Inc. All rights reserved.

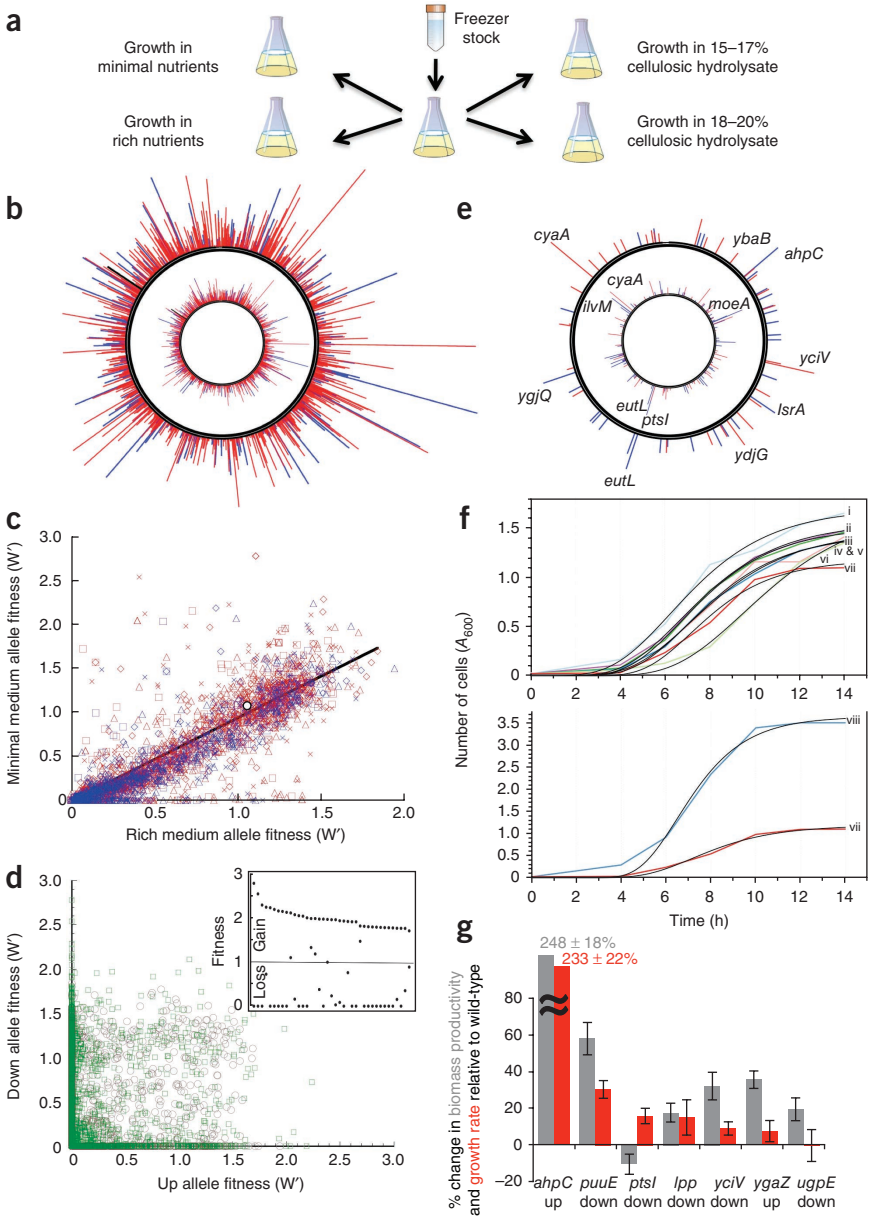

Figure 5 Alleles identified during pooled growth in media and cellulosic hydrolysate. (**a**) TRMR alleles were recovered from frozen cultures and allowed to grow in a rich medium, minimal medium or cellulosic hydrolysate. (**b**) Allele frequencies after growth in media plotted by genomic location. Inner circle, rich; outer circle, minimal; blue, up allele; red, down allele; black, control allele frequency × 10. (**c**) Allele fitness in minimal medium plotted against fitness of the same allele in rich medium. Shapes describe the affected gene function as determined by clusters of orthologous groups: ◊, information storage and processing; □, cellular processes; △, metabolism; ×, poorly characterized; blue, up allele; red, down allele; black, control allele. Fitness trend was fit to a line shown in black ($R^2 = 0.748$). (**d**) The fitness of down alleles compared with the corresponding up alleles. Brown ○, rich medium; green □, minimal medium. For alleles that cluster toward either the *x* or the *y* axis, the up allele and the down allele report opposite effects. Inset shows fitness benefits ($W > 1$) of top 40 alleles for growth in minimal medium, and the fitness effects (usually detrimental, $W < 1$) of the orthogonal alleles. (**e**) Fitness ($\ln W$) plotted by genomic location of alleles isolated after growth in hydrolysate. Inner circle, 15–17% hydrolysate; outer circle, 18–20% hydrolysate; blue, up allele; red, down allele. Some alleles conferring high fitness are labeled. (**f**) Growth curves of isolated variants in cellulosic hydrolysate. Each growth curve is the average of three replicates. Curves are fit with a Gompertz function[50] (black). Alleles are denoted with roman numerals, as follows: (i) *puuE* down (pale blue), (ii) *yciV* down (purple), (iii) *ygaZ* up (green), (iv) *lpp* down (pink), (v) *ugpE* down (blue), (vi) *ptsI* down (pale green), (vii) wild-type MG1655 (red), (viii) *ahpC* up (blue). Error bars are minimal and are not shown for clarity. A_{600}, absorbance at 600 nm. (**g**) Percent change in biomass productivity and maximum growth rate for isolated variants grown in hydrolysate relative to *E. coli* MG1655 grown in hydrolysate. Biomass productivity (gray bars) is the area under each growth curve. Growth rate (red bars) is the maximum growth rate as calculated from the Gompertz function. Values are the average of three replicates; error bars denote s.d.

© 2010 Nature America, Inc. All rights reserved.

resulting in oxidative damage and eventual cell death[33]. We used TRMR to discover a previously unknown phenotype: decreased expression of *sodC*, which produces a superoxide-mediating enzyme[34], confers resistance to exogenous methylglyoxal, possibly by affecting superoxide concentrations in the periplasm.

Excess valine causes feedback inhibition of leucine and isoleucine biosynthesis, leading to inhibition of cell growth as these amino acids become scarce. Microarray results identified *ilvN* down as the allele conferring the best growth, and this genomic region has been indicated in several previous studies[35,36]. Unexpectedly, sequencing showed that the *leuL* down allele also could grow well on valine plates. The *leuL* down mutation would cause increased expression of the leucine biosynthesis operon *leuABCD* by circumventing the alleged transcription attenuation caused by leuL[37]. Mutations of this operon have not previously been associated with valine resistance. However, a recent attempt to increase production of noncanonical amino acids in engineered *E. coli* cells demonstrated that overexpression of *leuABCD* shifts metabolite pools from valine toward isoleucine and leucine[38].

Genome-scale quantitative growth phenotypes

To further demonstrate that TRMR performs well at the genome scale, we combined the up and down allele libraries and measured fitness in liquid cultures that contained rich or minimal nutrients (**Fig. 5a**). The liquid cultures were allowed to grow for an average of eight generations, before and after which aliquots of cells were plated for analysis of individuals or frozen for microarray analysis. Additionally, an aliquot of control cells (barcoded and kanamycin resistant; **Supplementary Notes**) was spiked into the culture at the start of selections. A known concentration of these control cells was used to assess the ability of barcode technology to measure allele concentrations during pooled growth. The control cells also serve as a wild-type standard with which the fitness of alleles can be compared.

Using barcode microarrays, we simultaneously tracked all of the alleles, which were reduced to approximately 2,500 alleles after growth selections (**Fig. 5b**). The numbers of control cells in the populations determined by microarray was not substantially different from estimates of control-cell numbers obtained from counting

kanamycin-resistant colonies. Microarrays revealed that the majority of alleles had similar growth phenotypes in both rich and minimal media (**Fig. 5c**, *x-y* diagonal). Noteworthy alleles that do not fit this trend are those that allow growth in the rich medium but are no longer observed in the minimal medium (**Fig. 5c**, alleles along *x* axis). Consistent with previous observations[19], many of these alleles consist of changes in the expression of genes involved in metabolism. Also of interest are those alleles that confer faster growth than that of the control cells in the minimal medium (a list of fitness values can be found in **Supplementary Table 5**).

These experiments also offer the first genome-wide glimpse of generally orthogonal expression alleles grown competitively in the same culture. We anticipated that if a particular up allele shows a fitness benefit, then the down allele is likely to show a negative effect on fitness, possibly being lost from the culture, and vice versa. This is often the case (see **Fig. 5d**, allele clustering toward the axes), providing further evidence that our synthetic cassettes are generally causing the intended effects at genome-wide loci. Exceptions such as improved growth resulting from both up and down expression alleles in the same environment may be due to secondary effects (such as increased transcription of multiple downstream genes) and require further investigation.

Mapping tolerance to lignocellulosic hydrolysate

We next applied TRMR to identify genes that improve tolerance to lignocellulosic hydrolysate derived from corn stover (provided by the US National Renewable Energy Laboratory). This class of feedstocks contains a variable array of growth inhibitors (known inhibitors include organic acids, aldehydes and phenolic-based compounds)[39,40]. To take hydrolysate variability into account, we measured growth of variants bearing our alleles in several mixtures of hydrolysate and minimal medium.

Microarray analysis of the alleles indicated that only a small subset of the population remained after each selection (**Fig. 5e**; see **Supplementary Table 6** for fitness values and gene ontology analysis). Many of the modifications that improved growth in lower concentrations of corn stover hydrolysate affected genes known to be involved in primary metabolism (*pgi* up, *eno* up and *tdcG* up), RNA metabolism (*rlmG* down, *rimM* up, *rsmE* down and *rrmA* down) and transport of sugars (*ptsI* down, *ptsI* up and directly downstream *crr*). Growth in higher concentrations of hydrolysate selected alleles related to secondary metabolism (*ispF* up and *dxs* up), vitamin metabolic processes (*nadD* up, *menD* up, *apbE* up, *pabC* up, *dxs* up and *ribB* up) and antioxidant activity (*ahpC* up, *tpx* up and *bcp* up). The down mutation of the adenylate cyclase gene (*cyaA*) conferred a growth advantage in every selection.

To confirm that the mutations conferred fitness advantages, we isolated seven alleles after the selections and characterized growth in hydrolysate relative to unmutated *E. coli*. (**Fig. 5f,g**). All seven alleles (*ahpC* up, *ugpE* down, *puuE* down, *ptsI* down, *ygaZ* up, *yciV* down and *lpp* down) yielded improvement in either growth rate or biomass productivity relative to the wild-type strain. Notably, the up allele of *ahpC* resulted in a large improvement. The *ahpC* gene and its downstream counterpart *ahpF* have not previously been identified as important for growth in hydrolysate. However, they have been implicated in resistance to organic solvents[41] and various oxidants[42,43], possibly indicating that during growth in cellulosic hydrolysate, reactive oxygen species in the form of peroxides and other oxidants are present or forming as a result of imbalances in metabolism[44]. In addition to identifying several important targets for future genome-engineering endeavors, many of which would have been difficult to predict *a priori*, these profiling studies shed light on general mechanisms of hydrolysate toxicity (such as the presence of oxidants) and growth advantage in hydrolysate (such as metabolism of preferred carbon sources).

DISCUSSION

We have described a new method for the genome-scale mapping of genes to traits and have shown that this method can increase the throughput of genetic studies by several orders of magnitude. Although some of the trait-conferring modifications we identified correspond to previously identified genomic regions, the majority would have been difficult to predict. Such unanticipated outcomes provide insight into many uncharacterized genes and, in some cases, into known genes with uncharacterized functions. We have already begun applying this method toward understanding a range of traits of importance in biotechnology, including improved growth in industrially relevant conditions and enhanced product formation.

We have designed TRMR to be easy to use and versatile. The molecular cloning procedures were accomplished within a week by a single researcher, with two additional days providing enough cells for 60 genome-wide selection and screening studies. Notably, data acquisition and analysis from TRMR is similar to genomics methods currently used by the yeast community and is amenable to a range of freely and commercially available software packages. The primary challenge to the broad dissemination of this method is the acquisition of oligonucleotide libraries, which will be overcome as DNA synthesis technologies continue to improve.

We envision that a broad range of additional studies could be performed using the basic TRMR platform described here by changing the targeting, functional or tracking design. For example, although the functional regions we used were promoters and translation sites, one might conceivably use sites associated with additional functions such as switches, oscillators or sensors[45]. Moreover, the TRMR approach is not limited to engineering or examining the *E. coli* genome. The design could be adapted for rapidly engineering yeast and a range of Gram-negative bacteria[23], provided the host has sufficient transformation and recombination capabilities. Additionally, TRMR may be carried out recursively, allowing for the accumulation of multiple beneficial mutations within a genome. Researchers could produce second- and third-generation recombinant cells by removing the antibiotic cassette between rounds of recombineering to allow isolation of cells containing an additional mutation, by using different antibiotic cassettes in the modular construction of the synDNA oligos so that different antibiotics could be used to isolate recombinants after each round of TRMR, or by eliminating altogether the need to isolate recombinants by relying on the increased efficiency of recombineering strategies such as those used in MAGE[24]. Integration of TRMR into directed-evolution programs would provide genome-scale construction and tracking of combinations of mutations, which would improve both the understanding and engineering of complex traits.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank D. Court (Center for Cancer Research, National Cancer Institute at Frederick, Maryland) for sharing plasmid pSIM5, C. Nislow and G. Giaever (University of Toronto, Ontario) for help with microarray analysis, A. Mohagheghi and M. Zhang (US National Renewable Energy Laboratories) for hydrolysate samples, M. O'Donnell for help in preparation of selective agar plates, Agilent for

© 2010 Nature America, Inc. All rights reserved.

access to the Oligonucleotide Library Synthesis product, and H. Marshall and the University of Colorado Microarray Facility for molecular barcode genotyping. The authors appreciate financial support provided by Shell, the Colorado Center for Biorefining and Biofuels (http://www.C2B2web.org) and the Colorado Energy Initiative (http://rasei.colorado.edu).

AUTHOR CONTRIBUTIONS

J.R.W. and R.T.G. conceived the study; J.R.W. designed and performed all experiments except for growth selections and allele confirmations in hydrolysate, which were conducted by P.J.R.; A.K.-F. aided J.R.W. in selection of targeting sequences and selection of barcode tags; A.K.-F. and P.J.R. assigned gene ontology terms; L.B.A.W. aided J.R.W. in selection design and microarray analysis; L.B.A.W. constructed circle plots; P.J.R., A.K.-F. and L.B.A.W. helped in manuscript preparation; J.R.W. and R.T.G. wrote the manuscript; R.T.G. supervised all aspects of the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Fox, R.J. *et al.* Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **25**, 338–344 (2007).
2. Turner, N.J. Directed evolution drives the next generation of biocatalysts. *Nat. Chem. Biol.* **5**, 567–573 (2009).
3. Winzeler, E.A. *et al.* Functional characterization of the *S. cervisiase* genome by gene geletion and parallel analysis. *Science* **285**, 901–906 (1999).
4. Fodor, S. *et al.* Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**, 767–773 (1991).
5. Blanchard, A.P., Kaiser, R.J. & Hood, L.E. High-density oligonucleotide arrays. *Biosens. Bioelectron.* **11**, 687–690 (1996).
6. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat. Biotechnol.* **17**, 974–978 (1999).
7. Cleary, M.A. *et al.* Production of complex nucleic acid libraries using highly parallel *in situ* oligonucleotide synthesis. *Nat. Methods* **1**, 241–248 (2004).
8. Ghindilis, A. *et al.* CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. *Biosens. Bioelectron.* **22**, 1853–1860 (2007).
9. Yu, D. *et al.* An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **97**, 5978–5983 (2000).
10. Murphy, K. Use of bacteriophage lambda recombination functions to promote gene replacement in *Escherichia coli*. *J. Bacteriol.* **180**, 2063–2071 (1998).
11. Zhang, Y., Buchholz, F., Muyrers, J. & Stewart, A.F. A new logic for DNA engineering using recombination in *Escherichia coli*. *Nat. Genet.* **20**, 123–128 (1998).
12. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M. & Davis, R.W. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* **14**, 450–456 (1996).
13. Cho, R.J. *et al.* Parallel analysis of genetic selections using whole genome oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **95**, 3752–3757 (1998).
14. Gill, R.T. *et al.* Genome wide screening for trait conferring genes using DNA microarrays. *Proc. Natl. Acad. Sci. USA* **99**, 7033–7038 (2002).
15. Lynch, M.D., Warnecke, T. & Gill, R.T. SCALEs: multiscale analysis of library enrichment. *Nat. Methods* **4**, 87–93 (2007).
16. Badarinarayana, V. *et al.* Selection analyses of insertional mutants using subgenic resolution arrays. *Nat. Biotechnol.* **19**, 1060–1065 (2001).
17. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
18. Ho, C.H. *et al.* A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat. Biotechnol.* **27**, 369–377 (2009).
19. Baba, T. *et al.* Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, 1–11 (2006).
20. Kitagawa, M. *et al.* Complete set of ORF clones of *Escherichia coli* ASKA library (a complete set of *E. coli* K-12 ORF archive): unique resources for biological research. *DNA Res.* **12**, 291–299 (2006).
21. Datsenko, K. & Wanner, B. One-step inactivation of chromosomal genes in *E. coli* K12 using PCR products. *Proc. Natl. Acad. Sci. USA* **97**, 6640–6645 (2000).
22. Ellis, H.M., Yu, D., DiTizio, T. & Court, D.L. High efficiency mutagenesis, repair, and engineering of chromosomal DNA using single-stranded oligonucleotides. *Proc. Natl. Acad. Sci. USA* **98**, 6742–6746 (2001).
23. Datta, S., Costantino, N., Zhou, X. & Court, D.L. Identification and analysis of recombineering functions from Gram-negative and Gram-positive bacteria and their phages. *Proc. Natl. Acad. Sci. USA* **105**, 1626–1631 (2008).
24. Wang, H. *et al.* Programming cells by multiplex genome engineering and accelerate evolution. *Nature* **460**, 894–898 (2009).
25. Lutz, R. & Bujard, H. Independent and tight regulation of transcriptional units in *Escherichis coli* via the LacR/O, the TetR/O and AraC/I1–I2 regulatory elements. *Nucleic Acids Res.* **25**, 1203–1210 (1997).
26. Shine, J. & Dalgarno, L. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl. Acad. Sci. USA* **71**, 1342–1346 (1974).
27. Kimura, M., Takatsuki, A., Yamaguchi, I. & Blasticidin, S. Deaminase gene from *Aspergillus terreus* (BSD): a new drug resistance gene for transfection of mammalian cells. *Biochim. Biophys. Acta* **1219**, 653–659 (1994).
28. Pierce, S.E. *et al.* A unique and universal molecular barcode array. *Nat. Methods* **3**, 601–603 (2006).
29. Baudin, A., Ozier-Kalogeropoulos, O., Denouel, A., Lacroute, F. & Culin, C. A simple and efficient method for direct gene deletion in *Saccharomyces cerevisiae*. *Nucleic Acids Res.* **21**, 3329–3330 (1993).
30. Markham, N.R. & Zuker, M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res.* **33**, W577–W581 (2005).
31. Defez, R. & de Felice, M. Cryptic operon for beta-glucoside metabolism in *Escherichia coli* K12: genetic evidence for a regulatory protein. *Genetics* **97**, 11–25 (1981).
32. Meijnen, J.P., de Winde, J.H. & Ruijssenaars, H.J. Engineering *Pseudomonas putida* S12 for efficient utilization of D-xylose and L-arabinose. *Appl. Environ. Microbiol.* **74**, 5031–5037 (2008).
33. Zhu, M.M., Skraly, F.A. & Cameron, D.C. Accumulation of methylglyoxal in anaerobically grown *Escherichia coli* and its detoxification by expression of the *Pseudomonas putida* glyoxalase i gene. *Metab. Eng.* **3**, 218–225 (2001).
34. Gort, A.S., Ferber, D.M. & Imlay, J.A. The regulation and role of the periplasmic copper, zinc superoxide dismutase of *Escherichia coli*. *Mol. Microbiol.* **32**, 179–191 (1999).
35. Sutton, A., Newman, T., Francis, M. & Freundlich, M. Valine-resistant *Escherichia coli* K-12 strains with mutations in the ilvB operon. *J. Bacteriol.* **148**, 998–1001 (1981).
36. Weinstock, O., Sella, C., Chipman, D.M. & Barak, Z. Properties of subcloned subunits of bacterial acetohydroxy acid synthases. *J. Bacteriol.* **174**, 5560–5566 (1992).
37. Wessler, S.R. & Calvo, J.M. Control of leu operon expression in *Escherichia coli* by a transcription attenuation mechanism. *J. Mol. Biol.* **149**, 579–597 (1981).
38. Sycheva, E.V. *et al.* Overproduction of noncanonical amino acids by *Escherichia coli* cells. *Microbiology* **76**, 712–718 (2007).
39. Chen, S.F., Mowery, R.A., Castleberry, V.A., van Walsum, G.P. & Chambliss, C.K. High-performance liquid chromatography method for simultaneous determination of aliphatic acid, aromatic acid and neutral degradation products in biomass pretreatment hydrolysates. *J. Chromatogr. A* **1104**, 54–61 (2006).
40. Mohagheghi, A. & Schell, D.J. Impact of recycling stillage on conversion of dilute sulfuric acid pretreated corn stover to ethanol. *Biotechnol. Bioeng.* **105**, 992–996 (2010).
41. Ferrante, A.A., Augliera, J., Lewis, K. & Klibanov, A.M. Cloning of an organic solvent-resistance gene in *Escherichia coli*: the unexpected role of alkylhydroperoxide reductase. *Proc. Natl. Acad. Sci. USA* **92**, 7617–7621 (1995).
42. Poole, L.B. Bacterial defenses against oxidants: mechanistic features of cysteine-based peroxidases and their flavoprotein reductases. *Arch. Biochem. Biophys.* **433**, 240–254 (2005).
43. Seaver, L.C. & Imlay, J.A. Alkyl hydroperoxide reductase is the primary scavenger of endogenous hydrogen peroxide in *Escherichia coli*. *J. Bacteriol.* **183**, 7173–7181 (2001).
44. Kohanski, M.A., Dwyer, D.J., Hayete, B., Lawrence, C.A. & Collins, J.J. A common mechanism of cellular death induced by bactericidal antibiotics. *Cell* **130**, 797–810 (2007).
45. Lu, T.K., Khalil, A.S. & Collins, J.J. Next-generation synthetic gene networks. *Nat. Biotechnol.* **27**, 1139–1150 (2009).
46. Datta, S., Constantino, N. & Court, D.L. A set of recombineering plasmids for gram-negative bacteria. *Gene* **379**, 109–115 (2006).
47. Pierce, S.E., Davis, R.W., Nislow, C. & Giaever, G. Genome-wide analysis of barcoded *Saccharomyces cerevisiae* gene-deletion mutants in pooled cultures. *Nat. Protoc.* **2**, 2958–2974 (2007).
48. Nour-Eldin, H.H., Hansen, B.G., Norholm, M.H.H., Jensen, J.K. & Halkier, B.A. Advancing uracil-excision based cloning towards an ideal technique for cloning PCR fragments. *Nucleic Acids Res.* **34**, e122 (2006).
49. Dean, F.B., Nelson, J.R., Giesler, T.L. & Lasken, R.S. Rapid amplification of plasmid and phage DNA using Phi29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095–1099 (2001).
50. Perni, S., Andrew, P.W. & Shama, G. Estimating the maximum growth rate from microbial growth curves: definition is everything. *Food Microbiol.* **22**, 491–495 (2005).

© 2010 Nature America, Inc. All rights reserved.

ONLINE METHODS

Strains, DNA and reagents. *Escherichia coli* MG1655 (wild type) was obtained from ATCC 700926. Genomic sequences were obtained from GenBank U00096.2, and gene annotation was from the Ecogene database version 2.20 (http://www.ecogene.org/). Pseudogenes and insertion elements were excluded from the protein-coding genes that were targeted. The kanamycin-resistant control strain (also called JWKAN) was constructed from *E. coli* ATCC 700926, with nucleotide 3,909,796 replaced with a barcoded kanamycin cassette[21] (**Supplementary Notes**). Up and down DNA cassettes were constructed using PCR and cloned into the pEM7/BSD plasmid (Invitrogen, **Supplementary Notes**). Oligonucleotide libraries were purchased from Agilent; all other oligonucleotides were purchased from Integrated DNA Technologies with standard desalting except where noted. The pSIM5 plasmid[46] was a gift from D. Court. All reagents were obtained from common commercial sources. All enzymes were from New England Biolabs except where noted. All sequencing was performed by Macrogen USA or Eurofins MGW Operon. Recipes and additional information can be found in **Supplementary Notes**.

Preparation of synthetic DNA and recombineering. A portion of the oligonucleotide library provided by Agilent (8,154 unique 189-mers) was amplified by two rounds of PCR. Products were treated with the USER enzymes (New England Biolabs), purified and ligated to the up cassette. Rolling-circle amplification, nuclease treatment and purification resulted in 8–10 μg synDNA. This procedure was also carried out in parallel to separately generate TRMR down synDNA. More details are available in **Supplementary Notes**.

E. coli cells containing the recombineering plasmid pSIM5 were grown in 800 ml SOB cultures at 30 °C and made recombineering proficient with minor modifications to reported methods[46]. Briefly, when cells reached an optical density at 600 nm of 0.7, flasks were transferred to water baths at 42 °C to induce the λRed enzymes for 15 min. Flasks were then transferred to an ice-water bath and cells were kept close to 4 °C for the remaining steps. Cells were collected by centrifugation and suspended with cold deionized water. Cell collection and washing was repeated once more, then cells were suspended to a final volume of 6.4 ml in water. Aliquots of cells (400 μl) were transformed in a 0.2-cm electrocuvette with approximately 1 μg of up or down synDNA and a pulse of 12.5 kV cm^{-1}. Transformation was carried out eight times to generate the up allele library and eight times to generate the down allele library. The cells from each transformation were recovered in 12 ml SOC medium for 1 h at 37 °C. Cells were collected by centrifugation and resuspended in 30 ml MA salts (**Supplementary Notes**). Centrifugation and resuspension was repeated twice more, with the final resuspension to a volume of 2 ml in MA salts. The up and down allele libraries were separately spread onto a total of 40 low-salt LB agar plates containing blasticidin-S (90 μg ml^{-1}) and allowed to grow at 37 °C for 22 h. Colonies were scraped from the agar plates and up and down allele libraries were each suspended in a total of 35 ml LB. Cells were collected by centrifugation and suspended to 3×10^9 cells per milliliter in LB medium containing 16% (vol/vol) glycerol and blasticidin-S (90 μg ml^{-1}). Aliquots of the up or down cell mixtures were stored at −80 °C.

Screens and selections. Freezer stocks were used to inoculate 50 ml low-salt LB medium containing 80 μg ml^{-1} blasticidin-S with 5×10^8 TRMR up cells and 5×10^8 TRMR down cells. This culture was allowed to grow with shaking at 37 °C to an optical density at 600 nm of 0.8. The cells were centrifuged at 4,500*g* for 6 min, decanted and suspended in 30 ml of MA salts. The cells were collected once more by centrifugation and suspended in MA salts to a concentration of 5×10^8 cells per milliliter. The JWKAN cells were added to a final concentration of 7.7×10^4 cells per milliliter. A 1.7 ml aliquot of the cell library (called the recovery culture) was frozen for microarray analysis, and the remainder was used for various growth selections.

Liquid selections were carried out with shaking at 37 °C in 600 ml of MOPS minimal medium containing 2 mM phosphate and 4% (wt/vol) glucose or in 600 ml LB medium. Each medium was inoculated with 2.4×10^8 cells from a recovery culture and allowed to grow to an optical density at 600 nm of 1.0–1.2. Cells were collected from each culture by centrifugation of 10-ml aliquots at 4,500*g* for 6 min, decanted and stored at −80 °C for microarray analysis. Growth results are the average of three array hybridizations.

Hydrolysate growth selections were carried out in various dilutions of hydrolysate in minimal media (15%, 16%, 17%, 18%, 19% and 20%). During selections, cell samples were taken for microarray analysis of populations, and cells were plated to isolate and identify individual alleles growing as colonies. Unique alleles from selections were identified and confirmed by PCR and studied for growth characteristics in hydrolysate. All growth curves were done in complete triplicate. More details are available in **Supplementary Notes**.

Growth on various selective agars was carried out by spreading a total of 0.7×10^8 cells of the allele mixtures recovered from freezer aliquots on five plates for each selective condition (salicin, D-fucose plus L-arabinose, valine, and methylglyoxal; plate recipes in **Supplementary Notes**). Plates were incubated at 37 °C until colonies were visible (1–3 d). Selection for *galK* down alleles was carried out on plates containing 2-deoxygalactose[9], and screens were carried out on MacConkey agar containing 1% (wt/vol) D-galactose. Screens of *lacZ* up alleles were carried out on LB agar plates containing 0.2% (wt/vol) glucose and 40 μg ml^{-1} X-gal. Screens of *lacZ* down alleles were carried out on LB agar plates containing 0.05% (wt/vol) IPTG and 40 μg ml^{-1} X-gal. Selections for control cells were carried out on LB agar plates containing kanamycin (30 μg ml^{-1}).

Microarray tracking. Genomic DNA was extracted from ~10^9 *E. coli* cells using Purelink Genomic Mini kit (Invitrogen). Barcode tags are amplified in 300 μl PCR reactions (final concentrations: 1× PCR buffer, 2.5 mM $MgCl_2$, 0.2 mM each dNTP, 1 μM each primer 5′-GTAGCACACGAGGTCTCT-3′ and Biotin-5′-TACGACTCACTATAGGGAGA-3′, 0.6 U μl^{-1} Taq polymerase and 0.5 μg genomic DNA or 30 pg synDNA). Reactions were cycled 25 times with an annealing temperature of 55 °C. Barcode tags were purified by agarose gel electrophoresis and extraction using the QIAquick gel extraction protocols (Qiagen, substitute buffer QX1 for QG). Tag purification was shown to reduce background hybridization. Microarray hybridizations to the Geneflex Tag4 16K V2 array (Affymetrix) were carried out according to published procedures[47] with the following modifications: 600 ng of purified tags (combined up tags and down tags) were hybridized along with ten tags (amplified and purified as above) included at known concentrations (0.5 pM to 10 nM). Intensity values are calculated for each tag after removal of replicate outliers and averaging of unmasked replicates using software (raw_file_maker.pl) that can be downloaded from http://chemogenomics.stanford.edu/supplements/04tag/download.html. Background hybridization was calculated from the average intensity of 1,642 unused tag probes; threshold intensity was set to background hybridization plus 2 s.d. The intensities of the ten spiked tags were used to calculate allele concentrations from array signals and correct for array saturation (**Supplementary Fig. 2**). Barcode frequencies were calculated by dividing barcode concentrations by the total concentration of all barcodes detected on the array.

© 2010 Nature America, Inc. All rights reserved.