Online Methods

Strains and media

All strains used in this study (**Supplementary Table 4**) were derived from *Escherichia coli* MG1655[13]. LB contained 0.1% Bacto Tryptone, 0.05% yeast extract, and 0.05% NaCl. Asparagine media contained M9 salts[14] supplemented with 2 g/L L-asparagine (Sigma), 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 10 μM thiamine, and micronutrients[15]. Sodium chloride and $FeSO_4$ were omitted from M9 salts and micronutrients, respectively. Glucose media was the same except that glucose (2 g/L) replaced asparagine. Media were supplemented with kanamycin (25 μg/mL) or chloramphenicol (20 or 25 μg/mL) as needed.

Construction of the Cml^R strain

In order to insert a GFP and Cml^R cassette simultaneously into the *lacZ* locus of strain MG1655, we first amplified a GFP reporter gene from pCMW5[16] and a Cml^R cassette from pKD3[17]. Then, we used a crossover PCR to link these two products and place them into the genome using the method of Datsenko and Wanner[18]. The primers used for the construction of this strain are provided (**Supplementary Table 5**). The GFP gene was not specifically used in this work.

Experimental evolution of strain ASN*

To start the experimental evolution, $\sim 1\times10^9$ washed, LB-grown, mid-exponential phase MG1655 Δ*lacZ*[5] cells were added to 50 ml asparagine media. Using serial transfers that kept the population size above $\sim 1\times10^7$, the culture was maintained for 39 days in early to mid exponential phase. During that time, the bulk population went through 90 generations. We shook the culture at 250 rpm at 37 °C.

Construction of strains to analyze the ASN* mutations

Using the method of Datsenko and Wanner[18], we placed antibiotic markers (kanamycin or chloramphenicol) next to each mutation location in both the ASN* and parental strains. Each marker replaced about 20 bases. For *sstT* and *lrp*, we placed the markers upstream of the genes with the promoter of the antibiotic resistance cassette pointing in the direction opposite of the genes in order to minimize polar effects. For *ansA*, we placed the marker downstream of the *ansA-pncA* operon.

To assemble the desired allele combinations, we first transduced the kanamycin-marked *ansA* alleles into the parental strain. Then, we removed the kanamycin markers using a FLP recombinase system[18]. Next, we transduced the *sstT* alleles using chloramphenicol markers. And finally, we transduced the *lrp* alleles using kanamycin markers. In addition to the desired alleles, the final strains all had kanamycin and chloramphenicol markers and a scar from the original *ansA* kanamycin marker. For comparison, the same markers and scar were put into the ASN* mutant. Sequencing confirmed that all strains had the desired alleles. We used P1*vir* phage for all transductions[4].

Sequences of primers used in strain construction and testing are in Supplementary Table 5.

P1*vir* lysate preparation

We prepared P1*vir* lysate as described previously[4]. In brief, we diluted (1:100) an overnight culture of the Tn5 kanamycin resistant library[5], which is in the parental background, into 250 ml of LB with 5 mM $CaCl_2$ and 0.2% glucose. After growing the culture with aeration at 37°C for 30 minutes, we added 2.5 ml of P1*vir* phage lysate (from MG1655) to the culture. We then continued incubation at 37˚C with aeration until the culture cleared. Next, we centrifuged the remains of the culture at 5525 g for 10 minutes to pellet the cell debris. In the end, we filtered the lysate through a 0.2 μm filter and stored it at 4°C.

Construction of a secondary library in the "evolved" background

We used a modified version of a previously published P1*vir* transduction protocol[4]. We pelleted cells from 25 ml of overnight, stationary phase culture of the evolved strain by centrifugation (5525 ×g, 15 min) and re-suspended them in 10 mL of LB with 5 mM $CaCl_2$ and 10 mM Mg SO_4. Then, in each of 24 microfuge tubes, we mixed 400 μl cells with 200 μl phage lysate from the parental strain library. We incubated the mixtures at 30ºC for 30 minutes without shaking. Then, we combined the reactions into two batches (12 reactions each) and added 12 ml of LB plus 10 mM sodium citrate to each batch. Then, we incubated the mixtures at 37ºC for 30 min without shaking and then pelleted the cells by centrifugation (15 min, 5525 ×g). We combined the pellets and resuspended them in 4 ml 1 M sodium citrate. To estimate the yield, we plated 1 μl of culture on an LB kanamycin plate. We then added the remaining culture to 250 ml LB plus kanamycin and shook it at 37 ºC for 10 h (until the culture reached mid-stationary phase). Finally, we pelleted the cells by centrifugation (15 min, 5525 ×g), resuspended them in 15-20 mL LB with 15% glycerol, and snap froze them with dry ice and ethanol.

Growth of secondary library under selective and non-selective conditions

In each experiment, we grew portions of the secondary P1*vir*-transduced transposon library in the presence and absence of selection. Selective and non-selective growth spanned the same number of generations.

Finding the distributions of markers across the genome (genetic footprinting)

We subjected samples of ~10^7 cells from both the population grown in selective conditions and the population grown in nonselective conditions to hybridization-based genetic footprinting to amplify the DNA adjacent to the transposons[5]. Samples from the selective and non-selective conditions were differentially labeled and hybridized to *E. coli* ORF arrays[5]. A gene's signal in each array channel

represented the frequency of mutants from the corresponding growth conditions that had transposon insertions in or near the gene.

We converted the hybridization signals from the selective growth and non-selective growth samples to depletion scores:

$$\text{Score}(g) = \frac{\text{hybridization signal of ' g' from nonselective growth}}{\text{hybridization signal of ' g' from selective gowth}},$$

where 'g' is an arbitrary gene.

Thus, loci that experienced more depletion from the selected population had higher scores. Depletion scores for all experiments in this work as well as all computational tools are available online at http://tavazoielab.princeton.edu/ADAM/ (also **Supplementary Software 1**).

Mutual-information based identification of adaptive loci

As ADAM spreads the signal from each adaptive mutation over multiple adjacent genes, neighborhoods of high depletion scores correspond to adaptive mutations. Direct examination of the depletion scores as a function of genome location (**Supplementary Fig. 3a**) typically indicated the regions in which functional mutations resided. Smoothing the data by taking a simple moving average, which emphasized regions of high depletion scores, typically allowed us to identify all of the true positives in a data set (**Supplementary Fig. 3b**). While easy and surprisingly effective, such techniques do not constitute a systematic approach for identifying the relevant genomic regions and suffer from a higher false positive rate than the computational method described below.

The core problem is the need to distinguish between the fitness effects of transposon disruptions and the linkage-based effects of adaptive mutations. The key difference between these two phenomena

lies not in the intensity of the scores but rather the number of consecutive genes that show high depletion scores. For example, in our Cml^R experiment, *lacI* had a depletion score comparable to that of *rfaQ* (2.23 vs 2.20); however, we identified *lacZ* as the site of mutation because in addition to *lacI*, a whole stretch of genes from *prpR* to *yaiP* showed depletion scores greater than 1.2 (see **Fig. 2a** in the main text). To capture these regions, we quantized the vector containing the depletion scores for all the genes into 4 bins: (i) the top 1% genes, (ii) the top 2-5% genes, (iii) the top 6-10% and (iv) the rest of the genes.

Then, we tiled the genome with spatial vectors of length 25 (see **Supplementary Fig. 4**). A spatial vector is a binary vector of length N (i.e., the total number of genes) in which 25 consecutive genes are set to '1' and all the rest are '0'. Each spatial profile overlaps with 24 of the genes in its neighboring vectors. The spatial profiles tile the whole genome.

Finally, we asked the question: which spatial profiles contain genes with higher depletion scores than expected by chance To answer this question, we used the notion of mutual information[19,20] to measure how informative a given spatial profile was about the depletion score categories:

$$MI(\text{spatial profile};\text{depletion score categories}) = \sum_{i=1}^{2}\sum_{j=1}^{4} P(i,j)\log\frac{P(i,j)}{P(i)P(j)}$$

where $P(i,j)$ is the fraction of genes whose spatial profile values are in the i^{th} state and whose depletion scores are in the j^{th} category, $P(i) = \sum_j P(i,j)$, and $P(j) = \sum_i P(i,j)$[20]. We tested the statistical significance of each spatial profile by comparing its *MI* (mutual information) value to those from 10,000 random shuffles of the depletion scores. We accepted as significant those spatial profiles whose *MI* values were higher than all of the randomly generated values.

Because the spatial profiles largely overlapped (**Supplementary Fig. 4**), we retained only the most informative profile from each region. To accomplish this, we considered the candidate spatial profiles in order of decreasing *MI* and used conditional information to remove profiles that did not satisfy the following with respect to each of the previously accepted spatial profiles:

$$\frac{MI(\text{spatial profile}; \text{depletion scores} \mid \text{an accepted spatial profile})}{MI(\text{spatial profile}; \text{an accepted spatial profile})} > 5.0$$

This equation compares the additional information provided by a new spatial profile, given an already accepted spatial profile, to the mutual information between the two spatial profiles and requires the ratio to be more than a certain threshold (5 in this case). Comparing the spatial profile being tested against each previously accepted spatial profile determines whether the candidate profile adds significant and independent information. In other words, we ensured that a spatial profile was both informative of the depletion score categories and also had little dependency with the previously accepted profiles[20]. The mutation sites had a high likelihood of residing close to the center of these significant regions near the maximal depletion scores. The tools for performing these analyses are available online at http://tavazoielab.princeton.edu/ADAM/.

Data presentation: Smoothing and filtering

Growth of the library under selective conditions caused some genomic regions to become effectively depleted of markers resulting in very low signals. Due to this lower bound on the hybridization signal, the depletion scores were sensitive to the original frequency of insertion events. The frequency of insertion events was more or less uniform across the genome; however, certain regions were "hotspots" or "cold spots" (**Supplementary Fig. 5**). For example, assume that growth of the library under selective conditions eliminates all markers near two genes and the array signal from the selected channel for both is the background value of say, 0.1. Further assume that the unselected

conditions did not alter the initial insertion frequency for the genes. If that initial insertion frequency for both genes was similar and gave a signal of 1, then the depletion score for both would be 10. If, however, one gene were in an insertion "hot-spot" with a signal of 10, then the depletion score would be 100. While the quantization method used to determine functional mutation locations is robust against such noise, the effects of the initial transposon insertion frequency distorted the plots of the depletion signal. In order to emphasize the effects of the selective conditions and deemphasize the effects of the initial transposon insertion frequency, when plotting the results, we filtered out the ~800 genes whose variance normalized hybridization signal (mean divided by standard deviation) in the unselected transposon library[5] was more than one standard deviation away from the genome-wide average.

After filtering, we smoothed each gene's score by taking a Gaussian-weighted average across the 15 neighboring genes on either side (**Supplementary Fig. 6**). This resulted in a smooth, bell-shaped signal around the site of each mutation, which was ideal both for presentation and for choosing candidate genes to search for the precise mutations. Note that these data manipulations did not affect the identification phase.

References

13. F. R. Blattner, G. Plunkett, 3rd, C. A. Bloch et al., *Science* **277** (5331), 1453 (1997).

14. F. M. Ausubel, R. Brent, R. E. Kingston et al., *Current protocols in molecular biology*. (Wiley Interscience, New York, NY, 1994).

15. F. C. Neidhardt, P. L. Bloch, and D. F. Smith, *J Bacteriol* **119** (3), 736 (1974).

16. S. Amini, H. Goodarzi, and S. Tavazoie, *PLoS Pathog* **5** (5), e1000432 (2009).

17. T. Baba, T. Ara, M. Hasegawa et al., *Mol Syst Biol* **2**, 2006 0008 (2006).

18. K. A. Datsenko and B. L. Wanner, *Proc Natl Acad Sci U S A* **97** (12), 6640 (2000).

19. T. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, Hoboken, NJ, 2006).

20. O. Elemento, N. Slonim, and S. Tavazoie, *Mol Cell* **28** (2), 337 (2007).

Supplementary File	Title
Supplementary Figure 1	The fitness effects of transposon insertion events
Supplementary Figure 2	Adaptive mutations underlying increased ethanol tolerance
Supplementary Figure 3	Identifying functional mutations using the ANS* data
Supplementary Figure 4	Identifying the mutated loci using the deletion scores
Supplementary Figure 5	Data filtering
Supplementary Figure 6	The Gaussian weights used for smoothing the data
Supplementary Table 1	Mutations identified in this study
Supplementary Table 2	Validating the adaptive mutations in ASN*
Supplementary Table 3	Validation of ETM adaptive mutations
Supplementary Table 4	Strains used in this study
Supplementary Table 5	Primers used for strain construction and verification
Supplementary Note 1	Adaptive mutations in experimental evolution of ethanol tolerance
Supplementary Software 1	ADAM computational tools

AOP

This array-based discovery tool creates linkage between functional mutations and selectable markers across a bacterial genome and can thus distinguish between adaptive and neutral mutations.

Issue

This array-based discovery tool creates linkage between functional mutations and selectable markers across a bacterial genome and can thus distinguish between adaptive and neutral mutations.