Global discovery of adaptive mutations

Hani Goodarzi[1,2], Alison K Hottes[1,2] & Saeed Tavazoie[1]

Although modern DNA sequencing enables rapid identification of genetic variation, characterizing the phenotypic consequences of individual mutations remains a labor-intensive task. Here we describe array-based discovery of adaptive mutations (ADAM), a technology that searches an entire bacterial genome for mutations that contribute to selectable phenotypic variation between an evolved strain and its parent. We found that ADAM identified adaptive mutations in laboratory-evolved *Escherichia coli* strains with high sensitivity and specificity.

Owing to their fast doubling time and intrapopulation diversity, bacteria are ideal organisms for investigating evolution on laboratory timescales[1]. Despite recent progress, however, discovering and interpreting the genetic changes underlying adaptive evolution remains challenging.

The problem of understanding how a parental strain of low fitness changes into an evolved strain of higher fitness can be broken down into two parts: finding all genetic differences between the two organisms and determining which differences are responsible for particular phenotypes. The former is being made feasible and affordable by the coming of age of whole-genome sequencing[2] and comparative genome resequencing technologies[3]. Unfortunately, although genetic tools have been developed to distinguish between neutral and adaptive mutations[4], the techniques remain labor-intensive. To overcome this limitation, we extended one of the classical techniques for mutation identification, linkage of a selectable marker to a functional mutation, into a powerful approach for the global profiling of adaptive mutations.

Our approach, array-based discovery of adaptive mutations (ADAM), uses parallel, genome-wide linkage analysis to simultaneously identify all mutated loci with direct fitness contributions (**Fig. 1**). ADAM requires three components, a library of selectable markers embedded in the DNA of the parental strain, a mechanism for transferring markers from the parental strain's library into the evolved strain in such a way that DNA from the parental strain adjacent to the marker replaces the corresponding DNA in the evolved strain and a method for measuring the frequency of markers throughout the genome. We first derived a collection of strains from the evolved strain, each with a selectable marker indicating where a segment of the evolved strain's DNA has been replaced with parental DNA. Usually the procedure swapped the evolved mutant's DNA for the identical parental sequence. In some cases, however, the DNA swap reverted one of the evolved strain's mutations. If the mutation was neutral, then the evolved mutant and the marked strain will be phenotypically identical. If, however, the reverted mutation was beneficial, then the marked strain will have lower fitness. We combined the marked strains and propagated them in selective conditions, causing the proportion of less-fit strains in the population, and consequently of markers near functional mutations, to decrease. We then used a single array hybridization to quantify the distribution of markers in the final, selected population relative to that of a population grown

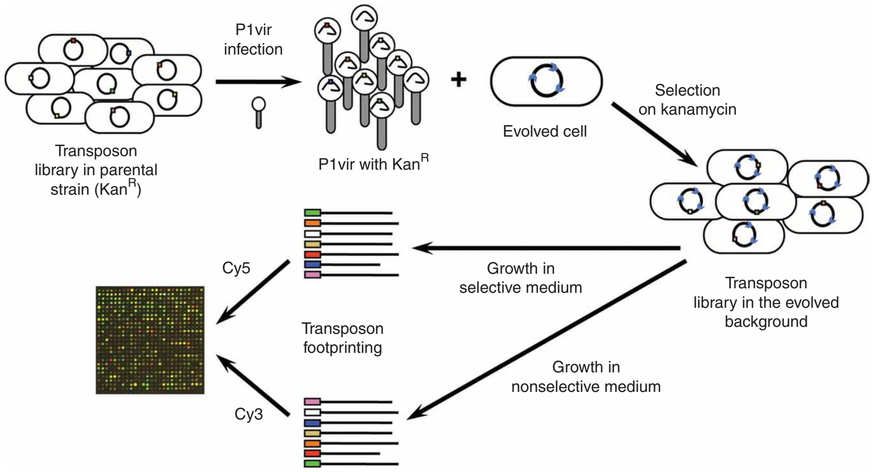

Figure 1 | Array-based discovery of adaptive mutations. The evolved strain is infected with P1vir lysate from a Kan^R transposon library in the parental background. Selection on kanamycin generates a secondary library in the evolved strain's background in which mutations near the transposon insertions have been corrected. After selective and nonselective growth of this library, the frequency of transposon insertion events in each locus is measured through a hybridization-based genetic footprinting approach. Genomic regions with a lower frequency of transposon insertions in the selected sample, relative to the nonselected sample, indicate the positions of functional mutations. A mutual information-based method is then used to identify genomic regions with functional mutations.

[1]Department of Molecular Biology & Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, USA. [2]These authors contributed equally to this work. Correspondence should be addressed to S.T. (tavazoie@genomics.princeton.edu).

RECEIVED 26 MARCH; ACCEPTED 20 JUNE; PUBLISHED ONLINE 13 JULY 2009; DOI:10.1038/NMETH.1352

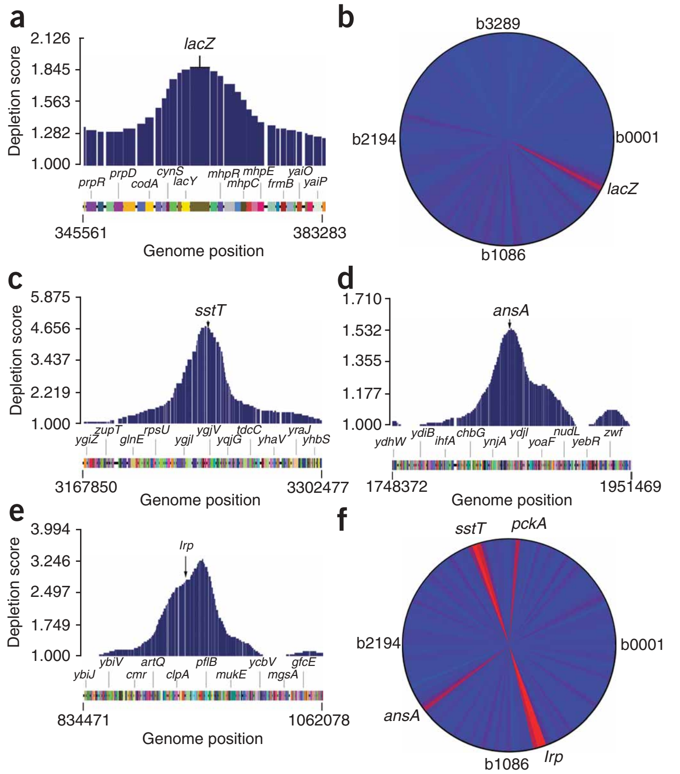

Figure 2 | Using ADAM to identify known and new mutations. (**a**) Smoothed depletion scores near the mutation (*lacZ*) in the Cml^R cassette–containing strain. (**b**) Average depletion score in slices of 25 genes across the genome for the Cml^R cassette–containing strain. Red denotes large depletion scores; blue denotes small depletion scores. The Blattner numbers indicate genome location. (**c**–**e**) Smoothed depletion scores in the regions near the identified mutations (*sstT* (**c**), *ansA* (**d**) and *lrp* (**e**)) for strain ASN*. (**f**) Average depletion score in slices of 25 genes across the genome for strain ASN*. For slices with identified mutations, the mutated locus is shown. Also given is the location of *pckA*, which occurs in a region with elevated depletion scores, but no mutations. The high depletion scores near *pckA* are limited to three genes and did not pass our information-theoretic statistical testing (**Supplementary Fig. 1**).

under nonselective conditions. Finally, we used a robust computational framework to pinpoint the likeliest locations for functional mutations.

As a proof of principle, we used ADAM to identify a single, known mutation of large effect. We chose as the 'evolved' strain, an MG1655 *Escherichia coli* strain with a chloramphenicol resistance (Cml^R) cassette inserted in the *lacZ* locus. The parental strain was identical, but lacked the Cml^R cassette.

We started with a transposon library in the parental strain[5], which provided a high-coverage suite of selectable markers (kanamycin resistance) across the genome. Transducing the markers into the Cml^R cassette–containing strain using P1vir phage[4] resulted in a library of ~5.0 × 10^5 transductants. As the markers brought with them DNA from the parental strain, many markers that recombined near the *lacZ* locus reverted the Cml^R cassette back to the wild-type allele, making the recipient chloramphenicol-sensitive.

Next we split the new library into two batches. We grew one part in rich medium with chloramphenicol to kill members lacking the Cml^R cassette, which decreased the number of strains with markers near the *lacZ* locus. We propagated the other part of the library for an equal number of generations (~7) without antibiotic to control for the general fitness effects of the transposon insertions. We then amplified the DNA adjacent to the transposon markers in each part of the library, differentially labeled the samples with Cy3 and Cy5 and hybridized them to an *E. coli* open reading frame array[5].

We defined a depletion score as the ratio of the marker frequencies in or near a gene after general growth versus phenotype-specific growth. Transposon insertions in loci close to *lacZ* were substantially depleted (**Fig. 2a**), but we did not observe notable depletion in any other region (**Fig. 2b**). A sensitive discovery of such genomic regions with high depletion scores (spanning tens of loci) was made possible through an information-theoretic computational framework (Online Methods). The computational tools used in this study are available online at http://tavazoielab.princeton.edu/ADAM/ and as **Supplementary Software 1**.

We next used ADAM to identify adaptive mutations in new, laboratory-evolved strains. First, we evolved MG1655 for fast growth in defined medium with asparagine as the sole carbon source and named the new, evolved isolate ASN*. As we created the original parental strain transposon insertion library in rich medium, the choice of a defined medium phenotype allowed us to test the technique in conditions in which we expected the transposon insertions themselves to severely alter fitness by disrupting some genes essential for growth in the medium[6].

As described above, we used P1vir phage to move a kanamycin-marked transposon library from the parental strain to the ASN* background. We split the library of evolved cells, and grew the two parts in glucose and asparagine media for 20 generations in exponential phase. Then we compared the marker distribution in the two cultures.

ADAM identified three regions with elevated depletion scores (**Fig. 2c–f** and **Supplementary Fig. 1**). Based on the functional annotations of the loci in each region, we selected candidate genes, which we then PCR-amplified and sequenced. We identified the underlying mutations as a single nucleotide insertion upstream of *sstT*, an IS2 insertion element integration upstream of *lrp* and a mismatch upstream of *ansA* (**Supplementary Table 1**).

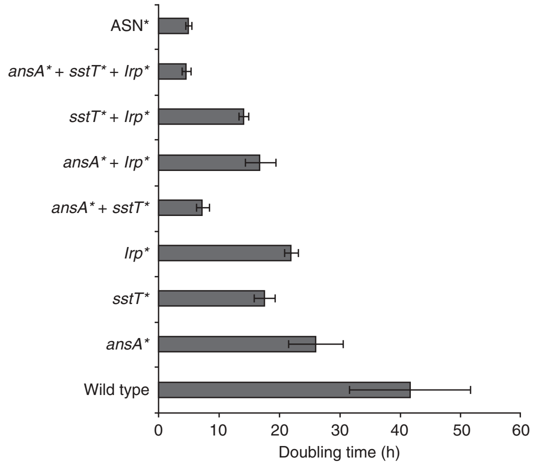

Figure 3 | Validation of discovered mutations in ASN*. The exponential phase doubling time of the wild-type strain, the evolved ASN* strain and strains with all combinations of the three identified mutations measured in asparagine medium. Shown are the mean and s.d. of three experiments. All strains had the same construction 'scars' and markers. The * indicates the presence of the evolved allele.

To find the strength of each mutation and to determine whether the three mutations collectively account for the growth rate of ASN* in asparagine medium, we generated a family of strains in the parental background that contain all allele combinations for the three loci and determined each strain's doubling time in asparagine medium (**Fig. 3**). The strain with all three mutant alleles was indistinguishable from ASN*, indicating that the three mutations are sufficient to explain the observed phenotype. Strains lacking any of the three mutant alleles grew more slowly than the ASN* strain, demonstrating that all three mutations are functional. However, loss of the *ansA* mutation, which had the smallest depletion score of the three, was more severe than loss of the *lrp* mutation. This discrepancy between depletion score and fitness effect may be due to a lower frequency of recombination within the *ansA* chromosomal neighborhood. Additional experiments showed that the beneficial mutations increase *ansA* and *sstT* expression and decrease *lrp* expression (**Supplementary Table 2**). Taken collectively, these data strongly suggest that ADAM discovered all functional mutations underlying ASN*'s high growth rate on asparagine medium.

To further test ADAM, we evolved an ethanol-tolerant strain (**Supplementary Note 1**). Using ADAM, we found four adaptive mutations in this new strain (**Supplementary Table 1**). Individually replacing each of the identified mutations with the wild-type copy resulted in strains less fit in ethanol-containing medium than the evolved strain (**Supplementary Table 3**). Furthermore, the magnitude of each locus's depletion score reflected the fitness effect of its mutation (**Supplementary Fig. 2**) suggesting that stronger mutations result in higher depletion scores. Additionally, we sequenced three regions with weaker scores that missed our statistical threshold. The absence of mutations in these regions demonstrates ADAM's high specificity.

ADAM can be modified for use in other bacteria. A random transposon library or a marked, whole-genome deletion collection[7,8] could serve as the selectable markers embedded in the DNA of the parental strain. In this work, we used P1vir phage to replace corresponding DNA between two strains, but suicide vectors or other generalized transducing phages could be substituted. The smaller the transferred DNA fragments, however, the more difficulty ADAM will have in identifying large adaptive mutations such as duplications and inversions. To measure marker frequency, we used a genetic footprinting technique, which could be modified for the specific markers used, to amplify the DNA adjacent to the markers and then quantified the distribution using in-house arrays[5]. For other organisms, marker densities could be determined using custom arrays or high-throughput sequencing.

ADAM could be used with whole-genome or comparative genome sequencing[3] to reveal both the exact locations of all mutations and the relevance of each to a phenotype of interest. Pinpointing the subset of beneficial mutations from among all the differences between two strains is desirable in many settings. First, during long experimental evolution experiments, strains often become mutators[9,10], and, even in the absence of hypermutation, drift commonly fixes some neutral mutations[11]. Second, many pathogenic bacteria have high mutation rates[12]. Hence, clinical evolutionary studies are likely to find many hitchhiker mutations. Third, when comparing two closely related natural isolates, only a fraction of the differences between them are likely to be important to any given phenotype. We showed that ADAM allows rapid and high-sensitivity profiling of adaptive mutations throughout the genome, a capacity crucial for studying the genetic basis of adaptation in native microbial ecologies, in the context of host-pathogen interactions and in the development of custom industrial strains.

METHODS

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

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

ACKNOWLEDGMENTS

We thank J. Liu for assistance in creating the ASN* strain. A.K.H. was supported by fellowship #08-1090-CCR-EO from the New Jersey Commission on Cancer Research. S.T. was supported by grants from the National Science Foundation Career Award (CAREER), Defense Advanced Research Projects Agency, National Institute of General Medical Sciences (P50 GM071508) and the National Institutes of Health Director's Pioneer Award (1DP10D003787-01).

AUTHOR CONTRIBUTIONS

H.G. conceived and designed the approach, performed experiments, analyzed the data and wrote the paper; A.K.H. performed experiments, analyzed the data and wrote the paper; S.T. wrote the paper.

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

1. Herring, C.D. *et al. Nat. Genet.* **38**, 1406–1412 (2006).
2. Shendure, J. & Ji, H. *Nat. Biotechnol.* **26**, 1135–1145 (2008).
3. Albert, T.J. *et al. Nat. Methods* **2**, 951–953 (2005).
4. Silhavy, T.J., Berman, M.L. & Enquist, L.W. *Experiments with Gene Fusions* (Cold Spring Harbor Press, Plainview, NY, 1984).
5. Girgis, H.S., Liu, Y., Ryu, W.S. & Tavazoie, S. *PLoS Genet.* **3**, 1644–1660 (2007).
6. Badarinarayana, V. *et al. Nat. Biotechnol.* **19**, 1060–1065 (2001).
7. Giaever, G. *et al. Nature* **418**, 387–391 (2002).
8. Jacobs, M.A. *et al. Proc. Natl. Acad. Sci. USA* **100**, 14339–14344 (2003).
9. Lenski, R.E., Winkworth, C.L. & Riley, M.A. *J. Mol. Evol.* **56**, 498–508 (2003).
10. Sniegowski, P.D., Gerrish, P.J. & Lenski, R.E. *Nature* **387**, 703–705 (1997).
11. Elena, S.F. & Lenski, R.E. *Nat. Rev. Genet.* **4**, 457–469 (2003).
12. Metzgar, D. & Wills, C. *Microbes Infect.* **2**, 1513–1522 (2000).

ONLINE METHODS

Strains and media. All strains used in this study (**Supplementary Table 4**) were derived from *Escherichia coli* MG1655 (ref. 13). LB contained 0.1% Bacto tryptone, 0.05% yeast extract and 0.05% NaCl. Asparagine medium contained M9 salts[14] supplemented with 2 g l^{-1} L-asparagine (Sigma), 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, 10 μM thiamine and micronutrients[15]. NaCl and $FeSO_4$ were omitted from M9 salts and micronutrients, respectively. Glucose medium was the same except that glucose (2 g l^{-1}) replaced asparagine. Media were supplemented with kanamycin (25 μg ml^{-1}) or chloramphenicol (20 or 25 μg ml^{-1}) as needed.

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

Experimental evolution of strain ASN*. To start the experimental evolution, $\sim 1 \times 10^9$ washed, LB-grown, mid-exponential phase MG1655 Δ*lacZ*[5] cells were added to 50 ml of asparagine medium. Using serial transfers that kept the population size above $\sim 1 \times 10^7$, the culture was maintained for 39 d in early- to mid-exponential phase. During that time, the bulk population went through 90 generations. We shook the culture at 250 r.p.m. at 37 °C.

Construction of strains to analyze the ASN* mutations. Using the method presented in ref. 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 to minimize polar effects. For *ansA*, we placed the marker downstream of the *ansA-pncA* operon.

To assemble the desired allele combinations, we first used the kanamycin markers to transduce the *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-resistance markers. And finally, we transduced the *lrp* alleles using kanamycin–resistance markers. In addition to the desired alleles, the final strains all had kanamycin- and chloramphenicol-resistance markers and a 'scar' from the original *ansA* kanamycin-resistance 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 P1vir phage for all transductions[4]. Sequences of primers used in strain construction and testing are listed in **Supplementary Table 5**.

P1vir lysate preparation. We prepared P1vir 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 min, we added 2.5 ml of P1vir 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 5,525*g* for 10 min 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 P1vir transduction protocol[4]. We pelleted cells from 25 ml of overnight, stationary phase culture of the evolved strain by centrifugation (5,525*g*, 15 min) and resuspended them in 10 ml of LB with 5 mM $CaCl_2$ and 10 mM $MgSO_4$. Then, in each of 24 microcentrifuge tubes, we mixed 400 μl cells with 200 μl phage lysate from the parental strain library. We incubated the mixtures at 30 °C for 30 min 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 37 °C for 30 min without shaking and then pelleted the cells by centrifugation (15 min, 5,525*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 °C for 10 h (until the culture reached mid-stationary phase). Finally, we pelleted the cells by centrifugation (15 min, 5,525*g*), resuspended them in 15–20 ml LB with 15% glycerol, and snap froze them on dry ice and ethanol.

Growth of secondary library under selective and nonselective conditions. In each experiment, we grew portions of the secondary P1vir-transduced transposon library in the presence and absence of selection. Selective and nonselective growth spanned the same number of generations.

Finding the distributions of markers across the genome (genetic footprinting). We subjected samples of $\sim 10^7$ 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 nonselective conditions were differentially labeled and hybridized to *E. coli* open reading frame 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 nonselective growth samples to depletion scores:

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

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/ (**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

doi:10.1038/nmeth.1352

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 dataset (**Supplementary Fig. 3b**). Although 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 is not in the intensity of the scores but rather the number of consecutive genes that show high depletion scores. For example, in our $\mathrm{Cml}^{\mathrm{R}}$ experiment, *lacI* had a depletion score comparable to that of *rfaQ* (2.23 versus 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 (**Fig. 2a**). To capture these regions, we quantized the vector containing the depletion scores for all the genes into four 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 (**Supplementary Fig. 4**). A spatial vector is a binary vector of length N (that is, 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:

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

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:

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

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 substantial 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. Owing 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, but certain regions were 'hot spots' 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 was in an insertion 'hot spot' with a signal of 10, then the depletion score would be 100. Although 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. 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 s.d.) in the unselected transposon library[5] was more than one s.d. 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.

13. Blattner, F.R. *et al. Science* **277**, 1453–1474 (1997).
14. Ausubel, F.M. *et al. Current Protocols in Molecular Biology* (Wiley Interscience, New York, 1994).
15. Neidhardt, F.C., Bloch, P.L. & Smith, D.F. *J. Bacteriol.* **119**, 736–747 (1974).
16. Amini, S., Goodarzi, H. & Tavazoie, S. *PLoS Pathog.* **5**, e1000432 (2009).
17. Baba, T. *et al. Mol. Syst. Biol.* **2**, 2006 0008 (2006).
18. Datsenko, K.A. & Wanner, B.L. *Proc. Natl. Acad. Sci. USA* **97**, 6640–6645 (2000).
19. Cover, T. & Thomas, J. *Elements of Information Theory* 2nd edn. (Wiley-Interscience, Hoboken, New Jersey, USA 2006).
20. Elemento, O., Slonim, N. & Tavazoie, S. *Mol. Cell* **28**, 337–350 (2007).

doi:10.1038/nmeth.1352