Published in final edited form as:
Nat Methods. 2009 August ; 6(8): 581–583. doi:10.1038/nmeth.1352.

Global Discovery of Adaptive Mutations

Hani Goodarzi[1,2], **Alison K. Hottes**[1,2], and **Saeed Tavazoie**[1]

[1]Department of Molecular Biology & Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA.

Abstract

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

Due to their fast doubling-time and intra-population diversity, bacteria are ideal organisms for investigating evolution on laboratory time-scales1. Despite recent progress, however, discovering and interpreting the genetic changes underlying adaptive evolution remains challenging.

Broadly speaking, 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 sequencing2 and comparative genome re-sequencing technologies3. Unfortunately, although genetic tools have been developed to distinguish between neutral and adaptive mutations4, the techniques remain labor intensive. To overcome this limitation, this study extends 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, termed ADAM for Array-based Discovery of Adaptive Mutations, employs 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 derive 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 evolved mutant's DNA is swapped for the identical parental sequence. In some cases, however, the DNA swap reverts one of the evolved

Correspondence to: Saeed Tavazoie e-mail: tavazoie@genomics.princeton.edu.
[2]These authors contributed equally.

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.

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 combine the marked strains and propagate 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 use a single array hybridization to quantify the distribution of markers in the final, selected population relative to that of a population grown under non-selective conditions. Finally, we use a robust computational framework to pinpoint the likeliest locations for functional mutations.

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

We started with a transposon library in the parental strain5, which provided a high-coverage suite of selectable markers (kanamycin resistance) across the genome. Transducing the markers into the Cml^R strain using P1 *vir* phage4 resulted in a library of ~5.0×10^5 transductants. Since 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 media with chloramphenicol to kill members lacking the Cml^R cassette, which decreased the number of strains with markers near the *lacZ* locus. The rest of the library was propagated for an equal number of generations (~7) without antibiotic to control for the general fitness effects of the transposon insertions themselves. We then amplified the DNA adjacent to the transposon markers in each part of the library, labeled the samples, and hybridized them to an *E. coli* ORF array5.

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), while notable depletion was not observed 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 (see Online Methods). The results and the computational tools used in this study are available online at http://tavazoielab.princeton.edu/ADAM/.

We next conducted more challenging experiments in which ADAM was used to identify adaptive mutations in novel, lab-evolved strains. First, we evolved MG1655 for fast growth in defined media with asparagine as the sole carbon source. As the original parental strain transposon insertion library was created in rich media, the choice of a defined media phenotype allowed us to test the technique in conditions where the transposon insertions themselves were expected to severely alter fitness by disrupting some genes essential in the media6.

As described above, P1 *vir* phage was used to move a kanamycin-marked transposon library from the parental strain to the new, evolved isolate, ASN*. The library of evolved cells was then split between glucose and asparagine media and grown for 20 generations in exponential phase. Then, the marker distribution in the two cultures was compared.

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).

To find the strength of each mutation and to determine if the three mutations collectively account for the growth rate of ASN* in asparagine media, 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 media (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 media.

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 media than the evolved strain (Supplementary Table 3). Furthermore, the magnitude of each locus's depletion score reflects 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 collection7,8 could serve as the selectable markers embedded in the DNA of the parental strain. In this work, we used P1*vir* 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 employed 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 arrays5. For other organisms, marker densities could be determined using custom arrays or high-throughput sequencing.

ADAM could be employed in concert with whole-genome or comparative genome sequencing3 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 mutators9,10, and, even in the absence of hypermutation, drift commonly fixes some neutral mutations11. Second, many pathogenic bacteria have high mutation rates12. 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 have shown 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Julia Liu for assistance in creating the ASN* strain. A.K.H. was assisted 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).

References

1. Herring CD, Raghunathan A, Honisch C, et al. Nat Genet. 2006; 38(12):1406. [PubMed: 17086184]
2. Shendure J, Ji H. Nat Biotechnol. 2008; 26(10):1135. [PubMed: 18846087]
3. Albert TJ, Dailidiene D, Dailide G, et al. Nat Methods. 2005; 2(12):951. [PubMed: 16299480]
4. Silhavy, TJ.; Berman, ML.; Enquist, LW. Experiments with gene fusions. Plainview, NY: Cold Spring Harbor Press; 1984.
5. Girgis HS, Liu Y, Ryu WS, et al. PLoS Genet. 2007; 3(9):1644. [PubMed: 17941710]
6. Badarinarayana V, Estep PW 3rd, Shendure J, et al. Nat Biotechnol. 2001; 19(11):1060. [PubMed: 11689852]
7. Giaever G, Chu AM, Ni L, et al. Nature. 2002; 418(6896):387. [PubMed: 12140549]
8. Jacobs MA, Alwood A, Thaipisuttikul I, et al. Proc Natl Acad Sci U S A. 2003; 100(24):14339. [PubMed: 14617778]
9. Lenski RE, Winkworth CL, Riley MA. J Mol Evol. 2003; 56(4):498. [PubMed: 12664169]
10. Sniegowski PD, Gerrish PJ, Lenski RE. Nature. 1997; 387(6634):703. [PubMed: 9192894]
11. Elena SF, Lenski RE. Nat Rev Genet. 2003; 4(6):457. [PubMed: 12776215]
12. Metzgar D, Wills C. Microbes Infect. 2000; 2(12):1513. [PubMed: 11099938]

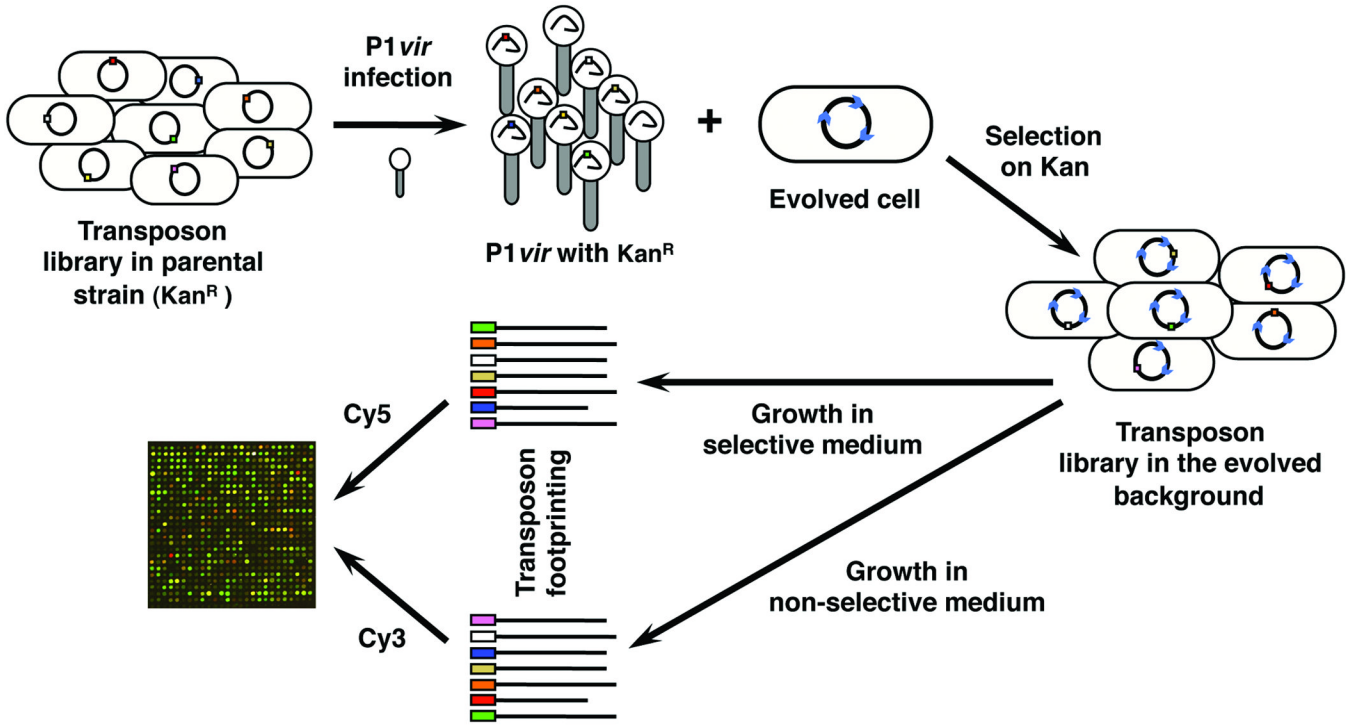

Figure 1. Array-based discovery of adaptive mutations

The evolved strain is infected with P1 *vir* 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. Following selective and non-selective 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 non-selected sample, indicate the positions of functional mutations. A mutual information based method (see Online Methods) is then used to identify genomic regions with functional mutations.

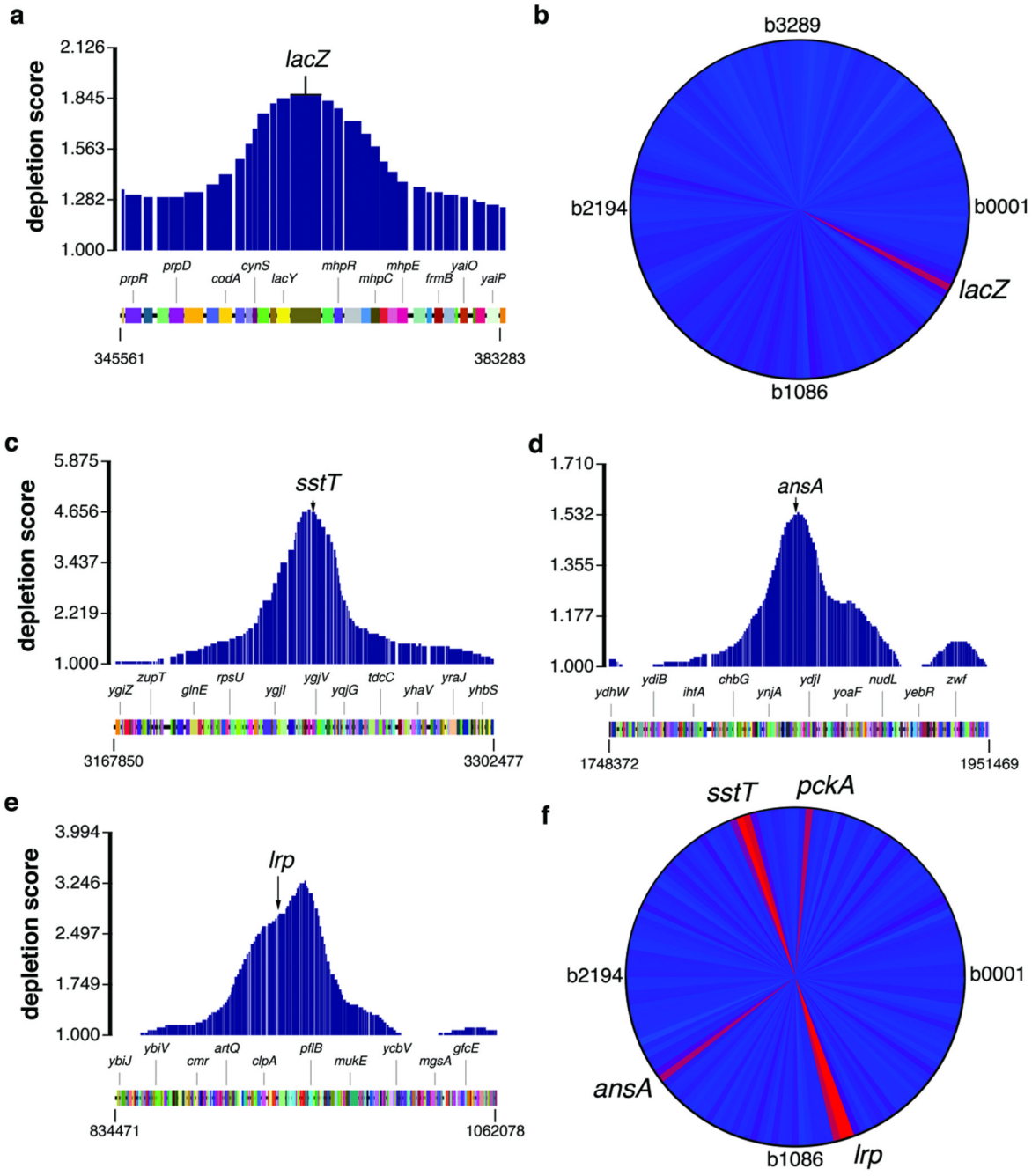

Figure 2. Using ADAM to discover known and novel mutations

(**a**) When mapping the location of a known chloramphenicol resistance cassette, the smoothed depletion scores peak near the *lacZ* locus where Cml^R is inserted. (**b**) The plot displays the average depletion score in slices of 25 genes across the genome. The *lacZ* locus shows strong signal, and all other loci contain a low, uniform background signal. (**c–e**) The smoothed depletion scores from applying ADAM to the ASN* strain peak near each of the identified loci (i.e., *sstT*, *ansA*, and *lrp*). (**f**) The depletion scores across the genome are low except for the regions of the three identified mutations and a region near *pckA*. The high score at *pckA* is limited to three genes and does not pass our information-theoretic statistical

testing (Supplementary Fig. 1). In (**a,c–e**), the horizontal axis depicts genome coordinates and genes. Data was smoothed to aid visualization (see Online Methods).

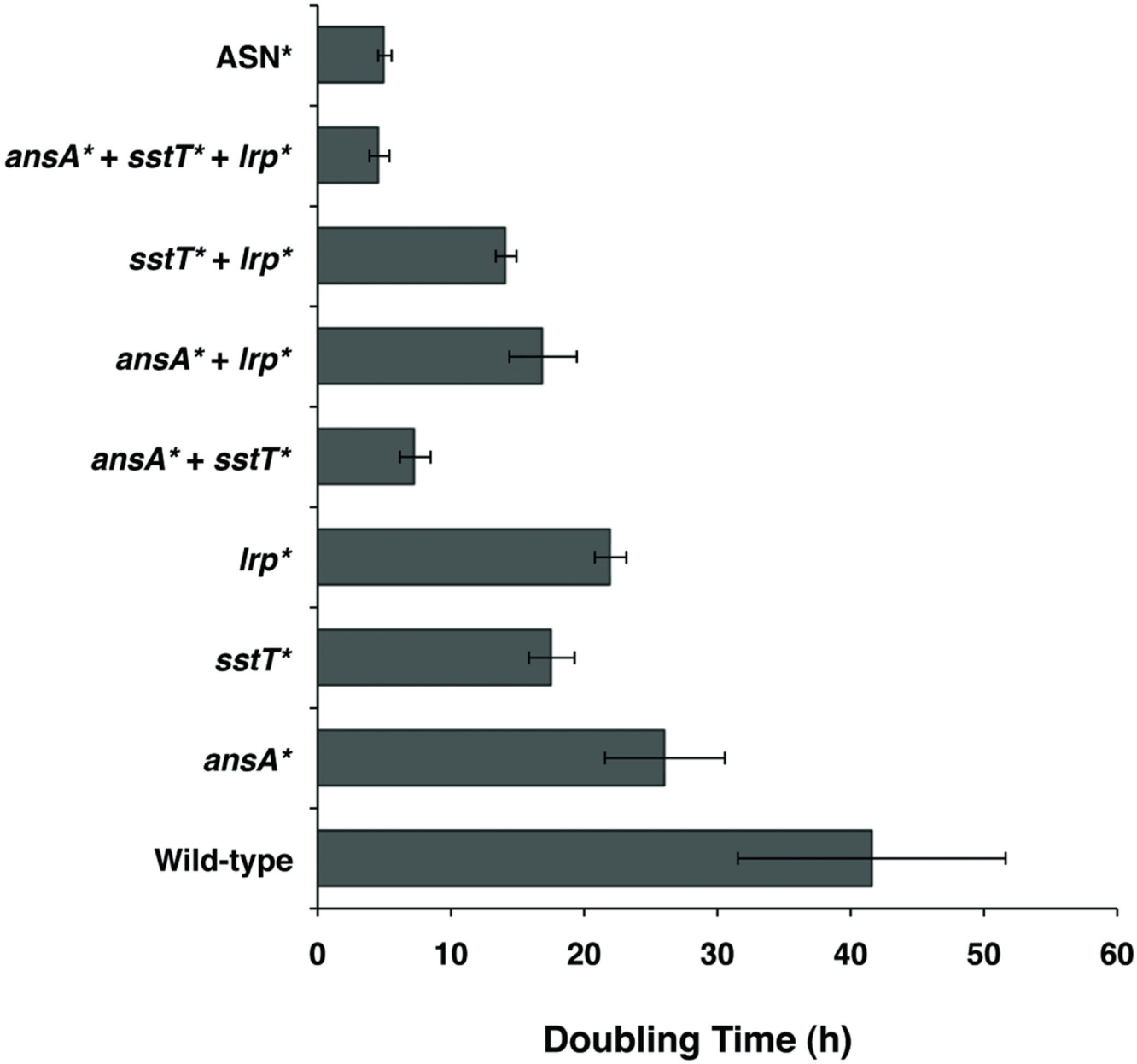

Figure 3. Validation of discovered mutations in ASN*

The exponential phase doubling time of the wild-type strain, the evolved ASN* strain, and strains with all combinations of the three identified mutations were measured in asparagine media. Shown are the mean and standard-deviation of three experiments. All strains had the same construction scars and markers. Here, *gene** indicates the presence of the evolved allele.