Coupling mutagenesis and parallel deep sequencing to probe essential residues in a genome or gene

William P. Robins[a], Shah M. Faruque[b], and John J. Mekalanos[a,1]

[a]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA 02115; and [b]Molecular Genetics Laboratory, International Centre for Diarrhoeal Disease Research, Mohakhali, Dhaka 1212, Bangladesh

Contributed by John J. Mekalanos, January 5, 2013 (sent for review December 15, 2012)

The sequence of a protein determines its function by influencing its folding, structure, and activity. Similarly, the most conserved residues of orthologous and paralogous proteins likely define those most important. The detection of important or essential residues is not always apparent via sequence alignments because these are limited by the depth of any given gene's phylogeny, as well as specificities that relate to each protein's unique biological origin. Thus, there is a need for robust and comprehensive ways of evaluating the importance of specific amino acid residues of proteins of known or unknown function. Here we describe an approach called Mut-seq, which allows the identification of virtually all of the essential residues present in a whole genome through the application of limited chemical mutagenesis, selection for function, and deep parallel genomic sequencing. Here we have applied this method to T7 bacteriophage and T7-like virus JSF7 of *Vibrio cholerae*.

Afforded by dramatic progress in DNA sequencing cost reduction and increased output that has grown at a rate exceeding that of Moore's law (1), the compilation of deposited sequences now provides a vast database for identifying proteins and motifs at an increasingly high resolution. This compilation is exemplified in the Pfam database; within 14 y of its inception, there are now more than 12,000 conserved protein families, some represented by over 100,000 sequences (2). Highly conserved residues have been documented that correspond to the core catalytic and active sites or protein–protein interaction surfaces (3). Programs such as SIFT (Sorting Intolerant From Tolerant) use amino acid conservation to predict tolerated from deleterious substitutions (4). However, residues that support the folding and basic structure of the protein may not be as conserved and thus may not be predicted to be essential by such in silico analyses. Such residues may control conformation of a protein only in the context of its own unique polypeptide sequence or in the milieu of a complex of coevolved interacting partners.

To better understand the contribution of each residue to the function of a protein, the specific sequence of a protein must also be understood within the evolved constraints imposed by its organism's biology rather than only by conserved sequences, motifs, or predicted structures. This understanding is often accomplished using mutagenesis and functional analysis. Approaches such as scanning alanine mutagenesis have provided significant insights to functionally important residues of proteins. However, the nonsaturating nature of such analyses, as well as the labor involved, have limited their usefulness to most biochemists and geneticists.

Here we define a unique highly parallel approach to defining functional residues of proteins based on their mutability alone. Our method (Mut-seq) takes advantage of the depth afforded by next-generation sequencing to characterize complex pools of mutated genomes and genes for functionality, and in doing so maps coding sequences for residues that show statistically high or low rates of mutability. In brief, we show that a large mutagenized population of viral genomes can be selected for growth fitness and then sequenced as a pool to define which amino acid residues can be changed to one or more other residues, and which cannot tolerate changes at all. The less or nonmutable residues are of special interest in that they may play pivotal roles in a protein's activity as contributors to an enzymes active site, essential functional motifs, or as structural elements or linkers between domains that confer proper protein folding and conformation. These insights into the essential residues contributing to the functionality of proteins may provide a new dataset useful to the development of small molecule inhibitors of essential proteins, and may also inform efforts to suppress drug resistance through evolved mutation.

Results

Sequencing Mutated Phage and Stringent Filtering of Reads to Identify Single-Base Substitutions. Mut-seq involves operationally the following steps: (*i*) mutagenesis of a genome or gene; (*ii*) recovery of a bank of mutagenized targets under a positive selection condition; (*iii*) deep sequencing of the entire bank; and (*iv*) alignment of sequence reads to identify and quantify base substitutions within the genomes or genes that represent mutations. We initially tested Mut-seq on bacteriophage T7 of *Escherichia coli*, a podophage with a genome size of ~40 kb and JSF7, an uncharacterized *Vibrio cholerae* podophage. We used the chemical mutagen hydroxylamine (HA), which specifically induces transitions of GC base pairs to AT base pairs. HA-treatment allows mutation of the phage genome but DNA is still packaged in the intact virion before genome internalization and replication in the cell. This specific mode of chemical mutagenesis allowed us to titrate the level of mutagenesis accurately, as well as provide a signature of induced mutations and separate these from sequencing errors. We generated and sequenced ~1.5 million randomly mutagenized plaque-forming units derived from a stock of 10 billion plaque-forming units. Because the mutagenized phage particles were recovered after growth on a bacterial host, we envisioned that only viable replication-proficient phages were sequenced. Deep sequencing of the DNA derived from these

Significance

In this work we present a technique called Mut-seq. We show that a very large population of genomes or genes can be mutagenized, selected for growth, and then sequenced to determine which genes or residues are probably essential. Here we have applied this method to T7 bacteriophage and T7-like virus JSF7 of *Vibrio cholerae*. All essential T7 genes have been previously identified and several DNA replication and transcription proteins have solved structures and are well studied, making this a good model. We use this information to correlate mutability at protein residues with known essentiality, conservation, and predicted structural importance.

Author contributions: W.P.R. designed research; W.P.R. performed research; W.P.R. and S.M.F. contributed new reagents/analytic tools; W.P.R. analyzed data; and W.P.R. and J.J.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The raw sequencing data in this paper have been deposited in the NCBI sequence read archive (SRA).

[1]To whom correspondence should be addressed. E-mail: john_mekalanos@hms.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1222538110/-/DCSupplemental.

mutagenized surviving phage progeny allowed us to map and count HA-induced mutations at every G/C position in the T7 genome, and thus measure the mutability across each protein coding sequence. In each of the four replicates, between 6.9% and 9.5% of 160–220 million total reads of 50-nt length were found to contain exactly one single-nucleotide substitution representing a prospective mutation. Stringent filtering was applied using CASAVA v1.8 quality scores (Q38) that predict accuracy 99.98% for the substitution and the flanking 11 nucleotides, further reducing the pool to only ~1% of original reads (Fig. 1). This filtering was imposed to remove reads with low-quality scores that may be erroneously counted as false-positive mutations. Within the pool, HA-induced mutations were mixed with other transition and transversion mutations. We attribute this finding to the significant depth of the sequencing coverage (200,000–500,000 per nucleotide), which was sufficient to detect even rare mutations introduced via amplification by the high-fidelity polymerases during PCR and flow-cell clustering, or via inaccuracies in the T7 DNA replication (5).

To ascertain whether the level of mutation was sufficient, we compared the frequency and total number of HA mutations in both

A

Replicate	Raw Reads	Mapped Reads (no SNPs)	Mapped Reads (1 SNP)	Total T7 positions with 1+ SNP	Filtered G/C → A/T HA SNPs (% all T7 G/C)	Invariant Mutants per Genome
T7_1A	160,585,684	125,993,720	15,616,051	39,063	242,499 (99%)	0
T7_1B	201,827,063	166,254,374	14,332,666	38,571	316,237 (99%)	0
T7_2A	216,704,229	186,562,335	15,034,520	39,665	348,354 (99%)	0
T7_2B	211,250,998	166,334,659	20,992,981	39,678	302,335 (99%)	0
T7_C1	19,571,457	18,220,262	577,011	39,063	4,730 (25%)	1
T7_C2	23,344,481	21,948,618	851,365	38,571	5,118 (27%)	5
T7_C3	21,663,728	19,807,098	735,392	39,665	4,732 (25%)	7
T7_C4	20,187,077	18,756,014	673,979	39,678	4,409 (23%)	5
T7_C5	24,823,118	22,704,498	736,584	39,063	6,415 (33%)	0
T7_C6	24,946,276	23,269,606	899,776	38,571	6,694 (35%)	5
T7_Ctotal	134,536,137	124,706,096	4,204,107	32,885	32,098 (70%)	3.8 (average)

B

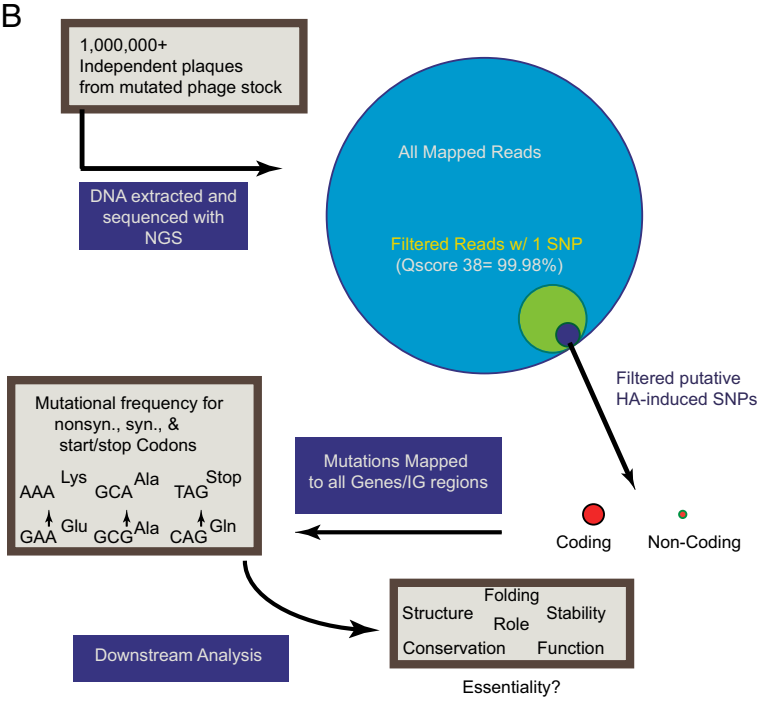

Fig. 1. Table of reads (*A*) and Mut-seq flowchart (*B*). Accounting of the mapped sequencing reads for each of the biological replicates from raw sequencing files including those mapped with and without mutations and the number of identified HA mutations (G/C to A/T substitutions). Biological replicates 1 and 2 were sequenced in duplicate as technical replicates (1A/B and 2A/B). The six genomes independently isolated from the mutant pool comprise samples C1 to C6. C1–C3 and C4–C6 were plated from the Bl21 and Bl21:DE3 outgrown strains, respectively. The total raw number of pooled C1-C6 reads (Ctotal) were comparable to each of the technical replicates in replicate 1 and 2 but exhibited a significantly reduced total number of putative H/A mutations. Identification of each invariant substitution in C1-C6 can be found in Dataset S1. (*B*) Flowchart illustrating the pipeline used to filter, map, and analyze mutations in this work. Circles that represent read pools are scaled to size to represent proportions.

mutated and nonmutated populations. The identity and quantity of base substitutions at every nucleotide position in the four replicates (1A/1B and 2A/2B) was compared with sequenced and mapped reads from six random independently isolated phages (C1–C6) from both mutant pools. These six phages provided a benchmark estimate for the number of mutations produced per unit genome. As expected, no mutations were found to be universally detected in the mutated pools, but we found an average of 3.8 mutants per sequenced individual phage genome, predicting approximately one mutation per 10 kb. The frequency of putative HA-induced substitutions was measured to be six- and ninefold increased in replicates 1 and 2, respectively, when compared in the pooled individual samples, verifying an increase in HA-induced mutations.

Measurement and Normalization of Synonymous, Nonsynonymous, and Stop-Codon Mutations. There are a restricted number of permitted amino acid changes induced by HA mutagenesis at G/C sites within the genetic code of individual amino acids. Some of these changes result in synonymous mutations and others create nonsynonymous substitutions, disrupt initiation sites, and introduce premature stop codons. To better compare frequencies of mutation between replicates with varying numbers of total reads and subtract the frequency of spontaneous mutations, a normalized mutability index (NMI) was implemented. This implementation was accomplished by multiplying the mutant count (MC) at a base position by a normalization factor derived from the ratio of total mapped reads (both those mapped with no substitutions and with only one substitution/mutation) measured in the mutated and nonmutated replicates, and then subtracting the background MC of base substitution at the same position in the nonmutated pool (see sample calculation in Fig. S1).

This implementation of a normalization is similar to the RPKM (reads per kilobase per million mapped reads) used for comparing gene-expression data measured in separate RNA-seq experiments with varying numbers of mapped read depth. However, in contrast to RPKM, to calculate and compare NMI values between two Mut-seq experiments, they must share a common nonmutated replicate of the same gene or genome.

The averaged NMI for all synonymous, nonsynonymous, and premature stop-codon substitutions in replicate pairs 1 and 2 were plotted against one another, and for each group were found to be similar. The correlation of the averaged NMI calculated between replicates 1A/B and 2A/B is graphed in Fig. 2 *A*–*C*; the ∼45° slope and intercept of plotted values demonstrates lack of bias in overall NMI value for either replicate. Fig. 2*A* shows the distribution of stop codons in essential genes and the corresponding average NMI value of the population in each replicate. As expected, the average threshold for nonsynonymous and synonymous mutations

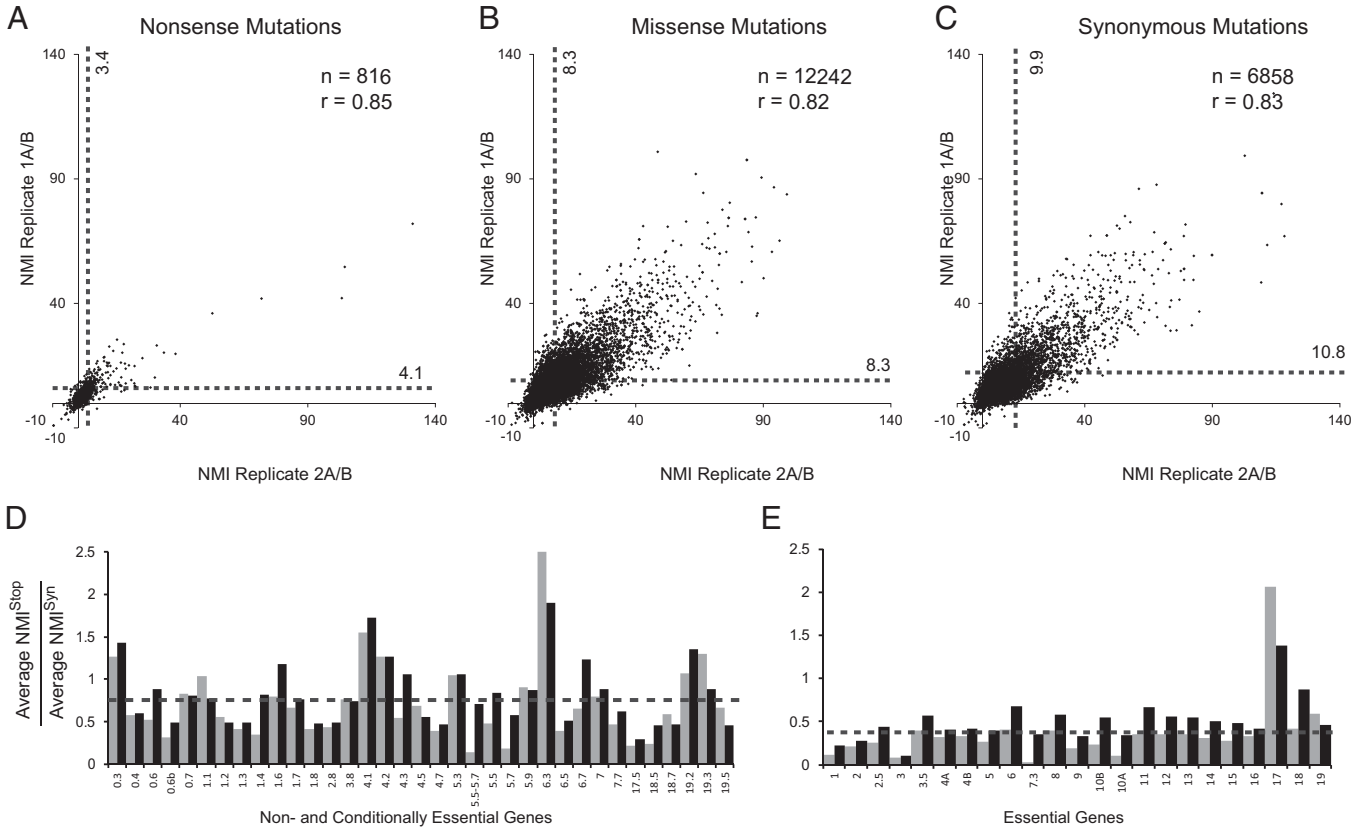

Fig. 2. Plotted NMI values for all nonsense, missense, and synonymous substitutions in 60 T7 genes to show correlation between biological replicates. The NMI of nonsense and synonymous substitutions can be used as ratios to predict gene essentiality. (*A*) The NMI value for each created premature stop codon in T7 essential genes from each replicate graphed against one another to show reproducibility. This graph includes genes *1, 2, 2.5, 3, 3.5, 4A/B, 5, 6, 7, 7.3, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18*, and *19*. The averaged NMI value for essential genes for each replicate set is also indicated. (*B*) The NMI value for each created nonsynonymous codon in T7 essential genes from each averaged biological replicate graphed against one another to show reproducibility. The averaged NMI value for essential genes for each replicate set is also indicated. (*C*) The NMI value for each created synonymous codon in T7 essential genes from each averaged biological replicate graphed against one another to show reproducibility. The averaged NMI value for essential genes for each replicate set is also indicated. (*D* and *E*) The graphed ratios of the average NMI for created premature stop to synonymous codons across all T7 genes for average of replicate 1A/B (light gray) and replicate 2A/B (black). These values are plotted separately for both nonessential (*D*) and essential genes (*E*). The average ratio is generally less than 0.4 for essential genes and increased and more varied for those nonessential.

(Fig. 2 *B* and *C*) indicated greater mutagenic permissiveness than that found for stop codons.

To assess detailed mutagenic depth on a gene-to-gene basis, the frequency of each substitution and resulting amino acid change, when applicable, was used to develop a mutational profile across the entire genome and 60 ORFs (Dataset S1). For residues in many essential genes, nearly 90% of nonsense mutations were found to have low NMIs less than 5, suggesting these indices may be useful for calculating essentiality of each gene. To evaluate known essential and nonessential genes, each of the 60 ORFs was compared by dividing the corresponding average NMI for premature stop-codon changes by the NMI for synonymous substitutions. This stop codon to permissive ratio was implemented as a relative metric for comparing the known essential to nonessential or conditionally essential T7 genes. The average ratio for all but one essential gene was found to be less than 0.43, whereas for other genes this increased to as much as 2.5 (Fig. 3 *D* and *E*). Only one essential gene differed. Gene *17* encodes for tail fiber and alone has been shown to complement defective gene *17* mutants *in trans* in liquid cultures (6), and therefore it seems likely that fibers released from lysed cells diffused and complemented defective fiberless *17* mutant phages.

Conserved Residues and Essential Residues Show Low Mutability in an Essential T7 replication Protein. Because trends in NMI values were found to correlate between replicates, we investigated the significance of NMI values at base positions that encode known essential and conserved residues. Information about essential residues and lethal mutations in T7 transcription and DNA replication proteins can be gathered from previous work. In addition, these enzymes have solved X-ray crystal structures. Here we investigated mutations in T7 gene *2.5* [T7 single-stranded binding (SSB) protein], gene *1* (T7 RNA polymerase), and gene *5* (DNA polymerase), three genes that fit these criteria. T7 SSB is a small protein homodimer that serves a strict structural role in stabilizing ssDNA. Using the solved X-ray crystal structure as a scaffold, many of the essential residues have been shown to be important for forming the DNA-binding cleft and stabilizing the dimer (7, 8).

The important residues in T7 gene *2.5* (SSB) were measured to have a significantly decreased NMI when the mutation produced a nonsynonymous codon. We matched the least mutable residues identified in all four replicates with known essential residues and those most conserved (bit score 3.5) within a group of 19 T7 SSB protein homologs (PHA00458 superfamily). Fig. 3*B* shows the conserved and essential amino acid residues in T7 gene *2.5* and its defined secondary structure prediction. Using this template, we mapped the least-mutable amino acid residues to known essential or conserved residues. The essential group was identified by Rezende et al. (7) as a set of 20 single amino changes in SSB shown to be lethal for T7 growth. Of the 13 essential amino acids that can be targeted by HA mutagenesis, 12 were shown to be nonmutable or least mutable, the exception being at the V168 residue. In the reference list, the V168F allele was shown to be lethal; however, the valine codon used and HA-induced transition limits this change to a more a similar isoleucine, which is likely a tolerated substitution. Furthermore, we indentified an expanded set of potentially disruptive or lethal mutations that alter residues proximal to those previously found to be essential (7). Together with known essential residues, some reside within the β-barrel domain, near DNA-binding domains, within protein loops, and within the C terminus.

To test the predicted essentiality of low NMI residue substitutions, the growth of a T7 phage-disrupted gene *2.5* with a *trxA* gene insertion was measured after complementation with six different nonsynonymous gp2.5 mutant genes or wild-type. A number of alleles were selected with NMI values ranging between −0.83 and 3.78. Three of six mutants impaired the efficiency of plating (EOP), two significantly (Fig. 2*C*), and four mutants yielded much

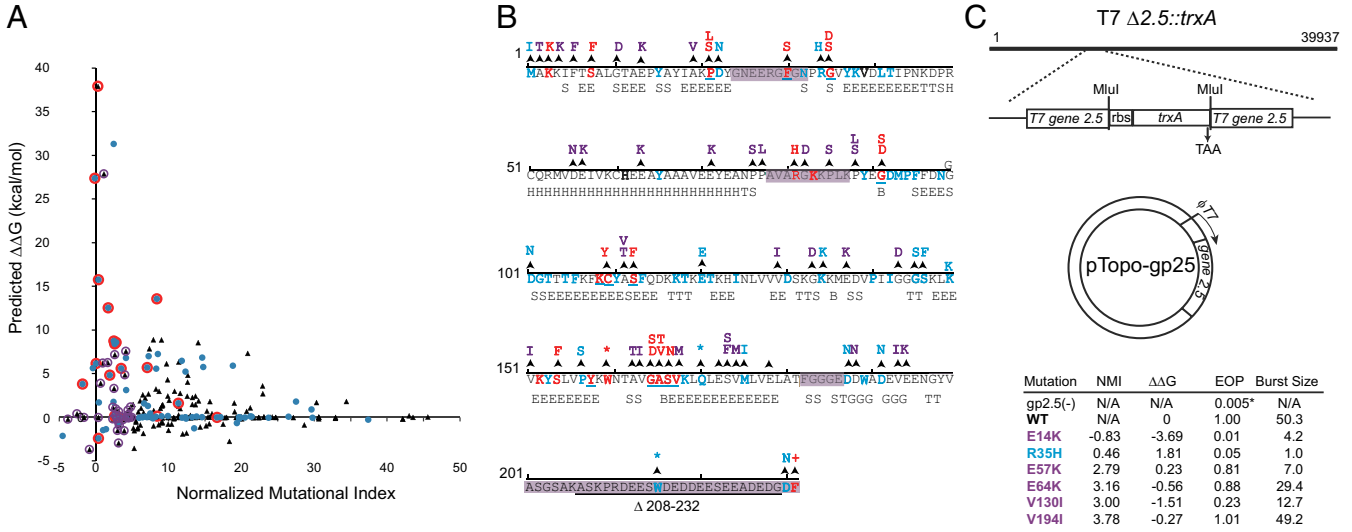

Fig. 3. The NMI correlates with both conserved and essential residues and substitutions that are predicted to effect protein stability. Additional essential residues predicted only by NMI can be shown to be deleterious to T7 growth. (*A*) Predicted ΔΔG and NMI values for all T7 gene *2.5* (SSB) positions averaged for 1A/1B and 2A/2B were plotted. By definition, all synonymous mutations had ΔΔG values of 0. Some nonconserved (black triangle) and conserved (filled blue circles) were also determined to be essential (open red circle) by prior work and marked accordingly. Recently identified least mutable positions are also indicated (filled purple circle). (*B*) Amino acid sequence and predicted secondary structure of T7 SSB specifying residues with low NMI values. Residues are colored if found to conserved in other T7-like SSB proteins (blue), essential (red), or newly identified (purple). If a position was found to be have a low NMI (<3) the rare substitution was noted above an arrow. Secondary structures are indicated DSSP codes (H, α-helix; B, β-bridge; E, extended strand; G, 3/10 helix; T, H-bonded turn; S, bend) and purple shading indicates residues missing from the 1JE5.PDB structure. Blank (no code) indicates possible loops and irregular elements. (*C*) Diagram showing the insertion of *trxA* gene and disruption of T7 gene *2.5*. The *trxA* gene is flanked 5′ by a Shine-Delgarno site and terminates with a TAA codon. To test complementation, wild-type T7 gene *2.5* and mutants are expressed downstream of the T7-RNAP promoter in pTopo-2.5. The EOP and burst size of complemented growth are compared with the observed NMI and predicted ΔΔG values. The low EOP measured in the absence of gene *2.5* is a result of recombination and reacquisition of gene *2.5* into the T7 genome.

smaller burst sizes (less than 13) than the complemented wild-type gene (~50). These results are consistent with our Mut-seq data for the low mutability of these residues because mutant viruses with these alleles would be expected to have a significant reduction in their efficiency of infection and burst size. Within the pool of genomes that survive mutagenesis, such mutants would be depleted, and thus fewer reads for mutations of this sort would be found. Indeed, this result was the case for the majority of "tolerated mutations" in residues that otherwise have low MNI values. These results are thus consistent with the underlying hypothesis that changing residues with low mutability will produce phages that are dramatically reduced in replication fitness.

Correlation of Predicted Structural Changes with a Subset of Residues That Exhibit Low NMI and Low MC in Other T7 Essential Genes with Solved Structures. To further investigate the expanded repertoire of mutants, we used in silico structural prediction to correlate NMI and change of total free energy (ΔΔG) between every HA-induced tolerated mutant and the wild-type protein using the PDB structure (Dataset S2). In this work we used FoldX (9), a protein-structure algorithm that uses empirical force fields to test the predicted change in free energy of every possible HA-directed mutation. It was expected that mutations in essential genes with predicted higher positive ΔΔG values should be generally deleterious to the protein conformation and also likely to negatively impact T7 growth, and thus selection would produce a specific low MC for a tolerated mutation. In the scope of this work, we applied this analysis to identify general trends between ΔΔG and NMI without attempting to interpret the consequence of individual changes or assigning quantitative importance.

There was a strong inverse correlation between predicted ΔΔG values and mutational depth; many of the mutations with ΔΔG values greater than +10 were least mutable (NMI < 5) (Fig. 3*A* and Dataset S2). A majority of the predicted "most disruptive" mutations included those previously identified as conserved or essential. In addition, we discovered substitutions within the expanded least-mutable set that were not predicted to disrupt structure. These substitutions include some of the mutants used for complementation and validated to be impaired for T7 growth.

To similarly examine mutability of other T7 proteins with available PDB structures, we applied this analysis using structures of T7 RNAP and T7 DNAP (Fig. 4 and Dataset S3). We expected the least-mutable residues to either interfere with protein structures or to interfere with catalysis or interactions with partner proteins. As with T7 SSB, there is a trend for residues with large ΔΔG values to exhibit reduced or minimal NMI values. A majority of the catalytic residues have NMI values less than 3 and when substitutions with high NMI were found in gene *5*, these mutations were not predicted to be structurally disruptive to the encoded protein.

Mutability Changes When T7 RNAP Is Provided *in Trans*. To examine changes in mutability in the apparent absence of selection, we used the T7 RNAP gene. The primary difference between replicate 1 and replicate 2 is that the mutated T7 phage pool was grown and plated separately on both lawns of *E. coli* Bl21 and Bl21 DE3, a strain expressing a copy of the gene *1* (T7 RNAP) from the chromosome. We found the difference in mutability for gene *1* was not striking, although there is some level of dissimilarity. The average NMI for substitutions in gene *1* that created premature stop codons and nonsynonymous mutations at essential residues were measured to be low in both replicates. These NMI values for predicted nonpermissive substitutions were found to be only 2× higher for mutant phages plated on Bl21 DE3 (Fig. 5), suggesting that these base substitutions were still deleterious or less permissive for growth when T7 RNAP was expressed *in trans*.

Nonmutable Regions in T7 RNAP Correspond to Essential Motifs and Residues. T7 RNAP is a single-subunit enzyme with well-defined catalytic domains. Like other polymerases, the structure of T7 RNAP has been described as a "cupped right-hand" and, accordingly, many of the relevant subdomains are aptly named as

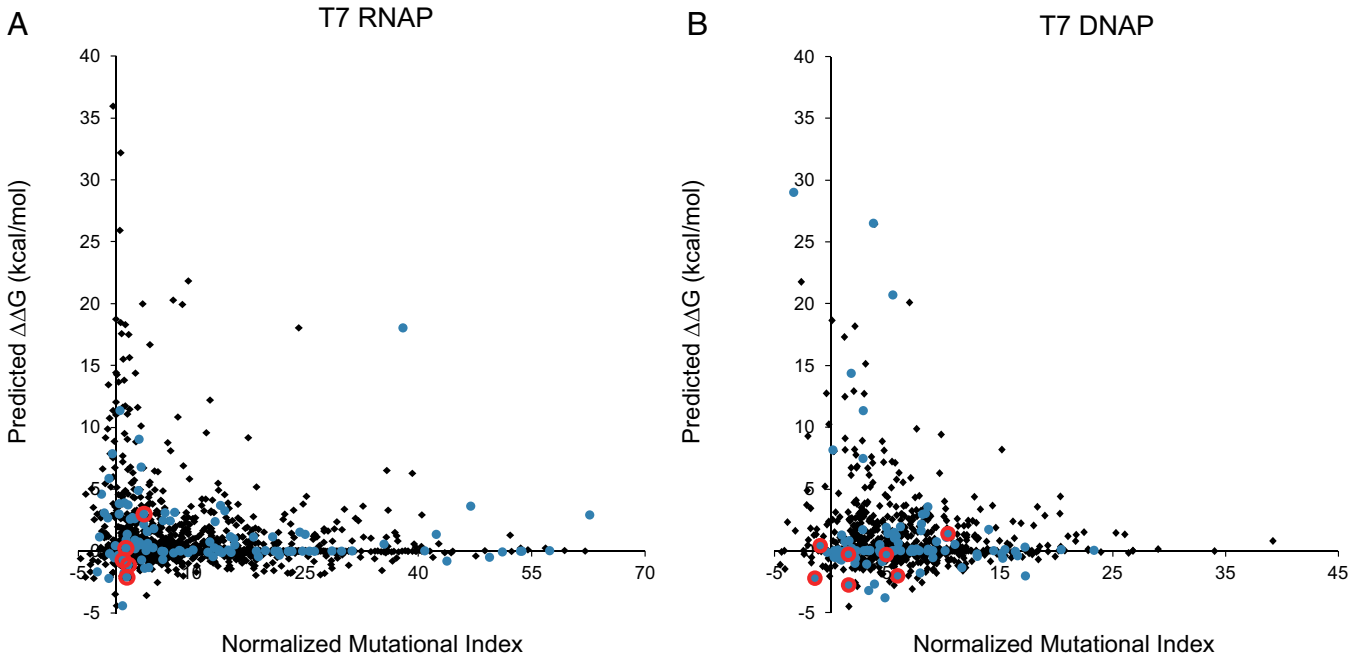

Fig. 4. Correlation of NMI and predicted ΔΔG for T7 DNAP and T7 RNAP. (*A*) Predicted ΔΔG and NMI values for all T7 gene *1* (T7 RNA polymerase) positions averaged for only 1A/1B were plotted. By definition, all synonymous mutations had ΔΔG values of 0. Some are nonconserved and conserved (filled blue circles) and also determined to be essential (open red circles) by prior work and marked accordingly. (*B*) Predicted ΔΔG and NMI values for all T7 gene *5* (T7 DNA polymerase) positions averaged for 1A/1B and 2A/2B were plotted. By definition, all synonymous mutations had ΔΔG values of 0. Some are nonconserved and conserved residues (filled blue circles) and also determined to be essential (open red circles) by prior work and marked accordingly.

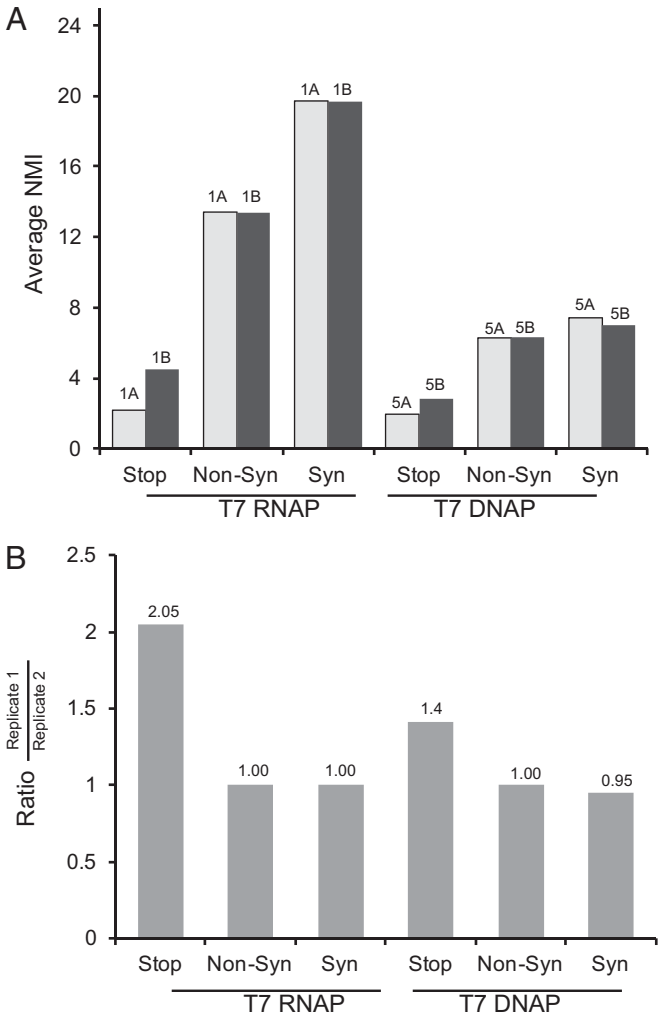

Fig. 5. Mutability of the T7 RNAP gene contrasts when complemented. (*A*) Comparing the average NMI for replicates 1 and 2 for stop, synonymous, and nonsynonymous mutations for gene *1* (T7 RNAP) and gene *5* (T7 DNAP). (*B*) The ratio of averaged stop, synonymous, and nonsynonymous NMI values between biological replicates 1A/B and 2A/B for gene *1*(T7 RNAP) and gene *5* (T7 DNAP). Replicates for gene *1* (T7RNAP) were expected to differ because the pool of phage for replicate 1A/B was selected on BL21; replicates 2A/B was selected on Bl21 DE3, a strain expressing T7 RNAP *in trans* from the host chromosome.

"palm," "thumb," and "fingers" (10). Residues in the finger and palm domains are folded into close proximity to form the catalytic pocket/active site and include positions known to be absolutely essential (11). Three motifs (A, B, C) are occupied by pivotal conserved and catalytic residues found other RNA and DNA polymerases (10, 12). In addition, the DX2GR motif, conserved in both DNA and RNA polymerases, occupies the palm domain and is shown to contact and stabilize the RNA: DNA hybrid (13). These motifs and other conserved residues scattered throughout the protein were implemented as benchmarks to further assess the mutational profile of the gene. Furthermore, 92% of substitutions that created nonsense mutations in gene *1* were found to have NMI values less than 5; thus, this value was chosen as a threshold for least permissive.

Mutations in conserved residues exhibited low mutability or were synonymous; this included specific residues in the ABC motifs that have been documented to be deleterious when mutated and that are known to be invariant in other DNA-dependent RNA polymerases (Fig. 6 *C–E*) (12). Some of these critical residues (K631 and Y639) cannot be mutated by HA because of their base composition, but the others were found to be among the least mutable positions in the protein. In the conserved DX2GR motif we found some residues exhibited similar levels of nonmutability (Fig. 6*A*). A decreased NMI was also measured in the last four C-terminus residues, referred to as the "foot" (Fig. 6*F*). Residues in this hydrophobic region are found to be flexible and contact residue D812 proximal to the active site. Evidence suggests these foot residues are important for magnesium-dependent catalysis and interaction with the incoming nucleotide and downstream DNA (14, 15). Notably, the NMI for synonymous mutations was found to be consistently increased compared with nonsynonymous mutations. Synonymous mutations comprised between 60% and 100% of those with NMI greater than 5. Conversely, only about 10% of synonymous changes exhibited low NMI (NMI < 5) values. Because some of these motif residues can be mutated to either class, these data provide a strong internal control for the observed skew in mutability between synonymous vs. nonsynonymous mutations.

One of the powerful applications we envision for Mut-seq is to confirm functions of gene products inferred by homology but for which no biochemistry is available. We reasoned that a similar NMI map across conserved residues of a heterologous gene that was selected for biological function would provide proof that that gene's product likely performed the same function as its biochemically characterized ortholog. To test this idea, we performed Mut-seq on *V. cholerae* phage JSF7, a podophage that encodes two T7 RNAP-like proteins. It is unclear why this T7-like phage possesses two polymerase genes, one positioned early as in T7 and the other positioned in the middle of its 46-kb genome after the genes for DNA replication proteins. When aligned to T7 RNAP, the coding sequences for ORF4 and ORF37 possess 17% and 28% amino acid homology to T7 RNAP, respectively (Fig. S2).

The observed MC for important residues was consistent with both phage RNAP polymerases being essential (Dataset S4). Base substitutions that produced nonsense mutations exhibited much lower counts than those that produced synonymous and nonsynonymous substitutions. Unexpectedly, we also found that G-to-A mutations were overrepresented significantly compared with C-to-T changes. We hypothesize that this G-to-A bias is a consequence of strand-specific DNA replication that results in only one of the strands being copied into the template used for DNA subsequent replication early during infection.

Active site and conserved motifs are to be found in both; however, the C-terminal foot motif is missing in ORF4. As shown in Fig. 6*G*, the landscapes of NMI values for residues in key motifs that are conserved between the T7 RNAP and both putative RNAPs of phage JSF7 were indeed very similar and included the well-characterized invariant T7 D537 and D812 positions and a number of neighboring residues shown to be important for enzyme activity in a number of biochemical studies (11, 13, 15, 16). This result provides strong genetic and biological evidence that the putative RNA polymerases are probably both active RNAPs, and that the conservation of amino acid sequence reflects the same biochemical constraints for functionality of these three heterologous enzymes.

Discussion

We have described a method, termed Mut-seq, which allows the very efficient identification of putative essential residues in genes and genomes. Other previous work has coupled chemical mutagenesis with phenotypic selection and deep sequencing to successfully identify individual residue mutations that impose defects in a selected pathway (17). The key attribute of this method is the ability to resolve, through deep sequencing and subtractive analysis, both deleterious as well as tolerated mutations, from simple mutational noise associated with PCR and other sequencing technologies. This process was done by comparing a mutated to a

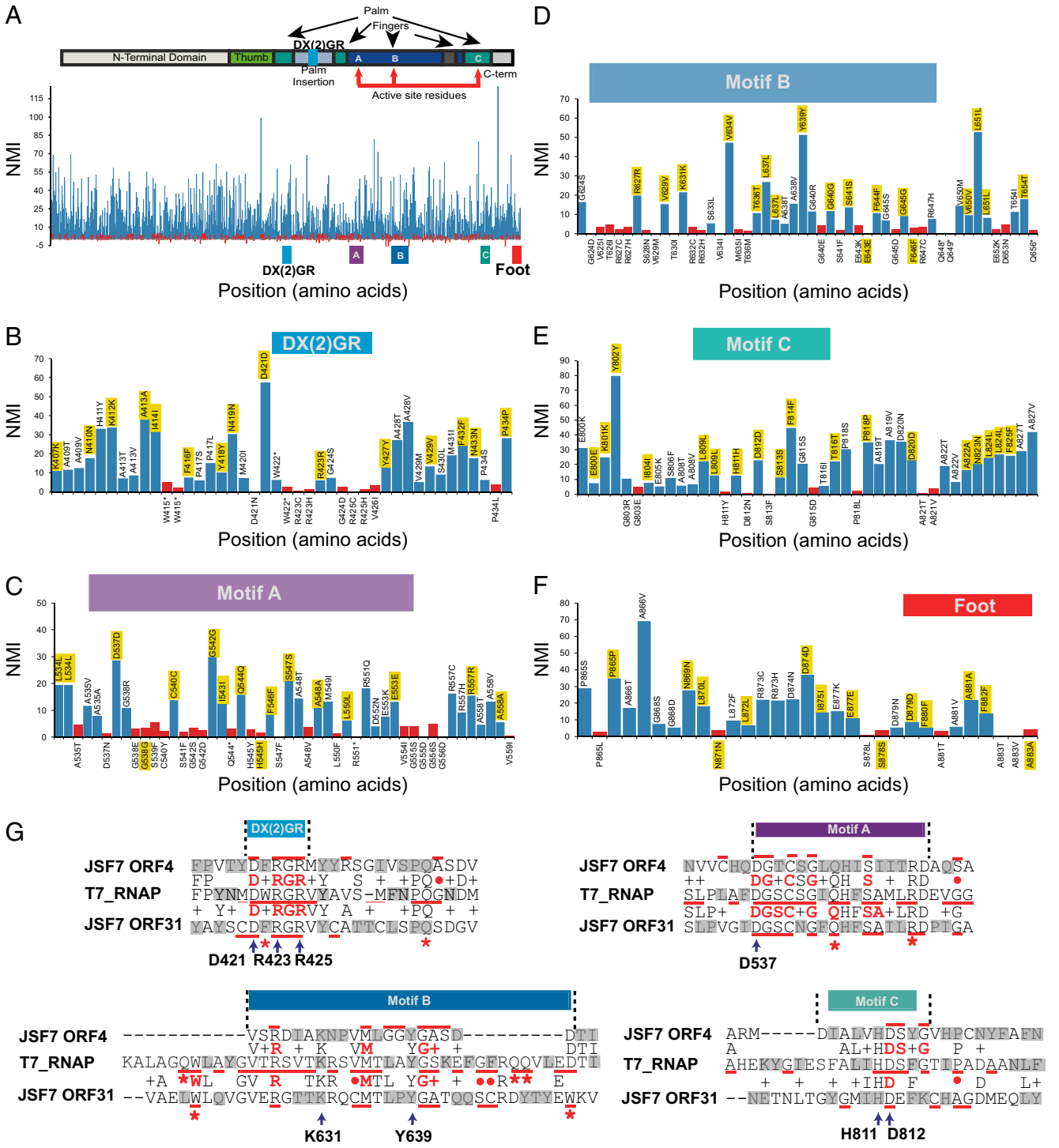

Fig. 6. Residues in T7 RNAP and two JSF7 RNAPs are exhibit low mutability in conserved domains and motifs. (*A*) Graphed NMI for all mutable positions across T7 RNAP and marked locations of important motifs and subdomains. Synonymous and nonsynonymous mutations proximal to the (*B*) DX2GR domain, (*C*) motif A, (*D*) motif B, (*E*) motif C, and the (*F*) foot domain. NMI values are bar graphed according to NMI values in blue unless the NMI value is less than 5 and then it graphed in red. Synonymous mutations are highlighted in yellow and are predominantly measured to be more mutable. (*G*) Both T7-like RNAPs found in JSF7 are aligned to T7 RNAP to show conservation of amino acid sequence and mutability. Conserved motifs are mapped. Residues and corresponding T7 RNAP positions known to significantly impair the enzyme activity when mutated, or T7 growth, are indicated by arrows and labeled. Residues that cannot be mutated to a nonsynonymous residue by HA are shaded. If a residue is measured to be nonmutable, it is underlined in red and when immutability is conserved, the conserved residue is colored red. If no conservation is apparent, this is replaced by a red dot. Premature stop codons are indicated as to separate from true nonsynonymous mutants.

nonmutated pool of targets and then imposing a quality-score filter to map and measure the frequency of specifically HA mutagen-induced mutations. We predict that recent approaches developed to address the level of mutational noise when applied to Mut-seq databases will allow even deeper mapping of true mutations that survive selection (18).

In applying Mut-seq it is important to achieve a high level of mutagenesis to confidently detect and measure the mutability of residues susceptible to mutagen-induced changes. Conceptually, we predicted that less than one mutation per gene or genome would ideally ensure that each mutation was being scored for its ability to permit or prevent function of the gene in the context of the biological selection imposed (e.g., phage growth). Here we exceeded one mutation per genome, which could have interfered with our objective to measure the effects imposed by each single residue change in the absence of other mutations. However, within a protein or genome, it seems highly unlikely that at 3.8 mutations per genome, suppressor mutants or synthetic lethal double or triple mutants significantly biased NMI values in this study.

It should be noted that infrequent substitutions that created synonymous mutations were measured to have very low NMI values, and some premature stop codons in essential genes had higher than expected NMI values that suggested permissiveness. We attribute some of these observed variations to the read-through of stop codons. The efficiency of translational termination for stop codons varies based upon the identity of the triplet and neighboring bases (19, 20). Why some synonymous codons in essential genes appear to be nonpermissive is less clear. Rigorous statistical analysis of stop and all codon use has been completed for bacteriophage T7 (21) and coupling this sort of methodology to the mutagenic frequency of each kind of codon change may help explain these discrepancies. We expect there are a number of stand-alone analyses that can be applied to generated Mut-seq datasets.

Although polarity is a complicating factor in prokaryotic translation-coupled transcription, it was not considered to be relevant here. Studies found that amber mutations did not appear to have a polar effect of T7 RNAP transcription of T7 DNA (22). Host transcription and intrinsic termination of early T7 genes is shown to be unaffected by polarity suppressors (23). Furthermore, host transcription of DNA from of the early promoters of T7 is antiterminating, and thus minimizes ρ-mediated termination during transcription of early genes (24).

For bacteriophage T7, the massive assembly of individual mutations that vary in frequency provides an additional resource for probing important and essential residues in proteins of interest. The increased panel of newly identified nonmutable and highly mutable residues in transcription and replication proteins may illuminate new targets for understanding requisite mechanisms. Similarly, there are nucleotide positions in intergenic regions that also exhibit immutability (Dataset S5) that may be useful for investigating *cis*-acting regulatory sequences. Many of these 60 proteins and motifs are conserved within genes present in other T7-like phages, and thus may provide a new resource for understanding the biology of this virus family. By probing the mutability of residues found in genes encoding two putative single-subunit RNA polymerases encoded by vibriophage JSF7, we also demonstrated that conservation of residues in encoded gene products reflects their essential functionality, even when gene products are evolutionarily distantly related.

To perform Mut-seq effectively, a strong positive selection is essential; this was achieved here because phage DNA must be ejected, transcribed, replicated, and packaged in an infectious virion to be recovered efficiently from a virus plaque. Clearly, one can apply Mut-seq to analyze other protein or viral targets (e.g., those needed for antibiotic resistance or bacterial growth) by simply using appropriate plasmid or virus expression systems that allow functional selection of target protein's function. Furthermore, one can apply Mut-seq to define tolerated mutations and nontolerated mutations in different selective environments (e.g., growth of an animal virus in tissue culture versus growth in an immunologically naive or immunized experimental animal). Such an analysis would likely contribute to our understanding of requirements for growth, tissue tropism in vivo, and escape of immune responses. The mining of databases that contain the diversity of polymorphisms found in HIV genomic sequences provides another example of how a Mut-seq database might be mined to define fitness landscapes for a mutagenized target gene or genomic sequence and in that way inform the design of better immunogens (25, 26). Moreover, besides mapping essential and nonessential residues in proteins of interest, Mut-seq databases may provide a new source of valuable information for small molecule drug design. By knowing in advance which residues of a target protein are mutable and which are not, we envision that crystallographers and chemists will be able to more confidently design small molecules that engage essential residues of the target while minimizing contacts with nonessential residues. Such an approach is likely to minimize the likelihood of evolved drug resistance through the mutation of nonessential amino acid residues in the target protein.

Another useful application of Mut-seq will be in the design of live-attenuated viral vaccines, one of the most historically efficient means of producing safe, immunogenic, and protective immunoprophylactics (27–33). The precision that Mut-seq allows identification of mis-sense mutations that reduce viral fitness but do not block replication, would likely allow investigators to deduce combinations of mutations that should show a desired combined level of attenuation. By application of genome synthesis methods (34–37), designers could construct a mutant viral genome that carries a combination of fitness-reducing mutations identified by Mut-seq analysis; such a mutant virus would be predicted to have a fleetingly low chance of reversion to wild-type. Even incremental steps in the reversion of such a virus could be further monitored quantitatively by Mut-seq analysis of the progeny of this genetically engineered attenuated virus after growth in the host. Finally, by using Mut-seq to define essential genes and residues for growth in an experimental host animal with those required for growth on cell lines in vitro, investigators should be able to define virulence and host-specific fitness genes that do not alter in viral fitness for manufacture in vitro. Thus, Mut-seq should see applications in the design of better live-attenuated viral and perhaps bacterial vaccines.

Methods

Strains, Phage, and Plasmids. *E. coli* strains BL21[*fhuA2 (lon) ompT gal)dcm] ΔhsdS*] and BL21 DE3 [*fhuA2 (lon) ompT gal (λ sBamHIo ΔEcoRI-B int::(lacI::PlacUV5::T7 gene1) i21 Δnin5) (dcm) ΔhsdS*] were purchased from New England Biolabs. *E. coli* strains HMS157 (F-, recB21 recC22 sbcA5 endA gal thi sup) (38) and JW5856 [F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, rph-1, ΔtrxA732::kan, Δ(rhaD-rhaB)568, hsdR514] (39) have been previously described. Bacteriophage T7 was a kind gift from Ian Molineux (University of Texas at Austin, Austin, TX). The stock of T7 used for this study has two point mutations, a G-to-T mutation at 15094 (Gene 4 A248S) and an A-to-G mutation at 29258 (Gene 16 K312G). These mutations were discovered during sequencing and appropriate changes were made to the reference sequence during mapping.

To prepare purified phage particles from liquid cultures, T7 phage was added to rapidly shaking flasks containing well-aerated 150-mL cultures of exponentially growing Bl21 or Bl21:DE3 cultures in LB at 30 °C. Upon complete lysis and clearing, lysates were made 1 M NaCl and cell debris was removed by centrifugation. To purify phage from plaques grown in soft agar, cold 30 mL LB was added to the soft agar overlay of 150-mm plates, top agar was scraped, allowed to sit for 15 min at 4 °C, and then LB was collected with scraped top soft agar, vortexed, and centrifuged to removed bacteria and agar. Phage was precipitated overnight on ice by addition of 8% (wt/vol) PEG, resuspended in 8 mL cold 100 mM NaCl, 50 mM Tris pH 8.0, and purified by ispoynic density gradient centrifugation in CsCl.

The T7 phage possessing the gene *2.5* disruption (T7Δ2.5::*trxA*) was constructed by digesting T7 DNA with MluI and then ligating a MluI-site flanked PCR product of *E. coli trxA* gene linked to an upstream Shine-Delgarno and disrupting gene *2.5*. The gene was not removed and replaced by *trxA* as before (38) in the case gene *2.5* mutants were to be recombined back to replace *trxA* to test complementation in the context of phage-directed expression. Ligated DNA was transfected into competent HMS157 pTopo-2.5 using the calcium chloride-shocked cells (40) and plaques were tested and purified. When the phage with the disrupted gene *2.5* was confirmed, it was subsequently propagated on BL21 pTopo-2.5.

The plasmid pTopo-2.5 was constructed by amplifying the gene from T7 DNA and cloning it under T7 promoter control in pTopo-2.1 (Invitrogen). The corresponding set of cloned point mutants were cloned into pTopo-2.1 and sequenced to confirm accuracy and directionality to the T7 promoter.

Burst Size and EOP U T7 Gene *2.5* Mutants. Burst size and EOP measurements for T7 gene *2.5* mutants was completed by infecting set of *E. coli* BL21 strains transformed with the corresponding set of mutant and wild-type pTopo-T72.5 plasmids. To determine EOP, the T7Δ2.5::trxA phage was plated on each a cells expressing each gene *2.5* mutant and by dividing the permissive titer when plated on Bl21 pTopo-2.5(WT). For phage burst size measurements, each complementing strain was grown in liquid at 30 °C and 5 mL (5×10^8 cells/mL) were infected with T7Δ2.5:*trxA* at a multiplicity of 0.1. At 16 min, cells were diluted 10,000× into 5 mL 30 °C LB and a 1.0 mL aliquot was vortexed with chloroform to kill infected cells, then centrifuged at 13,500 × *g* for 2 min using an Eppendorf 5424 micro centrifuge, and the supernatant was titered on Bl21 pTopo2.5(WT) to measure unabsorbed phage. The remaining infected culture was maintained shaking at 30 °C and the aliquots were taken and titered on BL21 pTopo2.5(WT) at increments of 20, 30, 35, 40, 45, and 50 min postinfection and incubated overnight at 30 °C in tryptone top agar. Phage titers increased between 30 and 45 min and then plateaued at 50 min. The titer at 20 min minus that unabsorbed was used to calculate the number of infecting particles. The burst per infected cell was calculated by dividing the titer measured at 50 min postinfection by the initial infecting titer.

Mutation of Phage. Purified T7 phages were treated with HA to mutate the genome in vitro before infection at a frequency that maximized both mutation and yield of infectious particles. HA permeates the viral capsid and modifies the 4-carbon of the cytosine pyrimidine ring via an addition of a hydroxyamino group. This process generates a distinct class of G:C to A:T transitions (41–43). We prepared HA stocks as done previously and recommended for other phage mutagenesis protocols (44). To a sterile tube, 0.33 g of HA and 560 μL of 4 M NaOH is brought to 2.5 mL to make a 2 M HA (pH 6.0) solution. As done previously, we treated the phage with several concentrations of HA between 0.1 M and 1.0 M HA for 24 h at 4 °C, dialyzed to 100 mM NaCl, 50 mM Tris pH 8.0, and plated to select for a pool that exhibited a reduction in titer of about 2–3 $\log_{10}$. This result is reported to correspond close to one mutation per unit length genome (44). Furthermore, in this reduction, a significant portion of phages are also inactivated by HA because it cleaves peptide bonds between asparagine and glycine (45). Although absent in the abundant virion capsid subunits, there are between five and eight Asn-Gly dipeptides in two of the internal core virion proteins and those that assemble into the tail tube and fibers. Of the phages that do infect and produce infective centers, one might expect each of these plaques originating from a viable wild-type or permissive mutant genome packed in an intact virion. To maximize independent mutations, phage stocks were treated with the mutagen and plated at a concentration chosen to generate a dense lawn of separated, individual plaques to avoid recombination between unrelated phage. Mutated phages were plated and selected on BL21 or BL21 DE3 on a 150-mm plate at a density of about 75,000 1.0-mm plaques per plate. Phage from 15 150-mm plates were pooled and purified. Because 10 billion phages were treated and only about 0.1% recovered, each recovered phage possessed an average of 3.8 mutations. At this density, given ~150,000,000 × 50-nt reads, every one of the 19,329 G/C residues would be represented by about 36 independent mutations, providing ample opportunity to sample every HA-induced residue change.

Library Preparation and Sequencing. CsCl-purifed phage was dialyzed overnight in 1 L 4 °C 50 mM Tris•Cl, 1 mM EDTA pH 8.0. DNA was extracted from dialyzed T7 phage using a one-third volume of Tris•Cl (50 mM) equilibrated mixture of phenol:chloroform:isoamyl alcohol (25:24:1) pH8.0. We vortexed this mixture to disrupt phage virions and incubated at 50 °C (20 min) to separate the top aqueous phase-containing nucleic acid from the protein interface and the phenol:cholorform:isoamyl. This process was done in triplicate and DNA was precipitated in cold ethanol, washed twice with 80% ethanol, dried, and dissolved in water. DNA was sheared at 4 °C using sonication (QSonica Q800R) for 30 min at 60% amplitude to produce an average fragment size of ~200 bp. Libraries were built using NEBNext DNA Library Mastermix kit protocol and amplified with standard multiplex Illumina primers. Sequencing was achieved using 50 cycles (single end) on the Illumina HiSEq 2000 system and analyzed using the CASAVA 1.8.2 Illumina Data analysis pipeline.

Aligning Reads and Mapping Substitutions. Reads were mapped to the complete reference T7 nucleotide sequence or to the nucleotide sequence of each JSF7 RNAP gene (ORF4 and ORF31) and then filtered using CLC-Genomics Workbench 4.8. First, all perfectly matching 50-nt reads (those lacking mismatches) were mapped and separated. Remaining reads that mapped and aligned with only one mismatch were kept and the quality scores were used to further vet base substitutions with greater confidence. A CASAVA 1.8 quality score of 38 (of 40) or higher was applied as a cutoff for the mismatch and the flanking 11 nt on either side. The identity and quantity of each single mutation at each position was tabulated. This table of counts was cross-referenced to all possible HA-induced mutations (G/C sites) to map total substitution counts at each position. No mutations were mapped to either of the 115 terminal bases as they are repetitive and not unique. Because the 50-nt read length anchors some of 165-nt repeats to unique adjacent sequences, the unique region of the T7 genome was extended to include these regions.

In the JSF7 mutant pool we found a large proportion of G positions to be changed, whereas C-position substitutions were largely underrepresented. This artifact may be explained by strand-specific replication of one strand during the phage infection (i.e., rolling circle) and all second-strand replication occurring on only newly synthesized DNA. Thus, G-to-A transitions can be explained directly by HA-mutagenesis, whereas the C-to-T change would normally require this mutation to be introduced indirectly during replication of second strand. Here Mut-seq has identified the first strand replicated and we have only included G-specific mutational changes for JSF7 in the analysis pipeline.

Identification of Phage Mutations and Mapping Residues Changes in Each Protein. Each annotated T7 gene was extracted from GenBank accession V01146 as a FASTA nucleotide file. Genes *6B* and *10B* are both products of translational frame-shifts and these FASTA file adjustments were made to represent the coding sequence, accordingly. The identity of every nucleotide position was examined and for each single G or C, the corresponding HA-induced A or T change was introduced into a new FASTA file. These FASTA files were translated into amino acid FASTA format, aligned to the reference FASTA with BLASTP, and a table of all synonymous and nonsynonymous changes was recorded. The complete list of synonymous and nonsynonymous residue substitutions for every possible HA-induced mutation provided a reference for the substitutions mapped in each replicate.

FoldX in Silico Structural Prediction. Multiple PDB files for each of the T7 SSB, RNAP, and DNAP proteins are available from the Research Collaboratory for Structural Bioinformatics PDB protein data bank. Within the scope of this work, a single PDB; 1JE5.PDB (T7 SSB), 1QLN.PDB (T7 RNAP), and 1T7P.PDB (T7 DNAP) were selected to measure the ΔΔG for each mutation. To run simulations of all possible HA-induced substitutions as a batch, all possible HA-induced mutations, with the exception of premature stop codons, at residues that are present within each PDB file were tested in a standalone FoldX package (FoldX v3.0 beta 5.1). The total predicted ΔΔG was tabulated for each possible mutated residue and then compared with NMI values.

ACKNOWLEDGMENTS. The authors thank Steve Lory and Ian Molineux for thoughtful comments regarding the work presented in this manuscript. This study was supported by Grant 2R01GM068851-09 from the National Institute of General Medical Sciences.

1. Wetterstrand KA (2012) DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program. Available at http://www.genome.gov/sequencingcosts/. Accessed December 1, 2012.
2. Punta M, et al. (2012) The Pfam protein families database. *Nucleic Acids Res* 40(Database issue):D290–D301.
3. Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. *J Mol Biol* 257(2):342–358.
4. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 31(13):3812–3814.
5. Cline J, Braman JC, Hogrefe HH (1996) PCR fidelity of pfu DNA polymerase and other thermostable DNA polymerases. *Nucleic Acids Res* 24(18):3546–3551.
6. Kemp P, Garcia LR, Molineux IJ (2005) Changes in bacteriophage T7 virion structure at the initiation of infection. *Virology* 340(2):307–317.
7. Rezende LF, Hollis T, Ellenberger T, Richardson CC (2002) Essential amino acid residues in the single-stranded DNA-binding protein of bacteriophage T7. Identification of the dimer interface. *J Biol Chem* 277(52):50643–50653.
8. Hyland EM, Rezende LF, Richardson CC (2003) The DNA binding domain of the gene 2.5 single-stranded DNA-binding protein of bacteriophage T7. *J Biol Chem* 278(9):7247–7256.
9. Schymkowitz J, et al. (2005) The FoldX Web server: An online force field. *Nucleic Acids Res* 33(Web Server issue):W382–W388.
10. Sousa R, Chung YJ, Rose JP, Wang BC (1993) Crystal structure of bacteriophage T7 RNA polymerase at 3.3 A resolution. *Nature* 364(6438):593–599.

11. Bonner G, Patra D, Lafer EM, Sousa R (1992) Mutations in T7 RNA polymerase that support the proposal for a common polymerase active site structure. *EMBO J* 11(10):3767–3775.
12. Delarue M, Poch O, Tordo N, Moras D, Argos P (1990) An attempt to unify the structure of polymerases. *Protein Eng* 3(6):461–467.
13. Imburgio D, Anikin M, McAllister WT (2002) Effects of substitutions in a conserved DX(2)GR sequence motif, found in many DNA-dependent nucleotide polymerases, on transcription by T7 RNA polymerase. *J Mol Biol* 319(1):37–51.
14. Lykke-Andersen J, Christiansen J (1998) The C-terminal carboxy group of T7 RNA polymerase ensures efficient magnesium ion-dependent catalysis. *Nucleic Acids Res* 26(24):5630–5635.
15. Gardner LP, Mookhtiar KA, Coleman JE (1997) Initiation, elongation, and processivity of carboxyl-terminal mutants of T7 RNA polymerase. *Biochemistry* 36(10):2908–2918.
16. Tunitskaya VL, Kochetkov SN (2002) Structural-functional analysis of bacteriophage T7 RNA polymerase. *Biochemistry (Mosc)* 67(10):1124–1135.
17. Nguyen BD, Valdivia RH (2012) Virulence determinants in the obligate intracellular pathogen *Chlamydia trachomatis* revealed by forward genetic approaches. *Proc Natl Acad Sci USA* 109(4):1263–1268.
18. Schmitt MW, et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci USA* 109(36):14508–14513.
19. Tate WP, et al. (1995) Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. *Biochem Bell Biol* 73(11–12):1095–1103.
20. Poole ES, Brown CM, Tate WP (1995) The identity of the base following the stop codon determines the efficiency of in vivo translational termination in *Escherichia coli*. *EMBO J* 14(1):151–158.
21. Sharp PM, Rogers MS, McConnell DJ (1984-1985) Selection pressures on codon usage in the complete genome of bacteriophage T7. *J Mol Evol* 21(2):150–160.
22. Studier FW (1972) Bacteriophage T7. *Science* 176(4033):367–376.
23. Kiefer M, Neff N, Chamberlin MJ (1977) Transcriptional termination at the end of the early region of bacteriophages T3 and T7 is not affected by polarity suppressors. *J Virol* 22(2):548–552.
24. Sedgwick WT (1915) American achievements and American failures in public health work. *Am J Public Health (N Y)* 5(11):1103–1108.
25. Dahirel V, et al. (2011) Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. *Proc Natl Acad Sci USA* 108(28):11530–11535.
26. Allen TM, et al. (2005) Selective escape from $CD8^+$ T-cell responses represents a major driving force of human immunodeficiency virus type 1 (HIV-1) sequence diversity and reveals constraints on HIV-1 evolution. *J Virol* 79(21):13239–13249.
27. Vignuzzi M, Wendt E, Andino R (2008) Engineering attenuated virus vaccines by controlling replication fidelity. *Nat Med* 14(2):154–161.
28. Hanley KA (2011) The double-edged sword: How evolution can make or break a live-attenuated virus vaccine. *Evolution (N Y)* 4(4):635–643.
29. Plotkin SA (2009) Vaccines: The fourth century. *Clin Vaccine Immunol* 16(12):1709–1719.
30. Weyer J, Rupprecht CE, Nel LH (2009) Poxvirus-vectored vaccines for rabies—A review. *Vaccine* 27(51):7198–7201.
31. Amanna IJ, Slifka MK (2009) Wanted, dead or alive: New viral vaccines. *Antiviral Res* 84(2):119–130.
32. Anonymous (1999) From the Centers for Disease Control and Prevention. Ten great public health achievements—United States, 1900–1999. *JAMA* 281(16):1481.
33. Anonymous (1999) From the Centers for Disease Control and Prevention. Impact of vaccines universally recommended for children—United States, 1900–1998. *JAMA* 281 (16):1482–1483.
34. Tian J, Ma K, Saaem I (2009) Advancing high-throughput gene synthesis technology. *Mol Biosyst* 5(7):714–722.
35. Liu Y, et al. (2012) Whole-genome synthesis and characterization of viable S13-like bacteriophages. *PLoS ONE* 7(7):e41124.
36. Yang R, et al. (2011) Chemical synthesis of bacteriophage G4. *PLoS ONE* 6(11):e27062.
37. Matzas M, et al. (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28(12):1291–1294.
38. Kim YT, Richardson CC (1993) Bacteriophage T7 gene 2.5 protein: An essential protein for DNA replication. *Proc Natl Acad Sci USA* 90(21):10173–10177.
39. Baba T, et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol Syst Biol* 2:2006–2008.
40. Benzinger R (1978) Transfection of Enterobacteriaceae and its applications. *Microbiol Rev* 42(1):194–236.
41. Freese E, Bautz E, Freese EB (1961) The chemical and mutagenic specificity of hydroxylamine. *Proc Natl Acad Sci USA* 47:845–855.
42. Schuster H (1961) The reaction of tobacco mosaic virus ribonucleic acid with hydroxylamine. *J Mol Biol* 3:447–457.
43. Franklin RM, Wecker E (1959) Inactivation of some animal viruses by hydroxylamine and the structure of ribonucleic acid. *Nature* 184:343–345.
44. Villafane R (2009) Construction of phage mutants. *Methods Mol Biol* 501:223–237.
45. Bornstein P, Balian G (1977) Cleavage at Asn-Gly bonds with hydroxylamine. *Methods Enzymol* 47:132–145.