Supplementary Data

I. Algorithm to select mutation mapping probes for resequencing.

A custom algorithm for analyzing mapping array data that are generated as the first step in the CGR process was developed and optimized: 1) to identify possible SNPs for analysis in the resequencing array used in the second step of CGR; and 2) to remove large deletions or segments with many clustered point mutation differences from consideration as these are not effectively assayed using our resequencing arrays. The algorithm first produces a global threshold that identified deletions or complex differences that can be omitted from the resequencing array, and then uses local variances to rank potential SNP. Oligonucleotides to examine these sites are added to the resequencing array, prioritized according to the rank of the corresponding possible SNP until the array’s capacity is reached. Since most possible SNPs are mapped by one to four probes, between 29 and 50 bases are resequenced on each strand. This involves 232 to 400 custom oligonucleotides per SNP or up to 1200 SNPs resequenced per array (assuming an average of 40 bases studied for each possible SNP). If analysis of more sites is needed, more resequencing arrays are generated. Such capacity was not important in the present study because rather few mutations (<10) were expected per strain analyzed, although our inclusion of all even remotely possible SNP sites ensured discovery of the great majority of bona fide SNPs. This great capacity does become important when analyzing closely related bacterial isolates from Nature, which may often have diverged at hundreds of positions.

Deletions and tight clusters of point mutation differences, which are not suitable for resequencing, are identified based on global medians and standard deviations of sets of forward and reverse probe intensity ratios (reference/test) and then defining a threshold for exclusion of sites of DNA sequence difference from the resequencing array. This threshold is determined separately for forward and reverse probes as follows: 1) any data point more than 4 times the global median value is removed; 2) the 80th percentile and standard deviation of the remaining data are calculated; 3) any value 3.5 standard deviations above the 80th percentile is removed; 4) the 80th percentile and standard deviation are calculated again for the remaining data; 5) the threshold is set at 3.5 standard deviations above the 80th percentile of this final data set. If the local median (calculated below) for any data point falls above this threshold, that site is excluded from the resequencing array.

Next, probe positions are ranked so that probes representing the most likely SNP loci are added first to the resequencing array. To rank probes the 80th percentile and standard deviation of all probes within 1,800 bases is calculated, and number of standard deviations each probe is above or below the 80th percentile is determined. The difference between the 80th percentile and the standard deviation is calculated and used as the rank score. If the local rank score for corresponding forward and reverse probes at the same position are 3.5 standard deviations above the 80th percentile of the local data window, 10,000 is added arbitrarily to the rank score, thereby ensuring its automated inclusion in the resequencing array until the capacity is reached.
II. Algorithm for selection of resequencing oligos.

An algorithm was developed to design probes for resequencing arrays that optimizes the oligonucleotide length, mismatch position, and melting temperature. The algorithm accepts input from the user to specify the minimum and maximum probe lengths and target T<sub>m</sub> for all probes on the array. The target T<sub>m</sub> of the probe is then divided in half, and the portion on each side of the mismatch is varied (within specified length parameters) to reach half of the specified target T<sub>m</sub>, calculated as follows:

\[
\text{Probe } T_m = 5*(G_n+C_n) + 1*(A_n+T_n)
\]

where \(G_n\) is the number of Gs, \(C_n\) is the number of Cs, and so-on for the other bases in the probe. This is an empirical modification of the Wallace rule\(^1\), made to better reflect surface probe behavior. This results in array probes with similar T<sub>m</sub>s and with mismatches near the thermodynamic center of each oligonucleotide, where they are most destabilizing. For the current study, oligos were between 29 and 39 nucleotides in length, and had T<sub>m</sub>s of 72°C.

III. Algorithms for SNP “uniqueness” testing.

Because the entire genome is hybridized to each resequencing array, it is important to know potentially cross hybridizing sites that could result in false positive identification of a SNP. The term uniqueness is used here to describe the likelihood of hybridization of each resequencing probe to another site, and thereby interference with mutation identification. This is determined by calculating sequence similarities between each resequencing probe and all sites in the genome sequence. Each probe is checked for uniqueness in two ways. First, a weighted measure of how often perfect matches for increasing numbers of central bases in each probe are found in the genome is calculated: the “Nmer frequency”. Each such perfect match contributes to the Nmer frequency score, with shorter oligos contributing less than longer oligos according to the following equation:

\[
\text{Nmer frequency score} = \sum [\text{N-mer frequency}]^* [0.75^\text{(29mer-Nmer length)/2}]
\]

The sum of the frequency scores for central Nmers from 19, 21, 23, 25, 27, and 29mers is calculated for each Probe calling a SNP. The higher the Nmer frequency score, the more likely it is that the probe will also hybridize elsewhere in the genome. Any SNP being called by a probe with an Nmer frequency above zero is considered a likely false positive.

For the second uniqueness test a “sequence dissimilarity” score is calculated by comparing the 29 bases spanning the SNP to every possible 29 base window in the reference genome. Mismatches at the end of a probe tend to disrupt hybridization less than centrally located mismatches, and hence are weighted differently. The weight value vs. position used in the current study is listed below:

| 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 1  | 1  | 1  | 2  | 2  | 2  | 2  | 2  | 2  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 4  | 2  | 2  | 2  | 2  | 1  | 1  | 1  |

The dissimilarity score for each possible 29mer in the genome is calculated by adding up the weight vectors for each base not matching its corresponding base in the 29 bases surrounding a SNP. If the lowest dissimilarity score for every 29mer in the genome is greater than 10, the SNP position is considered unique. For example, if the 29mer that most closely matches the 29 bases
surrounding a SNP contained 3 mismatches between positions 8 and 22, that SNP position would have a dissimilarity score of 12 (4+4+4), and be sufficiently dissimilar to be considered unique, whereas six mismatches in positions 1 to 6 would not meet the uniqueness threshold of 10 (1+1+1+2+2+2 = 9). SNP sites that pass this criterion (threshold value >10) are given a passing scores of 1, while SNP sites that fail (threshold value <10) are given a failing score of 0. SNP sites that fail this criterion are considered likely false positives.

1 Wallace, R.B.; Shaffer, J.; Murphy, R.F.; Bonner, J.; Hirose, T.; Itakura, K. Nucleic Acids Res. 6, 3543 (1979)