Methods Online

Bacterial strains, growth and media

The *S. pneumoniae* strain used in this study (AC2394) is an acapsular derivative of TIGR4, a serotype 4 clinical isolate. *S. pneumoniae* was grown in Todd Hewitt broth supplemented with yeast extract (THY) and 5 µl/ml Oxyrase (Oxyrase, Inc) or on Sheep's blood agar plates at 37^{o}C in a 5% CO_2 atmosphere. Where appropriate, blood plates contained 4 µg/ml chloramphenicol (Cm) and/or 200 µg/ml Spectinomycin (Sp).

Construction of marked deletions

Gene knockouts of the five query genes *ccpA*, *regR*, *malR*, *malX* and SP_1683 were constructed by replacing the entire coding sequence with a Cm resistance cassette. Single gene and double gene knockouts that were used to validate single gene fitness and genetic interactions with *ccpA* were constructed in the same manner as the query genes but instead an Sp resistance cassette was used (see **Supplementary Table 6** for the strains used in this study). Construction of PCR products for gene replacement and transformation of *S. pneumoniae* were performed as described previously [16,30].

Transposon library construction

The *magellan5* mini-transposon is a derivative of the Himar1 Mariner transposon that contains a spectinomycin resistance marker. The mini-transposon is carried on a plasmid (pR412), which contains an ampicillin resistance marker and can be propagated in *E. coli*. A single nucleotide was changed in both inverted repeats of *magellan5* with the Quickchange multi site directed mutagenesis kit (Stratagene). The inverted repeat sequence is changed from 5' – agaccggggacttatcaGccaacctgt -3' to 5' -agaccggggacttatcaTccaacctgt -3' thereby introducing an Mme *I* restriction site (the mutation is in uppercase and the Mme *I* site is underlined). The *magellan5* mini-transposon with the Mme *I* site is from now on referred to as *magellan6*.

Library construction was essentially performed as previously described [31]. In short, plasmid DNA harboring *magellan6* was purified from *E. coli* with the Qiagen mini plasmid preparation kit (Qiagen). *S. pneumoniae* DNA was isolated with the Qiagen Blood and Tissue kit from 3 ml of an exponentially growing culture. *In vitro magellan6* transposition reactions were carried out with purified MarC9 transposase, 1 µg of *S. pneumoniae* target DNA and 1 µg of *magellan6* plasmid DNA. Reactions were incubated for 1 hour at 30°C, inactivated for 20' at 72°C, ethanol precipitated and resuspended in gap repair buffer [50 mM Tris (pH 7.8), 10 mM $MgCl_2$ 1 mM DTT, 100 nM dNTPs and 50 ng BSA]. Repair of transposition product gaps was performed with *E. coli* DNA ligase overnight at 16°C after which repaired transposition products were transformed into naturally competent *S. pneumoniae*. The following day transformants were scraped off of blood plates and pooled into libraries of approximately 25,000 transformants/library for the wild type strain and 10,000 – 25,0000 transformants/library for the query strains (in total 6 libraries were screened for wild type and three libraries for each query strain). To separate transformants from any residual agar, cells were grown in THY for a maximum of 2 hours after which libraries were harvested. From a fraction of the library, DNA was isolated (t_1) while another fraction was used to seed 10 ml of THY. Libraries were grown in THY for approximately seven generations to late exponential growth phase after which bacteria were harvested and DNA was isolated (t_2). The exact expansion of each library from t_1 to t_2, was determined by plating dilutions of the culture from both time points on blood agar plates. Note that both the transformation procedure and the 2 hour growth to separate transformants from residual agar may place a limit on the dynamic range of mutants that we were able to identify by Tn-seq. Specifically, mutants with fitness values less than 0.5 were not identified in our study presumably due to their outcompetition during these early steps. It is likely that changes to the protocol that reduce the number of generations prior to selection at t_1, will increase the dynamic range.

Sample preparation and Illumina sequencing

DNA from two time points, pre-selection (t_1) and post-selection (t_2), was digested for 2.5 hours at 37^{o}C with Mme *I* (NEB). To prevent re-ligation in the following step the 5' phosphate group was removed with Calf intestinal alkaline phosphatase (NEB) after which DNA was phenol/chloroform extracted, ethanol precipitated and dissolved in H_2O. Next an adapter was ligated on the overhang left by Mme *I*, overnight at 16^{o}C after which DNA was again phenol/chloroform extracted, ethanol precipitated and dissolved in H_2O. A PCR is performed with the adapter-ligated samples as template and with primers of which one is complementary to the adapter and the other to the mini-transposon inverted repeat sequence. Both primers also contain Illumina specific sequences that have complementarity to two specific primers in the Illumina flow cell and enable eventual sequencing. Although it is essential that the adapter sequence contains one of the Illumina specific sequences, the rest of the adapter can vary. Therefore it is possible to include a barcode sequence in the adapter that enables mixing of samples in a single flow cell lane and subsequent separation after sequencing. The resulting PCR product is 160 bp in length and is amplified with the following parameters: 95^{o}C for 30 s, followed by 22 cycles of 10 s at 95^{o}C, 25 s at 55^{o}C and 15 s at 72^{o}C, followed by 10 min at 72^{o}C (through Illumina sequencing of PCR products amplified with 19, 22 and 25 cycles we determined that the PCR remained linear at least up to 25 cycles). The PCR product is gel purified, dissolved in H_2O and sequenced on an Illumina Genome Analyzer II according to the manufacturers protocol (Illumina). One to three libraries from a single time point were sequenced in a single flow cell lane. Following 30 sequencing cycles, raw data is extracted, split into different samples based on the 6-nucleotide-barcode sequence and stripped from the barcode and four nucleotides of the adapter sequence. This resulted in 6-10x10^{6}, 20 base *S. pneumoniae* specific reads per flow cell lane.

Fitness calculation

Following sequencing, 20 bp reads are mapped to the *S. pneumoniae* TIGR4 genome with the program Bowtie [32]. Bowtie parameters (-m 1 –n 1 –best) were set such that reads could contain a single mismatch but were only allowed if they mapped to a unique location. If mapping to multiple sites was possible the read was excluded from the analyses. In addition, sequences from the first two wild type libraries were also analyzed with the program CLC genomics workbench v2.1.1 (CLC bio) and the program MAQ [33], which confirmed the Bowtie analyses. In addition, sequences from all libraries were visually checked with the program CLC genomics workbench. Due to the short read length the error rate was low and on average only 5% of the reads had to be

discarded for two reasons: i) 4.5% of these 'reads' could not be mapped to a single location. These reads map to 119 endogenous transposon related genes and 11 other repeated sequences (see **Supplementary Table 7** for removed genes), ii) the remaining reads (0.5%) did not map anywhere and were categorized as junk sequences. For the two time points (t_1 and t_2) the number of reads at each location are recorded; on average 82 reads were mapped per insertion/time point. Since an insertion with a low number of reads that slightly fluctuates over time could influence the data disproportionately, only insertions with eight or more reads at one time point were included in the analyses. This resulted in excluding 2.5% of the insertions. Lowering the limit to a minimum of five reads or setting the limit higher to a minimum of 15 reads did not significantly alter the results, which suggests that eight is a robust limit. Subsequently the data is normalized by equalizing the total number of sequenced reads per time point. The number of reads per time point never varied more than 10% and therefore the normalization factors for all data sets are small and lie between 1.03 and 1.10. The change over time in the number of reads at a specific location is then used to calculate fitness. Thus for each insertion, fitness W_i, is calculated by comparing the fold expansion of the mutant relative to the rest of the population with the following equation [34,35]:

$$W_i = \frac{\ln[N_i(t_2) * d / N_i(t_1)]}{\ln[(1 - N_i(t_2)) * d / (1 - N_i(t_1))]}$$

In which $N_i(t_1)$ and $N_i(t_2)$ are the frequency of the mutant in the population at the start and at the end of the experiment, respectively, and d (expansion factor) represents the growth of the bacterial population during library selection. The expansion factor was directly measured for each library by plating appropriate dilutions of the bacterial library at t_1 and t_2 and dividing the number of bacteria counted at each time point (d = number of bacteria at t_2/ number of bacteria at t_1).

After fitness is determined for each insertion, data is normalized against a set of genes that are inactive in *S. pneumoniae* and have no fitness effect (see **Supplementary Table 2**). These genes are degenerate transposon related sequences and disrupting them with a transposon will not result in a fitness effect. Since their fitness is therefore equal to that of the wild type background strain in which the transposon library is generated they can be used as a reference point to which the rest of the library is normalized. To achieve this, fitness is calculated over all these genes and their average fitness is subsequently set to 1. The same factor is used to normalize the rest of the dataset, thus setting all fitness values relative to 1. The normalization factors used for all datasets

were small and lie between 0.981 and 1.014. This shows that fitness for the reference genes was already close to 1, which underscores the fact that these genes are neutral. Note that in cases where a query strain was used as the background strain, the average of the reference genes was set to the fitness of the query strain (*ccpA* = 0.84, *regR* = 1.0, SP_1683 = 0.98 , *malX* = 1.04 and *malR* = 1.05). Finally all of the insertions in a gene are used to calculate a gene's average fitness and standard deviation. To further control for deviations in fitness due to insertions with small numbers of reads a weighted average is used. This means that insertions with 50 or less reads receive a proportional lower weight. This slightly increases correlations between replicates and it slightly lowers standard deviation. A generally used method to determine fitness is the competition index where the change in mutant to wild type ratio over time is determined [36,37]. The advantage of using the equation above and including the expansion factor is that W_i now represents the actual growth rate per generation, which makes fitness independent of time, more accurate and enables comparisons between experiments.

To determine whether fitness effects significantly differed from wild type and were advantageous or disadvantageous, three requirements had to be fulfilled: i) fitness was calculated from at least ten transposon insertions (genes with less than ten insertions are excluded), ii) fitness had to deviate from wild type fitness (1) by at least 4%, and iii) fitness had to be significantly different from wild type in a one sample *t*-test with Bonferroni correction. All significant deviations therefore had a *P* value < 0.00003 (*P* value /number of comparisons). Genes were listed as "possibly essential" if no insertions were recorded or less than three insertions were recorded and the gene size was ≥400bp.

1x1 Competition assays and single strain growth

For 1x1 competitions bacteria were recovered from blood agar, subcultured in THY and grown at 37°C for 1.5 hrs to an OD_{600} ~0.1 - 0.2. Subsequently two strains were mixed in a 1:1 ratio. Mutants were always competed against their background strain; transposon insertions and strains with a single gene drug marker replacement were competed against the wild type strain, while double mutants (which all had *ccpA* deleted) were competed against the *ccpA* mutant. Competitions were grown for approximately 7 generations to late exponential growth phase and ratios of the competing strains were determined at the start and at the end of the competition by plating appropriate dilutions on blood agar plates with selective antibiotics. Competitions were performed no less than 5 times.

Single strain growth was performed in THY no less than four times. For each strain, OD_{600} measurements were taken every half hour and were used to calculate doubling times for 10

mutant strains and the wild type.

Genetic interactions

Genetic interactions are defined as a deviation from the multiplicative model, which states that if a strain deleted for gene *i* has a fitness W_i and a strain deleted for gene *j* has a fitness W_j, then the double mutant strain W_{ij} is expected to have the fitness W_i x W_j (note that at least one strains' fitness (W_i or W_j) needs to deviate significantly from 1). Because we wanted to minimize the number of false positives, genetic interactions had to fulfill three requirements: i) fitness was calculated from at least ten transposon insertions (genes with less than ten insertions are excluded), ii) expected fitness and observed fitness had to deviate by at least 10%, and iii) significant interactions had to pass the student's *t*-test with Bonferroni correction for multiple testing, which approximately lowers the level of a significant *P* value to 0.05/n (n=the number of comparisons). All significant interactions therefore had a *P* value < 0.00003. Significant genetic interactions could be further split into seven categories [23,24]; synergistic (W_i=W_j<W_{ij}), partial masking (W_i<W_j<W_{ij}), masking (W_i<W_j=W_{ij}), partial suppression (W_i<W_{ij}<W_j), suppression (W_i=W_{ij}<W_j), antagonistic (W_{ij}<W_i≤W_j) and co-equal (W_i=W_j=W_{ij}). Interactions were visualized with the program Cytoscape [38] and assigned to a category based on a student's *t*-test with Bonferroni correction for multiple testing. In addition, arrows in the networks depict the hypothesized direction of the interaction; for the categories partial masking and masking, W_j is the source and W_i is the target, which means that the gene with the highest fitness (W_j) is masking the effect of the gene with the lower fitness (W_i). For the categories partial suppression and suppression, W_i is the source and W_j is the target, here the gene with the lowest fitness (W_i) is suppressing the effect of the gene with the highest fitness (W_j). The source is thus the node where the arrow originates and the target where the arrow ends.

Supplementary File	Title
Supplementary Table 1	Transposon insertion hot and cold spots
Supplementary Table 2	Wild type library fitness values.
Supplementary Table 3	GSEA enrichment analysis
Supplementary Table 4	Genetic interactions and fitness values for five query genes
Supplementary Table 5	Genetic interactions for five query genes were split into seven different categories.
Supplementary Table 6	Strains and plasmids used in this study
Supplementary Table 7	Repeated genes removed from analyses