Supplementary Material

Supplementary Methods

Collection and processing of nasopharyngeal swabs

At each study visit, NP samples were collected on sterile rayon swabs by trained study nurses and transported in Stuart media directly to the clinical microbiology laboratory. They were plated within 8 hours directly onto blood agar and cultured at 37°C for 48 hours in 5% carbon dioxide. *S. pneumoniae* isolates were identified by colony morphology, Gram stain, catalase testing and optochin susceptibility. Antimicrobial susceptibility testing was performed by E-test [E-test strips; BioMerieux Inc] and breakpoints were interpreted according to 2011 CLSI guidelines. Isolates were stored in tryptic soy broth (TSB) with 20% glycerol at -80°C for further testing.

Serotype determination

Quellung reaction, when required, was performed with typing sera prepared in the CDC Streptococcus Lab and Statens Serum Institut.

Rooting the SCD Phylogenetic tree

To root the phylogenetic tree, a gene content dendrogram and heatmap was constructed as described in the online methods, but with the inclusion of several additional Streptococcal species (data not shown). These non-pneumococcus Streptococcal species were used to root the partitioned SNV and gene-content-based phylogenetic tree.

Gene Presence/Absence and Phylogentic Analysis

Gene content-based hierarchical clustering was performed by treating each strain as a binary vector coded by presence or absence of a member of each ortholog group. Pairwise Euclidean distance was used to construct the distance matrix. Only ortholog groups that had genes present in at least two of the samples were used in the hierarchical clustering. This subset was used to reduce the contribution of noisy singleton columns to the distance calculation.

Samtools and several scripts were used to identify the variable positions in the 1376 genes that constitute the core genome (Li et al., 2009). Genes were classified as core if they were present in at least 98% of the samples. An alternate allele was called at a particular position if there were at least 5 reads covering the position, the alternate allele was supported by at least one read from each strand, and at least 80% of the overlapping aligned reads supported the alternate base. Positions were masked with 'N' if the above condition was not met for either the reference allele or the alternate allele.

Phylogenetic analyses were performed using GARLI (Zwickl, 2006) In the partitioned SNV-gene presence/absence case, one partition consisted of 10000 selected variable positions from the 1376 core genes. The GTR model was fit to the data, allowing rates to vary according to a discrete Gamma distribution with 4 rate categories. An estimated proportion of sites were assumed invariant. The remaining partition consisted of a binary presence/absence matrix, representing presence/absence of a member of each ortholog group (column) in each strain (row). Only ortholog groups with genes present in at least 2 strains were included. This data was treated as discrete morphological data, allowing gain/loss of genes along the underlying tree. This is consistent with the highly recombinatory nature ofS. pneumoniae. In addition to the partitioned analysis, each partition was also analyzed separately. For each analysis, 100 bootstrap replicates were performed. A majority-rule consensus (MRC) tree was computed using Dendropy (Sukumaran and Holder, 2010). Branches were collapsed if they had less than bootstrap support lower than 0.50.

To determine the best-fitting phylogeny, the mean bootstrap support across all internal nodes was determined for each of the three trees. This was performed on the 0.01 MRC topology in order to include low-support clades in the calculation. These would otherwise be excluded in the 0.50 MRC phylogenies and would result in inflation of the mean bootstrap value in phylogenies with many multifurcations.

Determining similarity between strains isolated from single patients.

In several cases, multiple sequenced isolates were obtained form a single patient. To determine whether such strains were more similar than expected by chance, a distance simulation was performed using the partitioned phylogenetic tree. For all n samples isolated from the same patient (n >1), a null distribution of distances was created by randomly selecting n taxa 1000 times from the SCD strains and measuring the branch lengths of the shortest subtree connecting these taxa on the portioned tree. For each set of n isolates, the null hypothesis of no significant similarity between the strains was tested using this null distance distribution. P-values below 0.05 were considered significant.

Assessing Significance of taxa class label association

To assess the degree of association between taxa of the same class (contemporary IPD SCD vs contemporary NP SCD and contemporary vs historical NP SCD), a sample of 1000 trees from the posterior distribution of the partitioned phylogenetic analysis was used in the tip label-permutation based method BaTS(Parker et al., 2008). Since p-values were all reported as zero due to the relatively small number of permutations (n=1000), the degree of random association was further quantified using the ratio of the mean observed parsimony score and the mean parsimony score of the randomly permuted sample labels from the posterior distribution of trees(Parker et al., 2008).

Use of Tn-seq to calculate gene fitness values

The Tn-seq experimental procedure and data analysis were performed essentially as described(van Opijnen et al., 2009; van Opijnen and Camilli, 2010, 2012). Briefly, six independently constructed libraries were generated, each containing from 1500 to 4000 *magellan6* mini-transposon insertion strains encoding spectinomycin resistance. Each library was split up into multiple starter cultures, and stored at −20°C in 15% glycerol.

For each mouse experiment, a starter culture was thawed and grown for 2–3 h in Todd Hewitt broth supplemented with yeast extract (THY) and 5 µL/mL Oxyrase (Oxyrase, Inc) to reach the exponential growth phase. DNA was isolated from a fraction of the culture (t_1), and

the rest of the culture was concentrated in PBS and used to infect groups of mice. A total of 6 libraries were used for the infections. In the first round, library 1 was recovered from 3 WT and 3 SCD mice, library 2 was recovered from 3 WT and 2 SCD mice, and library 3 was recovered from 2 WT and 3 SCD mice. In the second round of experiments library 4 was recovered from 5 WT and 6 SCD mice, library 5 was recovered from 8 WT and 6 SCD mice, and library 6 was recovered from 4 WT and 6 SCD mice. Mice were injected with 1×10^5 CFUs of bacteria in a volume of 100 µL via intraperitoneal injection. Bacterial loads in the blood were quantified at 3, 6, 9, 12, 15 and 18 hours post challenge to approximate bacterial replication. A terminal bleed was collected once mice were moribund, typically at 18–24 hours post infection, and was plated on TSA plates supplemented with 150 µg/mL spectinomycin. The next day (i.e., after a 12-hour incubation), bacteria were scraped off of plates, and DNA was isolated (t_2). The *in vivo* expansion of each library from t_1 to t_2 was determined after measuring the bacterial load in each mouse at several time points.

DNA from pre-selection (t_1) and post-selection (t_2) were prepared for Illumina sequencing. The six libraries were harvested from moribund mice on selective media. Finally, the samples were multiplexed using six different barcoded adapters (one barcode per library) and sequenced in a single-flow cell lane on an Illumina HiSeq 2000 instrument. After 30 sequencing cycles, raw data was extracted, split into different samples on the basis of the 4-nucleotide-barcode sequence, and then the barcode and adapter sequences were stripped from each sequence (read).

After sequencing, reads were mapped to the *S. pneumoniae* TIGR4 genome by the Bowtie program using parameters (-m 1 –n 1 –best)(Langmead et al., 2009). If mapping to multiple sites was possible, then the read was excluded from the analyses. On average, 8% of the reads had to be discarded for 2 reasons: i) 6% could not be mapped to a single location, mapping to multiple sites, such as endogenous transposon-related genes or other repeated sequences; ii) the remaining 2% did not map anywhere and were categorized as junk

sequences. Insertions that mapped to a location in the last 10% of a gene were removed from the analysis to minimize the influence of truncated but functional genes. This procedure resulted in 5-15 × 10^6 *S. pneumoniae*–specific reads per flow cell lane.

The number of reads at each location was recorded at t_1 and t_2. On average, 250 reads were mapped per insertion/time point. Because insertions with a very low number of reads that slightly fluctuate over time can disproportionately influence the data, only insertions with 15 or more reads at t_1 were included in the analyses. Subsequently, the data was normalized by equalizing the total number of sequenced reads per time point (normalization factors for all data sets are small and lie between 0.92 and 1.06). The change over time in the number of reads at a specific location was then used to calculate fitness. Thus, for each insertion, fitness (W_i) is calculated by comparing the fold-expansion of the mutant to that of the rest of the population by using the following equation(van Opijnen et al., 2006):

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

$N_i(t_1)$ and $N_i(t_2)$ represent 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. After fitness was calculated for each insertion, values were normalized against a set of neutral genes: those having no fitness effect in *S. pneumoniae*. This group of genes consists of pseudo genes and degenerate transposon-related sequences. The same factor was then used to normalize the rest of the dataset and express all fitness values relative to the WT background. Finally, all of the insertions in a gene (except for the last 10% of the coding sequence) were used to calculate a gene's average fitness and standard deviation. A weighted average was used to further control for deviations in fitness that were due to insertions with small numbers of reads. Therefore, insertions with 50 or fewer reads received a proportionally lower weight than did those with more reads.

Because we were determining bacterial fitness during infection, we expected that an infection bottleneck effect would lead to stochastic loss of insertion mutants. To enable accurate fitness value calculations, this random loss of mutants had to be accounted for in the analysis. The bottleneck for each mouse was calculated by determining the proportion of insertion mutants that were lost from the neutral set of genes. Because these genes have no effect on fitness, we assume that all insertions lost from this set are lost due to stochastic processes that occur during or shortly after the infection procedure. To remove the bottleneck effect from each gene's fitness value, the same proportion of insertions mutants that disappeared during *in vivo* selection were removed from each gene's total number of insertions. The resulting set of insertions was then reanalyzed and fitness was recalculated. The W_i of each gene after these corrections represents the growth rate per generation and enables direct comparisons between experiments(van Opijnen and Camilli, 2012).

To determine whether a gene's fitness between WT and SCD mice significantly differed, 3 requirements had to be fulfilled: i) fitness had to be composed of at least 3 data points, ii) fitness had to deviate by at least 20%, and iii) fitness had to be significantly different in a one-sample *t*-test with Bonferroni correction for multiple testing.

Mutant Construction

For a subset of the genes with fitness defects identified by Tn-seq, stable chromosomal mutations were generated by PCR-based overlap extension(Iannelli and Pozzi, 2004). Briefly, regions upstream and downstream of the target region were PCR-amplified and spliced onto an erythromycin resistance cassette. The final PCR product was transformed into the pneumococcus, replacing the targeted region with the antibiotic resistance cassette. To confirm the mutation, primers outside of the transformed region were used for PCR and subsequent region sequencing. The primers used to generate the knockouts and other information about the knockout strains are detailed in Supplementary Table S4.

Generation of Sickle Cell Mice

All experiments involving animals were performed with prior approval of and in accordance with guidelines of the St. Jude Institutional Animal Care and Use Committee (Protocol #250). The St. Jude laboratory animal facilities have been fully accredited by the American Association for Accreditation of Laboratory Animal Care. Laboratory animals were maintained in accordance with the applicable portions of the Animal Welfare Act and the guidelines prescribed in the DHHS publication, Guide for the Care and Use of Laboratory Animals. Lethally irradiated, 8-week-old female C57B/J6 mice (Jackson Labs, Bar Harbor, ME) were transplanted as described previously with 2×10^6 bone marrow cells from either BERK SCD mice or WT mice(Pestina et al., 2009). Baytril (enrofloxacin; 2.27% solution, Bayer) was administered as antimicrobial prophylaxis for 3 weeks after transplantation. The sickle phenotype was confirmed by hemoglobin cellulose acetate electrophoresis of red cell lysates 100 days posttransplant(Persons et al., 2003). Complete blood counts were determined to ascertain whether the white blood cell number, hematocrit, hemoglobin and red blood cell distribution were equivalent to those of the sickle donor. Only mice having full engraftment as confirmed by hemoglobin electrophoresis and blood counts at 12 weeks transplantation were used for further studies.

Fitness Challenges

To confirm the fitness values obtained by Tn-seq, the respective gene replacement mutants were tested in competition with the parental TIGR4 for mouse challenge in a sepsis model of infection in both C57/Bl6 and C57/Bl6 sickle cell mice. A total of 1000 CFUs of the TIGR4 and mutant strain were mixed and intraperitoneally injected as a single injection. Blood was collected, and the numbers of WT and mutant CFUs per mL of blood were determined by performing serial dilution and replica plating

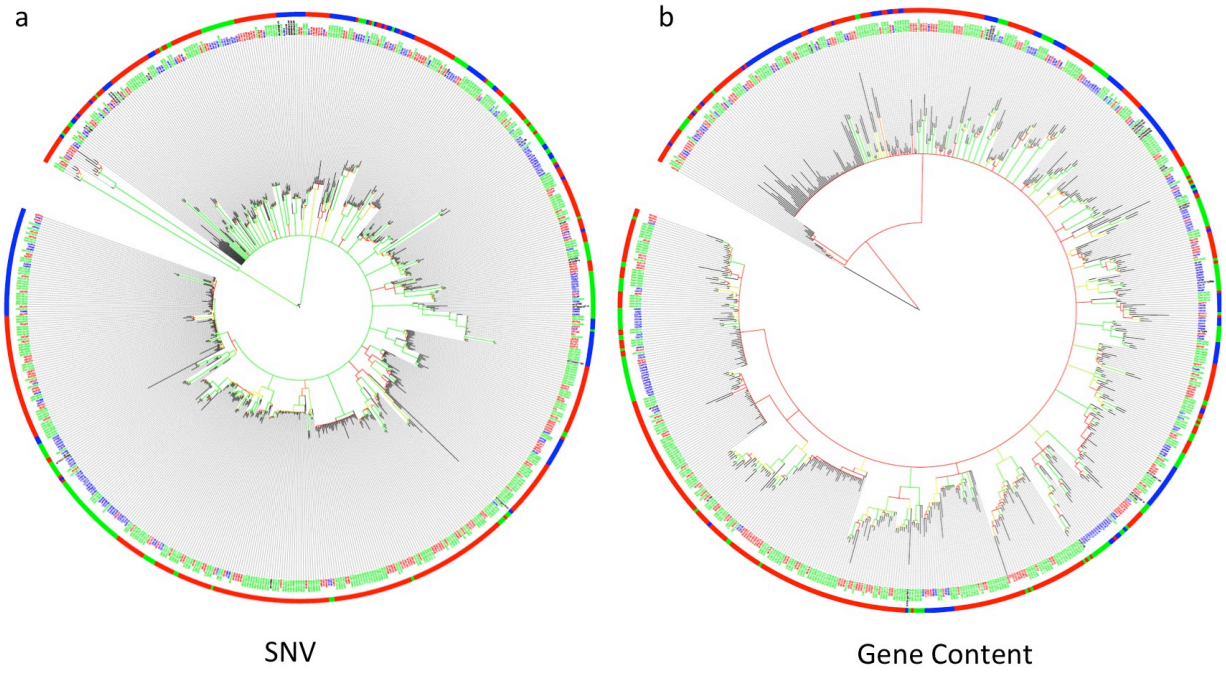
Purification of Sp1449 and PiaA. The coding sequence for Sp1449 was amplified from TIGR4 and cloned into the pET28b cloning vector in BL21-DE3 cells. Cultures were grown to OD_{600} = 0.5 and induced with 0.07 mM IPTG overnight at 23˚C. Bacterial pellets were lysed with

Bugbuster (Novogen) reagent according to manufacturers protocols. CppA was purified on His-Selected Nickel Affinity Gel following manufacturers protocol for native conditions. PiaA was expressed and purified as previously described (Brown et al., 2001). Protein was dialyzed using Pierce Slide-A-Lyzer dialysis cassette overnight with sterile PBS. Dialyzed protein was stored at -80˚C in a 10% glycerol solution until further use. Purity of protein was determined to be >95% by visualization on a 10% Bis-Tris gel stained with Simply Blue Safestain (Invitrogen).

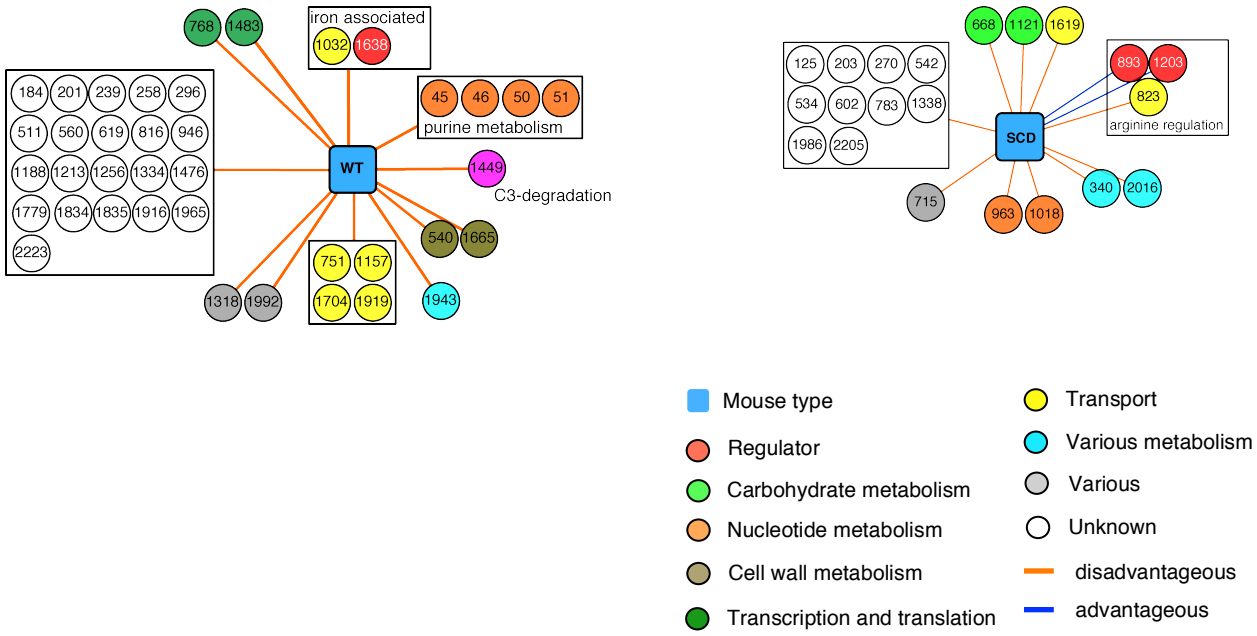
Vaccination of Mice. Purified CppA or PiaA (10 ug/mouse) was mixed with adjuvant (Alhydrogel) in PBS for 1 hour at 4˚C immediately prior to injection. Eight to ten week old C57/bl6 or SCD mice vaccinated by intraperitoneal injection of 10 ugs of CppA or PiaA with 10 ugs Alyhydrogel adjuvant in a total volume of 100 uL. Mice were boosted twice at two week intervals. Control mice received adjuvant only at the same dosing schedule. One week following the final boost, serum was collected and antibody titer against CppA and PiaA determined by ELISA. Both wild type and SCD mice vaccinated with recombinant proteins developed high antibody titers (Fig S7). Mice were challenged two weeks after the final boost with $1x10^7$ CFUs of TIGR4 in a volume of 25 uL via intranasal challenge. Mice were monitored every six hours for the development of disease.

Mouse Challenge

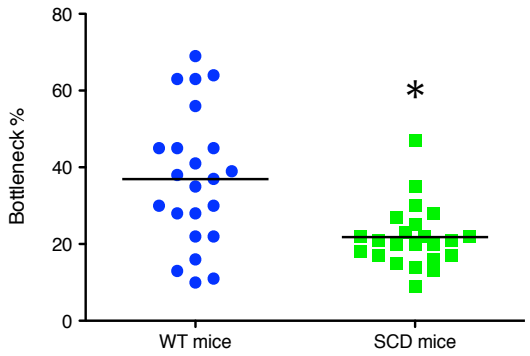
All mice were maintained in BSL2 facilities, and all experiments were done while the mice were under inhaled isoflurane (2.5%) anesthesia. For survival studies, bacteria were introduced by intranasal administration of 10^7 CFU of bacteria in PBS (25 µL), a model which effectively recapitulates the progression of disease from nasopharyngeal colonization, to pneumonia, and finally to the development of sepsis and meningitis (Orihuela et al., 2004). Mice were monitored daily for signs of infection, and differences in time-to-death among the mice were compared via Mantel-Cox log rank test. Bacterial burden in blood was enumerated by serial dilution and counting on TSA blood agar plates.



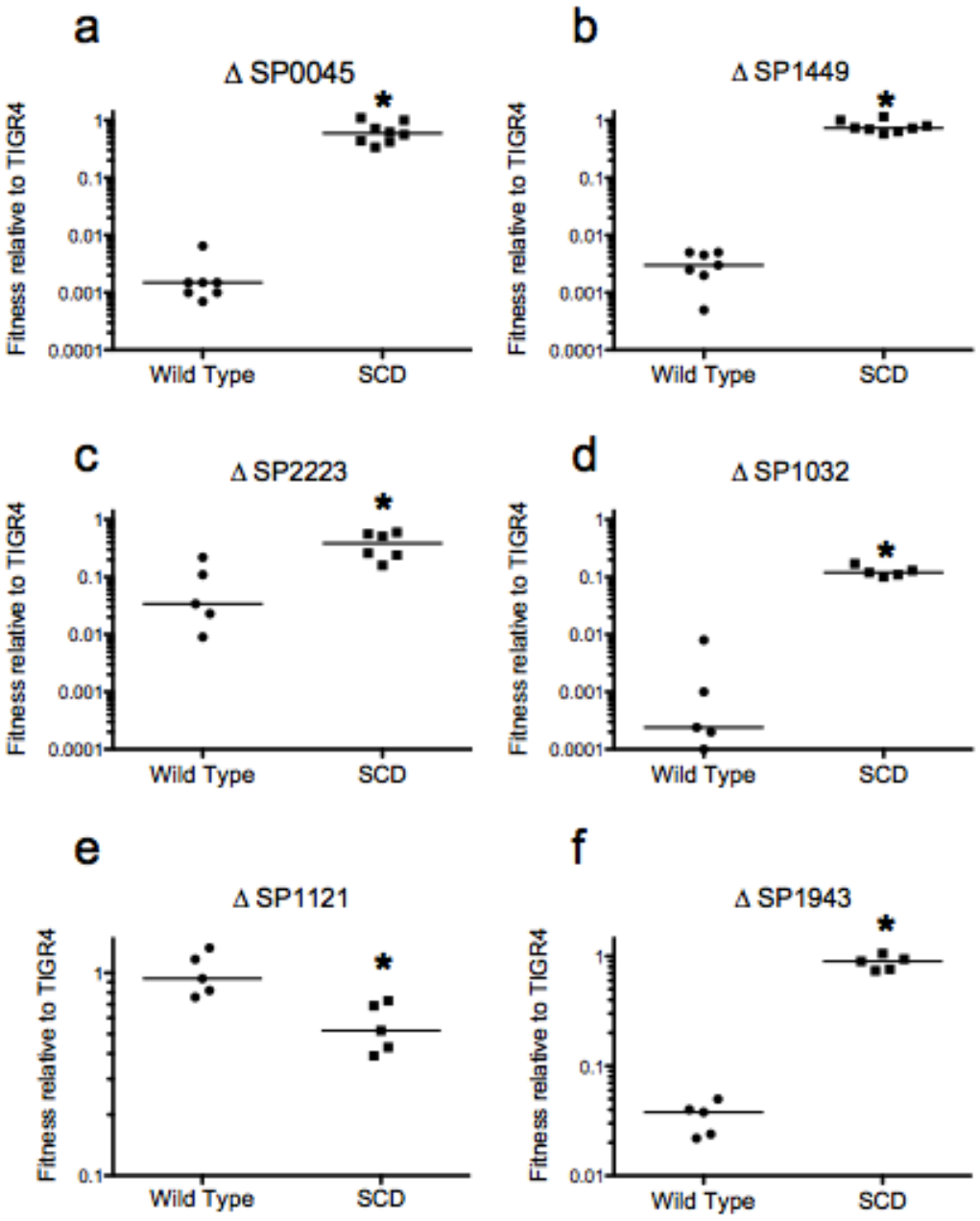
Supplementary Figure S1. Phylogeny of strains based on SNVs or gene content from analysis detailed in Figure 2. The historical SCD (blue text), contemporary SCD (red text), contemporary nasopharyngeal cohort (green text), and genbank strains (black text). Colors of the branches reflects bootstrap values (red to green representing 0.65-1.0). Outer circle reflects strain serotype, serotypes included in PCV7=green, PCV13=blue, and on-vaccine coverage=red. This phylogeny was based on either SNVs (a) or gene content (b).



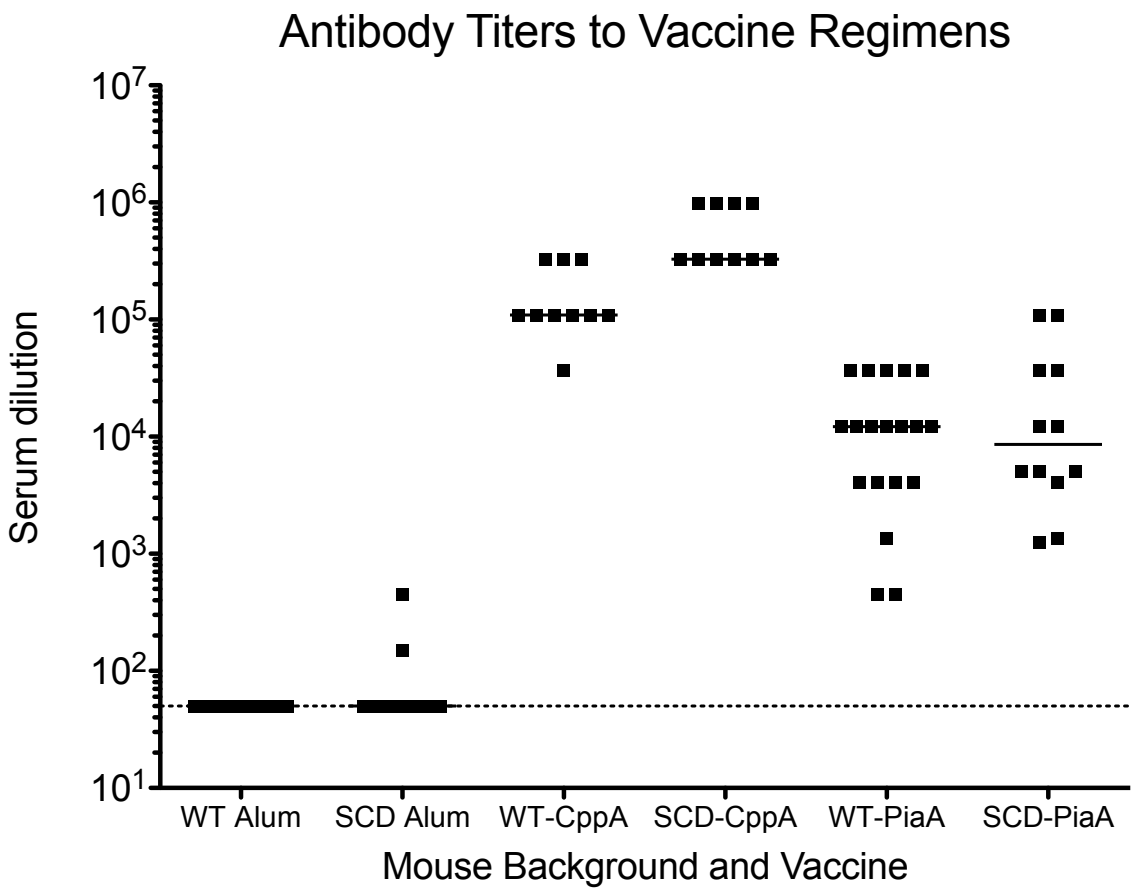
Supplementary Figure S2. Genes important for full virulence in wild type (WT) and SCD mice as identified in Table 2. WT and SCD mice impose both overlapping and non-overlapping requirements for specific pneumococcal genes and pathways to enable full virulence. The non-overlapping genes are depicted. Genes whose loss decreases virulence (disadvantageous) are connected by the orange lines. Conversely, genes whose loss increases virulence (advantageous) are connected by the blue lines. The predicted functional roles of the genes are indicated based on the colored circles.



Supplementary Figure S3. Population bottlenecks for Tn-seq in wild type (WT) and SCD mice from the challenge experiment in Table 2. The indiscriminate clearing of bacteria was calculated for each mouse (Supplementary methods). A lower value indicates reduced initial clearance capacity for pneumococci. * $p<0.0001$ by Mann-Whitney test; each data point represents an individual mouse and the medians are shown by the lines. Data pooled from five independent mutant libraries.



Supplementary Figure S4. Confirmation of host-specific fitness differences from predictions in Table 2. Six genes with a significantly different fitness between wild type (WT) and SCD murine hosts by Tn-Seq were reassessed using deletion mutation by gene replacement. Equivalent CFUs of TIGR4 and the deletion mutant were co-inoculated and the respective amounts of each strain after bacteremia at 24 hours was enumerated. A fitness value of 1 indicates that equivalent numbers of the TIGR4 and mutant strain were recovered. Each data point represents an individual mouse and the medians are shown by the lines. *p<0.01 by Mann-Whitney test.



Supplementary Figure S5. Antibody titers of Wild Type and SCD mice following CppA and PiaA vaccination as detailed in Figure 5. Wild type and SCD mice were vaccinated with recombinant CppA or PiaA (prime followed by two boosts) and bled one week following final boost. Antibody titers were determined by ELISA. Data represent reciprocal dilution of serum with signal above background. Each data point represents one mouse. Dotted line represent limit of detection.

Supplementary Table S2. Significant differences in pneumococcal gene fitness between wild type (WT) and SCD mice. [a] Gene number in *S. pneumoniae* TIGR4, [b] Mean Tn-seq fitness calculated for WT and SCD mice,[c] Number of insertions in a gene, [d] Gene name, [e] significant difference ****p<0.0002, ***p<0.002 (one sample *t*-test with Bonferroni correction for multiple testing). A subset of these with clinical correlation is detailed in Table 2.

Locus[a]	WT fitness ± sem[b]	N[c]	SCD fitness ± sem[b]	N[c]	Name[d]	Sig[e]
SP0045	0.20 ± 0.05	67	0.71 ± 0.05	71	putative phosphoribosylformylglycinamidine synthase	****
SP0046	0.28 ± 0.09	26	0.66 ± 0.07	25	amidophosphoribosyltransferase	***
SP0050	0.00	32	0.50 ± 0.08	28	formyltransferase / IMP cyclohydrolase	****
SP0051	0.14 ± 0.12	7	0.77 ± 0.07	6	phosphoribosylamine-glycine ligase	****
SP0125	1.26 ± 0.03	7	1.04 ± 0.02	6	hypothetical protein	****
SP0184	0.19 ± 0.16	5	1.04 ± 0.04	9	hypothetical protein	****
SP0201	0.57 ± 0.13	20	0.92 ± 0.03	18	hypothetical protein	***
SP0203	0.90 ± 0.09	9	0.43 ± 0.18	7	hypothetical protein	***
SP0239	0.39 ± 0.23	8	0.97 ± 0.04	7	conserved hypothetical protein	***
SP0258	0.78 ± 0.02	7	1.00 ± 0.02	6	hypothetical protein	****
SP0270	0.95 ± 0.01	6	0.58 ± 0.17	6	hypothetical protein	***
SP0296	0.00	5	0.65 ± 0.17	9	hypothetical protein	***
SP0340	1.25 ± 0.05	14	1.02 ± 0.01	9	autoinducer-2 production protein	****
SP0511	0.75 ± 0.05	8	1.04 ± 0.03	7	hypothetical protein	****
SP0534	1.24 ± 0.03	7	1.02 ± 0.01	6	hypothetical protein	****
SP0540	0.58 ± 0.18	14	0.98 ± 0.08	14	BlpN protein	***
SP0542	1.09 ± 0.02	15	0.72 ± 0.09	13	hypothetical protein	***
SP0560	0.48 ± 0.19	9	0.93 ± 0.03	7	hypothetical protein	***
SP0602	1.13 ± 0.06	9	0.72 ± 0.19	8	pep27 protein	***
SP0619	0.44 ± 0.15	18	1.00 ± 0.10	18	conserved hypothetical protein	***
SP0668	0.99 ± 0.02	6	0.78 ± 0.03	6	glucokinase	****
SP0715	1.46 ± 0.02	13	1.16 ± 0.03	12	lactate oxidase	****
SP0751	0.43 ± 0.20	14	1.00 ± 0.07	18	branched-chain amino acid ABC transporter protein	***
SP0768	0.65 ± 0.19	11	1.06 ± 0.08	12	ribosomal RNA large subunit methyltransferase N	***
SP0783	1.20 ± 0.03	7	0.95 ± 0.02	6	conserved hypothetical protein	****
SP0816	0.58 ± 0.23	9	0.99 ± 0.05	7	hypothetical protein	***
SP0823	1.34 ± 0.05	6	0.84 ± 0.02	6	amino acid ABC transporter, permease protein	****
SP0893	0.67 ± 0.25	6	1.12 ± 0.06	13	putative transcriptional repressor	***
SP0946	0.31 ± 0.15	16	0.87 ± 0.04	18	conserved hypothetical protein	***
SP0963	0.95 ± 0.03	11	0.52 ± 0.20	8	dihydroorotate dehydrogenase, electron transfer subunit	***
SP1018	1.37 ± 0.03	8	1.12 ± 0.01	6	thymidine kinase	****
SP1032	0.43 ± 0.17	23	1.04 ± 0.09	23	iron-compound ABC transporter	***
SP1121	1.11 ± 0.05	31	0.70 ± 0.08	35	1,4-alpha-glucan branching enzyme	****
SP1157	0.00	13	0.96 ± 0.13	17	voltage-gated chloride channel family protein	****
SP1188	0.00	9	0.49 ± 0.18	7	hypothetical protein	***
SP1203	0.47 ± 0.26	7	0.97 ± 0.05	6	putative transcriptional repressor	***
SP1213	0.15 ± 0.14	9	0.78 ± 0.17	7	conserved domain protein	***
SP1256	0.44 ± 0.22	9	1.09 ±0.05	7	conserved hypothetical protein	***
SP1318	0.60 ± 0.09	39	0.98 ± 0.04	36	v-type sodium ATP synthase, subunit G	***
SP1334	0.78 ± 0.03	4	1.01 ± 0.03	6	conserved hypothetical protein	****
SP1338	1.27 ± 0.03	7	1.03 ± 0.01	6	hypothetical protein	****
SP1449	0.00	35	0.60 ± 0.09	30	C3-degrading proteinase	****
SP1476	0.67 ± 0.16	7	1.05 ± 0.02	6	hypothetical protein	***
SP1483	0.39 ± 0.25	8	0.94 ± 0.17	6	ATP-dependent RNA helicase, DEAD/DEAH box family	***
SP1619	1.04 ± 0.08	23	0.72 ± 0.13	16	PTS system, IIA component	***
SP1638	0.31 ± 0.25	4	0.97 ± 0.07	6	iron-dependent transcriptional regulator	***
SP1665	0.57 ± 0.27	8	1.03 ± 0.04	7	ylmE protein	***
SP1704	0.69 ± 0.13	20	1.02 ± 0.04	16	ABC transporter, ATP-binding protein	***
SP1779	0.00	8	0.50 ± 0.16	12	hypothetical protein	***
SP1834	0.51 ± 0.19	7	1.02 ± 0.03	6	hypothetical protein	***
SP1835	0.35 ± 0.21	5	0.99 ± 0.09	9	hypothetical protein	***
SP1916	0.00	9	0.82 ± 0.21	7	PAP2 family protein	***
SP1919	0.45 ± 0.25	6	1.08 ± 0.09	8	ABC transporter, permease protein	***
SP1943	0.18 ± 0.11	14	0.68 ± 0.12	16	acetyltransferase, GNAT family	***
SP1965	0.45 ± 0.22	5	0.96 ± 0.03	9	hypothetical protein	***
SP1986	1.38 ± 0.04	7	0.94 ± 0.03	6	hypothetical protein	****

SP1992	0.48 ± 0.14	18	1.00 ± 0.03	16	cell wall surface anchor family protein	***
SP2016	0.99 ± 0.03	27	0.65 ± 0.10	28	nicotinate-nucleotide pyrophosphorylase	***
SP2205	1.11 ± 0.04	19	0.80 ± 0.03	19	DHH subfamily 1 protein	****
SP2223	0.19 ± 0.12	17	0.88 ± 0.10	21	conserved hypothetical protein	****

Supplementary Table S3. List of all reference strains used in the analysis in Figures 2 and 4 and their corresponding accession numbers.

Reference Strain	**Accession**
Streptococcus mitis B6	NC_013853.1
Streptococcus oralis Uo5	NC_015291.1
Streptococcus pseudopneumoniae IS7493	NC_015875.1
Streptococcus pyogenes A20	NC_018936.1
Streptococcus pyogenes Alab49	NC_017596.1
Streptococcus pyogenes MGAS10750	NC_008024.1
Streptococcus pyogenes MGAS6180	NC_007296.1
Streptococcus pyogenes SF370	NC_002737.1
Streptococcus sanguinis SK36	NC_009009.1
Streptococcus pneumoniae 670-6B	CP002176
Streptococcus pneumoniae 70585	CP000918
Streptococcus pneumoniae AP200	CP002121
Streptococcus pneumoniae ATCC 700669	FM211187
Streptococcus pneumoniae CGSP14	CP001033
Streptococcus pneumoniae D39	NC_008533.1
Streptococcus pneumoniae G54	CP001015
Streptococcus pneumoniae Hungary19A-6	CP000936
Streptococcus pneumoniae INV104	FQ312030
Streptococcus pneumoniae INV200	FQ312029
Streptococcus pneumoniae JJA	CP000919
Streptococcus pneumoniae OXC141	FQ312027
Streptococcus pneumoniae P1031	CP000920
Streptococcus pneumoniae R6	AE007317
Streptococcus pneumoniae SPN032672	FQ312039
Streptococcus pneumoniae SPN033038	FQ312042
Streptococcus pneumoniae SPN034156	FQ312045
Streptococcus pneumoniae SPN034183	FQ312043
Streptococcus pneumoniae SPN994038	FQ312041
Streptococcus pneumoniae SPN994039	FQ312044
Streptococcus pneumoniae ST556	CP003357.1
Streptococcus pneumoniae Taiwan19F-14	CP000921
Streptococcus pneumoniae TCH8431/19A	CP001993
Streptococcus pneumoniae TIGR4	AE005672

Supplementary Table S4. Oligonucleotides used to generate mutants used for challenge experiments in Figure 5.

Mutant	Oligo name	Sequence
SP0045	BM45UPF	GCTCAAGTAATACGAAAGG
	BM45UPR	GTTTGCTTCTAAGTCTTATTTCCcctttatttcagctcttgc
	BM45DNF	GAGTCGCTTTTGTAAATTTGGCAGATTTTCTAATAGATAG
	BM45DNR	ccgcaaatcccatagccgcgc
SP1943	BM1943UPF	GCAAGAAGTCCATAGTGTCC
	BM1943UPR	GTTTGCTTCTAAGTCTTATTTCCgaagctcctctaacaatttc
	BM1943DNF	GAGTCGCTTTTGTAAATTTGGGGAATGGTGCTAGCTTTAC
	BM1943DNR	gctctgccaaataaatcttc
SP1449	BM1449UPF	GAGCTGTATCCAGTGGCTATG
	BM1449UPR	GTTTGCTTCTAAGTCTTATTTCCgcctgataaccagcaatac
	BM1449DNF	GAGTCGCTTTTGTAAATTTGGGAAGAACAAATCGAGGCAGG
	BM1449DNR	ggccactcgttcctgacaac
SP1032	JR138	GTTGGTGTGCTGCTTACAATATCCATTTTAATAATGG
	JR139	gtttgcttctaagtcttatttccAAAAACTCCTTAAACATATTTCAAGTCTATTGTATCG
	JR140	AAAACAAATAAACCTAGGCATAATTTTTATAATCTGCC
	JR141	CATAAATAAGTACTGGACCCCCGATTATTGCAGTAATTATCCC
SP1121	JR142	CCGCCATCTTTATTAGGTGTCCATGGAGCATCATCATAGTCCA
	JR143	gtttgcttctaagtcttatttccGGATAATAGAGAAGCATTAAAAACCTTTATGACGGG
	JR144	AAATCAATATTTTTTCAAAAAATTGCGAAAACGCC
	JR145	GATGCAGATTTGGAACATGCTGCCAAGCAAATTGTTGCGGG
SP2223	JR150	GAAAATTTAGATTTACTTGATGAATTGGTAACATCAGC
	JR151	gtttgcttctaagtcttatttccCCCTTCTTTCTAGTGTCTATTATAATAGGTTACTACGATGGG
	JR152	AGGAAAAACGAATGAAAAAAGAACAAATTCCCAATCTCTTAACAATAGG
	JR153	CCACACGTTGCTTTTGGCCACCTGATAGACGCGC

Supplementary References

Brown, J.S., Ogunniyi, A.D., Woodrow, M.C., Holden, D.W., and Paton, J.C. (2001). Immunization with components of two iron uptake ABC transporters protects mice against systemic Streptococcus pneumoniae infection. Infect Immun *69*, 6702-6706.

Iannelli, F., and Pozzi, G. (2004). Method for introducing specific and unmarked mutations into the chromosome of Streptococcus pneumoniae. Mol Biotechnol *26*, 81-86.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology *10*, R25.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078-2079.

Orihuela, C.J., Gao, G., Francis, K.P., Yu, J., and Tuomanen, E.I. (2004). Tissue-specific contributions of pneumococcal virulence factors to pathogenesis. J Infect Dis *190*, 1661-1669.

Parker, J., Rambaut, A., and Pybus, O.G. (2008). Correlating viral phenotypes with phylogeny: accounting for phylogenetic uncertainty. Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases *8*, 239-246.

Persons, D.A., Hargrove, P.W., Allay, E.R., Hanawa, H., and Nienhuis, A.W. (2003). The degree of phenotypic correction of murine beta -thalassemia intermedia following lentiviral-mediated transfer of a human gamma-globin gene is influenced by chromosomal position effects and vector copy number. Blood *101*, 2175-2183.

Pestina, T.I., Hargrove, P.W., Jay, D., Gray, J.T., Boyd, K.M., and Persons, D.A. (2009). Correction of murine sickle cell disease using gamma-globin lentiviral vectors to mediate high-level expression of fetal hemoglobin. Molecular therapy : the journal of the American Society of Gene Therapy *17*, 245-252.

Sukumaran, J., and Holder, M.T. (2010). DendroPy: a Python library for phylogenetic computing. Bioinformatics *26*, 1569-1571.

van Opijnen, T., Bodi, K.L., and Camilli, A. (2009). Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods *6*, 767-772.

van Opijnen, T., Boerlijst, M.C., and Berkhout, B. (2006). Effects of random mutations in the human immunodeficiency virus type 1 transcriptional promoter on viral fitness in different host cell environments. J Virol *80*, 6678-6685.

van Opijnen, T., and Camilli, A. (2010). Genome-wide fitness and genetic interactions determined by Tn-seq, a high-throughput massively parallel sequencing method for microorganisms. Curr Protoc Microbiol *Chapter 1*, Unit1E 3.

van Opijnen, T., and Camilli, A. (2012). A fine scale phenotype-genotype virulence map of a bacterial pathogen. Genome research *22*, 2541-2551.

Zwickl, D.J. (2006). Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. (The University of Texas at Austin), pp. 115.