Genome-wide essential gene identification in *Streptococcus sanguinis*

Ping Xu1,2,3, Xiuchun Ge1, Lei Chen1, Xiaojing Wang1,*, Yuetan Dou1,*, Jerry Z. Xu1, Jenishkumar R. Patel1, Victoria Stone1, My Trinh1, Karra Evans1, Todd Kitten1,2,3, Danail Bonchev2 & Gregory A. Buck2,3

1Philips Institute of Oral and Craniofacial Molecular Biology, Virginia Commonwealth University, Richmond, Virginia, United States of America; 2Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, Virginia, United States of America; 3Department of Microbiology and Immunology, Virginia Commonwealth University, Richmond, Virginia, United States of America.

A clear perception of gene essentiality in bacterial pathogens is pivotal for identifying drug targets to combat emergence of new pathogens and antibiotic-resistant bacteria, for synthetic biology, and for understanding the origins of life. We have constructed a comprehensive set of deletion mutants and systematically identified a clearly defined set of essential genes for *Streptococcus sanguinis*. Our results were confirmed by growing *S. sanguinis* in minimal medium and by double-knockout of paralogous or isozyme genes. Careful examination revealed that these essential genes were associated with only three basic categories of biological functions: maintenance of the cell envelope, energy production, and processing of genetic information. Our finding was subsequently validated in two other pathogenic streptococcal species, *Streptococcus pneumoniae* and *Streptococcus mutans* and in two other gram-positive pathogens, *Bacillus subtilis* and *Staphylococcus aureus*. Our analysis has thus led to a simplified model that permits reliable prediction of gene essentiality.

The search for essential genes has long been a challenge. An essential gene is defined as one whose loss is lethal under a certain environmental condition. The identification of essential genes in bacteria promises to (i) identify critical genes and pathways for controlling pathogenic bacteria by identifying potential targets for antimicrobial drug development; (ii) reveal the minimal gene set for living organisms and to shed light on the origin of life; and (iii) reveal bacterial relationships during evolution. Several genome-wide mutant libraries of model microbes have been constructed. Although these libraries are invaluable for research in systems biology they show inconsistent essential gene results, even for closely related strains. This lack of consensus has prevented reliable prediction of essential genes or pathways in species that have not yet been examined.

To obtain a reliable account of essential genes, we turned to an opportunistic pathogen, *Streptococcus sanguinis* strain SK36, after completing its genome sequence. The streptococci encompass a large group of important human pathogens. Many *Streptococcus* species are responsible for infectious diseases, such as pneumonia, bacteremia, strep throat, rheumatic fever, scarlet fever, meningitis, infective endocarditis, and dental caries. *S. sanguinis* has long been recognized as a principal causative agent of infective endocarditis, and its virulence factors have been the subject of a number of investigations. It also initiates biofilm formation on tooth surfaces. The complete genome sequence provides an opportunity to greatly advance our understanding of this organism by enabling the construction of a comprehensive set of genome-wide mutants. *S. sanguinis* is in many ways an ideal candidate for such a study. The SK36 chromosome is 2.39 Mb and contains 2270 putative protein-coding genes, far fewer than that of most microbes used in previous genome-wide gene replacement mutagenesis studies; e.g., *Acinetobacter baylyi* with ~3310 genes, *Bacillus subtilis* with ~4100 genes, *E. coli* with ~4300 genes, and *S. cerevisiae* with ~6000 genes. More importantly, *S. sanguinis* is highly competent. In our laboratory, up to 20% of *S. sanguinis* cells can be transformed by a simple method using ≤50 ng of DNA. Mutants of non-essential genes are therefore readily obtained, facilitating identification of essential genes. We report here identification of essential genes in *S. sanguinis* and a simplified picture of gene essentiality in streptococci and other bacteria that has emerged from our findings.

**Results**

**Generation of *S. sanguinis* mutants.** We set out to identify the essential genes of *S. sanguinis* by systematic gene replacement. Taking advantage of the completed *S. sanguinis* genome, we designed a PCR method (Figure S1A) to
Figure 1 | Flowchart of mutant construction. Procedures for creation and identification of essential genes. Wt, wild type. *Wild-type and mutant band sizes expected to differ by $\geq 100$ bp.

Precisely replace target genes with the aphA-3 kanamycin resistance (Km') gene. The cassette used initially lacked a promoter to avoid potential dysregulation of downstream genes, which could complicate subsequent use of the mutant library in functional studies. To facilitate translation, we provided ribosome binding sites (RBS) for the aphA-3 and adjacent downstream S. sanguinis genes. To ensure efficient homologous recombination, we created long (~1 kb) flanking sequences upstream and downstream of each targeted gene. Flanking sequences were limited to 1 kb based on two considerations: (i) the total length of the amplicon (~3 kb; 1 kb for the aphA-3 gene plus 2 kb of flanking sequences) allowed ready PCR amplification; and (ii) the feasibility of sequencing flanking regions by Sanger’s method from both ends (~1 kb). To test our approach, we randomly selected 10 and then 96 genes for mutagenesis by double cross-over homologous recombination. We successfully created 10 and 93 mutants, respectively, suggesting that the approach was efficient. We then designed and synthesized primers for replacement of 2266 of the 2270 ORFs in S. sanguinis, excluding only those four ORFs contained entirely within larger ORFs. In total, we synthesized over 10,000 primers in the construction of the 2266 gene replacement constructs (Table S1). These constructs were created by high-throughput PCR and purification in a 96-well format. Three PCR reactions were required for each replacement and one or more PCR reactions for mutant confirmation (Figure S1A and Figure 1). Over 9,000 PCR fragments were amplified for mutant creation and confirmation. To preclude false identification of genes as essential due to low transformation frequencies, we performed experiments to optimize the transformation efficiency of S. sanguinis SK36. Using our optimized method, up to 2x10^8 mutant colonies could be obtained from 10^7 bacterial cells for non-essential gene transformations.

In the initial round of replacements, we found that many of the genes for which mutants could not be obtained were annotated as essential genes associated with specific pathways. The distribution of essential genes in specific categories based on KEGG pathway maps was examined. The categories of translation, carbohydrate metabolism, nucleotide metabolism, and replication and repair had the greatest number of essential genes, with over 30 each (Figure S2). The categories with the highest percentage of essential genes were translation (76%), transcription (67%), protein folding, sorting and degradation (57%), and glycan biosynthesis (50%).

We next assigned genes to biochemical pathways based on their KEGG annotations in KEGG (see Figure 2 for an overview and Table S4 for a detailed analysis). The essential genes were clustered in many pathways: (i) glycolysis, (ii) pentose phosphate pathway (PPP), (iii) peptidoglycan biosynthesis, (iv) terpenoid backbone biosynthesis, (v) glycerophospholipid metabolism, (vi) glycerolipid metabolism, (vii) fatty acid biosynthesis, (viii) nucleotide biosynthesis, (ix) metabolism of cofactors and vitamins including folate biosynthesis, (x) energy metabolism (production of ATP, NADH and NADPH), (xi) DNA replication, (xii) transcription, (xiii) protein biosynthesis, (xiv) GTP-binding proteins and (xv) cell division.

These results were in keeping with expectation, given our current understanding of bacterial metabolism. Among the metabolic
pathways containing essential genes, glycolysis and pyruvate oxidation from α-D-glucose 6-phosphate to acetyl-CoA play a pivotal role because they generate energy and provide input molecules for other essential metabolic pathways. Terpenoid backbone biosynthesis feeds into the pathway for synthesis of peptidoglycan, which is the main component of the cell wall. Glycerophospholipid metabolism in conjunction with glycerolipid produces lipoteichoic acid (LTA), a component of the cell wall in gram-positive bacteria. Phospholipids biosynthesized from glycerophospholipid metabolism and fatty acid biosynthesis compose the cell membrane. The nucleotides synthesized from PRPP derived from the pentose phosphate pathway (PPP) or from nucleosides imported from the growth medium supply materials for DNA replication and transcription. Several vitamins and cofactors including folate, biotin, pantothenate, nicotinate, riboflavin, SAM and FeS clusters are involved in the pathways above. ATP and NADH/NADPH produced from glycolysis, pyruvate oxidation and PPP provide energy.

We then analyzed essential genes by network functions using systems biology. All essential genes and their related functions were linked together in pathways and studied as a whole. This picture was simplified dramatically when it became apparent that all of these pathways could be linked to three basic biological functions: maintenance of the cell envelope, energy production and processing of genetic information. Remarkably, we found only three essential genes (none of which have been assigned exact functions) that could not be linked to these three functions (SSA_0575, SSA_0800, and SSA_1903; Table S4). SSA_0575 is annotated as a haloacid dehalogenase-like hydrolase, SSA_0800 as a glutamine amidotransferase and SSA_1903 as a conserved hypothetical protein.

Identification of additional essential genes in minimal medium.

From our analysis, we found that the three functions of maintenance of the cell envelope, energy production and processing of genetic information, and no others, are essential to S. sanguinis. Based on the genome sequence, we predicted that S. sanguinis possesses de novo synthesis pathways for all amino acids and nucleotides from glycolysis and PPP (Figure 3A). We identified 96 genes responsible for biosyntheses of 19 amino acids starting from glycolysis, leaving only the undetermined gene responsible for L-asparagine biosynthesis from L-aspartate (Figure 3A). We identified 96 genes responsible for biosyntheses of 19 amino acids starting from glycolysis, leaving only the undetermined gene responsible for L-asparagine biosynthesis from L-aspartate (Figure 3A). Two pathways each were present for biosyntheses of L-glutamate, L-serine and L-glycine and L-threonine. Yet, in our initial screen, we found few essential genes related to
amino acid synthesis or nucleotide synthesis from PRPP to UMP/GMP/AMP, suggesting amino acids and nucleotide precursors were provided by the rich brain heart infusion (BHI) medium. To examine this possibility, we compared the growth of strains with mutations in genes responsible for amino acid and nucleotide de novo synthesis in a chemically defined medium (CDM) with that in BHI, as described previously27 (Figure 3B; Table S5). CDM contains only 3 amino acids (L-glutamate, L-cysteine and L-leucine) and lacks nucleosides and nucleobases. All of these mutants grew to levels similar to SK36 in BHI medium (Figure 3B). However, the growth of the mutants was significantly lower than that of SK36 in CDM medium, with the following exceptions: (i) the mutants whose deleted genes were involved in the synthesis of the amino acids that were supplied in CDM medium (Glu, Cys and Leu); (ii) the mutants possessing alternative pathways (Ser, Gly and Thr); and (iii) the mutants whose replaced genes had paralogs or isozymes. Although dramatically reduced, some of the mutants still grew to low levels, perhaps benefitting from carry-over of nutrients from the inocula. Similar results were obtained with 23 mutants associated with nucleotide biosynthesis from PRPP and/or L-glutamine (Figure 3B). The combined results suggest that S. sanguinis does indeed obtain required amino acids and nucleotide precursors from the BHI medium. Provision of nutrients by BHI appears to explain another result that initially appeared discrepant with our model. We predicted that SSA_1201, SSA_1202 and SSA_1033 would be essential because they were required for biosynthesis of CoA from pantothenate, but mutants of these three genes were obtained (Figure 2). We found, however, that these mutants exhibited minimal or undetectable growth in CDM, suggesting pantetheine-4P is provided by BHI (Table S5).

Examination of paralogs and isozymes with double-knockout mutants. In our analysis, we also found that some genes predicted to be critical in one of these three functions were not identified as essential. We considered that their functions may be performed by genes encoding paralogs or isozymes with similar functions. To examine this possibility, we identified gene paralogs based on protein sequence homology (Table S3). We also examined the genome

Figure 3 | Growth of S. sanguinis mutants related to amino acid and nucleotide biosyntheses in BHI and CDM media. (A) amino acid and nucleotide biosynthesis pathways in which mutants grew (black arrow) or did not grow (red arrow) in CDM medium. Components of the CDM medium are indicated in green. PPP, pentose phosphate pathway, dashed arrow, undetermined gene responsible for asparagine biosynthesis. Note the two pathways for serine, glycine, threonine, glutamate and AICAR biosynthesis. ASA, L-aspartate 4-semialdehyde; CM, chorismate; E4P, D-Erythrose 4-phosphate; Hser, homoserine or o-phospho-L-homoserine; PRFAR, phosphoribulosyl-formimino-AICAR-phosphate. (B) growth of mutants in BHI and CDM media. Relative growth was represented by mutant OD_{450} / SK36 OD_{450} after 2d anaerobic incubation.
annotation and literature for potential isoforms. When we re-examined the linked essential genes in the network, we found paralogs or isoforms in every case in which a series of linked essential genes was interrupted by a “non-essential” gene in the same pathway (Figure 2; for example, paralogs SSA_0791/SSA_1494 involved in peptidoglycan biosynthesis, isoforms SSA_0578/SSA_2195 involved in NAD\(^+\)/NADP\(^+\) biosynthesis). This suggested that these non-essential genes might substitute for one another in performing essential functions. To test this hypothesis, we selected four pairs of paralogous or isoform genes for double-gene knockouts to examine essentiality: SSA_0791/SSA_1494, SSA_1827/SSA_2168, SSA_0578/SSA_2195, and SSA_0352/SSA_1188. As a control, double mutants were created by combining the replacement of each gene of interest with a replacement of the SSA_0169 gene, which is a hypothetical gene with no known function\(^{22}\). As we anticipated, double mutants could not be constructed for the gene pairs SSA_0791/SSA_1494, SSA_0578/SSA_2195, or SSA_0352/SSA_1188, whereas SSA_0169 double mutants were readily obtained (Table 1). Unexpectedly, the double mutant SSA_1827/SSA_2168 was viable. This may result from the existence of an alternative pathway (composed of SSA_0049/SSA_0050, SSA_0287 and SSA_1826) in glycerolipid metabolism.

**Identification of essential genes in other streptococcal species.** To validate the hypothesis that essential genes are associated with these three functions, we examined representative genes in two more pathogenic streptococcal species. We first selected 8 genes in *S. pneumoniae* of the essential genes with homologs in *S. sanguinis* matched *S. pneumoniae* essential genes, with the rest matching non-essential *S. pneumoniae* genes. The vast majority of *S. sanguinis* essential genes (202 of 218) had orthologs in all of the other 48 streptococcal genomes (Table S6A). Thus, in agreement with expectation, most of the genes we identified as essential were highly conserved within streptococci.

**Comparison with essential genes in other gram-positive bacteria.** We then compared essential genes of *S. pneumoniae* with those in other species that have been examined experimentally. All essential genes of the 13 bacterial species presently included in the Database of Essential Gene (DEG)\(^{33}\) were collected and searched against the *S. pneumoniae* protein database by BLASTP to find homologs (Table S7). The sequence comparison suggested that there were many differences. In the other species, about half (33 to 66%) of the essential genes with homologs in *S. pneumoniae* matched *S. pneumoniae* essential genes, with the rest matching non-essential *S. pneumoniae* genes.

To find the reason for this inconsistency, we analyzed more carefully the other two gram-positive species in the database, *Staphylococcus aureus* NCTC8325 and *B. subtilis* 168 (Figure 4 and Table S8), that were subjected to genome-wide screens\(^{12,22}\) in similar nutrient-rich media (BHI and LB). Some essential genes in *S. pneumoniae* had homologs identified as non-essential in *S. aureus* and

### Table 1 | Double-gene knockouts in *S. sanguinis*.

| Double mutant | Donor   | Product Name                                      | Acceptor | CFU   | Gel | Essential |
|---------------|---------|--------------------------------------------------|----------|-------|-----|-----------|
| Ssx_0791::1494| SSA_0791| UDP-N-acetylglucosamine 1-carboxyvinyltransferase| Ssx_1494 | 0     | yes |           |
| Ssx_0578::2195| SSA_0578| nicotinic acid mononucleotide adenyllytransferase| Ssx_2195 | 0     | yes |           |
| Ssx_0352::1188| SSA_0352| ribonuclease Hill                                 | Ssx_1188 | 0     | yes |           |
| Ssx_2168::1827| SSA_2168| NAD(P)H-dependent glycerol-3-phosphate dehydrogenase| Ssx_1827 | >5000 | 1-band | No        |
| Ssx_0169::1494| SSA_0169| hypothetical protein                              | Ssx_1494 | >5000 | 1-band | No        |
| Ssx_0169::2195| SSA_0169| hypothetical protein                              | Ssx_2195 | >5000 | 1-band | No        |
| Ssx_0169::1188| SSA_0169| hypothetical protein                              | Ssx_1188 | >5000 | 1-band | No        |
| Ssx_0169::1827| SSA_0169| hypothetical protein                              | Ssx_1827 | >5000 | 1-band | No        |

All mutants with double crossover replacement are named Ssx. The number indicates the replaced gene.

### Table 2 | Confirmation of essential genes in *S. pneumoniae* and *S. mutans*.

| SP_ID | Product Name                                      | CFU* | Gel† | Essential | SMU_ID* | CFU* | Gel† | Essential |
|-------|--------------------------------------------------|------|-----|-----------|---------|------|-----|-----------|
| SP_0382| diphosphomevalonate decarboxylase                 | 0    | yes |           | SMU_937 | 0    | yes |           |
| SP_0383| diphosphomevalonate kinase                       | 0    | yes |           | SMU_938 | 0    | yes |           |
| SP_0384| isopentenyl pyrophosphate isomerase              | 0    | yes |           | SMU_939 | 0    | yes |           |
| SP_1514| FOFL ATP synthase subunit C                     | 2    | 2-band | yes | SMU_1534 | 1 | 2-band | yes |
| SP_1513| FOFL ATP synthase subunit A                     | 2    | 2-band | yes | SMU_1533 | 0 | yes |           |
| SP_1511| FOFL ATP synthase subunit delta                  | 6    | 2-band | yes | SMU_1531 | 1 | 2-band | yes |
| SP_0262| phosphatase cytidylyltransferase                 | 0    | yes |           | SMU_1785 | 0 | yes |           |
| SP_0261| undecaprenyl pyrophosphate synthase              | 0    | yes |           | SMU_1786 | 0 | yes |           |
| SP_0489| PAP2 family protein                              | 242  | 1-band | no | SMU_1702c | 874 | 1-band | no |

* Gene identifiers. IDs in the same row indicate orthologous genes
† Number of transformants obtained from attempt to mutagenize the gene indicated
‡ Appearance of PCR products resulting from amplification of the target gene locus from selected transformants
B. subtilis, and vice versa although many essential genes in S. sanguinis were shared with these two species. However, we found that most of the differences could be explained by gene paralogs, isozymes, or alternative pathways. As shown in Figure 4, B. subtilis synthesizes peptidoglycan from UDP-N-acetylmuramoyl-L-alanyl-D-glutamate via an alternative pathway using meso-2,6-diaminopimelate rather than L-lysine that results in the genes responsible for producing meso-2, 6-diaminopimelate from L-aspartate 4-semialdehyde via lysine biosynthesis pathway being essential. S. aureus cross-links peptidoglycan via the essential femXAB operon rather than the streptococcal murMN. The essential mevalonate pathway in the terpenoid backbone pathway in S. sanguinis and S. aureus is replaced by an essential MEP/DOXP pathway in B. subtilis. In S. sanguinis, LTA and wall teichoic acid (WTA, a component of cell wall in gram-positive bacteria) are likely synthesized by the same set of enzymes because both have identical components in S. sanguinis and repeating units with identical structures in S. pneumoniae, whereas S. aureus and B. subtilis use a set of enzymes different from LTA synthesis to produce wall teichoic acid, and the genes encoding them are essential. S. sanguinis lacks a respiratory chain and undergoes fermentation to produce acids such as lactate and acetate, and the F1Fo-ATPase uses ATP hydrolysis to pump intracellular protons out. This is indispensible for creating a protonmotive force for a variety of transport processes and to maintain pH homeostasis, which likely results in the genes encoding F1Fo-ATPase being essential in S. sanguinis. In contrast, S. aureus and B. subtilis use the electron transport chain to generate protonmotive force, and the genes associated with its biosynthesis (i.e. menaquinone) are essential. B. subtilis and S. aureus likely maintain cellular pH homeostasis via a Na+/H+ transporter, the genes for which are essential in B. subtilis and paralogous in S. aureus. S. sanguinis lacks this transporter. In genetic information processing, the set of essential genes in S. sanguinis is smaller than those in other two species, but their essential pathways are still identical.

Discussion

Accurate prediction of essential genes is important for identification of drug targets to combat the emergence of pathogens and antibiotic-resistant bacteria, especially for serious infectious agents for which there is no research model system available. In addition, the rational design of bacterial cells through synthetic biology, as is currently possible, requires an understanding of the minimal gene set for further advances. However, the lack of consensus among experimental essential gene lists has prevented prediction of essential genes in other species. In our essential gene comparisons, we found significant differences of essential genes identified for S. sanguinis and those identified in previous studies with other species. We suggest several possible explanations for these differences. (i) We would expect some different genes due to genetic inheritance among species. For example, the essential genes for cell wall composition in gram-positive bacteria are different from those in gram-negative bacteria. (ii) Identified essential genes using single-gene knockouts will vary when the bacteria being compared differ in protein paralogs, isozymes or alternative pathways for essential functions. These cases can be tested by creating double-gene knockouts, as we have demonstrated (Table 1). (iii) Differences in screening conditions may have caused differences in essential gene identification, although the most favorable environmental conditions for each bacterium were used in most screens. As we found for S. sanguinis above, the use of a minimal medium obviously results in the identification of additional essential genes. (iv) False-negative results may have been obtained in some cases due to partial gene inactivation, which is known to occur frequently in random insertion mutagenesis. (v) Some genes may have been falsely identified as essential because of genetic system limitations. Many bacterial species have low transformation frequencies. In these cases, even a moderate reduction in plating efficiency may result in lack of mutant recovery, resulting in categorization of the gene as “essential” in a large-scale, genome-wide mutagenesis. (vi) Finally, insufficient expression of the selection marker, as occurred with our promoterless Km+ gene, may result in false-positive identification of essential genes.

Many in silico methods have also been established to predict essential genes. These methods include ortholog identification, genomic intrinsic feature analysis, gene evolution rate, phylogenetic conservation, network analysis and machine learning based integrative approaches. Of these, essential gene prediction via phyletic conservation is the most commonly used. Our results indicate that the vast majority of S. sanguinis essential genes are indeed conserved.
in closely related species (Table S6A). We also found that essential genes had greater sequence conservation among the streptococci than non-essential genes. The average identity for the essential gene protein sequences was 79.71% while for the non-essential gene protein sequences, it was 58.96% (Table S6B). It should be noted, however, that phyletic conservation alone was a poor predictor of gene essentiality. We found 787 non-essential S. sanguinis genes with orthologs in all 48 streptococcal genomes. Inclusion of bacteria from other genera in our analysis would have reduced the number of conserved genes identified, likely increasing the specificity of the phyletic method for essential gene prediction. However, the sensitivity of this method then decreases, as we found a number of essential S. sanguinis genes with orthologs that can be detected in different genera by identity of annotated function, but not by sequence similarity (data not shown).

Using our findings, we believe that most essential genes in other bacteria can now be predicted based on their genome annotations. We have established a model of essential pathways (Figure 5), in which the essential genes are linked by crucial chemical compounds. Although essential genes may differ among species due to different final products or alternative pathways, we propose they will contribute to the three basic functions of maintenance of the cell envelope, energy production, and processing of genetic information. We used this model for predicting essentiality of paralogs and isozenymes by double-gene knockouts in S. sanguinis and for predicting essential genes in S. mutans, S. pneumoniae, B. subtilis and S. aureus. Our predictions were largely confirmed.

This model can also be used to explain many of the apparent inconsistencies observed previously among different organisms that were not predictable based on gene sequence conservation alone. Bacteria can be categorized into gram-negatives, gram-positives and Mycobacteria due to the differences in their cell envelope compositions. The greatest difference is that gram-negative bacteria contain an outer membrane with lipopolysaccharides while LTA and WTA are found in gram-positive bacteria and mycolyl-arabinogalactan-peptidoglycan complex in Mycobacteria. This leads to different sets of essential genes responsible for the respective envelope components. In the gram-negative bacteria E. coli14 and Pseudomonas aeruginosaa, the genes responsible for lipopolysaccharide biosynthesis have been demonstrated essential. Many genes responsible for biosynthesis of mycolyl-arabinogalactan-peptidoglycan complex were identified as being essential in Mycobacterium tuberculosis15. In B. subtilis and S. aureus16,17, the genes responsible for the biosynthesis of WTA and LTA were essential. Our results indicated most S. sanguinis orthologs of genes responsible for LTA biosynthesis in S. aureus were nonessential. Although this can be explained by the existence of paralogs (Figure 2), we cannot exclude the possibility of an alternative biosynthetic pathway for S. sanguinis LTA synthesis, as glycerophosphate residues in the LTA of S. sanguinis DSM 20567 may be substituted with D-alanine ester, α-D-gluco-syl and α-isomalto-oligosaccharide residues18.

Most bacteria possess a respiratory chain that is important for energy production and maintenance of redox balance. It has been demonstrated that the genes involved in synthesis of electron transport chain components, such as CoQ, menaquinone or heme, are essential in E. coli19, P. aeruginosaa, B. subtilis and S. aureus16,17. Although an important function of electron transport is generation of ATP via the F1Fo-ATPase, there was only one instance of an F1Fo-ATPase gene being identified as essential in previous studies performed with the latter two species16,17. One possibility is that these genes were missed. Most replacements of F1Fo-ATPase genes in S. sanguinis, S. pneumoniae, and S. mutans generated "double-band"

![Figure 5](image-url) | **Figure 5**: Deduced essential pathways based on S. sanguinis essential genes. Colors indicate the three functions associated with essential pathways, as in Figure 2. I, glycolysis and pyruvate oxidation; II, peptidoglycan biosynthesis; III, terpenoid backbone biosynthesis; IV, glycerolipid metabolism (LTA biosynthesis); V, WTA biosynthesis; VI, lipopolysaccharide biosynthesis; VII, MAP biosynthesis; VIII, phosphoglycerolipid metabolism; IX, fatty acid biosynthesis; X, electron transport chain; XI, pentose phosphate pathway; XII, nucleotide biosynthesis; XIII, genetic information processing including DNA replication, transcription, protein biosynthesis, GTP-binding proteins and cell division. Black arrow, all bacteria; black dashed arrow or oval, Mycobacteria only; green arrow, gram-negative bacteria only; blue arrow, gram-positive bacteria only; red arrow, Mycobacteria only; violet arrow, gram-positive bacteria only; orange arrow, gram-negative bacteria only; green arrow, gram-negative bacteria only; yellow arrow, Mycobacteria only; black arrow, all bacteria; black dashed arrow or oval, alternative pathways; blue arrow, gram-positive bacteria only; green arrow, gram-negative bacteria only; violet arrow, mycobacteria only; bold arrow, synthesis of cell envelope. CA, chorismate; Cyt, cytochrome; DAG, 1,2-diacylglycerol; DG-DAG, diglycerol-diacylglycerol; G3P, D-glyceraldehyde 3-phosphate; GPPGL, glycerophosphoglycerolipid; L-G3P, glycerol-3-phosphate; LPS, lipopolysaccharide; Malonyl-ACP(CoA), malonyl-ACP or malonyl-CoA; M-DAP, meso-2,6-diaminopimelate; MEV, mevalonate; MG-DAG, monoglycerol-diacylglycerol; PA, phosphatidate; PG, phosphatidylglycerolphosphate; R-5-P, D-Ribulose 5-phosphate; PPP, pentose phosphate pathway; PtGro, phosphatidylglycerol; UDP-MAG, UDP-N-acetylmuramoyl-L-allyl-D-glutamate; UDP, undecaprenyl diphosphate.
mutants (Table 2; Table S3). This would lead to identifying F1Fo-ATPase genes as nonessential in studies in which the presence of double-bands was not investigated. Nevertheless, the ability of these species to grow anaerobiocically indicates that ATP generation by F1Fo-ATPase is not essential, suggesting that electron transport is required for other purposes, such as secondary active transport. In S. sanguinis, all eight subunits of the F1Fo-ATPase were identified as essential. Since streptococci lack a respiratory chain, the essential function of the F1Fo-ATPase in these species is also not generation of ATP and is, instead, likely to be export of protons using energy from ATP hydrolysis. A previous analysis of essential genes in S. pneumoniae identified five of the eight F1Fo-ATPase subunit genes as essential.10,17,18,24 Here, we demonstrated the other three genes encoding F1Fo-ATPase components were also essential in S. pneumoniae, as were their orthologs in S. mutans (Table 2).

In conclusion, by choosing an ideal test organism and by employing exhaustive measures to avoid false-positive and false-negative identifications, we have reliably identified the essential genes of S. sanguinis. The validity of our findings is suggested by the virtually perfect association between our list of essential genes and the list of genes expected to be required for the three functions of cell envelope maintenance, energy production, and processing of genetic information. Although S. sanguinis is important in its own right, the relative ease with which these results can be used to identify essential genes in other prominent streptococcal pathogens including S. pyogenes and S. pneumoniae lends increased importance to this work. Moreover, our study suggests that with minimal additional effort, our results can be used to predict essential genes for most bacteria for which an annotated genome sequence is available.

It should be noted that although we examined all protein coding regions for their essentiality, it is well known that some non-coding regions are also essential if they contain DNA sequences for important biological functions. Such sequences include the chromosomal origin of replication, promoters, tRNAs, rRNAs and perhaps small RNAs.

Finally, although the focus of the current study was the identification of essential genes, the study has also produced an ordered, comprehensive library of non-essential gene mutants of S. sanguinis SK36. The design of the mutagenic constructs to (i) ensure near-complete deletion of each gene; (ii) retain expression signals for adjacent genes; and (iii) introduce a promoter only in cases where it was required, in combination with the care with which each mutant was characterized ensures that the library will be of great value. We are currently employing the mutants for a number of investigations of gene function involving conventional and systems biology approaches.

Methods

Bacterial strains and growth. S. sanguinis strain SK36 was grown at 37 °C in BHI broth (BD) as described previously.24 S. pneumoniae TIGR4 and S. mutans UA159 strains purchased from ATCC were grown in Todd Hewitt (TH; BD) broth plus 0.5% NaCl cultured overnight at 37 °C. Streptococci were then transferred into 96-well plates for growth with the overnight cultures being diluted with multichannel tips. The inocula were incubated for 2 days at 37 °C for the same conditions as above. Cultures were then mixed by pipetting several times with a multichannel pipette, and 200 μl cultures were transferred into 96-well plates for measuring OD450. The relative growth of each mutant was calculated as mutant OD450/SK36 OD450.

Gene function analysis. The pathway distributions of essential genes were analyzed via KEGG. KEGG annotations were downloaded to a local computer from /pub/kegg/. The assigned KO numbers and path numbers of S. sanguinis genes were extracted. Multiple KO numbers or path numbers from KEGG were often assigned for a single gene if it was involved in different pathways. To view the essential pathways, these genes were assigned to the pathway with the greatest percentage of essential genes. To study essential genes as a whole, as many essential genes as possible were linked together via pathways. Based on a product in one pathway being the substrate of another pathway, the essential pathways were linked together.

Gene replacement and mutant storage. The above linear PCR amplicons (~50 ng) were directly transformed into S. sanguinis as described previously.24 Allelic exchange mutants generated by double cross-over homologous recombination were selected by two-day microaerobic incubation on BHI agar plates containing 500 μg/ml kanamycin. For each replacement mutant, one colony was randomly picked and cultured in BHI with Km. To determine whether the mutant contained the expected gene replacement, colony PCR was performed using F1 and R3 primers. The PCR amplicon was examined by long agarose gel electrophoresis. The sequencing confirmation was performed using the Kmr primer, which binds to the Km’ cassette (Figure S1A). The mutations which gave rise to a DNA amplicon of the expected size by long gel electrophoresis and with the expected junction sequence determined from the Km’ primer or the T1 primer were collected as the final confirmed mutants. These were retained and cryopreserved in 30% glycerol at −80 °C.

The methods for construction of gene replacement amplicons with the promoter-containing Km’ cassette for S. mutans UA159 and S. pneumoniae TIGR4 were the same as for S. sanguinis. Transformation of S. mutans UA159 was conducted similarly to that of S. sanguinis except S. mutans CSP was used. Transformation of S. pneumoniae TIGR4 was performed as described by Bricker and Camilli24.

Microarray analysis. S. sanguinis cultures at late log phase were used for microarray analysis. RNA from each of three independent samples was isolated by RNeasy mini kit (Qiagen, Valencia, CA). Spotted microarray slides were obtained from the Pathogen Functional Genomics Resource Center at JCVI. The microarray was performed according to the manufacturer’s protocol. Each sample was divided into two parts and labeled separately by Cy5 and Cy3 dyes for microarray. The microarray data were analyzed using the programs SpotFind and Midos to obtain the expression ratio of each gene labeled with each dye. All ratios were within the range of 0.6 to 1.5, indicating consistency in the labeling and analysis. Additionally, the ratio of dye intensity to background in a microarray slide was obtained after SpotFind analysis. Absolute expression of each gene was represented by the average ratio of dye intensity to background for each slide. The microarray data have been deposited in the NCBI Gene Expression Omnibus (GEO) with record GSE25340.

Growth comparison of mutants in CDM and BHI media. Selected mutants were cultured overnight at 37 °C in 96-well blocks with 1 ml BHI broth under anaerobic conditions (10% CO2, 10% H2 and 80% N2 with a palladium catalyst). The overnight cultures were inoculated into 1 ml either CDM24 or BHI medium in 96-well plates by dipping with multichannel tips. The inocula were incubated for 2 days at 37 °C for the same conditions as above. Cultures were then mixed by pipetting several times with a multichannel pipette, and 200 μl cultures were transferred into 96-well plates for measuring OD450. The relative growth of each mutant was calculated as mutant OD450/SK36 OD450.

Comparative genomics. Comparative genomic analyses were performed to identify conserved proteins in S. sanguinis as previously described.24 We downloaded 48 other completed streptococcal genomes and their annotations from public databases. S. sanguinis proteins were compared to other streptococcal protein databases by BLASTP. Significant matches (E-value < 1e-5) were tabulated for each of these other streptococcal genomes. The S. sanguinis database was also BLASTP searched against itself to identify paralogs. Protein sequence conservation was calculated by percent amino acid identity of orthologs. Essential genes in Staphylococcus aureus NCTC
30. Trieu-Cuot, P., Klier, A. and Courvalin, P. DNA sequences specifying the
27. Ge, X.
28. Turner, L. S.
25. Di Filippo, S.
23. Chaudhuri, R. R.
22. French, C. T.
21. de Berardinis, V.
24. Xu, P.
8. Akerley, B. J.
5. Liao, B. Y., Scott, N. M. and Zhang, J. Impacts of gene essentiality, expression
14. Sassetti, C. M., Boyd, D. H. and Rubin, E. J. Genes required for mycobacterial
13. Kobayashi, K.
12. Jacobs, M. A.
11. Gerdes, S. Y.
6. Ji, Y., et al. Identification of critical staphylococcal genes using conditional
19. Gallagher, L. A.
7. Giaever, G.
3. Glass, J. I.
1. Gerdes, S. Y.
10. Thannasi, J. A. et al. Identification of 113 conserved essential genes using a high-
throughput gene disruption system in Streptococcus pneumoniae. Nucleic Acids
Res 30, 3152–62 (2002).
1. Gerdes, S. Y. et al. Experimental determination and system level analysis of essential
genes in Escherichia coli MG1655. J Bacteriol 185, 5673–84 (2003).
12. Jacobs, M. A. et al. Comprehensive transposon mutant library of Pseudomonas
aeruginosa. Proc Natl Acad Sci U S A 100, 14339–44 (2003).
13. Kobayashi, K. et al. Essential Bacillus subtilis genomes. Proc Natl Acad Sci U S A 100,
14675–83 (2003).
14. Sassetti, C. M., Boyd, D. H. and Rubin, E. J. Genes required for mycobacterial
growth defined by high density mutagenesis. J Bacteriol 186, 7926–35 (2004).
15. Song, J. H. et al. Identification of essential genes in Streptococcus pneumoniae by
allelic replacement mutagenesis. Mol Cells 19, 365–74 (2005).
16. Baha, T. et al. Construction of Escherichia coli K-12 in-frame, single-gene
knockout mutants: the Keio collection. Mol Syst Biol 2, 2006.0008 (2006).
17. Liberati, N. T. et al. An ordered, nonredundant library of Pseudomonas aeruginosa
strain PA14 transposon insertion mutants. Proc Natl Acad Sci U S A 103, 2833–8 (2006).
18. Gallagher, L. A. et al. A comprehensive transposon mutant library of Francisella
novicida, a bioweapon surrogate. Proc Natl Acad Sci U S A 104, 1009–14 (2007).
19. de Berardinis, V. et al. A complete collection of single-gene deletion mutants of
Acinetobacter baylyi ADP1. Mol Syst Biol 4, 174 (2008).
20. French, C. T. et al. Large-scale transposon mutagenesis of Mycoplasma pulmonis.
Mol Microbiol 69, 67–76 (2008).
21. Chandhuri, R. R. et al. Comprehensive identification of essential Staphylococcus
aureus genes using Transposon-Mediated Differential Hybridisation (TMDH).
BMC Genomics 10, 291 (2009).
22. Xu, P. et al. Genome of the opportunistic pathogen Streptococcus sanguinis. J Bacteriol
189, 3166–75 (2007).
23. Di Filippo, S. et al. Current patterns of infective endocarditis in congenital heart
disease. Heart 92, 1490–95 (2006).
24. Paik, S. et al. Identification of virulence determinants for endocarditis in
Streptococcus sanguinis by signature-tagged mutagenesis. Infect Immun 73,
6064–74 (2005).
25. Ge, X. et al. Identification of Streptococcus sanguinis genes required for biofilm
formation and examination of their role in endocarditis virulence. Infect Immun 76,
2551–9 (2008).
26. Turner, L. S. et al. Development of genetic tools for in vivo virulence analysis of
Streptococcus sanguinis. Microbiology 155, 2573–82 (2009).
27. Kohenrander, E. P. and London, J. Adhere today, here tomorrow: oral bacterial
adherence. J Bacteriol 175, 3247–52 (1993).
28. Ge, X. et al. Identification of Streptococcus sanguinis genes required for biofilm
formation and examination of their role in endocarditis virulence. Infect Immun 76,
2551–9 (2008).
29. Trieu-Cuot, P., Klier, A. and Courvalin, P. DNA sequences specifying the
transcription of the streptococcal kanamycin resistance gene in Escherichia coli and Bacillus subtilis. Mol Gen Genet 198, 348–52 (1985).
30. Kanehisa, M. et al. The KEGG resource for deciphering the genome. Nucleic Acids
Res 32 Database issue D277–D280 (2004).
31. Xu, P. et al. The genome of Cryptosporidium hominis. Nature 431, 1107–12 (2004).
32. Zhang, R. and Lin, Y. DEG 5.0, a database of essential genes in both prokaryotes
and eukaryotes. Nucleic Acids Res 37(Database issue) D455–D458 (2009).
33. Briggs, G. H. and Holmgren, K. Ecology and nature of immunoglobulin A1 protease-
producing streptococci in the human oral cavity and pharynx. Infect Immun 31,
686–73 (1981).
34. Trieu-Cuot, P. and Courvalin, P. Nucleotide sequence of the Streptococcus
faecalis plasmid gene encoding the 3’5’-aminolucinolide phosphotransferase
type III. Gene 23, 331–41 (1983).
35. Lukomsik, S. et al. Nonpolar inactivation of the hypervariable streptococcal
inhibitor of complement gene (sic) in serotype M1 Streptococcus pyogenes
significantly decreases mouse mucosal colonization. Infect Immun 68, 535–42 (2000).
36. Bricker, A. L. and Camilli, A. Transformation of a type 4 encapsulated strain of
Streptococcus pneumoniae. FEMS Microbiol Lett 172, 131–5 (1999).

Acknowledgements
This work was supported by the National Institutes of Health grant R01DE018138 (PX) and,
in part, by Virginia Commonwealth University Presidential Research Incentive Program (PRIP) 144602-3 (PX). We thank the DNA Core Facility at Virginia Commonwealth University for ABI Big Dye sequencing. We thank the Pathogen Functional Genomics Resource Center at JCVI for providing S. sanguinis spotted microarray slides.

Author contributions
PX conceived the main idea and the conceptual design of the experiments. PX, XG, LC, XW, YD, JX, JP, VS, MT and KE constructed mutants; PX designed primers and performed bioinformatics analysis; XG established the high throughput transformation system and performed microarray and gene functional analysis; XW performed double-knockout and mutant growth comparison; PX, XG, LC, JX, TK and DB analyzed the data. PX, XG, TK, DB and GB wrote the manuscript.

Additional information
Supplementary information accompanies this paper at http://www.nature.com/
scientificreports

Competing financial interests: The authors declare no competing financial interests.

License: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/

How to cite this article: Xu, P. et al. Genome-wide essential gene identification in
Streptococcus sanguinis. Sci. Rep. 1, 125; DOI:10.1038/srep00125 (2011).