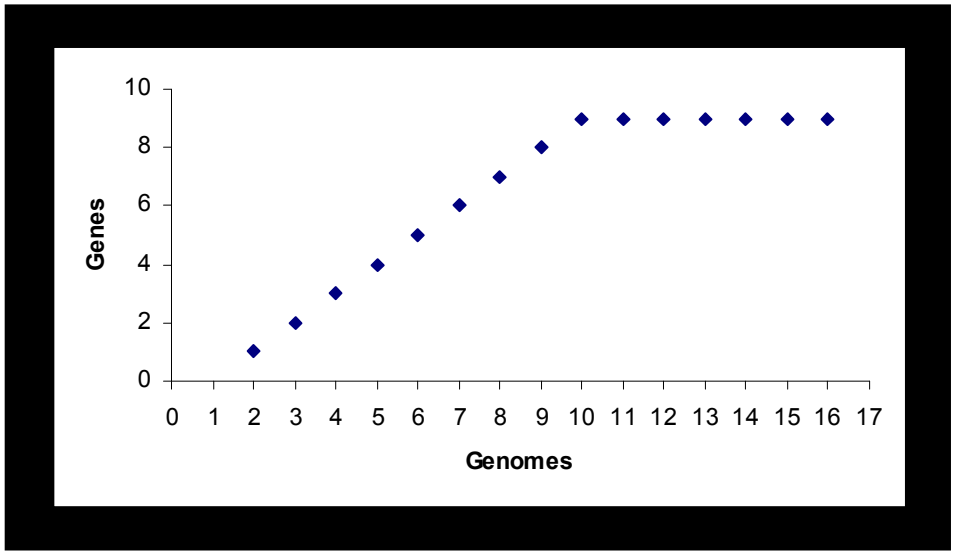
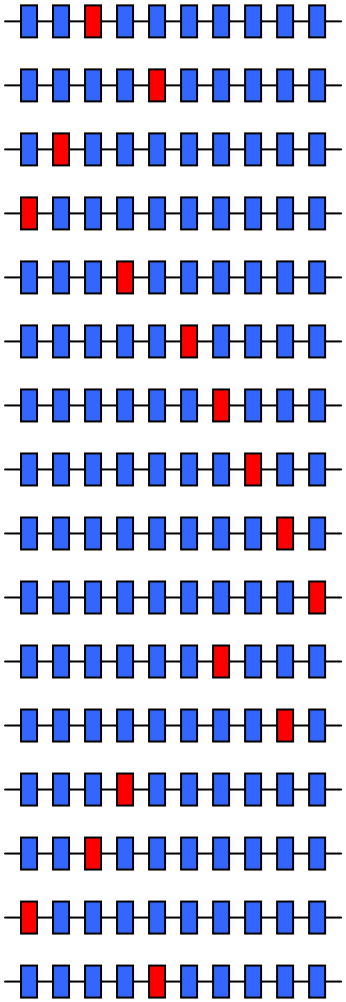
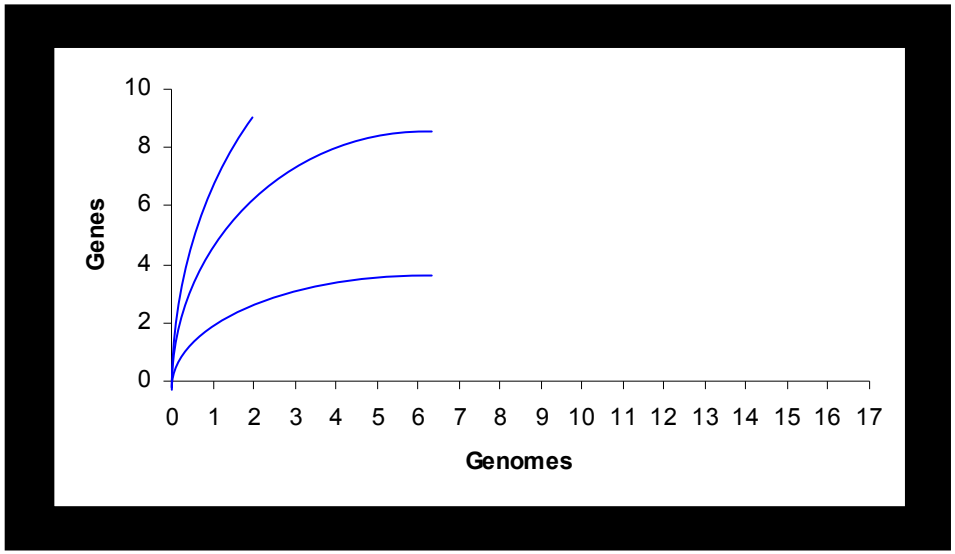
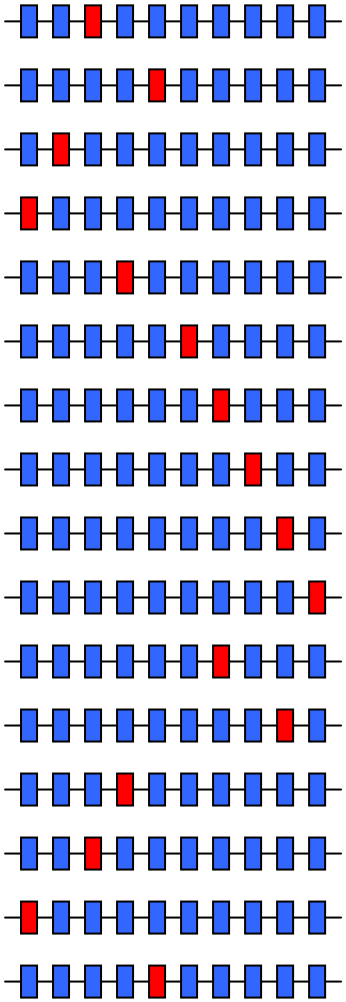
Παν-Genome





Pan-Genome





Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

1. Pan-genome >>> single bacterium's genome
2. Means of diversity: Genetic recombination & large non-core set of genes
3. 17 Streptococcus pneumoniae strains: CDSs grouped into 3,170 orthologous gene clusters
4. Of which 1,454 (46%) were conserved among all 17 strains
5. 1,716 (54%) not found in all strains
6. Each strain's genome: 21 and 32% noncore genes
7. Finite pan-genome model predicts:
    1. S. pneumonia pan-genome > 5,000 orthologous clusters
    2. 99% of the orthologous clusters (frequencies of >0.1) if 33 representative genomes are sequenced

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

8. S. pneumonia: model organism for the study of bacterial transformation
9. System for the uptake of DNA from the environment that allows for extensive recombination
10. 6 Streptococcus agalactiae: **~20%** of the genes not shared among all strains
11. 13 Haemophilus influenzae: **~50%** of the genes are conserved among all strains
12. Pan-genome pool: each member contributes to and draws genes from
13. Phylogenetic-tree building: sum|gn,i - gn,k|

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

14. Core and supragenome values differ depending on the order in which the strains are added (??)
15. Decay in the number of new clusters and stabilization of the number of core orthologous clusters at ~**1,400**
16. After a finite number of genomes is sequenced, the number of new orthologous clusters will be very low
17. All clusters:
    1. estimated supragenome size is **5,117** clusters and the core set contains **27%** of the clusters
    2. **90%** of the supragenome is predicted to be identified after **142** strains are sequenced
18. Clusters (frequencies >= 0.1):
    1. estimated supragenome size drops to **2,979** clusters where **46.5%** are core
    2. **95%** of the supragenome is predicted to be identified after sequencing of **17** strains and **99%** after sequencing of **33** strains

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

19. Unique genes:
    1. 62% hypothetical
    2. 5.4% phage
    3. 33% wide range of proteins including putative transporters, transcriptional regulators
20. Strains from the same geographical location may be more similar?
21. This enormous genetic diversity calls attention to the need for markers of human virulence phenotypes and highlights the potential difficulty associated with this task
22. S. pneumonia strains are presently categorized based on capsule type and MLST
23. The capsular serotype is an important virulence factor and affects the ability of pneumococci to cause invasive disease

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

24. Even within the same capsular type, virulence is highly related to the genetic background of the strains
25. Serotype, MLST type, and/or genetic background may correlate, but in other cases, they do not
26. Since pathogenesis is probably a consequence not only of capsular type but also of multiple other genes, MLST type and serotype alone are not ideal markers for the disease phenotype of S. pneumoniae strains
27. May not be universal: 8 Bacillus anthracis with no new genes uncovered after analysis of only 4 genomes
28. This degree of variation is intrinsic to naturally transforming bacteria such as H. influenzae and S. pneumoniae, which undergo extensive DNA recombination events
29. Both bacteria exist exclusively in the human mucosa, where they form biofilms (foster genomic plasticity?)

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

TABLE 2. Summary of CDSs and orthologous gene clusters for 17 *S. pneumoniae* strains

Gene category	No. of orthologous clusters (% of total)	No. of CDSs (% of genes)
Core	1,454 (46)	30,070 (73)
Distributed	1,140 (36)	9,679 (23)
Unique	576 (18)	576 (1)
Excluded by size		1,120 (3)
Total	3,170 (100)	41,445 (100)

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

TABLE 3. Numbers of CDSs and orthologous clusters for individual *S. pneumoniae* strains

Strain name	Genome size (kb)	No. of CDSs	No. of orthologous clusters	No. of unique clusters	% Noncore clusters[a]
INV200	2,200	2,259	1,850	28	21.4
PAT6420135	Unknown	2,500	1,913	10	24
R6	2,038	2,274	1,925	3	24.5
D39	2,000	2,304	1,940	3	25
CGSSp18BS74	2,033	2,354	1,955	13	25.6
CGSSp3BS71	2,027	2,331	1,960	7	25.8
CGSSp11BS70	2,044	2,336	1,986	25	26.8
TIGR4	2,160	2,410	1,993	1	27
INV104B	2,200	2,508	2,012	74	27.7
OXC141	2,200	2,663	2,014	67	27.8
CGSSp9BS68	2,090	2,397	2,021	37	28
CGSSp23BS72	2,053	2,411	2,022	41	28.1
CGSSp19BS75	2,069	2,432	2,031	40	28.4
CGSSp6BS73	2,119	2,434	2,056	55	29.3
CGSSp14BS69	2,084	2,763	2,068	61	29.7
23F	2,200	2,428	2,069	43	29.7
TIGR670-6B	2,100	2,641	2,157	68	32.6

[a] Calculated as (total clusters in the strain − 1,454 core clusters)/(total clusters in the strain).

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

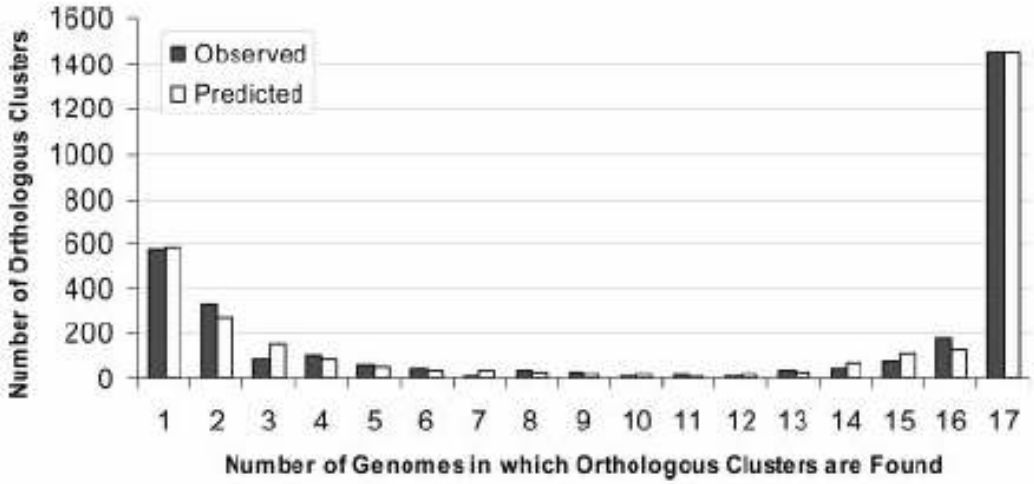


FIG. 1. Histogram of the number of observed and predicted (by the finite supragenome model) orthologous gene clusters that are present in a given number of genomes. There were 1,454 orthologous clusters observed in all strains (core); 1,140 distributed among more than one strain, but not all; and 576 in only one strain.

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

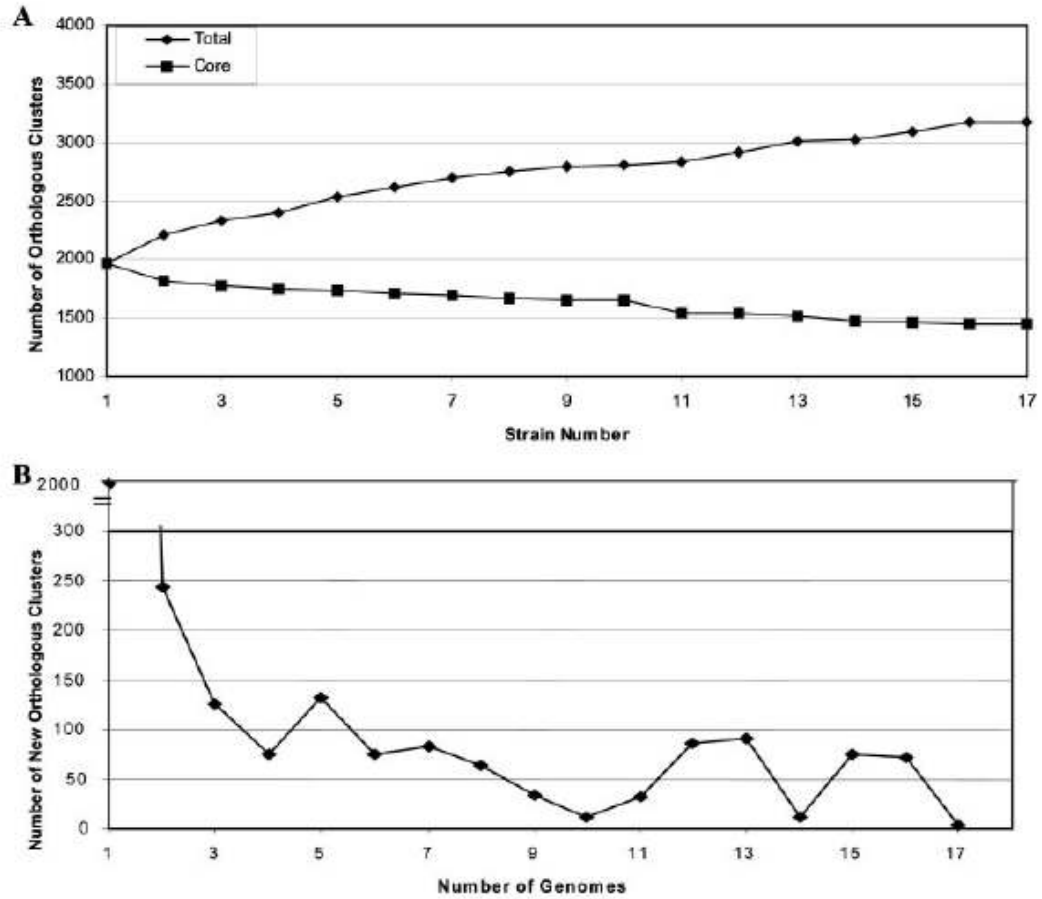


FIG. 4. (A) Plot of the numbers of total and core observed orthologous clusters as a function of the number of strains sequenced. (B) Plot of the number of new observed orthologous clusters as a function of each genome. Numbers were calculated first for two strains and then iteratively for strains added one by one.

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

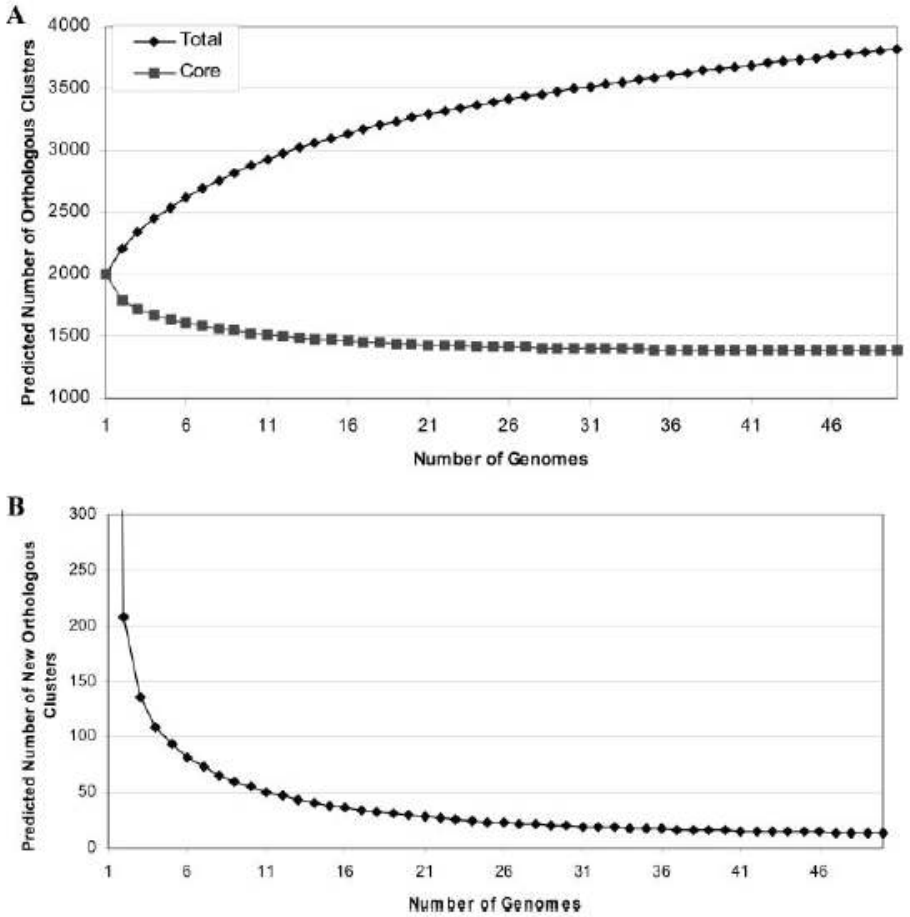


FIG. 5. Predictions using the finite supragenome model. (A) Plot of the numbers of total and core predicted orthologous clusters as functions of the number of strains sequenced. (B) Plot of the number of new predicted orthologous clusters as a function of each genome sequenced. Numbers were calculated first for two strains and then iteratively for strains added one by one.

Comparative Genomic Analyses of Seventeen Streptococcus pneumonia Strains: Insights into the Pneumococcal Supragenome

N. Luisa Hiller et al

TABLE 4. Predicted coverage of the *S. pneumoniae* supragenome using the finite supragenome model

Population frequency	Supragenome coverage (%)	No. of strains sequenced
≥0.1	90	11
≥0.1	95	17
≥0.1	99	33
All	90	142

Group B Streptococcus: global incidence and vaccine development

Atul Kumar Johri et al

1. Public health challenge: develop vaccines effective against infectious diseases that have global relevance
2. Vaccines against serotypes of group B Streptococcus (GBS) in the United States and Europe are not optimally efficacious against serotypes in other parts of the world
3. Group B Streptococcus (GBS or Streptococcus agalactiae): Gram-positive, opportunistic pathogen, colonizes the gastrointestinal and genitourinary tracts of up to 50% of healthy adults
4. Causes pneumonia, septicaemia and meningitis in neonates, is responsible for significant morbidity in pregnant women and the elderly, and is a serious cause of mortality in immunocompromised adults
5. Human isolates of GBS express a capsular polysaccharide (CPS), a major virulence factor that helps the microorganism evade host defence mechanisms
6. Isolates of GBS can be divided into nine CPS serotypes: Ia Ib, II, III, IV, V, VI, VII and VIII
7. Clinical trials of conjugate vaccines prepared with purified CPS types Ia, Ib, II, III and V have demonstrated that these preparations are safe and immunogenic
8. Not unexpectedly, these preparations do not offer protection against other GBS serotypes, such as type VIII

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

9. It is now possible to determine the complete genome sequence of a pathogen in a short period of time and at a relatively low cost and screen the inclusive set of potential proteins encoded by microorganisms in search of vaccine candidates (reverse vaccinology)
10. This genome-wide in silico prediction process typically targets approximately 10–25% of all genome-encoded proteins and necessitates high-throughput cloning and recombinant protein expression for target validation
11. Reverse vaccinology is being used against: streptococci, Chlamydiae spp., staphylococci, Plasmodium falciparum and bioterrorism associated agents including Yersinia pestis
12. GBS disease:
    1. In neonates is divided into early onset and late-onset disease. In early onset disease (the first 6 days of life), the neonate is usually infected by exposure to GBS before or during the birth process
    2. Some early onset infections can occur when the neonate is exposed to GBS during passage through the birth canal, but most early onset infections are probably caused by ascending movement of the organism from the maternal genital area through ruptured membranes into the amniotic fluid
    3. Here, the organism multiplies and ultimately colonizes the respiratory tract of the fetus
    4. As a consequence, pneumonia can develop and the bacteria can disseminate in to the bloodstream causing septicaemia
    5. Bloodstream dissemination allows the bacteria to reach multiple anatomical sites, where subsequent tissue penetration can result in meningitis and osteomyelitis

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

13. This disease progression indicates that GBS has to adhere, invade and transcytose several epithelial and endothelial cell barriers to cause disease
14. India 1999 neonatal GBS bacteraemia: 0.17 per 1,000 live births;Native Indians living inSouth Africa 1991: 2.6 per 1,000 live births
15. The principal difficulty in developing globally effective GBS vaccines is the existence of several serotypes with different geographical distributions and the heterogeneous cross-reactivity between serotypes — a vaccine suitable for Asian or European populations might not be suitable for African populations
16. Capsular polysaccharide-based vaccines: experiments demonstrating the protective nature of polysaccharide-specific antibody can be traced back to the 1930s when it was reported that protection against GBS infections in mice could be achieved by using CPS-specific polyclonal rabbit serum
17. These trials demonstrated the safety of the antigen but also highlighted the need to improve the immunogenicity of the CPS, as only 60% of the recipients of the type III CPS vaccine showed significant IgG responses
18. Consequently, to improve efficacy, the first GBS conjugate vaccines were prepared with serotype III. Conjugate vaccines based on all nine serotypes were prepared and tested pre-clinically, although there is little or no cross protection between serotypes

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

19. For 95% population coverage in Europe or North America, 5 serotypes are needed (Ia, Ib, II, III and V); however for other regions (e.g. Japan) would not be appropriate, owing to a different distribution of serotypes
20. Protein-based vaccines: Until quite recently, only a limited number of GBS proteins were investigated as potential vaccine candidates, including C protein complex, Rib, Sip and C5a peptidase
21. Only a few proteins are conserved at the gene level in most of the GBS isolates
22. Genomics allows antigen candidates to be identified on the basis of sequence conservation in different serotypes and strains of a given pathogen, and by predicting the surface exposure of a protein
23. Comparing S. agalactiae, S. pyogenes and Streptococcus pneumoniae ~50% of the genes are homologous, indicating substantial overlap in the virulence mechanisms used by these pathogens

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

24. Although, the sequence of a single genome does not reflect how genetic variability drives pathogenesis within a bacterial species, the identification of universal GBS vaccine candidates by multigenome analysis is promising
25. In this approach the genome sequences of 8 GBS strains belonging to different serotypes were compared: 1,811 genes (~80% of each genome) were shared by all strains — the *core* genome — and 765 genes were not present in all strains — the *variable* genome
26. Using in silico analysis, genes encoding putative surface-associated and secreted proteins were identified from these two sub-genomes
27. 589 proteins were identified (396 ***core*** genes and 193 ***variable*** genes), of which 312 were successfully expressed, purified and used to immunize mice. A combination of 4 proteins, Sip (core subgenome), and 3 other surface-associated proteins (variable subgenome), elicited protection in infant mice and their combination proved highly protective against a large panel of GBS strains, including all circulating serotypes

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

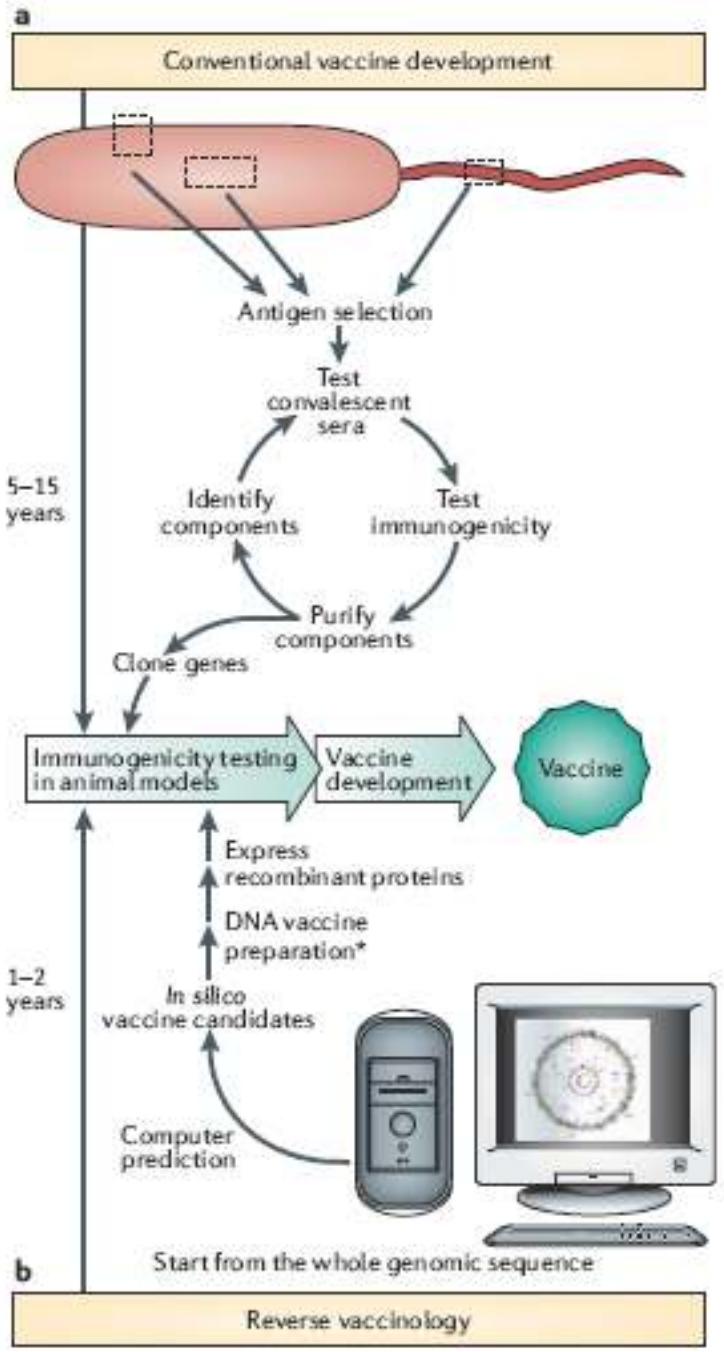
28. Many bacterial virulence factors and antigens are only expressed ***in vivo***, thus experimental approaches focused on ***in vitro*** grown bacteria can overlook important protective antigens
29. Proteomic approaches: Proteomics, in conjunction with genomic approaches, provides interesting insights into microbial pathogenesis at an organism level
30. There are several reasons for focusing on the analysis of proteins:
    1. first, the level of mRNA expression frequently does not represent the amount of active protein in a cell
    2. second, the gene sequence does not give any information on posttranslational modifications that could be essential for protein function and activity
    3. third, genome analysis does not provide information on dynamic cellular processes

Note:

Post-translational modification: The enzymatic processing of a polypeptide chain after translation from messenger RNA and after peptide-bond formation has occurred. For example, glycosylation, acylation)

Multilocus sequence type (MLST): An unambiguous procedure for characterizing isolates of bacterial species using the sequences of internal fragments of seven housekeeping genes. For each housekeeping gene, the different sequences that are present in a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the seven loci define the allelic profile, or sequence type)

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al



Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

Table 1 | **Current status of GBS vaccine research and development**

Vaccine target	Advantages/ approach	Limitations
Capsular carbohydrate		
Unmodified polysaccharide vaccine (type III serotype)	Phase I trials indicated that the vaccine was safe and well tolerated[67]	Only 60% of the recipients showed an immune response; Requirement to improve immunogenicity of the CPS
Conjugate polysaccharide vaccine	Type III serotype: increase in immunogenicity when coupled to an immunogenic protein (tetanus toxoid (TT)); Conjugate vaccine with all nine currently identified GBS serotypes (Ia, Ib, II, III, IV, V, VI, VII and VIII) prepared and tested preclinically[11,68,108,109]	Capsular conjugate vaccines of this type need to be multivalent in order to provide sufficient coverage against prevalent serotypes
Conjugate bivalent polysaccharide vaccine	Bivalent vaccine (GBS type II-TT and type III-TT) combined and administered; Well tolerated	Further testing is warranted to investigate immune interference when more than two GBS CPS conjugate vaccines are simultaneously administered[10]
Conjugate multivalent polysaccharide vaccine	Proposed that effective GBS vaccine in the United States includes five major serotypes (Ia, Ib, II, III and V); It is anticipated that multivalent vaccines will include each conjugate vaccine prepared separately[10]	Formulation of a GBS conjugate vaccine for use in the United States might not be effective in other regions[110]
Proteins		
C5a peptidase	Present on all strains and serotypes of GBS; Little or no antigenic variability; Capable of inducing antibodies that are opsonically active[73]; Immunization induces serotype-independent protection	Progress as a potential vaccine is unknown
β-Component of the C protein	Elicits protective immunity in animal models[111]	This protein is only present in a minority of strains that cause infection (~20%)
LmbP	Expressed by most GBS strains	Progress as a potential vaccine is unknown[84]
Sip	Present on all GBS strains; Induces protective antibodies; Recombinant SIP protein protected mice infected with numerous GBS strains[72]	Biological function is not well understood; No recent reports of progress towards the development of a vaccine[14,83,84]
LrrG	Highly conserved protein antigen that induces protection[85]	Progress as a potential vaccine is unknown

CPS, capsular polysaccharide; GBS, group B *Streptococcus*; LmbP, laminin binding protein; Sip, surface immunogenic protein.

Group B Streptococcus: global incidence and vaccine development
Atul Kumar Johri et al

Table 2 | **The application of genomic/proteomic technologies to identify potential GBS vaccine candidates**

Method/technique	Vaccine candidates identified	Characteristics	Limitations
Comparative genome analysis[16]	Sip, CAMP factor, R5 protein, Enolase, Hyaluronidase, Haemolysin/cytolysin (cylE)	Identifies conserved genes; Detects putative virulence factors	None observed
Multiple genome screening[19]	Three cell-wall surface-anchor proteins (components of a pilus-like structure), Sip	Identifies conserved genes; Detects putative virulence factors	None observed
STM[101]	LmbP, a permease, Hyaluronate-associated proteins, Clp protease homologue	Identifies genes that are essential for virulence, based on a negative-selection method; Direct selection for antigenicity	Genes that are essential for growth are not identified
Proteomic approach[22]	Ornithine carbamoyltransferase, Phosphoglycerate kinase, Non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase, Purine nucleoside phosphorylase, Enolase, Glucose-6-phosphate isomerase	Identifies cell-surface-expressed proteins using 2-dimensional electrophoresis (2DE) and MALDI-MS	Proteins expressed only *in vivo* are not identified

CAMP, Christie, Atkins and Munch-Petersen; Clp protease, class III heat shock protein; GBS, group B Streptococcus; LmbP, laminin binding protein; MALDI-MS, matrix-assisted laser desorption/ionization mass spectrometry; Sip, surface immunogenic protein; STM, signature tag mutagenesis.

Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition

Tristan Lefebure et al

1. 26 Streptococcus genome
2. Streptococcus genomes: extreme levels of evolutionary plasticity, with high levels of gene gain and loss
3. **S. agalactiae:** large pan-genome with little recombination in its core-genome
4. **S. pyogenes:** smaller pan-genome with much more recombination of its core-genome
5. Core-genome recombination was evident in all lineages (18% to 37% of the core-genome judged to be recombinant)
6. **S. agalactiae** normally behaves as a commensal organism that colonizes the genital or gastrointestinal tract of healthy adults, but it can cause life threatening invasive infection in susceptible hosts, such as newborns, pregnant women, and nonpregnant adults with chronic illnesses
7. **S. pneumoniae** is the leading cause of human bacterial infection worldwide (carried also asymptomatically)
8. **S. mutans** is implicated as the principal causative agent of human dental caries (tooth decay)

Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition

Tristan Lefebure et al

9. **S. thermophilus** is a non-pathogenic, food microorganism, widely used in the dairy product industry
10. **S. suis** is responsible for a variety of diseases in pigs, including meningitis, septicemia, arthritis, and pneumonia. It causes occasional cases of meningitis and sepsis in humans
11. Number of protein coding genes per genome in species of Streptococcus is relatively similar (ranging from 1,697 to 2,376) but the gene composition is much more variable
12. Based on the gene content table obtained by OrthoMCL three strains of S. agalactiae, S. pyogenes or S. thermophilus share about 75% of their genes, and 3 different species of Streptococcus share only around 50%
13. Pan-genome: Even with 26 genomes is still open
14. In contrast the core-genome: plateau around 600 genes

Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition

Tristan Lefebure et al

15. Approximately unbiased (AU) test: 39 out of 260 genes rejected the concatenated tree
16. A small group of genes rejected most and at the same time, their trees were rejected by most of the genes
17. The majority of genes did not reject the concatenated tree and only a small subset of genes proposed significantly different trees

Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition

Tristan Lefebure et al

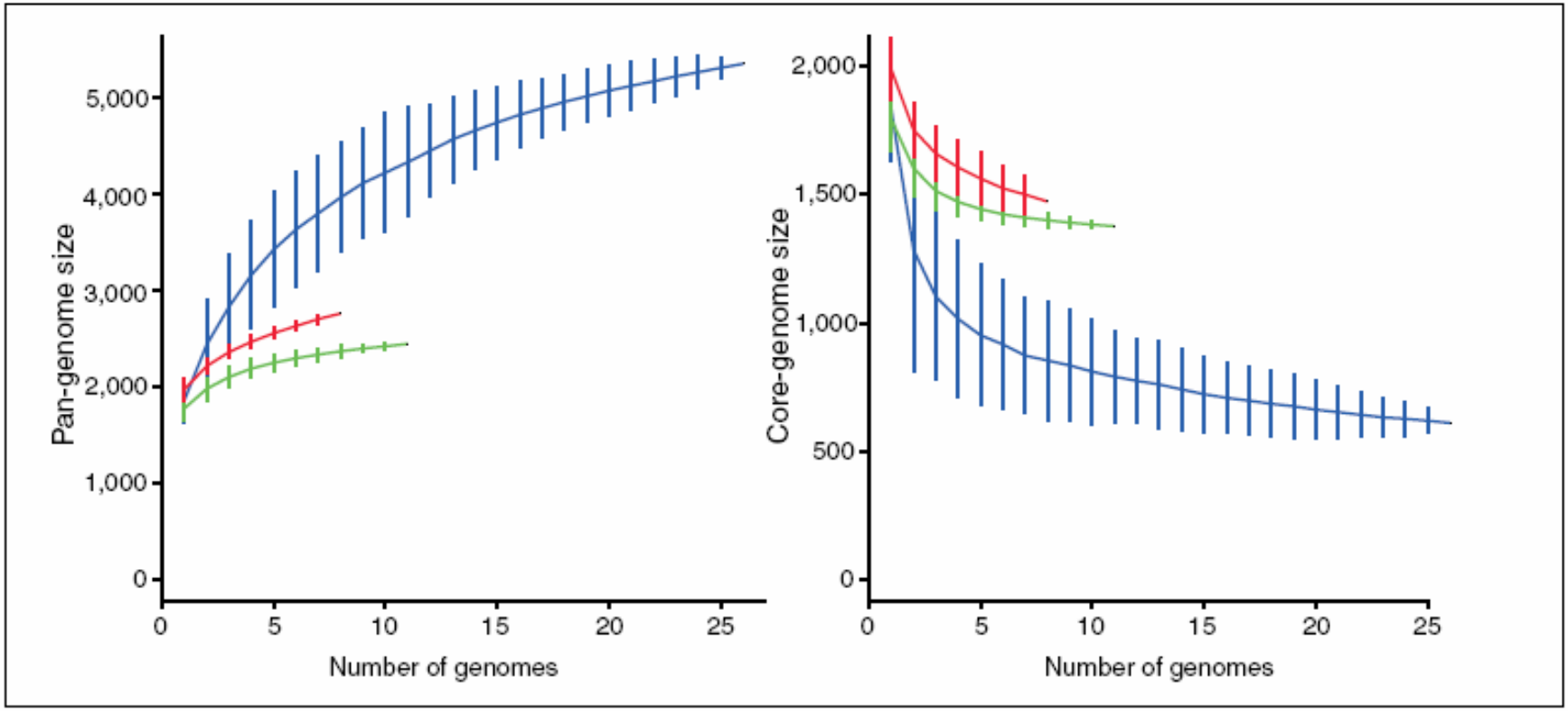


Figure 2

Accumulation curves for the total number of genes (left) or the number of genes in common (right) given a number of genomes analyzed for the different species of *Streptococcus* (in blue), the different strains of *S. agalactiae* (in red) and *S. pyogenes* (in green). The vertical bars correspond to standard deviations after repeating one hundred random input orders of the genomes.

Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition

Tristan Lefebure et al

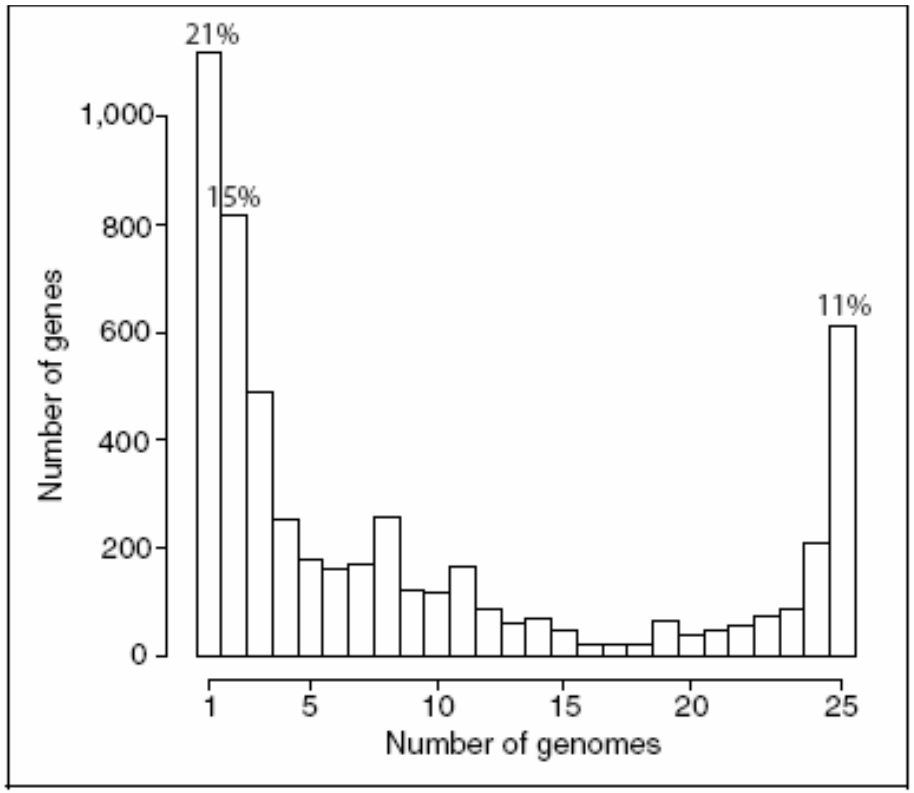


Figure 3

Frequency of genes within the 26 genomes included in this analysis. Genes present in a single genome represent lineage specific genes, while at the opposite end of the scale, genes found in all 26 genomes represent the *Streptoccocus* core-genome.

Identification of a Universal Group B Streptococcus Vaccine by Multiple Genome Screen

Domenico Maione et al

1. Group B Streptococcus (GBS) is a multi-serotype bacterial pathogen representing a major cause of life-threatening infections in newborns. To develop a broadly protective vaccine:
    1. analyzed the genome sequences of 8 GBS isolates
    2. cloned and tested 312 surface proteins as vaccines
    3. 4 proteins elicited protection in mice (their combination protective against a large panel of strains)
2. Protection also correlated with antigen accessibility on the bacterial surface
3. Clinical phase 1 and phase 2 trials of conjugate vaccines prepared with CPS from GBS types Ia, Ib, II, III, and V are safe and highly immunogenic in healthy adults
4. However they do not offer protection against serotypes that are prevalent in other parts of the world
5. This analysis: core genome of 1811 genes (80% of each genome), variable genome of 765 genes
6. Computer algorithms: select within the two sub-genomes genes encoding putative surface-associated and secreted proteins. Predicted surface-exposed proteins: 396 were core genes and 193 were variable genes. Of these 589 proteins, 312 were successfully expressed in Escherichia coli

Identification of a Universal Group B Streptococcus Vaccine by Multiple Genome Screen

Domenico Maione et al

7. Each purified soluble protein was next used to immunize groups of adult female mice. At the end of the immunization schedule, these were mated, and the resulting offspring were challenged with a dose of GBS calculated to kill 80 to 90% of the pups
8. This systematic screening identified 4 antigens capable of significantly increasing the survival rate among challenged infant mice
9. Each antigen elicited protection against more than one strain but not against all strains
10. As expected, whenever the corresponding gene was absent from the challenge strain, the antigen was not protective. However, in a few cases, protection was not conferred even though the challenge strain carried the antigen-coding gene

Identification of a Universal Group B Streptococcus Vaccine by Multiple Genome Screen

Domenico Maione et al

11. At least two major conclusions can be drawn from this work
    1. Multistrain genome analysis and screening constitute an effective new approach to identifying vaccine candidates that can provide broad protective activity when used in combination.

       Of the 4 antigens identified, none could be classified as universal because, in a fraction of GBS strains, either their coding gene was absent or their surface accessibility was negligible.

       The 4-antigen vaccine used in this work protected mice against 12 virulent strains belonging to all nine major GBS serotypes

       To estimate the strain coverage of the vaccine analyzed the surface expression of the 4 antigens on a total of 37 GBS isolates: at least one of the antigens was highly accessible to antibodies in 32 out of the 37 strains tested (87% of circulating strains) *assuming that these strains sufficiently reflect the variability in the population*
    2. The extent of surface accessibility of antigens may vary from strain to strain, even if the antigens coding genes are conserved

       Such variability may be due to:
       1. differences in gene expression
       2. antigen masking by other cellular components (e.g., CPS)
       3. protein degradation
       4. other factors. The strains should be selected not only because they carry the gene for the antigen under examination, but also in light of the amount of expression and accessibility of the antigen itself

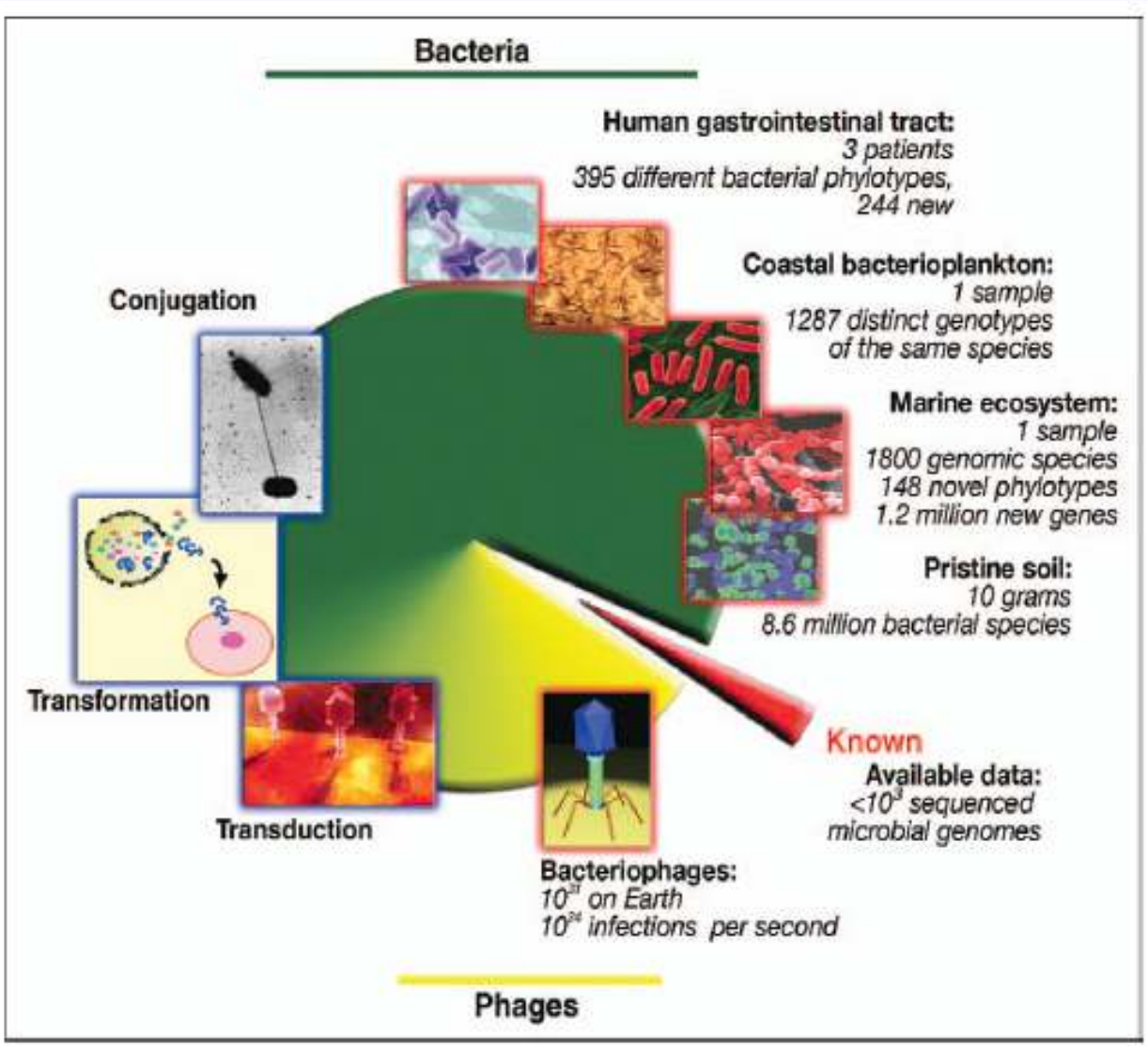
The microbial pan-genome
Duccio Medini et al

1. In some species new genes are discovered even after sequencing the genomes of several strains
2. A bacterial species can be described by its pan-genome (*core genome* + *dispensable genome)*
3. Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome
4. In theory some bacterial species will never be fully described, because new genes will be added to the genome of the species with each new genomic sequence
5. GBS pan-genome: 2713 genes (1806 core genome, 907 dispensable genome). The GBS pan-genome is predicted to grow by an average of 33 new genes every time a new strain is sequenced
6. Analysis on 5 Streptococcus pyogenes: asymptotic value of 27 specific genes for each new genome added

The microbial pan-genome
Duccio Medini et al

7. Different behavior: eight Bacillus anthracis isolates converge to zero after the addition of only a 4^{th} genome. Hence, the B. anthracis species has a 'closed' pan genome, and 4 genome sequences are sufficient to completely characterize this species
8. The importance of the mechanisms of lateral (or horizontal) gene transfer in evolutionary processes has been hotly debated in recent years but it is now generally accepted that enables an organism to quickly adapt to a changing environment
9. Such genes are continuously exchanged within and between bacterial species by three main processes:
    1. by transformation, when genetic material can be taken up from the environment
    2. by transduction, when the DNA is delivered by a virus
    3. by conjugation, when DNA is directly exchanged between cells
10. Transformation and conjugation require that the source and target organisms live in close contact, and bacteriophages might enable bacterial species populating different environments to exchange genetic material, which often contains genes that are crucially important for pathogenesis
11. Global population of phages: estimated to be 10^{31} causing an average of 10^{23} infections per second, it is easy to conclude that the global pool of genes present in the microbial world might be in the order of billions

The microbial pan-genome
Duccio Medini et al



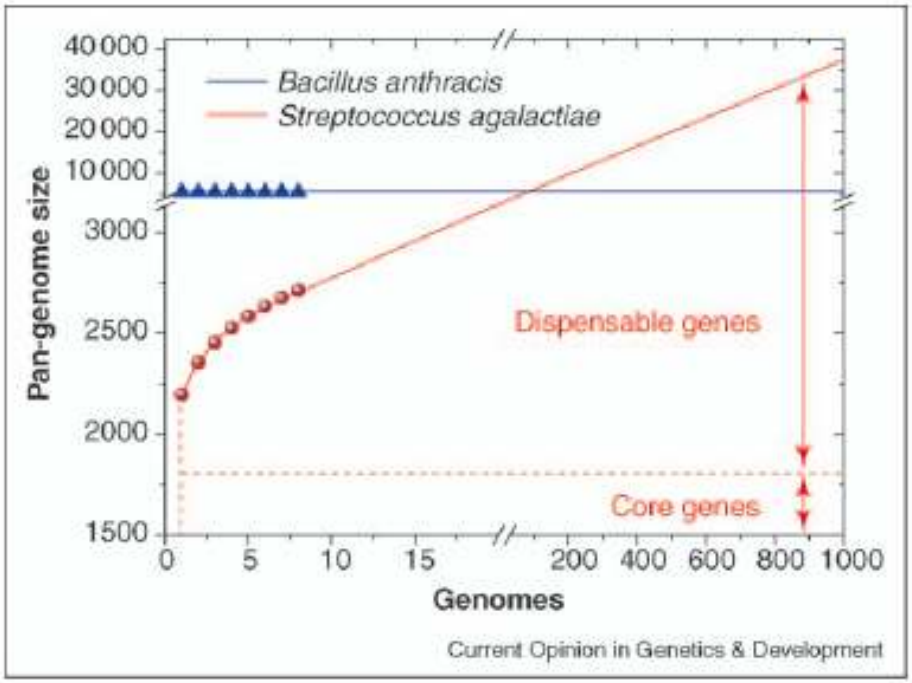
The microbial pan-genome
Duccio Medini et al

12. Serotypes and sequence types do not correlate with genomic diversity. Classical methods to catalogue bacterial species are based on knowledge convenient phenotypic traits. The most popular is the agglutination of bacterial cells by specific antisera against the capsular polysaccharide surrounding many pathogens
13. For a variety of encapsulated bacteria, this method has been widely used for epidemiology studies and vaccine design, assuming that all strains belonging to the same serogroup are similar
14. Multilocus enzyme electrophoresis (MLEE) and multilocus sequence typing (MLST), which are based on the detection of variability associated with housekeeping genes, were applied to several bacterial species and led to the classification of strains into 'clonal complexes' and sequence types, respectively
15. For GBS strains genomic diversity does not segregate with serotypes or MLST sequence-types
16. In fact, the analysis revealed that, often, isolates belonging to different serogroups are more closely related than are isolates of the same serogroup
17. Capsular specificity genes are present in the dispensable genome, which is exchanged freely between strains with different genetic background
18. By contrast, the genes used to determine the MLST type belong to the core genome, and they do not pick up similarities present in the dispensable genome, which often are linked to pathogenic features

The microbial pan-genome
Duccio Medini et al

19. Streptococci, Meningococci, H. pylori, Salmonellae and E. coli are likely to have an open pan-genome
20. B. anthracis, Mycobacterium tuberculosis and Chlamydia trachomatis live in isolated niches with limited access to the global microbial gene pool (closed pan-genome)
21. Buchnera aphidicola, an endosymbiont of aphids:no chromosome rearrangements, duplications or horizontal gene transfer in the past 50 million years
22. This example shows that the criteria used to define microbial species might be inconsistent with the genetic information. In the future, we will need to consider how to handle these inconsistencies

The microbial pan-genome
Duccio Medini et al



The microbial pan-genome
Duccio Medini et al

Table 1

Number of genomes sequenced in different bacterial species.

Species with sequenced genome(s)	Number of species (% of the total)	Number of genomes sequences per species
Streptococcus agalactiae, *Bacillus anthracis*, *Burkholderia mallei*	3 (1.2%)	8
Burkholderia pseudomallei	1 (0.4%)	7
Staphylococcus aureus, *Streptococcus pyogenes*	2 (0.8%)	6
Salmonella enterica, *Escherichia coli*, *Bacillus cereus*, *Chlamydophila pneumoniae*, *Haemophilus influenzae*, *Listeria monocytogenes*, *Xylella fastidiosa*	7 (2.8%)	5
Prochlorococcus marinus, *Buchnera aphidicola*, *Burkholderia cenocepacia*, *Ehrlichia ruminantium*, *Legionella pneumophila*, *Pseudomonas syringae*, *Streptococcus thermophilus*, *Yersinia pestis*	8 (3.2%)	3
Streptococcus pneumoniae, *Mycobacterium tuberculosis*, *Neisseria meningitidis*, *Bacillus licheniformis*, *Bifidobacterium longum*, *Campylobacter jejuni*, *Chlorobium phaeobacteroides*, *Corynebacterium glutamicum*, *Haemophilus somnus*, *Helicobacter pylori*, *Lactococcus lactis*, *Leptospira interrogans*, *Mycoplasma genitalium*, *Pseudomonas aeruginosa*, *Shigella flexneri*, *Staphylococcus epidermidis*, *Synechococcus elongates*, *Thermus thermophilus*, *Tropherymaa whipplei*, *Vibrio vulnificus*, *Xanthomonas campestris*	21 (8.3%)	2
Various species	211 (83.3%)	1

Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing

Mariagrazia Pizza et al

1. Neisseria meningitidis is a major cause of bacterial septicemia and meningitis. Sequence variation of surface-exposed proteins and cross-reactivity of the serogroup B capsular polysaccharide with human tissues have hampered efforts to develop a successful vaccine
2. To overcome these obstacles, the entire genome sequence of a virulent serogroup B strain (MC58) was used to identify vaccine candidates. A total of 350 candidate antigens were expressed in Escherichia coli, purified, and used to immunize mice. The sera allowed the identification of proteins that are surface exposed, that are conserved in sequence across a range of strains, and that induce a bactericidal antibody response, a property known to correlate with vaccine efficacy in humans
3. Meningococcal meningitis and sepsis are devastating diseases that can kill children and young adults within hours despite the availability of effective antibiotics. The diseases are caused by Neisseria meningitidis, a Gram-negative, capsulated bacterium that has been classified into five major pathogenic serogroups (A, B, C, Y, and W135) on the basis of the chemical composition of distinctive capsular polysaccharides
4. In the 1960s, vaccines consisting of purified polysaccharide antigens were developed against four (A, C, Y, and W135) of the five pathogenic serogroups. These vaccines are highly effective in adults but are not efficacious in infants and young children, the age groups mostly exposed to disease
5. Currently, there are no vaccines available for prevention of serogroup B N. meningitides (MenB) disease, which is responsible for 32% of all meningococcal disease in the United States and for 45% to >80% of the cases in Europe

Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing

Mariagrazia Pizza et al

6. The use of capsular polysaccharide as the basis of a vaccine for prevention of MenB diseases has been problematic. The MenB capsular polysaccharide is identical to a widely distributed human carbohydrate [a(238)N-acetyl neuraminic acid or polysialic acid], which, being a self-antigen, is a poor immunogen in humans. Furthermore, use of this polysaccharide in a vaccine may elicit autoantibodies
7. We identified 570 such ORFs and, by means of the polymerase chain reaction (PCR), we amplified and cloned the DNA sequences of these hypothetical genes in Escherichia coli to express each polypeptide
8. We obtained successful expression with 350 ORFs (61%). More specifically, 70 predicted lipoproteins, 96 predicted periplasmic proteins, 87 predicted inner membrane proteins, and 45 predicted outer membrane proteins; there were 52 proteins with uncertain prediction
9. Proteins with more than one hydrophobic trans-membrane domain had the highest rate of expression failure
10. The recombinant proteins were purified and used to immunize mice. Immune sera were then tested in enzyme-linked immunosorbent assay (ELISA) and fluorescenceactivated cell sorter (FACS) analyses to detect proteins that were present on the surface of a set of MenB strains selected to represent the diversity of invasive strains within the natural population of this species

Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing

Mariagrazia Pizza et al

11. Of the 85 proteins found to be strongly positive in at least one of the above assays, we selected for further studies seven representative proteins (genome-derived Neisseria antigens; GNA) that were positive in all three assays and whose genes were not predicted to be phase variable
12. Each of the proteins raised an immune response that induced complement-mediated bactericidal activity
13. To test the suitability of these proteins as candidate antigens for conferring protection against different MenB strains and not just against the homologous strain, we used a collection of strains isolated worldwide and over many years to investigate whether the new candidate molecules were conserved and accessible to antibodies
14. The results suggest that these proteins may induce immunity against most strains of MenB and, possibly, against the other pathogenic strains of N. meningitides

Identification of Vaccine Candidates Against Serogroup B Meningococcus by Whole-Genome Sequencing

Mariagrazia Pizza et al

11. Of the 85 proteins found to be strongly positive in at least one of the above assays, we selected for further studies seven representative proteins (genome-derived Neisseria antigens; GNA) that were positive in all three assays and whose genes were not predicted to be phase variable
12. Each of the proteins raised an immune response that ind[illegible]ed complement-mediated bactericidal activity
13. To test the suitability of these proteins as candidate anti[illegible] for conferring protection against different MenB strains and not just against the homologous strain, we u[illegible]ollection of strains isolated worldwide and over many years to investigate whether the new candidate m[illegible]were conserved and accessible to antibodies
14. The results suggest that these proteins may induce imm[illegible]t most strains of MenB and, possibly, against the other pathogenic strains of N. meningitides

Phase variation is an immune evasion technique employed by various types of bacteria, including Salmonella species. It involves the switching of surface antigens, to evade specific adaptive immune system responses.

Salmonella use this technique to switch between different types of the protein flagellin. As a result, flagella with different structures are assembled. Once an adaptive response has been mounted against one type of flagellin, or if a previous encounter has left the adaptive immune system ready to deal with one type of flagellin, switching types renders previously high affinity antibodies ineffective against the flagella.

Reverse vaccinology
Rino Rappuoli

1. Biochemical, serological and microbiological methods have been used to dissect pathogens and identify the components useful for vaccine development. Although successful in many cases, this approach is time-consuming and fails when the pathogens cannot be cultivated in vitro, or when the most abundant antigens are variable in sequence
2. Now genomic approaches allow prediction of all antigens, independent of their abundance and immunogenicity during infection, without the need to grow the pathogen in vitro
3. This allows vaccine development using non-conventional antigens and exploiting non-conventional arms of the immune system. Many vaccines impossible to develop so far will become a reality
4. Since the process of vaccine discovery starts in silico using the genetic information rather than the pathogen itself, this novel process can be named reverse vaccinology
5. The conventional approach to vaccine development uses two methods: first, attenuation of pathogens by serial passages in vitro to obtain live-attenuated strains to be used as vaccines, and second, identification of protective antigens to be used in non-living, subunit vaccines

Reverse vaccinology
Rino Rappuoli

6. In order to identify the components of the pathogen suitable for vaccine development, the pathogen is grown in laboratory conditions and the components building the pathogen are first identified one at a time, by biochemical, serological or genetic methods
7. The identification of protective antigens that could be potential vaccine candidates involves separating each component of the pathogen one by one. This approach is time-consuming and allows the identification only of those antigens that can be purified in quantities suitable for vaccine testing
8. Since the most abundant proteins are most often not suitable vaccine candidates, and the genetic tools required to identify the less abundant components may be inadequate or not available at all, this approach can take years or decades
9. For the bacterial and parasitic pathogens studied to date, the maximum number of potential vaccine antigens identified during a century of vaccine development is usually less than ten
10. This conventional method also means that vaccine development is not possible when the pathogen cannot be grown in laboratory conditions. An exception to this has been the hepatitis B vaccine where the pathogen, although unable to grow in vitro, could be recovered in large quantities from the plasma of infected people

Reverse vaccinology
Rino Rappuoli

11. Once a suitable antigen is identified, it needs to be produced in large scale, often by growing the pathogen itself. Cloning of the gene coding for the antigen is often necessary in order to better characterize and produce the identified antigen(s)
12. Finally, the new molecule can enter vaccine development. Although successful in many cases, this approach took a long time to provide vaccines against those pathogens for which the solution was easy and failed to provide a solution for those bacteria and parasites that did not have obvious immunodominant protective antigens
13. The reverse approach to vaccine development takes advantage of the genome sequence of the pathogen. The genome sequence provides at once a catalog of virtually all protein antigens that the pathogen can express at any time. This approach starts from the genomic sequence and, by computer analysis, predicts those antigens that are most likely to be vaccine candidates

Reverse vaccinology
Rino Rappuoli

Comparison of conventional and genomic approaches to vaccine development.

	Conventional vaccinology	Reverse vaccinology
Essential features		
	Most abundant antigens during disease	All antigens immunogenic during disease
	Antigens immunogenic during disease	Antigens even if not immunogenic during disease
	Cultivable microorganism	Antigens even in non-cultivable microorganisms
	Animal models essential	Animal models essential
	Correlates of protection useful	Correlates of protection very important
		Correct folding in recombinant expression important
		High-throughput expression/analysis important
Advantages		
	Polysaccharides may be used as antigens	Fast access to virtually every single antigen
	Lipopolysaccharide-based vaccines are possible	Non-cultivable microorganisms can be approached
	Glycolipids and other CD1-restricted antigens can be used	Non abundant antigens can be identified
		Antigens that are not immunogenic during infection can be identified
		Antigens that are transiently expressed during infection can be identified
		Antigens not expressed *in vitro* can be identified
		Non-structural proteins can be used
Disadvantages		
	Long time required for antigen identification	Non proteic antigens cannot be used (polysaccharide, lipopolysaccharides, glycolipids and other CD1-restricted antigens)
	Antigenic variability of many of the identified antigens	
	Antigens not expressed *in vitro* cannot be identified	
	Only structural proteins are considered	

The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates

David A. Rasko et al

1. Whole-genome sequencing has been skewed toward bacterial pathogens as a consequence of the prioritization of medical and veterinary diseases
2. It is becoming clear that in order to accurately measure genetic variation within and between pathogenic groups, multiple isolates, as well as commensal species, must be sequenced
3. Comparison of 17 E. coli genomes, 8 of which are new, resulted in identification of ~2,200 genes conserved in all isolates. We were also able to identify genes that were isolate specific
4. Fewer strain-specific genes were identified than anticipated, suggesting that each isolate may have independently developed virulence capabilities
5. Pangenome calculations indicate that E. coli genomic diversity represents an open pangenome model containing a reservoir of more than 13,000 genes, many of which may be uncharacterized but important virulence factors
6. This bacterium can be grown readily, and its genetics are easily manipulated in the laboratory, making it a common workhorse and one of the best studied prokaryotic model organisms

The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates

David A. Rasko et al

7. E. coli isolates can cause serious illness in humans and are associated with at least six distinct disease presentations that result in billions of dollars in lost work time and doctor and hospital visits each year
8. Diarrheagenic E. coli strains are well known from recent outbreaks in the United States; however, a significant proportion of E. coli isolates cause disease outside the intestinal tract, and these isolates are known as extraintestinal pathogenic E. coli (ExPEC)
9. The ExPEC isolates cause a range of diseases in humans, including urinary tract infections and neonatal meningitis
10. The diarrheagenic pathogenic variants (pathovars) of E. coli are also diverse in terms of their clinical presentation, age groups affected, and associated virulence factors
11. Five distinct clinical groups of diarrheagenic isolates have been identified: enteroaggregative E. coli (EAEC), enterohemorrhagic E. coli (EHEC), enteropathogenic E. coli (EPEC), enteroinvasive E. coli (EIEC), and enterotoxigenic E. coli (ETEC)
12. Examination of the number of genes in each of the 17 E. coli genomes revealed that the isolates have a genome size of 5,020
13. The “conserved core” genome size (the genes that are highly conserved in all 17 isolates) is 2,344

The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates

David A. Rasko et al

14. The exponential decay model suggests that the number of conserved core genes is approaching an asymptote with the comparison of 17 genomes
15. The model is based on the median number of conserved genes in each of the permutations of genome comparisons and predicts that the number of core genes in E. coli is approximately 2,200 genes
16. The number of unique genes ranges from ~20 in the laboratory-adapted isolates to more than 300 depending on the genome used as the reference
17. The large deviation from the mean is indicative of the high degree of variation within E. coli, as well as the fact that not all isolates have similar clinical presentations
18. The pangenome of E. coli is considered open. An open species pangenome indicates that the species is still evolving by gene acquisition and diversification
19. The open pangenome of the species E. coli indicates that continued sequencing should result in identification of ~300 novel genes per genome

The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates

David A. Rasko et al

E. coli strain	Pathovar[a]	No. of genes	No. of conserved genes[b]	No. of unique genes[c]	No. of group-specific genes[d]
K-12	Commensal	4,238	2,312	21	11
W3110	Commensal	4,384	2,324	31	11
HS	Commensal	4,433	2,321	94	11
E24377A	ETEC	5,111	2,318	246	5
B7A	ETEC	5,357	2,446	256	4
EDL933	EHEC	4,863	2,327	27	122
Sakai	EHEC	5,497	2,337	208	126
CFT073	ExPEC/UPEC	5,589	2,341	308	56
F11	ExPEC/UPEC	5,198	2,412	203	46
UTI89	ExPEC/UPEC	5,176	2,288	109	45
536	ExPEC/UPEC	4,734	2,347	134	46
APEC01	ExPEC/Avian	4,561	2,295	158	3[d]
E22	EPEC	5,575	2,343	274	5
E110019	EPEC	5,586	2,397	219	4
B171	EPEC	5,330	2,382	234	6
Ec042	EAEC	4,899	2,305	308	3
101-1	EAEC	4,812	2,356	155	4

[a] Pathovar assignment is based on previous identification of the strain (see references).

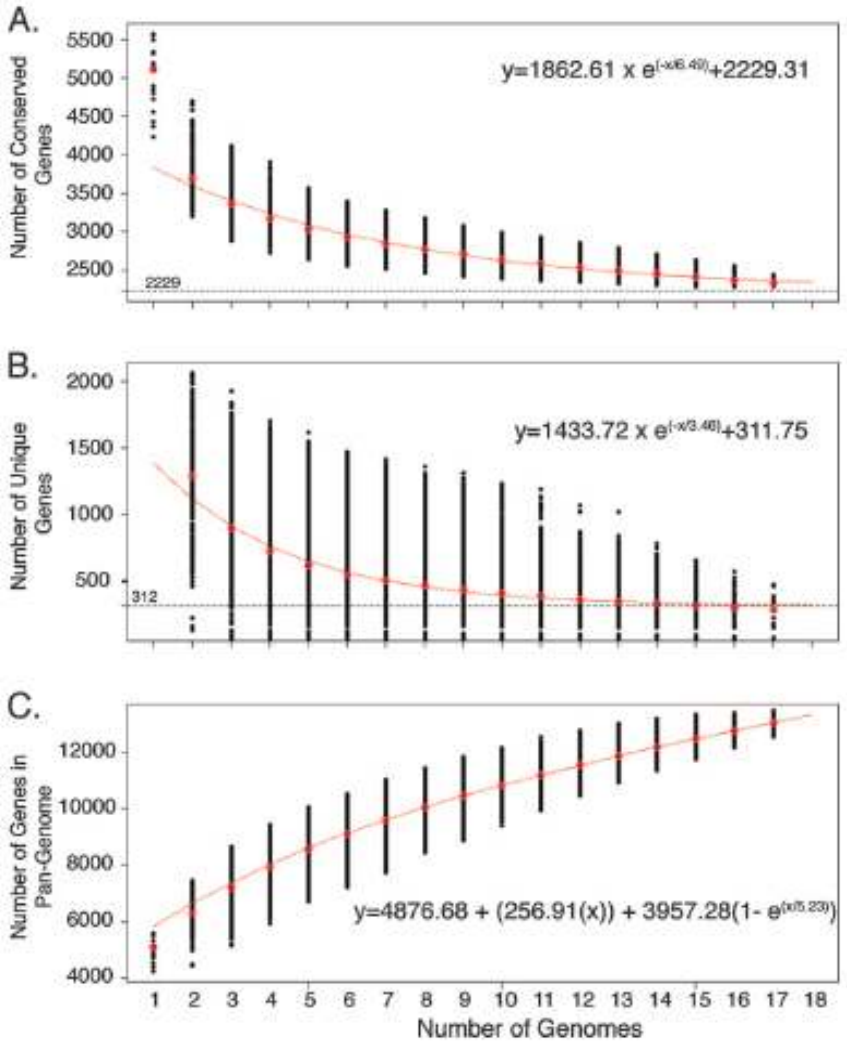
[b] Conserved genes are genes whose BSR is ≥0.8 in all isolates.

[c] Unique genes are genes whose BSR is <0.4 in all other isolates tested.

[d] Group-specific genes are genes whose BSR is >0.8 in the members of the group but <0.4 in all other isolates.

The Pangenome Structure of Escherichia coli: Comparative Genomic Analysis of E. coli Commensal and Pathogenic Isolates

David A. Rasko et al



Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome

Herve Tettelin et al

1. Generated the genomic sequence of six strains representing the five major disease-causing serotypes of Streptococcus agalactiae, the main cause of neonatal infection in humans
2. Analysis of these genomes and those available in databases showed that the S. agalactiae species can be described by a pan-genome consisting of a core genome shared by all isolates, accounting for ~80% of any single genome, plus a dispensable genome consisting of partially shared and strain-specific genes
3. Mathematical extrapolation of the data suggests that the gene reservoir available for inclusion in the S. agalactiae pan-genome is vast and that unique genes will continue to be identified even after sequencing hundreds of genomes
4. The most recent definition of a bacterial species comes from the pregenomic era. In 1987, it was proposed that bacterial strains showing ~70% DNA-DNA reassociation and sharing characteristic phenotypic traits should be considered to be strains of the same species

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome

Herve Tettelin et al

5. Each strain pair was compared by means of the following:
    1. a Smith and Waterman protein search on all of the predicted proteins by using the SSEARCH program (version 3.4)
    2. a DNA search of all of the predicted ORFs of a strain against the complete DNA sequence of the other strain, by using the FASTA program (version 3.4)
    3. a translated protein search of all of the predicted proteins of a strain against the complete DNA sequence of the other strain, by using the TFASTY program (version 3.4)
    4. A gene was considered conserved if at least one of these three methods produced an alignment with a minimum of 50% sequence conservation over 50% of the protein/gene length

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome

Herve Tettelin et al

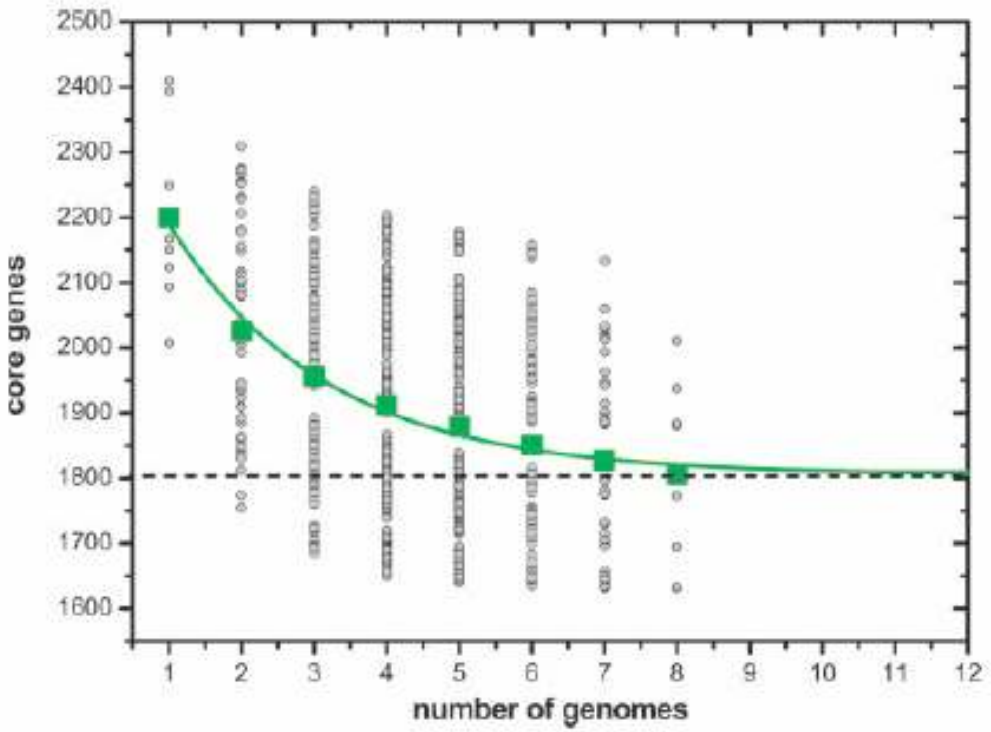


Fig. 2. GBS core genome. The number of shared genes is plotted as a function of the number n of strains sequentially added (see *Materials and Methods*). For each n, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations. Squares are the averages of such values. The continuous curve represents the least-squares fit of the function $F_c = \kappa_c \exp[-n/\tau_c] + \Omega$ (see Eq. **1** in *Supporting Text*) to data. The best fit was obtained with correlation $r^2 = 0.990$ for $\kappa_c = 610 \pm 38$, $\tau_c = 2.16 \pm 0.28$, and $\Omega = 1{,}806 \pm 16$. The extrapolated GBS core genome size Ω is shown as a dashed line.

Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: Implications for the microbial pan-genome

Herve Tettelin et al

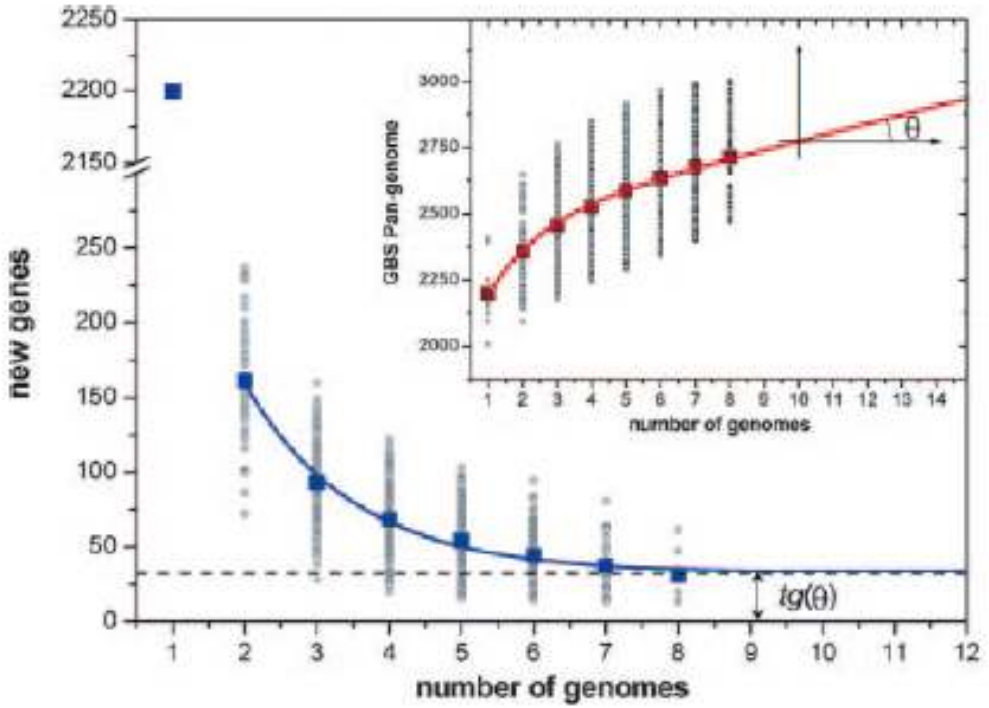


Fig. 3. GBS pan-genome. The number of specific genes is plotted as a function of the number n of strains sequentially added (see *Materials and Methods*). For each n, circles are the $8!/[(n-1)!\cdot(8-n)!]$ values obtained for the different strain combinations; squares are the averages of such values. The blue curve is the least-squares fit of the function $F_s(n) = \kappa_s \exp[-n/\tau_s] + tg(\theta)$ (see Eq. 2 in *Supporting Text*) to the data. The best fit was obtained with correlation $r^2 = 0.995$ for $\kappa_s = 476 \pm 62$, $\tau_s = 1.51 \pm 0.15$, and $tg(\theta) = 33 \pm 3.5$. The extrapolated average number $tg(\theta)$ of strain-specific genes is shown as a dashed line. (*Inset*) Size of the GBS pan-genome as a function of n. The red curve is the calculated pan-genome size