

Reductive evolution in Streptococcus agalactiae and the emergence of a host adapted lineage

Isabelle Rosinski-Chupin, Elisabeth Sauvage, Barbara Mairey, Sophie Mangenot, Laurence Ma, Violette da Cunha, Christophe Rusniok, Christiane Bouchier, Valérie Barbe, Philippe Glaser

▶ **To cite this version:**

Isabelle Rosinski-Chupin, Elisabeth Sauvage, Barbara Mairey, Sophie Mangenot, Laurence Ma, et al.. Reductive evolution in Streptococcus agalactiae and the emergence of a host adapted lineage. BMC Genomics, 2013, 14 (1), pp.252. 10.1186/1471-2164-14-252 . pasteur-00818470

HAL Id: pasteur-00818470

https://pasteur.hal.science/pasteur-00818470

Submitted on 27 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

BMC Genomics

RESEARCH ARTICLE Open Access

Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage

Isabelle Rosinski-Chupin[1,2*], Elisabeth Sauvage[1,2], Barbara Mairey[3], Sophie Mangenot[3], Laurence Ma[4], Violette Da Cunha[1,2], Christophe Rusniok[2,5], Christiane Bouchier[4], Valérie Barbe[3] and Philippe Glaser[1,2]

Abstract

Background: During host specialization, inactivation of genes whose function is no more required is favored by changes in selective constraints and evolutionary bottlenecks. The Gram positive bacteria *Streptococcus agalactiae* (also called GBS), responsible for septicemia and meningitis in neonates also emerged during the seventies as a cause of severe epidemics in fish farms. To decipher the genetic basis for the emergence of these highly virulent GBS strains and of their adaptation to fish, we have analyzed the genomic sequence of seven strains isolated from fish and other poikilotherms.

Results: Comparative analysis shows that the two groups of GBS strains responsible for fish epidemic diseases are only distantly related. While strains belonging to the clonal complex 7 cannot be distinguished from their human CC7 counterparts according to their gene content, strains belonging to the ST260-261 types probably diverged a long time ago. In this lineage, specialization to the fish host was correlated with a massive gene inactivation and broad changes in gene expression. We took advantage of the low level of sequence divergence between GBS strains and of the emergence of sublineages to reconstruct the different steps involved in this process. Non-homologous recombination was found to have played a major role in the genome erosion.

Conclusions: Our results show that the early phase of genome reduction during host specialization mostly involves accumulation of small and likely reversible indels, followed by a second evolutionary step marked by a higher frequency of large deletions.

Keywords: *Streptococcus agalactiae*, Host-adaptation, Non-homologous recombination, Gene inactivation, Virulence

Background

Comparative genomics of strains belonging to species with a large spectrum of hosts, such as *Staphylococcus aureus*, have highlighted two main evolutionary trends linked to the adaptation to a new host: acquisition of new functions through lateral gene transfer facilitating colonization of new niches and gene loss associated with the host specialization [1-3]. These trends were also recognized when considering the emergence of highly virulent host-specialized pathogens from bacterial species with a broader host range. For instance massive gene losses were associated to the emergence of human or equine pathogens such as *Salmonella enterica* sv Typhi and Paratyphi, *Bordetella pertussis* and *parapertussis*$_{hom}$ and *Burkholderia mallei* that respectively derive from *Salmonella* Typhimurium, *Bordetella bronchiseptica* and *Burkholderia pseudomallei* [4-9]. During these transitions, it was postulated that gene inactivation and deletions were probably favored by genetic drift and evolutionary bottlenecks. In line with this model the host-specialized pathogens often showed a much higher number of insertions sequences (IS) than their parental strains and IS expansion was proposed to be largely responsible for gene deletions and genome rearrangements observed in these species [10].

Streptococcus agalactiae also referred to as Group B streptococcus (GBS) is a Gram-positive bacterium that has emerged as a leading cause of neonatal infections during the sixties and represents an increasing cause of infections in the elderly and in adults with underlying diseases [11-13]. As a commensal it colonizes the

* Correspondence: ichupin@pasteur.fr
[1]Unité de Biologie des Bactéries Pathogènes à Gram Positif, 28 rue du Docteur Roux, Paris, Cedex 15 75724, France
[2]CNRS UMR 3525, Paris, France
Full list of author information is available at the end of the article

BioMed Central

© 2013 Rosinski-Chupin et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

digestive and genitourinary tracts of up to 30% of the human adult population [14]. However, *S. agalactiae* was initially described as an animal pathogen causing mastitis in ruminant [15]. Since the 70's, *S. agalactiae* was found to be responsible for epidemic events of invasive diseases in fish farms, leading to a mortality of up to 30% [16-18]. Cases of infection were also reported for other aquatic poikilotherms such as frogs [19] and aquatic mammals such as dolphins [20]. How GBS is able to adapt to its different hosts remains poorly understood. The genetic diversity of GBS populations has been studied using different methods including multilocus sequence typing (MLST) [21], which led to the recognition of different clonal complexes (CC). Some of these clonal complexes display host preference. For instance, CC67 is essentially associated with the bovine host and the hypervirulent sequence type (ST) 17 strains are mainly isolated from humans. However incidentally strains belonging to human-associated clonal complexes are also isolated from bovines [22] suggesting that relationships between clonal complexes and host specificity are not so strict. Further analysis of the complete genome sequences of eight isolates of human origin and one of bovine origin has highlighted the composite organization of *S. agalactiae* genomes with a conserved backbone (representing the core genome of the species) and a dispensable genome composed of genomic islands that are highly variable between the different strains [23-26].

The ST261 strain 2-22 (or ATCC 51487) was initially isolated as responsible for several epidemics in fish farms in Israel [27,28]. This strain, which showed a restricted metabolic pattern, thermosensitivity and lack of ß hemolytic and CAMP activities was first classified as a different species, *Streptococcus difficile,* but proved later to be a genuine serotype Ib *S. agalactiae* strain [17]. *S. agalactiae* strains were repeatedly isolated from fish infections and found to cluster into two main groups [20,29]. The first group corresponds to strains belonging to the clonal complex 7, also displaying strains isolated from human and bovine hosts. The other strains share two or more common MLST alleles with strain 2-22 and are classified in ST246, 257, 259, 260, 552 and 553. As ST261 strains, these STs were until now never isolated from humans [29]. In addition, a third group of strains belonging to clonal complex 283 has recently been described in fish and humans [29]. The draft genome sequences of one ST260 strain, strain STIR-CD-17, and two ST7 strains, strain ZQ0910 and GD201008-001, respectively isolated from disease outbreaks affecting farmed tilapia in Honduras and Nile Tilapia in China were recently published [30-32].

To decipher the phylogenomic relationships between *S. agalactiae* strains isolated from fish or other poikilotherm animals and strains isolated from human or bovine, we analyzed the genomic sequence of seven isolates belonging to ST260-261 and ST6-7, including the strain 2-22. Comparative analysis confirmed that the two groups of GBS strains responsible for fish epidemic diseases are distantly related. We found that adaptation to fish does not involve any specific function compared to human CC7 isolates. Conversely, specialization to the fish host of the ST260-261 strains was associated with massive gene inactivation and deep remodeling of metabolic and regulatory networks that we also characterized at the transcriptome level. This genome reduction likely occurred through RecA independent recombination.

Results and discussion

Fish ST7 strains are closely related to human strains

We first compared the gene content of *S. agalactiae* strains isolated from fish but grouped into the same CC as the human strain A909 by sequencing the genome of strains CF01173 and SS1014, isolated in USA and UK respectively (Table 1). Whole genome sequence comparison showed that strain CF01173 differed from strain A909 by only 389 SNPs (Additional file 1: Table S2). CF01173 was also closely related to the recently described strains ZQ0910 and GD201008-001 [31,32] isolated from diseased fish in China, which differ by only 105 and 100 SNPs respectively. In contrast the ST6 strain SS1014 was more distant (3484 SNPs), and proved to be closer to the ST6 strain H36B (689 SNPs) isolated from human [26] (Additional file 1: Table S1 and Table S2). Analysis of SNP distributions along the genome sequence showed a uniform distribution when strains of the same ST were compared (Figure 1), with a mean polymorphism of 0.1-0.2 SNP per 1000 nt except in the sequence of an inserted prophage. In contrast alignment of ST6 versus ST7 strains revealed a mosaic pattern of regions of low polymorphism (0.1-0.2 SNP per 1000 nt) interrupted by several regions of higher polymorphism (5-15 SNP per 1000 nt on average) that were probably gained by recombination with distantly related GBS strains, as previously suggested [33]. One of these regions corresponds to the capsule locus that encodes a serotype Ia capsule in ST7 strains and a serotype Ib capsule in ST6 strains. Therefore ST6 and ST7 strains probably shared a common ancestor and recently diverged by recombination with other GBS strains, modifying the capsular serotype between both ST.

Gene content was similar between strains CF01173 and A909 except for 13 genes that were disrupted in CF01173 and five in A909 (Additional file 2: Table S3). Seven genes were specifically disrupted in strain SS1014. In addition, we found that the three ST7 strains CF01173, ZQ0910 and GD201008-001 isolated from fish shared one short genomic island that was absent from other GBS and probably resulted from lateral gene transfer. This genomic island encodes proteins 70-95% identical with proteins of *Streptococcus anginosus,* including a protein with a

Table 1 Characteristics of the different strains used in the study

Strain	Origin	Infection	MLST	Serotype	Geographical origin	Genome size (kbp)	Ref.
2-22	trout	meningitis	ST261	Ib	Israel	1,839	[27]
SS1218	frog	-	ST261	Ib	Louisiana	1,797	[34]
05-108A	tilapia (*Oreochromis* sp.)	meningitis	ST260	Ib	Honduras	1,801	[35]
90-503	Hybrid striped bass	meningitis	ST260	Ib	Louisiana	1,753	[35]
SS1219	frog	-	ST260	Ib	Taiwan	1,798	[34]
SS1014	striped bass	-	ST6	Ib	USA	2,016	[34]
CF01173	trout	-	ST7	Ia	GB	2,027	[36]
A909	human	-	ST7	Ia	-	2,128	[26]

LPXTG-motif cell-wall anchor domain (GBS1173_1788). Sharing of this island in addition to the low number of SNPs between strains CF01173, ZQ0910 and GD201008-001 suggests that the three strains have a common recent ancestor that propagated worldwide. In contrast, emergence of the ST6 strain SS1014 can be considered as an independent event.

Overall, the low level of polymorphism with human strain indicates that CC7 strains infecting fish recently diverged from strains isolated in humans and bovines. Since at least two independent events of emergence were observed, this suggests that CC7 strains might be more amenable to fish colonization/infection than other GBS clonal complexes isolated from humans or bovines. In agreement with this hypothesis, it was recently shown that a *S. agalactiae* ST7 strain isolated from human was able to cause disease in Nile tilapia [37].

ST260-261 strains form an independent lineage which underwent reductive evolution

To explore the phylogenomics relationships between the second group of strains isolated from fish, and CC7

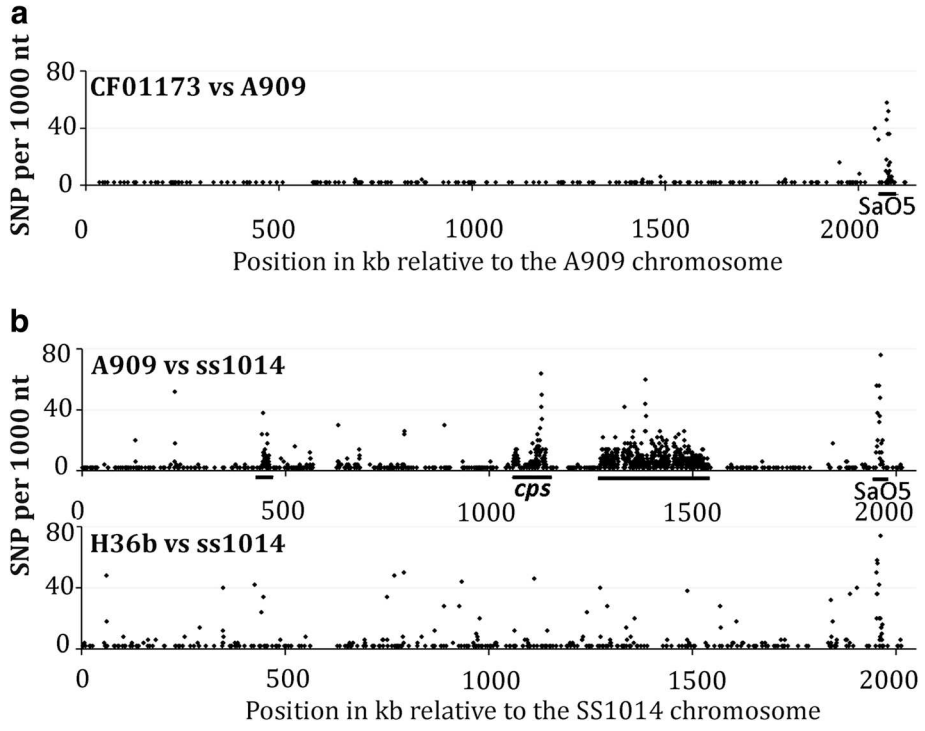


Figure 1 SNP distribution along the genome sequences of CC7 strains. (**a**) The number of SNPs (y axis) per 1000 nt in strain CF01173 (ST7) versus strain A909 (ST7) is plotted according to the position of the corresponding fragment on the strain A909 genome (x axis). The goodness of fit to a uniform distribution was tested using a Chi square test on 19 class intervals of 100,000 nt, excluding regions of prophage SaO5, and differences in SNP distribution were found to be not significant ($p = 0.20$). (**b**) The numbers of SNPs (y axis) per 1000 nt in strains A909 (ST7) and H36B (ST6) versus strain SS1014 (ST6) are plotted according to the position of the corresponding fragment on the strain SS1014 genome (x axis). Regions of higher polymorphism, including the capsule locus (*cps*) and prophage SaO5, are underlined.

strains, we sequenced the genome of five strains belonging to ST260 and ST261. While the genomes of four of these strains were obtained as draft sequences, the genome of strain 2-22 was sequenced to completion. Strain 2-22 genome consists of a single circular chromosome of 1,838,867 bp (Figure 2a); this is 10 to 25% smaller than the genome sizes of other sequenced GBS strains, which range from 2,065 kb (ST17 human strain COH1) to 2,456 kb (ST67 bovine strain FSLS3-026). The G + C content (35.5%) is similar to that of other GBS strains. Compared to human GBS genomes, the genome of strain 2-22 lacks one rDNA cluster and 9 tRNA genes (71 tRNA genes and 6 rDNA clusters versus 80 and 7 respectively in other GBS genomes). The deletion of this rDNA cluster is associated with the translocation of a 150 kb genomic region probably caused by recombination between flanking ribosomal RNA operons, as also observed in *Salmonella* Typhi [38,39]. Except for this region the genome of strain 2-22 is syntenic to the genomes of human strains (Figure 2b). The four other strains have a similar genome size as strain 2-22 (Table 1).

Whole-genome sequence comparison of the five strains showed that they clustered into two distinct subgroups correlating with the MLST classification (Additional file 1: Table S2). The ST261 strain SS1218 isolated from frog in Louisiana differed from the strain 2-22 by only 30 SNPs. Strains 90-503 isolated in Louisiana in 1990 and 05-108A isolated in 2005 in Honduras, with 49 SNPs can be considered as variants of the same clone. Strain SS1219 isolated from frog in Taiwan diverged from the 2 former ST260 strains by 130 SNPs. On average, the ST260 strains showed 3,100 SNPs with ST261 strains. ST260-261 strains were also related to the ST552 strain Sa20-06 (3,400 and 1,700 SNPs respectively). In contrast, ST260-261 displayed 15,000 SNP with CC7 strains. Analysis of the SNP distribution along the genome sequences of ST260 and ST261

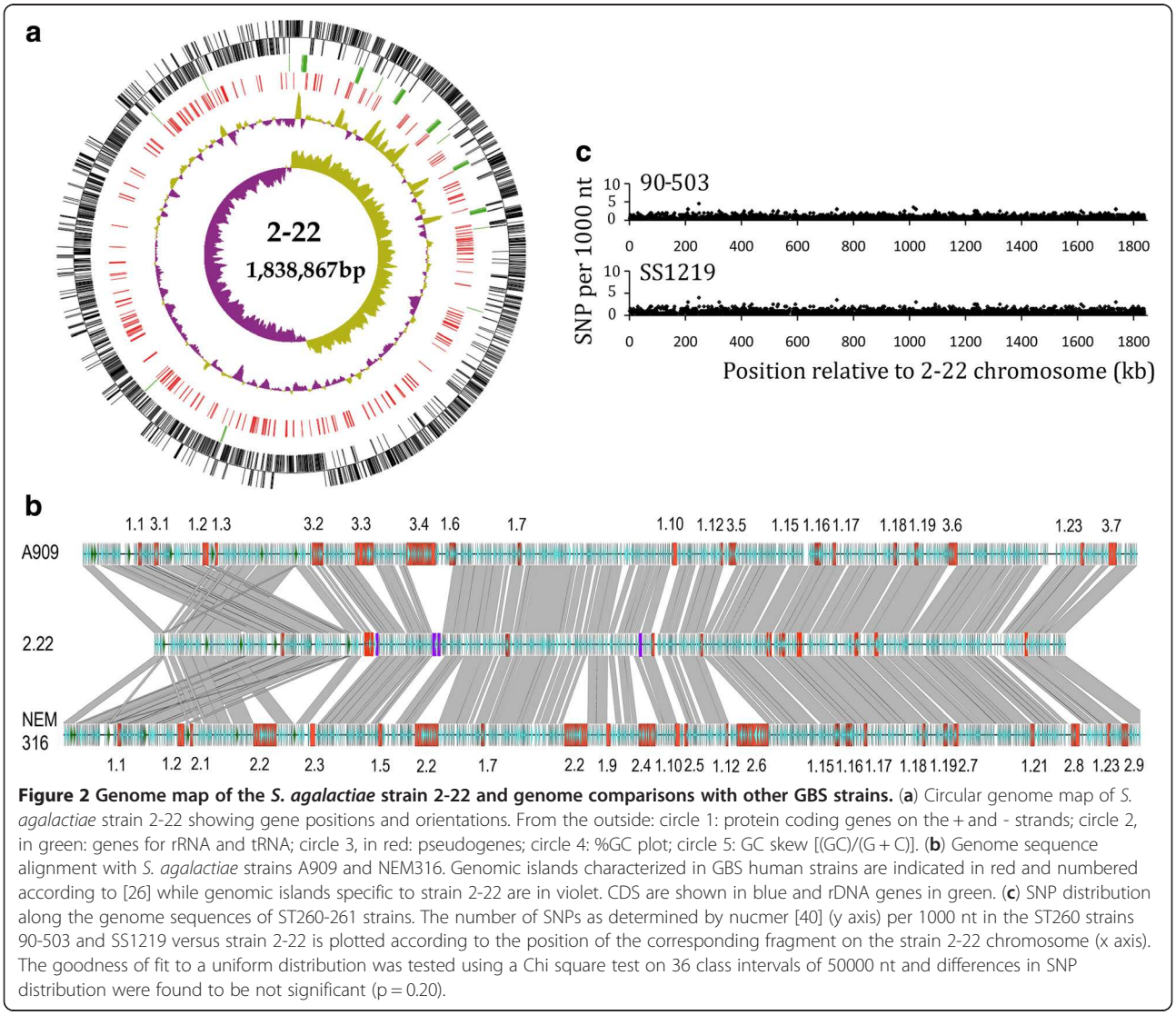


Figure 2 Genome map of the *S. agalactiae* strain 2-22 and genome comparisons with other GBS strains. (**a**) Circular genome map of *S. agalactiae* strain 2-22 showing gene positions and orientations. From the outside: circle 1: protein coding genes on the + and - strands; circle 2, in green: genes for rRNA and tRNA; circle 3, in red: pseudogenes; circle 4: %GC plot; circle 5: GC skew [(GC)/(G + C)]. (**b**) Genome sequence alignment with *S. agalactiae* strains A909 and NEM316. Genomic islands characterized in GBS human strains are indicated in red and numbered according to [26] while genomic islands specific to strain 2-22 are in violet. CDS are shown in blue and rDNA genes in green. (**c**) SNP distribution along the genome sequences of ST260-261 strains. The number of SNPs as determined by nucmer [40] (y axis) per 1000 nt in the ST260 strains 90-503 and SS1219 versus strain 2-22 is plotted according to the position of the corresponding fragment on the strain 2-22 chromosome (x axis). The goodness of fit to a uniform distribution was tested using a Chi square test on 36 class intervals of 50000 nt and differences in SNP distribution were found to be not significant ($p = 0.20$).

strains revealed a uniform pattern of 3 SNPs per kb (Figure 2c), suggesting that no recombination occurred in this lineage.

A phylogenetic analysis of strains of human, bovine and fish or frog origins confirmed that ST260-261-552 strains constitute a distinct lineage. Separation of this lineage from other *S. agalactiae* clonal complexes, including CC7 strains, was probably ancient, pre-dating the separation between the three strains of human origin (NEM316, A909 and 2606V/R) (Figure 3). While the comparison of the whole genome sequence showed the same mean identity between ST260-261 strains and the three human strains, a higher proportion of nucleotides was found to align with the genome of strain A909 (97.15%) than with other GBS genomes (2603 V/R: 95.64%, NEM316: 95.43%, FSLS3-026: 92.27%) (Additional file 1: Table S1).

The low proportion of nucleotides that did not align with the genomes of human GBS strains suggested that ST260-261 strains encode only few specific functions. Indeed, only four genomic islands were characterized in strain 2-22, that represented 25 kb in total (in violet on Figure 2b). However these four regions essentially contained pseudogenes and one of them was absent from the ST260 strains. We also identified in ST260-261 strains eleven to twenty copies of ISSag1, an insertion sequence previously described in the genomes of human isolates and other streptococci [42] (Additional file 3: Table S4). In addition, the genomes of ST260-261 strains contained 10 regions categorized as genomic islands in human *S. agalactiae* genomes [26], but shared by most GBS characterized so far (shown for strain 2-22 in Figure 2b). They did not contain any Integrative and Conjugative Elements (ICE) or prophages and only two inactivated copies of integrases were identified. They also lacked two mobile genetic elements encoding important virulence loci: the cis-mobilizable element encoding the major surface antigen Alpha like protein/Rib [43] and the composite transposon coding for the C5A peptidase and the laminin binding protein [44]. In total the difference in genomic island content between ST260-261 strains and GBS strains isolated from humans or bovines accounted for about 70-80% of the genome reduction (210-230 kb). An interesting exception was the genomic island 3.2 that was previously described only in strains A909 (ST7) and H36B (ST6) and therefore is also shared by CC7 strains isolated from fish. This conservation is in favor of a role in fish colonization. This GI encodes two phosphotransferase systems (PTS) for galactitol, sugar ABC transporters and genes for galactose utilization (*GBS222_0398-0414* in strain 2-22) (Additional file 4: Figure S1). Sharing of this genomic island explained the higher number of aligned nucleotides between strains 2-22 and A909.

Genome annotation predicted 1547, 1568, 1560 and 1569 protein coding genes for strains 2-22, 90-503, SS1219 and 05-108A respectively, which is significantly less compared to 2096 in NEM316, 1990 in A909 and 2135 in 2603V/R. Furthermore, 190-220 pseudogenes were identified in each strain (Additional file 5: Table S5) compared to 27 to 41 in the human strains. This revealed a massive reduction of the functional genome during the time-course of adaptation to fish in the ST260-261 lineage.

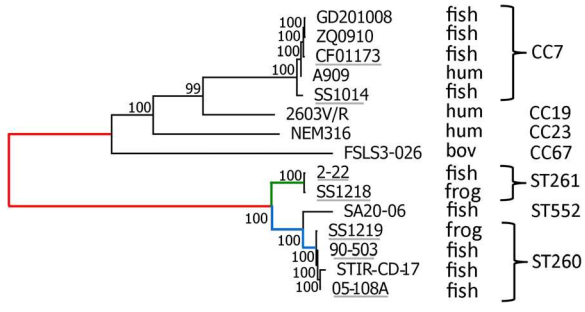


Figure 3 Phylogenetic relationships between *S. agalactiae* strains isolated from fish/frog, bovine and human. A maximum likelihood phylogenetic tree was constructed using the MEGA 5.04 software [41] based on 25,597 concatenated SNPs from the core genome of GBS strains (1.34 Mb) of which regions recombined between ST6 and ST7 strains have been removed. The bootstrap values resulting from 1000 replicates are indicated. Strains sequenced in this work are underlined. The tree branches that are further analyzed in Figure 4 are colored (red: N_{fbh}-$A_{260+261}$ branch; green: $A_{260+261}$-A_{261} branch; blue: $A_{260+261}$-A_{260} branch). A phylogenetic analysis using the Neighbor-Joining method gave the same topology for the tree (not shown).

Reductive evolution is an ongoing process in the ST260-261 lineage

To get more insights into the evolution of the ST260-261 lineage, we further analyzed the nucleotide changes leading to gene disruption and genome size reduction compared to human strains. Strain A909, belonging to CC7, was used as a reference to compare the genome sequences of the five ST260-261 strains. We also aligned the genomes of two human and one bovine isolates to identify nucleotide changes specific to the fish lineage. Genomic islands and insertion sequences were excluded from the analysis as well as thirty sequences whose evolution involved both insertions and deletions of nucleotides. In these conditions, sequence alignment between strain 2-22 and A909 genomes revealed 621 simple insertion/deletion events ranging from 1 to 10,095 nt (Figure 4a). Using the sequences of human strains as outgroups, we found that, among these indels, 160 corresponded to deletions and 60 to insertions that specifically occurred in the ST261 sublineage, after ST260-261 divergence. The mean size of deletions largely exceeded that of insertions since only 300 nucleotides were gained while 47,000 were lost. Deletions, insertions and nucleotide replacements were respectively

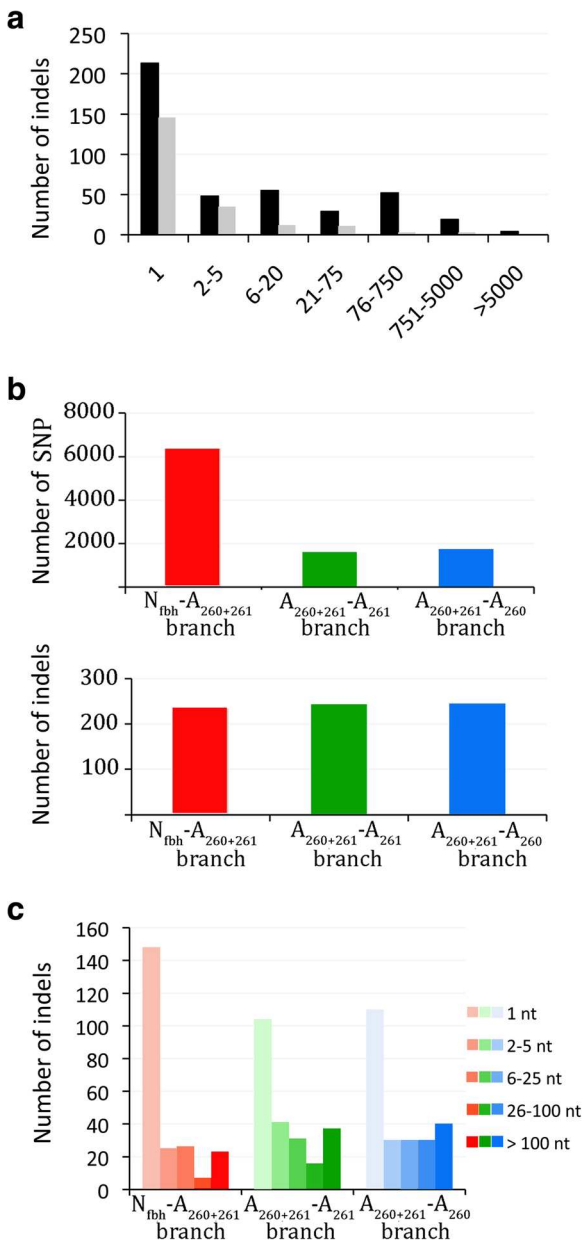


Figure 4 Ongoing reductive evolution in the ST260-261 lineage. (**a**) Distribution of insertions/deletions between strains 2-22 and A909, according to their sizes in nucleotides. Black bars correspond to sequences missing in the 2-22 strain compared to A909 and gray bars to sequences missing in A909 compared to strain 2-22. Only indels mapped in the core genome sequences are represented. **b**) Comparison of the number of indels and SNPs that occurred before (N_{fbh}-$A_{260+261}$ branch) and after sublineage separation ($A_{260+261}$-A_{261} and $A_{260+261}$-A_{260} branches). Indels and nucleotide changes relative to the A909 sequence were deduced from the comparison of the 15 strains shown in Figure 3. The mutation events were positioned on the phylogenetic tree according to a parcimonious approach. (**c**) Size distribution of the deletion events identified on the N_{fbh}-$A_{260+261}$, $A_{260+261}$-A_{261} and $A_{260+261}$-A_{260} branches respectively. Colors (red, green and blue) correspond to the branch colors of the tree presented in Figure 3.

responsible for 76, 21 and 19 gene disruptions. In addition, 15% of the indels led to minor protein size modifications (less than 20%) and their consequences on protein functions were more difficult to predict. Finally 30% of the indels were characterized in intergenic regions. Altogether mutations leading to gain or loss of nucleotides represented approximately 12% of total mutations after the divergence from ST260 strains and constituted the major process of gene disruption. This evolution was not specific to the ST261 sublineage as a similar number of insertions and deletions occurred in the ST260 sublineage accounting for 86 gene disruptions (Figure 4b).

In addition, 235 indels and 6363 SNPs relative to strain A909 sequence were common to the five ST260-261 strains but were not observed in A909 *vs* other GBS strains. Forty per cent of these indels were intragenic and led to 91 gene disruptions (Figure 4c). Finally, 34 genes disrupted by a small indel in one sublineage were deleted in the other, suggesting that gene disruption could be a preliminary step to gene loss and genome reduction by secondary longer deletions. As a consequence some loci were observed under different states of decay in the two sublineages. This is the case for instance for the *cyl* locus, the *CRISPR2* locus and the *pil2* pilus locus (Additional file 4: Figure S2). Furthermore, the average size of the indels was found to increase following the divergence of the two sublineages (Figure 4c) compared to the common branch. Therefore, while the process of reductive evolution was already evident in the ancestor of the ST260-261 strains, it was even more pronounced after divergence of the two sublineages.

To better evaluate the specificity of the evolutionary process occurring in ST260-261 strains, we compared the core genome sequence of strains A909 and NEM316, two strains of human origin differing by approximately 11,000 SNP. We detected 314 indels, which represents half the number of indels in the 2-22 *vs* A909 comparison. These indels were essentially short and as much as 80% of them occurred in intergenic regions while less than 4% led to gene inactivations. Therefore, gain or loss of nucleotides was also frequent in other *S. agalactiae* strains, but most indels in coding sequences were probably eliminated by purifying selection. Fitting with this hypothesis, we observed one indel per 5-10 SNPs during the recent evolution of ST7 strains where purifying selection is probably not effective. A similar proportion of indels *vs* SNP was also reported for two strains of *S. enterica* sv Paratyphi [4].

Finally, to gain further insights into the mechanism of the observed insertions and deletions, we analyzed the sequences flanking the indels in strain 2-22. Among the 358 indels of 1 nt, 210 (60%) occurred in homopolymeric tracts longer than four nucleotides and likely resulted from DNA polymerase slippage during replication. Interestingly, 136 of the 193 (70%) indels larger

than four nucleotides apparently also involved recombination between repeated sequences. The median size of the repeats was 8 nucleotides, with 8 sequences larger than 20 nucleotides, the largest being 254 nt long (Figure 5a). Since the threshold of repeat length for RecA-dependent homologous recombination is 23-27 nt [45], most of the indels probably occurred by RecA independent recombination. It has been shown that, under laboratory conditions, the efficiency of illegitimate recombination is highly dependent on the size of the repeats [46] and inversely dependent on the distance between repeats [47-49]. Our results suggest that RecA independent recombination between repeats of moderate sizes may also lead to long deletions. Furthermore we did not find any evidence that deletions of long sequences may depend on longer repeats (Figure 5b). This may indicate that the lower efficiency of illegitimate recombination between short repeats is buffered by their higher frequency in the genome. Some of the large deletions might also correspond to several consecutive shorter ones.

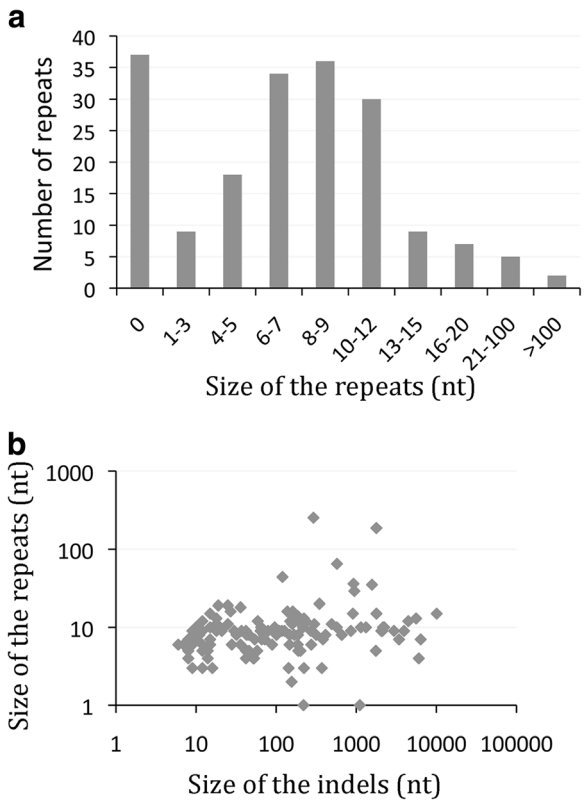


Figure 5 Characteristics of the repeats identified at the extremities of insertions/deletions. (**a**) Size distribution of the repeats flanking the 193 indels larger than 4 nucleotides identified in the genome alignment of strains 2-22 and A909. (**b**) Scatter plot (in log scale) showing the size of the indel in nucleotides (x axis) and the length of the flanking repeat in nucleotides (y axis) for the 193 indels larger than 4 nucleotides.

From this analysis we propose that accumulation of small and easily reversible indels is common in *S. agalactiae*, mostly occurring as a consequence of RecA independent recombination. Most of these indels are probably removed from the population due to purifying selection during long-term evolution. Alternatively they may be compensated by a second loss or gain of nucleotides. In the ST260-261 lineages host specialization led to the relaxation of the negative selection on genes that became dispensable, allowing the accumulation of larger deletions.

Functional adaptation to a host restricted way of living

ST260-261 strains differ from other GBS strains by their host specificity and high pathogenicity. As mentioned, this lineage has only few specific genes compared to other strains and most of them are carried by genomic islands in the process of being eliminated. In particular, we did not identify any specific surface proteins or putative virulence factors. Therefore, the specific properties of the ST260-261 strains more likely rely on functions shared with human isolates and on the loss of some functions. The 1432 genes common to the two sublineages were grouped into functional categories according to KEGG classification (Figure 6a). Compared to functional classification of A909 genes, this analysis revealed a dramatic drop in the proportions of genes involved in mobile and extrachromosomal functions (that reflects the loss of most genomic islands) and in cellular processes, including the pathogenesis subcategory (Figure 6b). The other major functions submitted to evolutionary erosion were energy metabolism, transport and binding, regulation and signal transduction (Figure 6b). Conversely the basal functions of the cell such as transcription and protein synthesis were nearly not affected.

Bacterial-host interactions and virulence factors

Surface components play an important role for tissue colonization and infection by mediating interactions between the pathogen and the host cells and evasion from immune defense. GBS strains possess two distinct polysaccharide antigens, the highly sialylated capsule polysaccharide and the group B-antigen. ST260-261 strains harbor the 16 genes involved in the type Ib capsule synthesis (GBS222_0990-1005) similarly to the ST6 strains H36B and SS1014. They also possess the 16 genes responsible for the synthesis of the group B antigen (GBS222_1160-1175). Only seven proteins with a LPXTG signal for cell-wall anchor are conserved (Table 2) including three important *S. agalactiae* virulence factors of human strains: the Fibrinogen-binding protein A, the Serine-rich protein Srr1 and the BibA/HvgA protein [50-53]. However, while in human strains these three proteins carry a variable number of a repeated motif, these repeats have

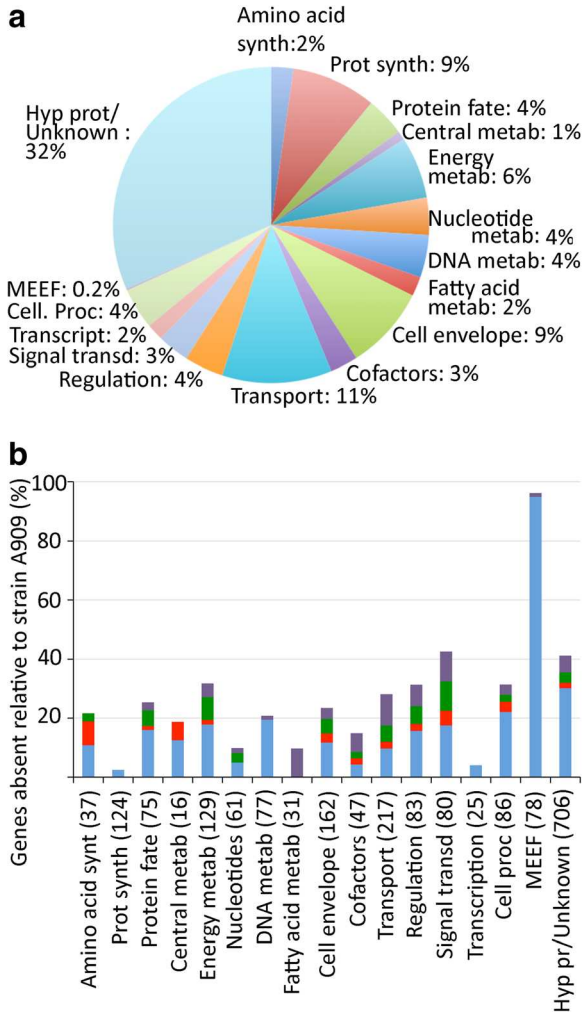


Figure 6 Functional classification of the genes conserved or inactivated during the evolution of the ST260-ST261 sublineages. (**a**) Functional classification of the genes of the ST260-261 strain core genome (KEGG classification). Only 1432 functional genes common to the ST260 and 261 sublineages were taken into account. MEEF stands for mobile and extrachromosomal element functions (**b**) Analysis of the genes present in strain A909 and inactivated (disrupted or deleted) in the ST260 and ST261 sublineages according to KEGG functional classification. In blue: Genes inactivated before the separation of the ST260-261 sublineages. In red: Genes inactivated by deletion in one sublineage and gene disruption in the other or by independent events in the two sublineages; in green and violet: genes inactivated in the ST261 or in the ST260 sublineage respectively. The number of genes in each functional category is indicated in brackets.

been lost in the ST260-261 strains. This might decrease their accessibility to host immune system and cellular receptors by tethering them to the cell wall.

Among the other proteins that are virulence factors in human GBS isolates, the five strains lack the C5A peptidase, the laminin binding protein and components of the pilus. The *cyl* locus has also been inactivated in the two sublineages (Additional file 4: Figure S2) in agreement with the absence of detectable hemolytic activity on blood agar plates. Another major test for *S. agalactiae* identification is the detection of CAMP activity [63,64]. While both ST260 and 261 strains were reported to be negative for the CAMP factor reaction [20] we found that the gene encoding the CAMP factor was disrupted only in ST261 strains. We experimentally confirmed that strain SS1219 was negative for the CAMP test detection, probably because of a lower level of gene expression (see below).

Finally, the five ST260-261 strains express a fibrinogen/fibronectin binding protein of the PavA family and a hyaluronidase that were proposed to be virulence factors in GBS or other *Streptococcus* species [62,65,66] and could also have a role in fish infection.

Altogether, approximately 60% of the genes for proteins involved in pathogenesis and considered as important virulence genes in human stains [67] were affected by genome reduction. In this context the capsule could be a major virulence factor in the fish host, as also observed for *Streptococcus iniae*, another streptococcus species pathogenic for fish [68,69].

Metabolism

Analysis of the missing functions revealed a profound remodeling of the metabolism of strains of this lineage. Numerous transport systems for carbon sources (ABC transporters and phosphoenolpyruvate/carbohydrate phospho-transferase systems) and enzymes for degradation of polysaccharides (amylase, extracellular pullulanase, enzyme for degradation of arbutin) were missing or inactivated, reflecting a reduced capacity to utilize diverse carbon sources. In addition ST260-261 strains seem unable to utilize glycerol and glycerol phosphate as genes encoding the glycerol kinase, the glycerol dehydrogenase and the glycerol-phosphate permease are missing or mutated. Fermentative pathways utilizing pyruvate/acetate conversion are also altered, as the phosphotransacetylase gene is missing. In mammals, GBS is primarily considered as a commensal of the digestive tract, an environment rich in diverse C-sources. In contrast, reduction in the catabolic capacities observed in ST260-261 strains is in favor of a transition to an obligate pathogen style-of-life.

Transport functions

Approximately 20% of the genes associated with transport systems were missing or inactivated. The targeted functions were mostly the import of nutrients, in particular of C-sources and the transport of inorganic and metal ions. The Na+/H + antiporter, the K + uptake permease were inactivated in both sublineages. Therefore ST260-261 strains may be affected in ionic exchange and

Table 2 Potential virulence factors in the ST260-261 strains

2-22	SS1219	90503	Gene	A909	Functional annotation	LPXTG[1]	Ref	Mechanism of pathogenicity
pseudo[2]	GBS1219_1647	GBS90503_1646	*cfa/cfb*	SAK_1983	CAMP factor	no	[54]	***Pore-forming Toxins***
GBS222_0988-1005	GBS1219_979-0996	GBS90503_0979-0996	*cps*	SAK_1246-1263	Capsule synthesis enzymes	no	[55]	***Immune evasion***
GBS222_0661	GBS1219_0659	GBS90503_0659	*sod*	SAK_0913	Superoxide dismutase	no	[56]	
GBS222_0308	GBS1219_0307	GBS90503_0307	*pbp1*	SAK_0222	Penicillin-binding protein 1	no	[57]	***Resistance to host antimicrobial peptides***
GBS222_1019	GBS1219_1007	GBS90503_1007	*pavA*	SAK_1277	fibronectin/fibrinogen binding protein	no	[58]	***Adherence and invasion***
GBS222_0877[3]	GBS1219_0872[3]	GBS90503_08728[3]	*fbsA*	SAK_1142	Fibrinogen-binding protein A	yes	[51,52]	
GBS222_1210[3]	GBS1219_1192[3]	GBS90503_1192[3]	*srr*	SAK_1493	Serin-rich repeat protein	yes	[59]	
GBS222_0182	GBS1219_0182	GBS90503_0182	*sip*	SAK_0065	Surface immunogenic protein	no	[60]	
GBS222_1669[3]	GBS1219_1663[3]	GBS90503_1662[3]		SAK_2002	Surface protein, BibA family		[61]	
GBS222_1026	GBS1219_1014	GBS90503_1014	*hylB*	SAK_1284	Hyaluronidase	no	[60,62]	
GBS222_0382	GBS1219_0382	GBS90503_0382		SAK_0502	surface protein	yes		***Cell-wall anchored proteins of unknown functions***
GBS222_0644	GBS1219_0646	GBS90503_0646		SAK_0896	surface protein	yes		
GBS222_1220	GBS1219_1202	GBS90503_1202		SAK_1503	serine-rich surface protein	yes		
GBS222_1221	GBS1219_1203	GBS90503_1203		SAK_1504	surface protein, Amidase family	yes		

[1]: proteins anchored to the peptidoglycan by an LPXTG C-terminal motif.
[2] : the gene has been inactivated.
[3] : amino-acid repeats have been lost in the corresponding protein.

would have a reduced capacity to maintain their homeostasis in the face of a changing external environment.

Transcriptional networks

Globally, gene disruption and deletion events affected 19 out of the 93 transcriptional regulators predicted in human strains (20%). As much as 13 out of the 21 two-component systems (TCS) (60%) found in GBS were inactivated, either in one (six TCS) or in both (seven TCS) sublineages (Additional file 6: Table S6). Interestingly, analysis of the genome sequence of strain 2-22 revealed that the Rgf TCS, which is involved in the control of virulence in the human ST17 strains [70], was associated with two putative bacteriocins with a double glycine leader peptide and with a bacteriocin export transporter (Additional file 4: Figure S3). This suggests that this TCS originates from a bacteriocin operon that has been partially deleted in human strains. Altogether our observations show that *S. agalactiae* strains adapted to fish may have a reduced capacity to respond to environmental changes compared to human strains and only eight TCS, including the two major systems, CiaRH and CovRS, may be sufficient to allow GBS adaptation to the different environments encountered in fish. Interestingly, both CiaRH and CovRS were also involved in the regulation of virulence genes in GBS human strains [71,72]. The higher virulence of ST260-261 strains might also be due to the deregulated expression of some virulence genes, as observed in the transition from local to systemic infections in Group A Streptococci [73,74].

Adaptation to fish is associated with broad changes in gene expression

As a first step to explore changes in gene expression linked to host adaptation, we performed a comparative analysis at the transcriptome level of strains A909 (ST7 human isolate), CF01173 (ST7 fish isolate), 2-22 (ST261 fish isolate) and SS1219 (ST260 frog isolate). 1389 genes present in the four strains and with 100% identity matches with probes of the array were taken into account in this analysis (Additional file 7: Table S7). Profound modifications in gene expression were observed in ST260-261 strains and to a lesser extent in strain CF01173 compared to strain A909 (Figure 7a). Although strains CF01173 and A909 are closely related, the expression of more than 130 genes varied by a two-fold factor between the two strains. In particular 40% of the genes involved in energy metabolism were expressed at a lower level in the fish isolate than in strain A909. In contrast expression of the gene encoding the virulence protein Srr1 was more than 20 fold increased compared to strain A909. Although up-regulation of the *srr1* gene

is not conserved in the ST260-261 fish isolates, in strain CF01173, it might facilitate penetration of the fish blood brain barrier, leading to an increased tendency to cause meningitis, as shown for the Srr1 protein of human isolates [59].

Both the number of genes for which expression was modified and the amplitude of these modifications were even larger in ST260-261 isolates (Figure 7a; Additional file 7: Table S7). For instance two genes encoding putative glyoxalases are expressed at 6-45 higher levels in ST260-261 strains than in strain A909, suggesting higher needs in methylglyoxal detoxification. Genes encoding the enzymes for the synthesis of arginine (arginosuccinate synthase and arginosuccinate lyase) were 10-15 fold more expressed. Four functional categories were more affected by variations in gene expression in both ST260-261 sublineages: metabolism, transport, regulation and signal transduction (Figure 7b). Interestingly these functional categories were also particularly affected by gene erosion, suggesting a remodeling of metabolic networks.

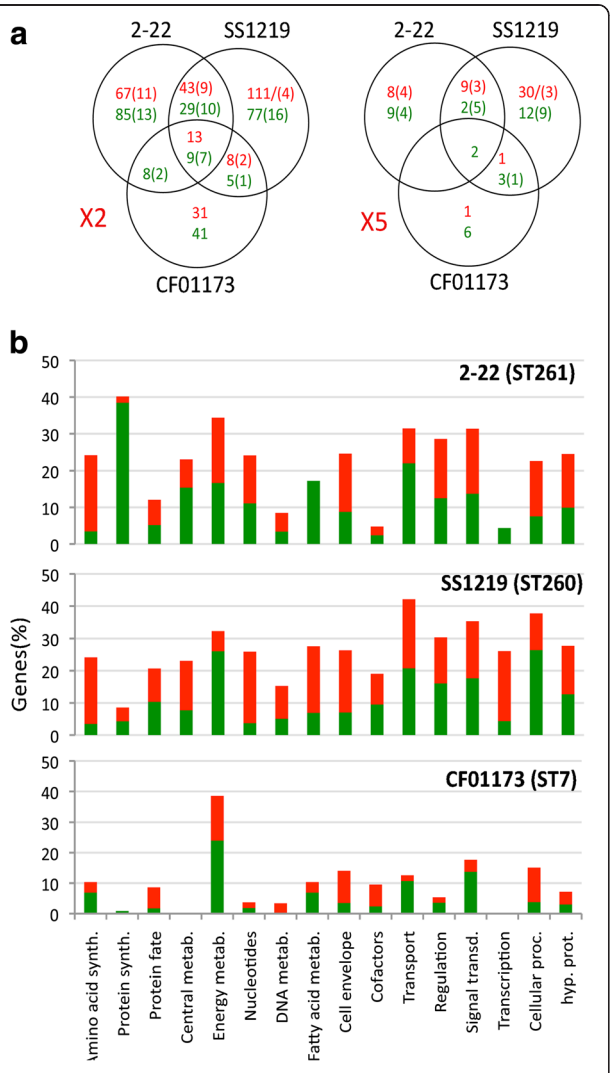


Figure 7 Genes differentially expressed between GBS isolates from fish and the human A909 strain. (**a**) Venn diagram showing the number of genes differentially expressed between the fish or frog isolates 2-22, SS1219 and CF01173 and strain A909; left panel: >2 fold variation; right panel: >5 fold variation. Numbers in green and red correspond to genes that are down- and up-regulated, respectively, in the fish isolates compared to strain A909. In brackets are the numbers of down- and up-regulated pseudogenes. (**b**) Proportions of up- (red bars) and down- (green bars) regulated genes, grouped by KEGG functional categories for the three fish or frog isolates relative to strain A909. Only the 1389 genes present in the four strains and with 100% identity matches with probes of the array were taken into account.

In addition to these global tendencies, ST260 and ST261 strains also harbor sublineage-specific variations in gene expression that probably result from differential gene decay and might lead to specific adaptations. For instance, in strain SS1219, genes encoding phosphoglycerate kinase, phosphoglyceromutase and pyruvate kinase involved in glycolysis and neoglucogenesis are expressed at a lower level than in strain A909. In parallel, genes coding for the arginine deiminase, ornithine carbamoylkinase and carbamate kinase involved in the deiminase pathway are up-regulated, suggesting a shift towards arginine fermentation as the main pathway for energy production in ST260 strains [75]. This up-regulation of the deiminase pathway is associated with a decreased expression of the negative regulator ArgR and the increased expression of the arginine/ornithine antiporter. In contrast, in the same growth conditions, upregulation of the genes for Hpr-kinase and CodY involved in catabolic repression and for lactate dehydrogenase suggests that strain 2-22 might generate energy essentially through glucose utilization and lactic fermentation. Iron import might be more efficient in ST260 than in ST261 strains since strain SS1219 expresses genes for iron ABC transporter and ferrichrome ABC transporter at higher levels than strain A909, while these genes are down-regulated or unchanged in strain 2-22. Both sublineages express lower amounts of ATP synthase, the enzyme responsible for ATP-proton motive force interconversion, than strain A909, indicating that they may be impaired in pH homeostasis in acidic conditions [76].

Since expression of pseudogenes is energetically costly and may generate deleterious products, we took advantage of the massive genome erosion in the ST260-261 lineage to look for evidence for negative selective pressure acting on pseudogene expression. Among the 184 pseudogenes represented by at least one probe on the array, only 22% were down-regulated compared to strain A909, whereas 64% did not present a significant change in mRNA levels and 14% were even up-regulated. Furthermore silencing of pseudogenes in the ST260-261 lineage did not seem to markedly increase with time. Indeed the proportion of down-regulated pseudogenes was

similar among pseudogenes arising before (28%) and after the differentiation of sublineages (16% and 25%). Altogether this indicates that no strong evolutionary pressure acts to silence pseudogene expression.

Although modifications in gene expression generally affected different genes in ST260-261 and CC7 strains, 29 genes were found to vary in the same way in the 3 strains isolated from fish. Interestingly, these 29 genes included the operon for the capsule synthesis, the genes for zoocin and hyaluronidase, two response regulators of TCS systems (*ciaR* and *relR*) and three targets of the CiaRH TCS that were up-regulated. Sixteen genes were downregulated or inactivated in the three fish strains compared to A909, among which 11 were involved in energy metabolism or in carbohydrate uptake. Whether these regulations reflect a common mechanism of adaptation to fish environment remains to be established. However, the higher expression of the genes for the capsule synthesis might favor resistance to environmental conditions and to fish immune system, as it has also been reported for *S. iniae* [68]. Similarly the high hyaluronidase expression may help *S. agalactiae* strains to break through fish tissues and be involved in virulence to fish [62].

Conclusions

Our results show that *S. agalactiae* strains leading to epidemic diseases in fish farms and cold blood animals belong to at least two distinct groups that differ by their strategies of host adaptation. CC7 strains have the potency to colonize and infect multiple hosts such as fish, human and cattle. From a genomic point of view, these CC7 fish strains are not distinguished from their human counterparts by any significant genomic island. However, contrasting with this genomic relatedness, large differences in gene expression were observed and could participate to the adaptation to the fish host. Conversely, our genome analysis indicates that strains of the ST260-261 complex diverged anciently from human and cattle strains and subsequently accumulated specific adaptations leading to the emergence of sublineages.

ST260-261 strains exhibit a striking pattern of genome reduction and we took advantage of the emergence of sublineages to reconstruct the different steps involved in this process. We found that accumulation of short indels can be observed all along the evolution of the GBS species, participating to strain-specific gene disruptions. Therefore even in the core genome of human GBS strains some genes are dispensable. Nevertheless the number of inactivated genes greatly increased during specialization of ST260-261 strains to fish. These gene inactivations mainly result from the ongoing accumulation of short indels, but a tendency to eliminate inactivated genes by deleting longer sequences is more noticeable in the sublineage specific branches. In contrast with what has previously been observed in the course of genomic reduction associated to host specialization or to intracellular symbiosis, deletion events are not correlated with an amplification of insertion sequences. Neither could complex genome rearrangements be noticed, suggesting that recombination between IS is not a general mechanism for genome reduction. Indeed our results point to non-homologous recombination as an alternative mechanism of genome reductive evolution.

Methods

Bacterial strains and growth conditions

S. agalactiae strains used in this study are described in Table 1. They were grown in Todd Hewitt medium (Difco) at 37°C except strains belonging to ST260-261 that were grown at 30°C.

Genome sequencing and assembly methods

To perform the complete sequence of strain 2-22 (ATCC 51487), a mix of capillary Sanger and 454 pyrosequencing (Roche, www.roche.com) was carried out until ~24-fold coverage. A single and a 3kb insert size mate-paired libraries, leading to around 23-fold coverage of 454 GSflx reads (11-fold coverage and 12-fold coverage respectively), were added to Sanger reads, which was derived from a 10 kb insert fragment size library. This library was constructed after mechanical shearing of genomic DNA and cloning of generated inserts into plasmid pCNS (pSU18-derived). Plasmid DNAs were purified and end-sequenced (3633 reads) by dye-terminator chemistry with ABI3730 sequencers (Applied Biosystems, Foster City, USA) allowing an approximately 1-fold supplementary coverage of the genome. The reads were assembled by Newbler (Roche) and validated via the Consed interface (www.phrap.org). A first finishing step was performed using primer walking of clones and polymerase chain reactions (PCRs) (92 and 119 additional reads respectively). Around 70-fold coverage of Illumina reads (36 bp) were mapped, using SOAP (http://soap.genomics.org.cn), for the polishing phase as previously described [77]. Remaining gaps were then covered using PCRs with Platinum Hi-fidelity (InVitrogen) and primers specific to gap edges. The PCR fragments were sequenced by primer walking. Order and orientation of the sequences contiguous to ribosomal DNA was determined using long range PCR with PlatinumHI-fi (Invitrogen).

The genomes of the six other strains isolated from fish or frog were sequenced on the Illumina GAIIx with 36-cycle single-end chemistry and coverages of 100-300×. Reads were assembled using Velvet [78] to produce draft sequences. The number of contigs in each sequence is detailed in Additional file 8: Table S8. Contigs were reordered by Mauve 2.3.1 software [79] using the strain 2-22 genome as a reference for ST260-261 strains and the strain A909 genome for CC7 strains. When possible,

overlapping ends of contigs were further assembled. Draft sequences were used to realign the reads using Bowtie [80], followed by alignment visualization with Tablet [81] to detect sequencing errors. Potential sequence ambiguities were only noticed for the ends of the contigs and therefore SNPs determined at ends of contigs were not considered as reliable.

Nucleotide sequence accession number

The *S. agalactiae* 2-22 complete genome sequence is available from DDBJ/GenBank/EMBL under accession number FO393392. Draft genome sequences of strains 90-503, SS1219, CF01173, SS1014, SS1218 and 05-108A are available under accession numbers CAPZ01000001-CAPZ01000082, CAQA01000001-CAQA01000070, CAQB 01000001-CAQB01000131, CAQC01000001-CAQC01 000099, CAUB01000001-CAUB01000070 and CATH01 000001-CATH01000087 respectively.

Annotation methods

Annotation of strain 2-22 was performed using the CAAT-box environment as previously described [23]. Briefly, coding sequences (CDS) were defined by combining Genmark predictions with visual inspection of each open reading frame (ORF) for the presence of a start codon with an upstream ribosome-binding site and blastp similarity searches on Uniref 90, Trembl and Swissprot databases. Function predictions were based on blastp similarity searches and on the analysis of motifs using the PFAM databases [82]. For the six other strains putative CDS were determined using Artemis [83]. Predicted CDS were tested in reciprocal pair-wise blastp comparison against NEM316, A909 and 2-22 sequences. ORFs longer than 100 codons or encoding a protein > 90% identity by blastp with a protein annotated in NEM316, A909 or 2-22 were retained for further analyses. For genes with an orthologous sequence in other *S. agalactiae* genomes, assignment of the first codon and annotation were transferred from the orthologous genes. The remaining genes were manually annotated. Open reading frames whose lengths differed by more than 20% from orthologous sequences in NEM316 or A909 sequences were considered as putative pseudogenes and individually analyzed. Differences caused by variant start predictions were eliminated, as well as pseudogenes resulting from a mutation localized in the first or in the last 50 nucleotides of a contig, considered as more error-prone.

Whole genome sequence comparisons

Determination of insertion/deletion events in ST260-261 strains compared to strain A909 [GenBank: NC_007432] was performed by combining blastn and nucmer analyses [40]. Each indel was individually inspected to differentiate possible misalignments from true indels and studied for the presence of repeat sequences at its extremities. We also aligned to A909 sequence the genome sequences of three *S. agalactiae* strains of human and bovine origins (NEM316, GenBank: [NC_004368]; 2603V/R [GenBank: AE009948] and FSL S3-O26/AEXT01000000 [GenBank: AEXT00000000] [24]). These alignments were used to produce tables of SNPs, indels and gaps relative to A909 genome sequence. Only indels that occurred internally to Velvet contigs in draft genome sequences and were repetitively observed for all strains of the same sequence type were taken into account. No significant difference in indel detection was noted using the whole-genome sequence of strain 2-22 or the Velvet-assembled contigs of the related ST261 strain SS1218. Indels relative to A909 core genome sequence were therefore classified into three categories related to strain phylogeny: i) detected in ST260-261 strains and in one or several other GBS strains, ii) observed for all strains belonging to the ST260-261 lineage but not detected in strains of bovine or human origins and iii) specific to one of the ST260-261 sublineages. For this last category, the allelic sequence identical to that of the human and bovine strains was considered to be the ancestral sequence, allowing to discriminate the mutated sequence and the nature of the mutation (insertion or deletion of nucleotides).

Phylogenetic analysis

The genome sequence alignment was used to generate a table of SNPs from the *S. agalactiae* core genome. Regions that were recombined between strains A909 and SS1014 were removed from the analysis, as well as sequence gaps. In total 25,597 polymorphic positions in a 1.34 Mb core genome sequence were used to generate a consensus phylogenetic tree by the Maximum Likelihood method based on the Tamura-Nei model with the MEGA 5.04 software [41]. A bootstrap consensus tree was inferred from 1000 replicates. The same tree was obtained using a Neighbor-Joining method and computation of evolutionary distances by the p-distance method.

SNP distribution

The SNP distribution between two genome sequences was inferred from genome alignment using nucmer and SNP were counted on a 1000 nt window. Indels were considered as unique mutation events and therefore counted as one SNP. The goodness of fit to a uniform distribution was tested using a Chi square test on class intervals of 50000 or 100000 nt as indicated in the figure legends.

RNA preparation and transcriptome analyses

S. agalactiae strains were grown in TH medium at 30°C (strains 2-22 and SS1219) or 37°C (strain CF01173). For

each transcriptome analysis, the reference strain A909 was grown at the same temperature as the tested strain. All bacterial cultures were harvested for RNA isolation at mid-exponential growth phase (OD 0.35-0.4). Total RNA was extracted as previously described [72]. RNA was prepared from three independent cultures and each RNA sample was hybridized twice to the microarrays (dye swap). RNA was reverse-transcribed with Superscript indirect cDNA kit (Invitrogen) and labeled with Cy5 or Cy3 (Amersham Biosciences) according to the supplier's instructions. The microarray contains 12889 45-60mer oligonucleotides designed on the predicted gene sequences of strains NEM316, A909 and COH1. The oligonucleotide design was carried out with the OligoArray server [84] (http://berry.engin.umich.edu/oligoarray/). The microarray was manufactured by Agilent Technologies. Only probes with 100% identity to A909, CF00173, 2-22 and SS1219 sequences were taken into account in the data analysis (ie 4401 probes on 1389 genes). Analysis was performed as described [85]. The transcriptome data are MIAME compliant and have been submitted to the ArrayExpress database maintained at www.ebi.ac.uk/microarray-as/ae/ under the Accession Numbers E-MEXP-3828, E-MEXP-3830 and E-MEXP-3829.

Additional files

Additional file 1: Table S1. Shows the results of genome sequence comparison between the seven GBS fish/frog isolates and human and bovine isolates. **Table S2** describes the number of SNPs between the genome sequences of GBS strains isolated from human and fish.

Additional file 2: Table S3. Lists the pseudogenes identified in strains CF01173 and SS1014.

Additional file 3: Table S4. Provides a list of the insertion sequences identified in the genome sequence of the seven GBS strains under study.

Additional file 4: Figure S1. Shows a comparison of genomic island 3.2 between strains 2-22, SS1219 and A909. **Figure S2** is a comparison of the organization of *pil2*, *cyl*, C5a peptidase and alpha-like protein loci in the ST260-261 strains versus strains isolated from human. **Figure S3** shows the organization of the *rgf* locus in ST260-261 and the sequences of the two putative bacteriocin-like peptides.

Additional file 5: Table S5. Lists the pseudogenes identified in ST260-261 strains.

Additional file 6: Table S6. Lists the two-component systems and transcription factors annotated in the genome sequences of ST260 and 261 strains.

Additional file 7: Table S7. Shows the results of Microarray expression analysis for strains 2-22 (ST261), SS1219 (ST260) and CF01173 (ST7).

Additional file 8: Table S8. Provides the characteristics of the Illumina reads and of contigs generated by Velvet.

Abbreviations

GBS: Group B Streptococcus; IS: Insertion sequence; GI: Genomic island; MGE: Mobile genetic element; ICE: Integrative conjugative element; MLST: Multilocus sequence typing; CC: Clonal complex; ST: Sequence type; TCS: Two-component system; CAMP factor: Christie Atkins Munch-Petersen factor; SNP: Single nucleotide polymorphism; PTS: Phosphotransferase system; CDS: Coding sequence; ORF: Open reading frame.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

IRC and PG conceived the study and wrote the article; IRC, PG and CR annotated the sequences; BM, SM, LM, CB and VB obtained the sequence data; EC prepared the biological material; EC, VDC and IRC performed and analyzed the microarray experiments; IRC analyzed the sequence data. All authors have read and approved the manuscript for publication.

Acknowledgements

We specially thank Drs J. Hawke, B. Beall and E. Duchaud for providing us the GBS strains used in the study. We thank Drs. E. Souche and P. Lechat for their help in bioinformatics analysis and Pr P. Trieu-Cuot and Dr. P.E. Douarre and R. Guérillot for helpful discussions. This work was supported by the French National Research Agency (grants ANR-08-GENM-027-001) and the LabEx project IBEID. The Institut Pasteur Genopole is a member of France Génomique (ANR10-IBNS-09-08).

Author details

[1]Unité de Biologie des Bactéries Pathogènes à Gram Positif, 28 rue du Docteur Roux, Paris, Cedex 15 75724, France. [2]CNRS UMR 3525, Paris, France. [3]CEA/IG/Genoscope, Evry, France. [4]Genomic Platform, Institut Pasteur, Paris, France. [5]Unité de Biologie des Bactéries Intracellulaires, Institut Pasteur, Paris, France.

Received: 31 December 2012 Accepted: 1 April 2013
Published: 15 April 2013

References

1. Ben Zakour NL, Guinane CM, Fitzgerald JR: **Pathogenomics of the staphylococci: insights into niche adaptation and the emergence of new virulent strains.** *FEMS Microbiol Lett* 2008, **289**(1):1–12.
2. Lowder BV, Guinane CM, Ben Zakour NL, Weinert LA, Conway-Morris A, Cartwright RA, Simpson AJ, Rambaut A, Nübel U, Fitzgerald JR: **Recent human-to-poultry host jump, adaptation, and pandemic spread of *Staphylococcus aureus*.** *Proc Natl Acad Sci USA* 2009, **106**(46):19545–19550.
3. Guinane CM, Penades JR, Fitzgerald JR: **The role of horizontal gene transfer in *Staphylococcus aureus* host adaptation.** *Virulence* 2011, **2**(3):241–243.
4. Holt KE, Thomson NR, Wain J, Langridge GC, Hasan R, Bhutta ZA, Quail MA, Norbertczak H, Walker D, Simmonds M, *et al:* **Pseudogene accumulation in the evolutionary histories of *Salmonella enterica* serovars Paratyphi A and Typhi.** *BMC Genomics* 2009, **10:**36.
5. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell DJ, Wain J, *et al:* **High-throughput sequencing provides insights into genome variation and evolution in *Salmonella* Typhi.** *Nat Genet* 2008, **40**(8):987–993.
6. Losada L, Ronning CM, DeShazer D, Woods D, Fedorova N, Kim HS, Shabalina SA, Pearson TR, Brinkac L, Tan P, *et al:* **Continuing evolution of *Burkholderia mallei* through genome reduction and large-scale rearrangements.** *Genome Biol Evol* 2010, **2:**102–116.
7. Parkhill J, Sebaihia M, Preston A, Murphy LD, Thomson N, Harris DE, Holden MT, Churcher CM, Bentley SD, Mungall KL, *et al*: **Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*.** *Nat Genet* 2003, **35**(1):32–40.
8. Moran NA, Plague GR: **Genomic changes following host restriction in bacteria.** *Curr Opin Genet Dev* 2004, **14**(6):627–633.
9. Stinear TP, Seemann T, Pidot S, Frigui W, Reysset G, Garnier T, Meurice G, Simon D, Bouchier C, Ma L, *et al*: **Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer.** *Genome Res* 2007, **17**(2):192–200.
10. Song H, Hwang J, Yi H, Ulrich RL, Yu Y, Nierman WC HSK: **The early stage of bacterial genome-reductive evolution in the host.** *PLoS Pathog* 2010, **6:**e1000922.
11. Farley M: **Group B streptococcal disease in nonpregnant adults.** *Clin Infect Dis* 2001, **33:**556–561.
12. Schuchat A: **Epidemiology of group B streptococcal disease in the United States: shifting paradigms.** *Clin Microbiol Rev* 1998, **11:**497.

13. Dermer P, Lee C, Eggert J, Few B: **A history of neonatal group B streptococcus with its related morbidity and mortality rates in the United States.** *J Pediatr Nurs* 2004, **19:**357–363.
14. van der Mee-Marquet N, Fourny L, Arnault L, Domelier A, Salloum M, Lartigue M, Quentin R: **Molecular characterization of human-colonizing *Streptococcus agalactiae* strains isolated from throat, skin, anal margin, and genital body sites.** *J Clin Microbiol* 2008, **46:**2906–2911.
15. Keefe G: ***Streptococcus agalactiae* mastitis: a review.** *Can Vet J* 1997, **38:**429–437.
16. Mian G, Godoy D, Leal C, Yuhara T, Costa G, Figueiredo H: **Aspects of the natural history and virulence of *S. agalactiae* infection in Nile tilapia.** *Vet Microbiol* 2009, **136:**180–183.
17. Vandamme P, Devriese L, Pot B, Kersters K, Melin P: ***Streptococcus difficile* is a nonhemolytic group B, type Ib Streptococcus.** *Int J Syst Bacteriol* 1997, **47:**81–85.
18. Wilkinson H, Thacker LG, Facklam RR: **Nonhemolytic group B streptococci of human, bovine, and ichthyic origin.** *Infect Immun* 1973, **7:**496–498.
19. Amborski RL, Snider TG 3rd, Thune RL, Culley DD Jr: **A non-hemolytic, group B streptococcus infection in cultured bullfrogs, *Rana catesbeiana*, in Brazil.** *J Wildlife Dis* 1983, **19:**180–184.
20. Evans JJ, Bohnsack JF, Klesius PH, Whiting AA, Garcia JC, Shoemaker CA, Takahashi S: **Phylogenetic relationships among *Streptococcus agalactiae* isolated from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infections in Japan.** *J Med Microbiol* 2008, **57:**1369–1376.
21. Jones N, Bohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, Glaser P, Rusniok C, Crook DW, Harding RM, *et al*: **Multilocus sequence typing system for group B streptococcus.** *J Clin Microbiol* 2003, **41:**2530–2536.
22. Sørensen UB, Poulsen K, Ghezzo C, Margarit I, Kilian M: **Emergence and global dissemination of host-specific *Streptococcus agalactiae* clones.** *MBio* 2010, **1:**e00178–10.
23. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, Zouine M, Couvé E, Lalioui L, Poyart C, *et al:* **Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease.** *Mol Microbiol* 2002, **45:**1499–1513.
24. Richards VP, Lang P, Pavinski Bitar PD, Lefebure T, Schukken YH, Zadoks RN, Stanhope MJ: **Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*.** *Infect Genet Evol* 2011, **11:**1263–1275.
25. Tettelin H, Masignani V, Cieslewicz M, Eisen J, Peterson S, Wessels M, Paulsen I, Nelson K, Margarit I, Read T, *et al*: **Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*.** *Proc Natl Acad Sci U S A* 2002, **99:**12391–12396.
26. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, *et al*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome".** *Proc Natl Acad Sci U S A* 2005, **102**(39):13950–13955.
27. Eldar A, Bejerano Y, Bercovier H: ***Streptococcus shiloi* and *Streptococcus difficile*: two new streptococcal species causing a meningoencephalitis in fish.** *Curr Microbiol* 1994, **28:**139–143.
28. Eldar A, Bejerano Y, Livoff A, Horovitcz A, Bercovier H: **Experimental streptococcal meningo-encephalitis in cultured fish.** *Vet Microbiol* 1995, **43:**33–40.
29. Delannoy CM, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, Turnbull JF, Zadoks RN: **Human *Streptococcus agalactiae* strains in aquatic mammals and fish.** *BMC Microbiol* 2013, **13:**41.
30. Delannoy CM, Zadoks RN, Lainson FA, Ferguson HW, Crumlish M, Turnbull JF, Fontaine MC: **Draft genome sequence of a nonhemolytic fish-pathogenic *Streptococcus agalactiae* strain.** *J Bacteriol* 2012, **194:**6341–6342.
31. Liu G, Zhang W, Lu C: **Complete genome sequence of *Streptococcus agalactiae* GD201008-001, isolated in China from tilapia with meningoencephalitis.** *J Bacteriol* 2012, **194:**6653.
32. Wang B, Jian J, Lu Y, Cai S, Huang Y, Tang J, Wu Z: **Complete genome sequence of *streptococcus agalactiae* ZQ0910, a pathogen causing meningoencephalitis in the GIFT strain of nile tilapia (*Oreochromis niloticus*).** *J Bacteriol* 2012, **194:**5132.
33. Brochet M, Rusniok C, Couvé E, Dramsi S, Poyart C, Trieu-Cuot P, Kunst F, Glaser P: **Shaping a bacterial genome by large chromosomal replacements, the evolutionary history of *Streptococcus agalactiae*.** *Proc Natl Acad Sci U S A* 2008, **105:**15961–15966.
34. Elliott JA, Facklam RR, Richter CB: **Whole-cell protein patterns of nonhemolytic group B, type Ib, streptococci isolated from humans, mice, cattle, frogs, and fish.** *J Clin Microbiol* 1990, **28:**628–630.
35. Olivares-Fuster O, Klesius PH, Evans J, Arias CR: **Molecular typing of *Streptococcus agalactiae* isolates from fish.** *J Fish Dis* 2008, **31:**277–283.
36. Michel C, Pelletier C, Boussaha M, Douet DG, Lautraite A, Tailliez P: **Diversity of lactic acid bacteria associated with fish and the fish farm environment, established by amplified rRNA gene restriction analysis.** *Appl Environ Microbiol* 2007, **73:**2947–2955.
37. Evans JJ, Klesius PH, Pasnik DJ, Bohnsack JF: **Human *Streptococcus agalactiae* isolate in Nile tilapia (*Oreochromis niloticus*).** *Emerg Infect Dis* 2009, **15:**774–776.
38. Helm RA, Lee AG, Christman HD, Maloy S: **Genomic rearrangements at *rrn* operons in *Salmonella*.** *Genetics* 2003, **165:**951–959.
39. Kothapalli S, Nair S, Alokam S, Pang T, Khakhria R, Woodward D, Johnson W, Stocker BA, Sanderson KE, Liu SL: **Diversity of genome structure in *Salmonella enterica* serovar Typhi populations.** *J Bacteriol* 2005, **187:**2638–2650.
40. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5:**R12.
41. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
42. Franken C, Brandt C, Bröker G, Spellerberg B: **ISSag1 in streptococcal strains of human and animal origin.** *Int J Med Microbiol* 2004, **294:**247–254.
43. Lachenauer CS, Creti R, Michel JL, Madoff LC: **Mosaicism in the alpha-like protein genes of group B streptococci.** *Proc Natl Acad Sci U S A* 2000, **97:**9630–9635.
44. Franken C, Haase G, Brandt C, Weber-Heynemann J, Martin S, Lämmler C, Podbielski A, Lütticken R, Spellerberg B: **Horizontal gene transfer and host specificity of beta-haemolytic streptococci: the role of a putative composite transposon containing *scpB* and *lmb*.** *Mol Microbiol* 2001, **41:**925–935.
45. Shen P, Huang HV: **Homologous recombination in *Escherichia coli*: dependence on substrate length and homology.** *Genetics* 1986, **112:**441–457.
46. Pierce JC, Kong D, Masker W: **The effect of the length of direct repeats and the presence of palindromes on deletion between directly repeated DNA sequences in bacteriophage T7.** *Nucleic Acids Res* 1991, **19:**3901–3905.
47. Bi X, Liu LF: ***recA*-independent and *recA*-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect.** *J Mol Biol* 1994, **235:**414–423.
48. Chedin F, Dervyn E, Dervyn R, Ehrlich SD, Noirot P: **Frequency of deletion formation decreases exponentially with distance between short direct repeats.** *Mol Microbiol* 1994, **12:**561–569.
49. Lovett ST, Gluckman TJ, Simon PJ, Sutera VA Jr, Drapkin PT: **Recombination between repeats in *Escherichia coli* by a *recA*-independent, proximity-sensitive mechanism.** *Mol Gen Genet* 1994, **245:**294–300.
50. Jonsson IM, Pietrocola G, Speziale P, Verdrengh M, Tarkowski A: **Role of fibrinogen-binding adhesin expression in septic arthritis and septicemia caused by *Streptococcus agalactiae*.** *J Infect Dis* 2005, **192**(8):1456–1464.
51. Schubert A, Zakikhany K, Pietrocola G, Meinke A, Speziale P, Eikmanns BJ, Reinscheid DJ: **The fibrinogen receptor FbsA promotes adherence of *Streptococcus agalactiae* to human epithelial cells.** *Infect Immun* 2004, **72**(11):6197–6205.
52. Pietrocola G, Schubert A, Visai L, Torti M, Fitzgerald JR, Foster TJ, Reinscheid DJ, Speziale P: **FbsA, a fibrinogen-binding protein from *Streptococcus agalactiae*, mediates platelet aggregation.** *Blood* 2005, **105**(3):1052–1059.
53. Schubert A, Zakikhany K, Schreiner M, Frank R, Spellerberg B, Eikmanns BJ, Reinscheid DJ: **A fibrinogen receptor from group B streptococcus interacts with fibrinogen by repetitive units with novel ligand binding sites.** *Mol Microbiol* 2002, **46**(2):557–569.
54. Lang S, Xue J, Guo Z, Palmer M: ***Streptococcus agalactiae* CAMP factor binds to GPI-anchored proteins.** *Med Microbiol Immunol* 2007, **196**(1):1–10.
55. Yim HH, Nittayarin A, Rubens CE: **Analysis of the capsule synthesis locus, a virulence factor in group B streptococci.** *Adv Exp Med Biol* 1997, **418:**995–997.

56. Poyart C, Quesnes G, Trieu-Cuot P: **Sequencing the gene encoding manganese-dependent superoxide dismutase for rapid species identification of enterococci.** *J Clin Microbiol* 2000, **38**(1):415–418.
57. Hamilton A, Popham DL, Carl DJ, Lauth X, Nizet V, Jones AL: **Penicillin-binding protein 1a promotes resistance of group B streptococcus to antimicrobial peptides.** *Infect Immun* 2006, **74**(11):6179–6187.
58. Holmes AR, McNab R, Millsap KW, Rohde M, Hammerschmidt S, Mawdsley JL, Jenkinson HF: **The *pavA* gene of *Streptococcus pneumoniae* encodes a fibronectin-binding protein that is essential for virulence.** *Mol Microbiol* 2001, **41:**1395–1408.
59. van Sorge NM, Quach D, Gurney MA, Sullam PM, Nizet V, Doran KS: **The group B streptococcal serine-rich repeat 1 glycoprotein mediates penetration of the blood-brain barrier.** *J Infect Dis* 2009, **199**(10):1479–1487.
60. Mello LV, De Groot BL, Li S, Jedrzejas MJ: **Structure and flexibility of *Streptococcus agalactiae* hyaluronate lyase complex with its substrate. Insights into the mechanism of processive degradation of hyaluronan.** *J Biol Chem* 2002, **277**(39):36678–36688.
61. Santi I, Scarselli M, Mariani M, Pezzicoli A, Masignani V, Taddei A, Grandi G, Telford JL, Soriani M: **BibA: a novel immunogenic bacterial adhesin contributing to group B streptococcus survival in human blood.** *Mol Microbiol* 2007, **63**(3):754–767.
62. Li S, Jedrzejas MJ: **Hyaluronan binding and degradation by *Streptococcus agalactiae* hyaluronate lyase.** *J Biol Chem* 2001, **276**(44):41407–41416.
63. Lang S, Palmer M: **Characterization of *Streptococcus agalactiae* CAMP factor as a pore-forming toxin.** *J Biol Chem* 2003, **278**(40):38167–38173.
64. Phillips EA, Tapsall JW, Smith DD: **Rapid tube CAMP test for identification of *Streptococcus agalactiae* (Lancefield group B).** *J Clin Microbiol* 1980, **12**(2):135–137.
65. Christie J, McNab R, Jenkinson HF: **Expression of fibronectin-binding protein FbpA modulates adhesion in *Streptococcus gordonii*.** *Microbiology* 2002, **148**(Pt 6):1615–1625.
66. Jedrzejas MJ, Mello LV, de Groot BL, Li S: **Mechanism of hyaluronan degradation by *Streptococcus pneumoniae* hyaluronate lyase. Structures of complexes with the substrate.** *J Biol Chem* 2002, **277**(31):28287–28297.
67. Maisey HC, Doran KS, Nizet V: **Recent advances in understanding the molecular basis of group B streptococcus virulence.** *Expert Rev Mol Med* 2008, **10:**e27.
68. Locke JB, Colvin KM, Datta AK, Patel SK, Naidu NN, Neely MN, Nizet V, Buchanan JT: ***Streptococcus iniae* capsule impairs phagocytic clearance and contributes to virulence in fish.** *J Bacteriol* 2007, **189**(4):1279–1287.
69. Miller JD, Neely MN: **Large-scale screen highlights the importance of capsule for virulence in the zoonotic pathogen *Streptococcus iniae*.** *Infect Immun* 2005, **73:**921–934.
70. Spellerberg B, Rozdzinski E, Martin S, Weber-Heynemann J, Lutticken R: **rgf encodes a novel two-component signal transduction system of Streptococcus agalactiae.** *Infect Immun* 2002, **70**(5):2434–2440.
71. Quach D, van Sorge NM, Kristian SA, Bryan JD, Shelver DW, Doran KS: **The CiaR response regulator in group B streptococcus promotes intracellular survival and resistance to innate immune defenses.** *J Bacteriol* 2009, **191:**2023–2032.
72. Lamy MC, Zouine M, Fert J, Vergassola M, Couve E, Pellegrini E, Glaser P, Kunst F, Msadek T, Trieu-Cuot P, *et al:* **CovS/CovR of group B streptococcus: a two-component global regulatory system involved in virulence.** *Mol Microbiol* 2004, **54**(5):1250–1268.
73. Walker MJ, Hollands A, Sanderson-Smith ML, Cole JN, Kirk JK, Henningham A, McArthur JD, Dinkla K, Aziz RK, Kansal RG, *et al:* **DNase Sda1 provides selection pressure for a switch to invasive group A streptococcal infection.** *Nat Med* 2007, **13**(8):981–985.
74. Sumby P, Whitney AR, Graviss EA, DeLeo FR, Musser JM: **Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity.** *PLoS Pathog* 2006, **2**(1):e5.
75. Cunin R, Glansdorff N, Piérard A, Stalon V: **Biosynthesis and metabolism of arginine in bacteria.** *Microbiol Rev* 1986, **50:**314–352.
76. Kakinuma Y: **Inorganic cation transport and energy transduction in *Enterococcus hira*e and other streptococci.** *Microbiol Mol Biol Rev* 1998, **62:**1021–1045.
77. Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, *et al:* **High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies.** *BMC Genomics* 2008, **9:**603.
78. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821–829.
79. Darling ACE, Mau B, Blattner FR, Perna NT: **Mauve: multiple alignment of conserved genomic sequence with rearrangements.** *Genome Res* 2004, **14**(7):1394–1403.
80. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
81. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D: **Tablet-next generation sequence assembly visualization.** *Bioinformatics* 2010, **26:**401–402.
82. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, *et al:* **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211–D222.
83. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation.** *Bioinformatics* 2000, **16:**944–945.
84. Rouillard JM, Herbert CJ, Zuker M: **OligoArray: genome-scale oligonucleotide design for microarrays.** *Bioinformatics* 2002, **18:**486–487.
85. Dramsi S, Dubrac S, Konto-Ghiorghi Y, Da Cunha V, Couvé E, Glaser P, Caliot E, Débarbouillé M, Bellais S, Trieu-Cuot P, *et al:* **Rga, a RofA-like regulator, is the major transcriptional activator of the PI-2a pilus in *Streptococcus agalactiae*.** *Microb Drug Resist* 2012, **18:**286–297.

doi:10.1186/1471-2164-14-252
Cite this article as: Rosinski-Chupin *et al.*: Reductive evolution in *Streptococcus agalactiae* and the emergence of a host adapted lineage. *BMC Genomics* 2013 **14**:252.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central