

Identifying evolutionarily conserved segments among multiple divergent and rearranged genomes

Bob Mau^{1,2,5}, Aaron E. Darling^{1,3}, and Nicole T. Perna^{1,4}

¹ Dept. of Animal Health and Biomedical Sciences,

² Dept. of Oncology

³ Dept. of Computer Science

⁴ The Genome Center of Wisconsin

University of Wisconsin - Madison

1656 Linden Dr., Madison, WI 53706, USA

⁵ To whom correspondence should be addressed: robertm@genome.wisc.edu

Abstract. We describe a new method for reliably identifying conserved segments among genome sequences that have undergone rearrangement, horizontal transfer, and substantial nucleotide-level divergence. A Gibbs-like sampler explores different combinations of sequence-based markers shared by the genomes under study. The sampler assigns each marker a posterior probability based on how frequently it participates in some collinear group of markers. Markers with high p.p. values are likely members of conserved segments. The method identifies both large-scale and local trends in segmental collinearity, providing suitable input for genome alignment and rearrangement history inference tools. Applying our method to genomes of four *Streptococci* reveals that rearranged segments in these organisms belong in two size categories: large conserved segments that are interrupted by a staccato of single gene or operon-size small segments. The rearrangement pattern of large segments is best explained by symmetric inversions about the origin of replication while the pattern of small segments is not.

1 Introduction

Nadeau and Taylor [1] introduced the concept of 'conserved segments' when comparing the genetic linkage maps of human and mouse. Conserved segments are homologous regions between genomes in which common genetic markers occur in the same order. Twenty years later, comparison of the completed human and mouse genomes found one third more large-scale rearrangements than predicted [2], as well as thousands of micro-rearrangements [3]. Similarly, the discovery of a large inversion between *E. coli* and *Salmonella typhimurium* by genetic analysis [4] spurred studies of genomic rearrangements in microbes. Pairwise comparisons of sequenced eubacterial genomes later confirmed that symmetric inversions about the origin and terminus of replication are common in this domain [5, 6]. Such pairwise comparisons are easy to implement but provide only

limited analytical power, whereas multiple genome comparison enables the application of more powerful phylogenetic methods.

We present a method to reliably identify conserved blocks of sequence among several genomes that have undergone rearrangement. Our approach to rearrangement identification relies on monotypic markers to suggest potential homology. Monotypic markers are genomic features that occur exactly once in each genome being compared. The order and orientation in which these markers appear can be written as signed permutations of integers. Applying breakpoint analysis [7] to these permutations separates the markers into disjoint subsets of collinear markers. The regions of DNA spanned by the markers inside a given subset form a locally collinear block, or LCB. Unlike the conserved segments originally described by Nadeau and Taylor, LCBs are based solely on sequence similarity and do not imply any type of common evolutionary history or biological significance. In particular, LCBs make no distinction between segments that are similar merely by chance and truly orthologous segments - segments whose similarity derives from a single locus in the most recent common ancestor (MRCA). Conserved segments are regions of strictly orthologous sequences [8] that may contain lineage specific sequence, but do not contain rearrangements of orthologous sequence. By using marker order rather than the chromosomal proximity of markers to assess segmental conservation, our method accommodates lineage specific lateral gene transfer. Previous analytical tools were either limited to closely-related taxa [9] or did not account for horizontal transfer events common in bacterial genomes [3, 10].

Our target data set for this work consists of a group of four *Streptococcus* species that have sufficiently diverged so that comparisons at the nucleotide level are not practicable. We have shown in earlier studies [9, 11] that multi-MUMs (multiple maximum unique matches) are simple and effective monotypic markers for finding genomic rearrangements, but their applicability is limited to closely related organisms. When comparing more distantly related genomes, exact nucleotide matches such as multi-MUMs fail to generate a sufficiently comprehensive set of monotypic markers. Various types of inexact matching algorithms have been designed with DNA sequence in mind [12–15], but BLAST [16] hits at the protein level remain one of the most sensitive and widely used pairwise alternatives. In order to compare three or more annotated genomes, we define a gene-based monotypic marker to consist of a single gene from each genome, where each gene is the reciprocal best BLAST hit of the other genes comprising the marker. Although our focus here is on gene-based marker sets, we present the algorithm in full generality.

2 Notation

We start with a collection of G genomes \mathcal{G} and M sequence-based markers \mathcal{M} such that each marker occurs once and only once in each $g \in \mathcal{G}$. One genome, denoted g_1 , is designated as a reference genome. In order to facilitate breakpoint determination, we assign an integer label to each marker based on the marker's

order in g_1 coordinates, *i.e.* the j^{th} marker in g_1 is labelled j . In other genomes, m_j may reside on the opposite strand, in which case that instance is labelled $-j$. Hence, $m_j = \pm j$, depending on its orientation in g_k relative to g_1 . Denote the G labels of m_j by $\phi_k(m_j) = \pm j$. A signed permutation ζ_k is constructed for each genome by sorting the M integers of $\phi_k(\mathcal{M})$ by their location in g_k . ζ_k thereby encodes the order and relative orientation of monotypic markers in g_k . Throughout, markers m are indexed by j or v , genomes g by k , and integer elements z in ζ_k will be indexed by i . When it is clear from context, the marker m_j is denoted by its label j for ease of exposition.

An adjacency in $\zeta = (z_1, \dots, z_i, \dots, z_M)$ exists whenever $z_{i+1} - z_i = 1$. Conversely, $z_{i+1} - z_i \neq 1$ indicates a breakpoint between markers i and $i + 1$. Breakpoints in ζ partition \mathcal{M} into locally collinear blocks, groups of consecutively increasing integers. For example, $\zeta = (1, 2, 3, 4, 5, -8, -7, -6, 10, 9, -13, -12, -11, 14)$ consists of six collinear blocks: $\{1, 2, 3, 4, 5\}$, $\{6, 7, 8\}$, $\{11, 12, 13\}$, and three singletons: $\{9\}$, $\{10\}$, and $\{14\}$. Extension to three or more genomes is carried out by stacking permutations on top of one another to create a G x M matrix of integer-valued permutation elements: $\mathbf{Z}_G(\mathcal{M}) = (\zeta_1, \dots, \zeta_G)$. The breakpoints of $\mathbf{Z}_G(\mathcal{M})$ are the union of breakpoints from each row ζ_k . In this framework, locally collinear blocks can be viewed as groups of consecutive integers that are present in either orientation in every row.

Our MCMC algorithm utilizes this matrix representation to determine LCBs “on the fly”. After locating the G instances of m in $\mathbf{Z}_G(\mathcal{M})$, the algorithm quickly identifies all surrounding collinear markers by scanning to the left and right of m in all rows (genomes) simultaneously until the first breakpoint is encountered in each direction.

The procedure can be formalized as follows. Let $z_k(m_j)$ be the unique occurrence of m_j in ζ_k . If $i(m_j, k)$ denotes the relative position of marker m_j in g_k (equivalently, in ζ_k), then $z_k(m_j) = z(k, i(m_j, k))$ in $\mathbf{Z}_G(\mathcal{M})$. In particular, $z_1(m_j) = z(1, j) = j$. Define a shift operator θ on $z_k(m)$ by $\theta^h(z_k(m)) = z(k, i(m, k) + \text{sgn}(\phi_k(m) * h))$. Hence, $\theta^h(z_k(m))$ is the h^{th} marker to the right or left of m in ζ_k . For $h > 0$, shift h units to the right in ζ_k whenever m is located on the same strand (*i.e.*, $\phi_k(m) = 1$) and h units to the left when m occurs on the opposite strand. Reverse directions for $h < 0$. Trivially, $\theta^0(z_k(m)) = z_k(m)$. The limits of collinearity about marker $m = m_j$ can be formulated as:

$$\begin{aligned} \text{Lend}(m : \mathbf{Z}(\mathcal{M})) &= j + \min_{u \leq 0} \{j + h = \text{sgn}(\phi_k(m))\theta^h(z_k(m)) \forall k, h : 0 \geq h \geq u\} \\ \text{Rend}(m : \mathbf{Z}(\mathcal{M})) &= j + \max_{u \geq 0} \{j + h = \text{sgn}(\phi_k(m))\theta^h(z_k(m)) \forall k, h : 0 \leq h \leq u\} \end{aligned}$$

In a sense, this can be viewed as a seed and extend method to identify LCBs. The set of markers $\{m_v : \text{Lend}(m_j : \mathbf{Z}_G(\mathcal{M})) \leq v \leq \text{Rend}(m_j : \mathbf{Z}_G(\mathcal{M}))\}$ comprises the longest uninterrupted stretch of collinear, co-oriented markers containing m . The need for explicit recognition of the dependence on $\mathbf{Z}_G(\mathcal{M})$ will become clear shortly.

First, we give an example consisting of fourteen markers in four genomes and find the LCB containing m_7 .

$$\mathbf{Z}_G(\mathcal{M}) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 1 & 2 & 3 & 4 & 9 & -8 & -7 & -6 & -5 & 10 & -13 & -12 & -11 & 14 \\ 1 & 2 & -5 & -4 & -3 & 14 & 6 & 7 & 8 & 9 & 11 & 12 & 13 & -10 \\ 6 & 7 & 8 & 9 & 1 & 2 & 3 & 4 & 5 & 10 & 11 & 12 & 13 & 14 \end{pmatrix}$$

For $m = m_7$, $z_1(m) = 7$, $z_2(m) = -7$, $i(7, 2) = 7$, $z_3(m) = +7$, $i(7, 3) = 8$, $z_4(4) = +7$ and $i(7, 4) = 2$. The first mismatch to the right of marker 7 occurs two markers to the left of -7 in g_2 , where $\text{sgn}(\phi_2(m)) \times \theta^2(z_2(m) = -7) = -1 \times 9 \neq 9$. Hence, $\text{Rend}(7) = 8$. Likewise, breakpoints in g_3 and g_4 occur two markers to the left of 7, so $\text{Lend}(7) = 6$. Consequently, $\text{LCB}(7) = \{6, 7, 8\}$.

3 A Pseudo-Gibbs Sampler

Not all markers help define evolutionarily conserved segments among genomes. On the contrary, some markers actually disrupt such segments. The problem is to identify those markers that optimally segregate into conserved segments. We attack the problem by surveying and assessing candidate subsets using Markov chain Monte Carlo technology. Each subset of \mathcal{M} can be represented by a vector of M zeroes and ones, where a one in the j^{th} position indicates that the j^{th} marker is included in the subset, denoted by \mathcal{M}^{in} . We call this representation a configuration. In the previous example, the configuration $(0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 0)$ corresponds to the subset $\mathcal{M}^{\text{in}} = \{6, 7, 8, 11, 12, 13\}$. Denote the collection of locally collinear blocks of \mathcal{M}^{in} by $\mathcal{L}(\mathcal{M}^{\text{in}})$ to emphasize the dependence on the set of included markers \mathcal{M}^{in} . Here, $\mathcal{L}(\mathcal{M}^{\text{in}})$ consists of two blocks: $(6, 7, 8)$ and $(11, 12, 13)$. Two different configurations on \mathcal{M} define distinct inclusion subsets $\mathcal{M}_1^{\text{in}}$ and $\mathcal{M}_2^{\text{in}}$. $\mathcal{L}(\mathcal{M}_1^{\text{in}})$ and $\mathcal{L}(\mathcal{M}_2^{\text{in}})$ are considered equal if their LCBs span the same intervals in every genome (*cf.* $\mathcal{M}^{\text{in}} = \{6, 8, 11, 13\}$ for configuration $(0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 1, 0)$).

We designed a pseudo-Gibbs sampler to explore the space of configurations in search of well-supported subsets as follows. Assign random variables X_j that map each m_j into a state of inclusion (1) or exclusion (0). Hence, the random vector $\mathbf{X}(\mathcal{M}) = (X_1(m_1), \dots, X_M(m_M))$ maps \mathcal{M} into $\mathbf{x} = (x_1, \dots, x_M)$, a configuration of size M . Initialize $\mathbf{X}^0(\mathcal{M}) = \mathbf{x}^0$ with a draw of M independent Bernoulli($\frac{1}{2}$) random variates. A Markov chain $(\mathbf{X}^0, \mathbf{X}^1, \dots, \mathbf{X}^n, \dots, \mathbf{X}^N)$ is run over the space of configurations as follows. Pick a marker at random and compute a score conditioned on the current configuration $\mathbf{X}^n = \mathbf{x}$. Then convert the score to a ‘‘conditional probability’’ to stochastically update $X_j^{n+1}(m_j)$. The specific formulae used are:

$$\text{Score}(m_j \mid \mathbf{X}^n = \mathbf{x}) = \sum_{v=L}^{j-1} w_v x_v + \max(\lambda(w_j - w_{\min}), 0) + \sum_{v=j+1}^R w_v x_v, \quad (1)$$

where $L = \text{Lend}(m : \mathbf{Z}_G(\mathcal{M}_n^{\text{in}}))$ and $R = \text{Rend}(m : \mathbf{Z}_G(\mathcal{M}_n^{\text{in}}))$ as defined above, w_m is a marker’s weight, and λ and w_{\min} down-weight the current marker.

$$\hat{p}_j = \frac{e^{\text{Score}(m_j)/c} - 1}{e^{\text{Score}(m_j)/c} + 1}, \text{ where } c > 0 \text{ is a scale parameter.} \quad (2)$$

$$\text{Sample } u \sim \text{Unif}[0,1]. \quad \text{If } u \geq \hat{p}_j, \quad X_j^{n+1} = 0, \quad \text{else } X_j^{n+1} = 1. \quad (3)$$

The score in (1) is the sum of the weights of all the collinear markers to the left and the right of m in the current configuration \mathbf{x} . When sets of markers consist of exact sequence matches, the weight w_m is simply the length of the match. For gene-based markers, calculation of w_m is complicated because every pair of reciprocal best BLAST hits generates a different BLAST bit score. We compute a gene-based marker’s weight w_m as the square root of the average bit score over all possible genome pairs. The square root transformation reduces the distributional skew of large scores in long genes. The formula for the update probability in (2) is the right half of a sigmoidal function.

For the analysis described here, LCBs consisting of one or two genes are not particularly illuminating. In the case of phylogenetic reconstruction based on rearrangements, they can lead to false inferences. Rather than summarily exclude such blocks, their frequency can be minimized by down-weighting the current marker in the score function. Subtracting a minimum weight offset w_{min} suffices for nucleotide based markers, but with gene-based markers, an additional multiplicative reduction is required (*i.e.* $\lambda < 1$).

The pseudo-Gibbs sampler iterates through these three steps tens of millions of times. From a random start \mathbf{x}_0 , the chain undergoes a burn-in period before entering well-supported configurations. These early realizations are discarded. As the Markov chain converges, some markers coalesce into collinear blocks because markers within a collinear block contribute to each other’s scores. Counters record the number of times each marker is updated and the number of times the update is a one.

When the Markov chain has completed its pre-assigned number of iterations, the relative frequencies of inclusion in \mathcal{M}^m are computed from the recorded counts for each m . Had we a bona fide Gibbs sampler, in which the conditional distributions were consistent with some joint probability distribution, the Hammersley-Clifford theorem [17] and standard Markov chain Monte Carlo theory [18] would guarantee that the relative frequency at each node converges to the appropriate marginal posterior probability.

$$\frac{\# \text{ of ones at } m_j}{\# \text{ of visits to } m_j} \rightarrow \pi_j(c) = \Pr(m_j \text{ present in } \mathcal{M}^{true} \mid \text{ scale parameter } c) \quad (4)$$

Although not a true Gibbs sampler, experimental evidence indicates that the pseudo-Gibbs sampler generates reproducible estimates of these marginals from random initial configurations – an empirical proof of convergence. Note that the dependence of the posterior probabilities on the scale parameter c is explicitly recognized in the conditional probability notation.

A second user-provided parameter is the probability threshold γ . An appropriate choice of γ is determined empirically once the sampler has run its course. Histograms of marginal posterior probabilities, such as the one below, suggest that most markers are either isolated (p.p. near 0) or part of a larger block (p.p. near one).

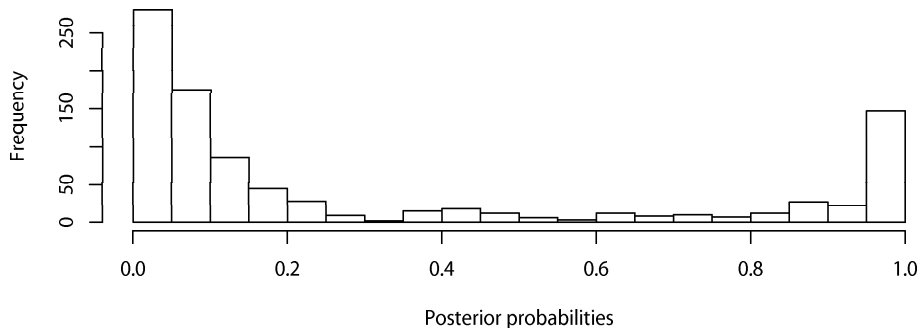


Fig. 1. Histogram of posterior probabilities using the medium resolution settings in Table 1 for a set of 938 gene-based markers present in four species. Discrimination is typically more pronounced with nucleotide-based markers.

The relative frequencies in (4) induce a stochastic ordering on \mathcal{M} . The stochastic ordering ranks markers by the strength of the evidence that each m joins with its neighbors to form a conserved segment. Recall that every configuration partitions \mathcal{M} into two disjoint subsets \mathcal{M}^{in} and its complement. Ranking $\{\pi_i(c)\}$ forms a family of partitions $\mathcal{M}^{in}(\gamma; c)$. Partitions of interest generally involve thresholds between 0.25 and 0.75. Hence, only a few configurations need actually be examined by the scientist.

Although ranking markers by their p.p. is a fairly robust procedure with respect to the scale parameter c , it is far from invariant. Large values of c omit small collinear blocks in favor of long blocks over large spans. Such runs are called low resolution. Conversely, high resolution (small c) runs identify small blocks that can disrupt larger low resolution blocks.

4 Results

Streptococcal strains are responsible for a wide range of diseases in humans. *S. pyogenes*, most commonly associated with “strep throat”, also causes pneumonia or rheumatic fever if untreated [19]. *S. agalactiae* is the leading cause of pneumonia and meningitis in newborns [20]. *S. pneumoniae* [21] also causes of pneumonia and meningitis, but has multiple phenotypes that distinguish it from the other two species. Finally, *S. mutans* [22] is responsible for a large percentage of tooth decay. More Streptococcal genomes (nine) have been sequenced than any other genus. Although detailed comparative analyses have been conducted within species [20, 23] little has been published about all four species beyond pairwise contrasts (see [24] Table 3, [6] Supplemental material, and [20] Figure 2). Curiously, these previous analyses show that the smallest genome, *S. pyogenes*, and the largest genome, *S. agalactiae*, are the two most closely related taxa. By contrast, *S. pneumoniae* is the most phylogenetically distant species.

We present an analysis of genome rearrangements among these diverse strains using our pseudo-Gibbs sampler. To generate a set of monotypic markers, a recip-

rocal best BLAST search was done between each pair of Streptococcal genomes, retaining only matches with an E-value <0.00005 covering at least 50% of both proteins. The distribution of common reciprocal best matches among the four taxa are shown in Figure 2. Although 968 genes common to the four genomes meet these criteria in all six paired comparisons, only 938 consist exclusively of one-to-one matches within each comparison. This reduced group of genes, from which putative gene duplications have been removed, forms a set of monotypic markers. Given our stringent match criteria, most scientists would categorize these genes as orthologs. We use these markers to investigate two complementary aspects of comparative genomics: identifying connected neighborhoods of orthologous genes and inferring ancestral genome architecture. These two problems demand different levels of resolution to identify rearranged segments.

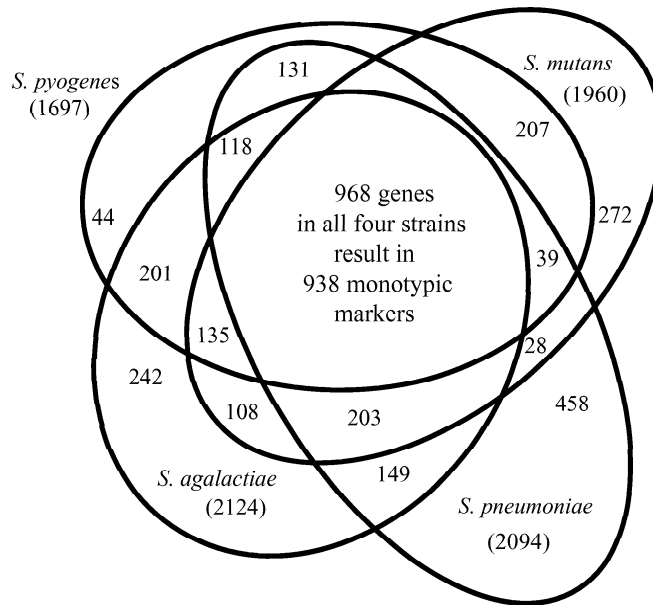


Fig. 2. This Venn diagram shows the partition of genes from all four Streptococcal genomes into 15 groups of mutual reciprocal best BLAST hits. *S. pneumoniae* has the most lineage-specific genes (458), while *S. pyogenes* has only 44 unique genes. Removing paralogous genes leaves 938 monotypic markers common to all four species.

The scale parameter c divides the score (1), affecting the size of detectable gene clusters. A smaller value of c generally improves sensitivity to small collinear segments but may introduce additional noise. The threshold parameter γ partitions markers into signal and noise: markers with $p.p. > \gamma$ are deemed signal while the rest are considered noise. As mentioned above, γ is chosen by inspecting the frequency distribution of posterior probabilities (see Figure 1).

Breakpoints in the selected set of signal markers define collinear segments of the genomes under study.

We begin with a search for clusters of orthologous genes. Since the introduction of clusters of orthologous groups, or COGs [25,26], the concept has been expanded to include connected gene neighborhoods [27,28]. Typically methods to construct gene neighborhoods start with triplets of genes, and inductively work their way up to larger connected neighborhoods. Rather than growing a neighborhood from a 'seed' COG, our method directly identifies neighborhoods as locally collinear blocks among the genomes. Statistical tests have been developed for assessing patterns of collinearity between two genomes [30,29], but they have not been extended to multi-genome analysis.

Our empirical approach permits us to indirectly modulate the minimum neighborhood size by adjusting γ and c . Table 1 shows a series of runs on the 938 Streptococcus markers using different parameter settings to achieve low, medium, and two high resolutions. In particular, observe that both 17-gene clusters are split into smaller LCBs at high resolution.

Number of Segments (LCBs)

Genes per segment	2	3	4	5	6	7	8	9	10	11	13	14	17	18	24	Total
Resolution parameters (c, γ, w_{min})																
Low (75,45%,20)				1	2	1	6	2	5	3	1		2	1	2	26
Medium (30,45%,8)			3	5	6	2	6	1	4	1	1		2	1	2	34
High-1 (20,50%,15)	1	4	20	7	7	2	6	2	3	1	2			1	2	57
High-2 (20,30%,15)	3	11	29	7	7	2	6	1	3	1	2	1			2	72

Table 1. Distribution of gene counts per collinear segment. Clusters are determined under three different conditions ranging in resolution from high to low. As the resolution increases, some large clusters split into smaller clusters by the emergence of a previously unnoticed intervening cluster.

The four runs in Table 1 use a score function where the current marker's weight is reduced 25 %. The role of λ in (1) is apparent when the medium resolution run in Table 1 is repeated without it (*i.e.* $\lambda = 1$) and the runs are compared. The number of LCBs jumps from 36 to 109, including 20 singletons and 15 pairs. Several large blocks are split into smaller segments, contributing to the increase while obscuring the underlying pattern of collinearity.

The same phenomenon can occur if γ is lowered. We present a particularly interesting example in Figure 3, magnified so genes can be represented as rectangles of varying length rather than points.

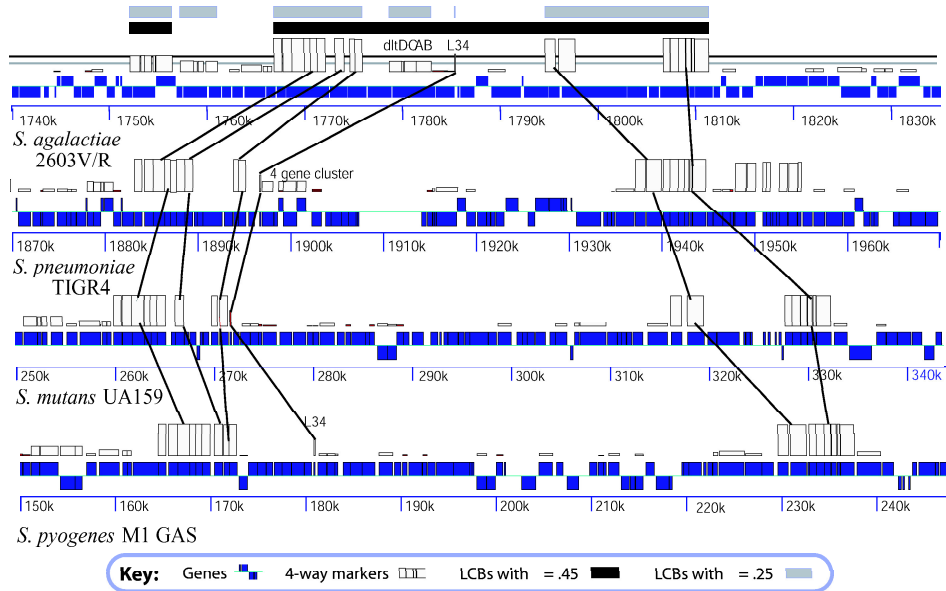


Fig. 3. A 100 kb region of *S. agalactiae* and the corresponding regions in the other three *Streptococcus* genomes. Meaning of rows within genome panel, starting at the bottom: genome coordinates, annotated genes (in black) with vertical position denoting transcriptional direction. Monotypic gene markers are drawn as white boxes with heights proportional to their posterior probability. Lines among the genomes connect clusters of markers that together comprise a large LCB. Other markers are outside the field of view in at least one genome and thus connecting lines can't be drawn. Black bars across the top denote LCBs formed from the medium resolution settings in Table 1. Lowering γ to 0.25 interjects small segments that break up the large black block.

The large black collinear region in Figure 3 merits special attention. The smallest gene contributing to this segment is the ribosomal protein L34. At the lower threshold, L34 becomes isolated by the group of genes immediately to its right in *S. pneumoniae*, labeled “four gene cluster”, and the *dltDCAB* operon to its left in *S. agalactiae*. Note that the largest genome, *S. agalactiae*, has the fewest number of lineage-specific genes within this segment.

We applied GRIMM [34] and MGR [35] to infer the ancestral genome organization of the four stains. Our initial analysis used high resolution collinear segments from Table 1. The rearrangement scenarios suggested by MGR on high resolution segments do not maintain replichore balance, indicating that some of the collinear segments in this data set may not have been rearranged by symmetric inversions (data not shown). We then examined the 26 segments that from the low resolution run. Markers that exceed the p.p. threshold. tend to cluster about the diagonals of the X-plots shown in Figure 4, a pattern consistent with multiple symmetric inversions.

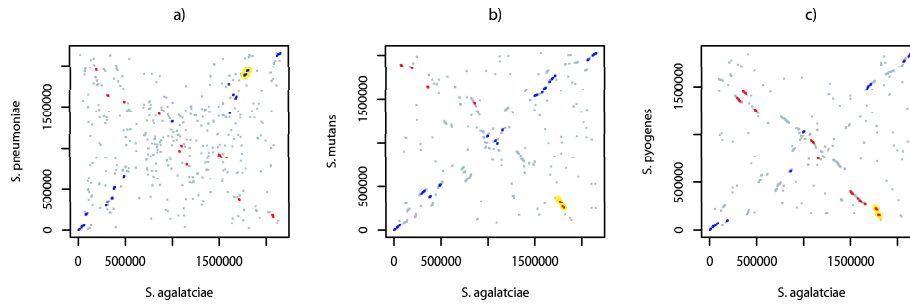


Fig. 4. X-plots of the 938 common orthologs, with each gene’s position in *S. agalactiae* on the horizontal axis plotted against the corresponding position in each of the three other strains. Many putative orthologs do not meet the p.p. threshold at low resolution, and are drawn in grey. Genes above the threshold are black. In (c), and to a lesser degree in (b), collinear segments near the center of the X-plot are visible in grey. This suggests that certain orthologs may be collinear in some genomes, but not in all four.

Using the 26 low-resolution segments, we ran GRIMM and MGR again. The result is a collection of 37 inversion events that maintain replicore balance among genomes (data not shown). Figure 5 shows a phylogenetic tree based on inversion events with branch labels giving the number of inversions per branch. Unlike some other comparisons [36], the frequency of genomic rearrangements between *Streptococci* appears to correlate well with the overall level of sequence divergence.

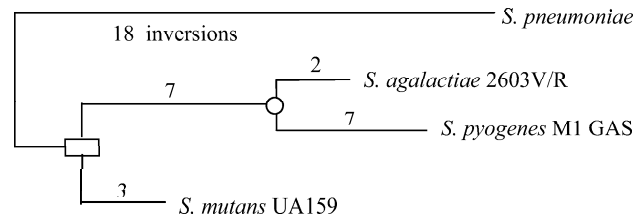


Fig. 5. Phylogeny of Streptococcal strains based on a parsimonious set of 26 genomic rearrangements of large segments, courtesy of MGR and GRIMM. The number of inversions along each lineage accompanies the branch. The circle denotes the ancestral genome of *S. pyogenes* and *S. agalactiae*; the rectangle is the ancestral genome of the circle and *S. mutans*.

The comparison among these four Streptococci allows us to infer the ancestral organization of the MRCA of *S. agalactiae*, *S. pyogenes*., and *S. mutans*, by using *S. pneumoniae* as an outgroup. A separate 3-way analysis could be conducted for the MRCA of *S. agalactiae* and *S. pyogenes* with *S. mutans* as the outgroup.

5 Discussion and Conclusion

By assigning posterior probabilities to monotypic markers our algorithm assists in making a “best guess” as to which LCBs constitute evolutionarily conserved segments among a group of genomes. Once identified, such conserved segments can be subject to further analyses such as multiple global alignment or phylogenetic inference of genome organization. Furthermore, the flexibility of our algorithm makes it well suited to comparisons of both eukaryotic and prokaryotic genomes. By using markers based on inexact protein or nucleotide sequence matches the algorithm accommodates significantly diverged genomes, and its ignorance of distance between markers allows it to be applied to genomes with significant lineage-specific content.

In eubacteria, the origin and terminus of replication divide a circular chromosome into two replichores of similar length. Equal sized replichores are thought to maximize efficiency of replication of the genome and an imbalance of more than 20% can be selected against [31]. It is currently believed that symmetric inversions are the predominant means of genome rearrangements in eubacteria [5, 6, 32]. GRIMM and MGR implement sorting by reversals for circular chromosomes without any constraint on the replichore sizes of ancestral intermediates. As such, these tools are not appropriate when small clusters of orthologous genes are included – if in fact they were translocated by some other means [33]. We stress that it is the scientist’s responsibility to judge whether such tools provide a useful analysis of a given bacterial data set.

If inversions are not responsible for all such “micro-rearrangements”, other evolutionary mechanisms must be. One explanation for the observed micro-rearrangements is transposition mediated by insertion sequences. An alternative explanation is parallel lateral gene transfer events, acting independently to introduce the same DNA to different loci in each lineage. This phenomenon is called convergent evolution. A related mechanism is serial evolution - a horizontal transfer of DNA into one lineage followed by a transfer from that lineage to a second one. A fourth possibility would be ancient gene duplication and subsequent loss of the original gene copy. Attributing the mechanism responsible for a particular micro-rearrangement remains an open problem.

6 Acknowledgements

We thank Guillaume Bourque and Glenn Tesler for help with MGR and GRIMM, and Elisabeth Tillier for a critical reading of a related manuscript. Funding for all authors was provided by NIH Grant GM62994-02. A.E.D. was supported in part by NLM Training Grant 5T15M007359-03.

References

1. Nadeau, J.H., Taylor, B.A.: Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc Natl Acad Sci USA* **81** (1984) 814-818

2. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al: Initial sequencing and comparative analysis of the mouse genome. *Nature* **420** (2002) 520-562.
3. Pevzner P, Tesler G: Genome rearrangements in Mammalian evolution: lessons from human and mouse genomes. *Genome Res* **13**(2003) 37-45.
4. Schmid MB, Roth JR: Selection and endpoint distribution of bacterial inversion mutations. *Genetics* **105**(1983) 539-557.
5. Eisen JA, Heidelberg JF, White O, Salzberg SL: Evidence of symmetric chromosomal inversions around the replication origin in bacteria. *Genome Biology* **1**(2000) 1-9.
6. Tillier ER, Collins RA: Genome rearrangement by replication-directed translocation. *Nat Genet* **26**(2000) 195-197.
7. Blanchette M, Kunisawa T, Sankoff D: Gene order breakpoint evidence in animal mitochondrial phylogeny. *J Mol Evol* **49**(1999) 193-203.
8. Fitch WM: Homology a personal view on some of the problems. *Trends Genet* **16**(2000) 227-231.
9. Darling ACE, Mau B, Blattner FR, Perna NT: Mauve: Multiple Alignment of Conserved Genomic Sequence with Rearrangements. *Genome Res* **14** (2004) 1394-1403.
10. Calabrese PP, Chakravarty S, Vision TJ: Fast identification and statistical evaluation of segmental homologies in comparative maps. *Bioinformatics* **19** (2003) Suppl 74-80.
11. Darling A, Mau B, Blattner FR, Perna NT: GRIL: Genome rearrangement and inversion locator. *Bioinformatics* **20** (2003) (122-124).
12. Buhler J: Efficient large-scale sequence comparison by locality-sensitive hashing. *Bioinformatics* **17** (2001) 419-428.
13. Ma, B, Tromp J, Li M: PatternHunter: faster and more sensitive homology search. *Bioinformatics* **18** (2002) 440-445.
14. Brudno M, Steinkamp R, Morgenstern B: The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. *Nucleic Acids Res.* (Web Server issue) (2004) W41-44.
15. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: Human-mouse alignments with BLAST². *Genome Res* **13**(2003) 103-107.
16. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25** (1997) 3389-3402.
17. Besag J: Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society Series B* **36** (1974) 192-236.
18. Tierney L: Markov chains for exploring posterior distributions. *Annals of Statistics* **22** (1994) 1701-1762.
19. Ferretti JJ, McShan WM, Ajdic D, Savic DJ, Savic G, Lyon K, Primeaux C, Sezate S, Suvorov AN, Kenton S et al: Complete genome sequence of an M1 strain of *Streptococcus pyogenes*. *Proc Natl Acad Sci U S A* **98** (2001) 4658-4663.
20. Tettelin H, Maignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD 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* **99** (2002) 12391-12396.
21. Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ et al: Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. *Science* **293** (2001) 498-506.

22. Ajdic D, McShan WM, McLaughlin RE, Savic G, Chang J, Carson MB, Primeaux C, Tian R, Kenton S, Jia H et al: Genome sequence of *Streptococcus mutans* UA159, a cariogenic dental pathogen. *Proc Natl Acad Sci U S A* **99** (2002) 14434-14439.
23. Smoot JC, Barbian KD, Van Gompel JJ, Smoot LM, Chaussee MS, Sylva GL, Sturdevant DE, Ricklefs SM, Porcella SF, Parkins LD et al: Genome sequence and comparative microarray analysis of serotype M18 group A *Streptococcus* strains associated with acute rheumatic fever outbreaks. *Proc Natl Acad Sci U S A* **99** (2002) 4668-4673.
24. Ferretti JJ, Ajdic D, McShan WM: Comparative genomics of streptococcal species. *Indian J Med Res* **119** (2004) Suppl 1-6.
25. Tatusov RL, Koonin EV, Lipman DJ: A genomic perspective on protein families. *Science* 1997, **278** (1997) 631-637.
26. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV: The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29** (2001) 22-28.
27. Rogozin IB, Makarova KS, Murvai J, Czabarka E, Wolf YI, Tatusov RL, Szekely LA, Koonin EV: Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res* **30** (2002) 2212-2223.
28. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV: Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol* **4** (2003) R55.
29. Hampson, S, McLysaght, A, Gaut, BS, Baldi, PF: LineUp: Statistical Detection of Chromosomal Homology with Application to Plant Comparative Genomics. *Genome Research* **13** (2003) 999-1010.
30. Durand, D, Sankoff, D: Tests for Gene Clustering. *Journal of Computational Biology*. **10** (2003) 453-482.
31. Guijo MI, Patte J, del Mar Campos M, Louarn JM, Rebollo JE: Localized Remodeling of the *Escherichia coli* Chromosome. The patchwork of segments refractory and tolerant to inversion near the replication terminus. *Genetics* **157** (2001) 1413-1423.
32. Ajana, Y., Lefebvre, J. F., Tillier, E., El-Mabrouk, N.: Exploring the set of all minimal sequences of reversals - An application to test the replication-directed reversal hypothesis. Second International Workshop, Algorithms in Bioinformatics (WABI 2002), *LNCIS* **2452**, R. Guigo and D. Gusfield eds. (2002) 300-315.
33. Lefebvre JF, El-Mabrouk N, Tillier ER, Sankoff D: Detection and validation of single gene inversions. *Bioinformatics* **19**, Suppl. 1, special issue, 11th International Conference on Intelligent Systems for Molecular Biology (2003) 190-196.
34. Tesler G: GRIMM: genome rearrangements web server. *Bioinformatics* **18** (2002) 492-493.
35. Bourque G, Pevzner PA: Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res* **12** (2002) 26-36.
36. Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry R: Genome sequence of *Yersinia pestis* KIM. *J Bacteriol.* **184** (2002) 4601-4611.