Taille des génomes

Exemple génomes procaryotes

Species	Genome size (Mb)
Bacteria	
Mycoplasma genitalium	0,58
Heamophilus influenzae	1,8
Escherichia coli K12	4,6
Bacillus subtilis	4,2
Pseudomonas aeruginosa PAO1	6,3
Archaea	
Nanoarchaeum equitans	0,49
Aeropyrum pernix	1,7
Natronomonas pharaonis	2,9
Sulfolobus solfataricus P2	3
Methanosarcina mazei	4,1

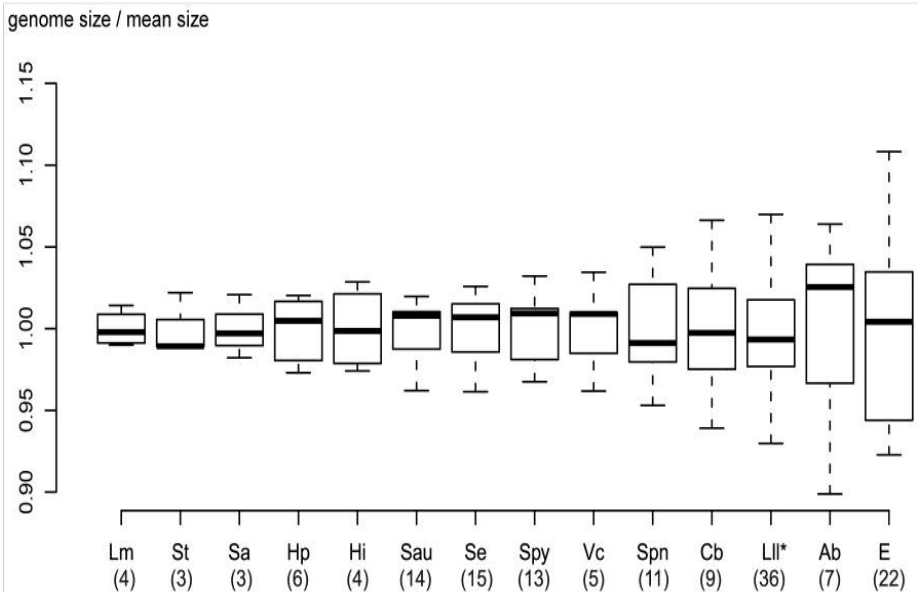
Exemple génomes eucaryotes

Species	Genome size (Mb)
Fungi	
Saccharomyces cerevisiae	12.1
Aspergillus nidulans	25.4
Protozoa	
Tetrahymena pyriformis	190
Invertebrates	
Caenorhabditis elegans	97
Drosophila melanogaster	180
Bombyx mori (silkworm)	490
Strongylocentrotus purpuratus (sea urchin)	845
Locusta migratoria (locust)	5000
Vertebrates	
Takifugu rubripes (pufferfish)	400
Homo sapiens	3200
Mus musculus (mouse)	3300
Plants	
Arabidopsis thaliana (vetch)	125
Oryza sativa (rice)	430
Zea mays (maize)	2500
Pisum sativum (pea)	4800
Triticum aestivum (wheat)	16 000
Fritillaria assyriaca (fritillary)	120 000

Taille des génomes

Cependant, au sein d'une même espèce bactérienne, la taille du génome peut varier considérablement :

Exemple parmi 42 souches d'*Escherichia coli*, on observe une variation de la taille du génome allant de 2,6 Mb à 5,7 Mb



Lm, *L. monocytogenes*; St, *S. thermophilus*; Sa, *S. agalactiae*; Hp, *H. pylori*; Hi, *H. influenzae*; Sau, *S. aureus*; Se, *S. enterica*; Spy, *S. pyogenes*; Vc, *V. cholerae*; Spn, *S. pneumoniae*; Cb, *C. botulinum*; Lll, *L. lactis* subsp. *lactis*; Ab, *A. baumannii*; Ec, *E. coli*.

Taille des génomes

Le nombre de gènes dans un génome varie moins que sa taille, cependant la corrélation avec la complexité de l'organisme n'est pas parfaite.

Species	Genome size (Mb)	Nombre approximatif de gènes
Mycoplasma genitalium (bactérie)	0,58	500
Heamophilus influenzae (bactérie)	1,8	1800
Escherichia coli K12 (bactérie)	4,6	4400
Saccharomyces cerevisaie (levure)	12	5800
Drosophila melanogaster (insecte)	180	14700
Caenorhabditis elegans (nematode)	103	20000

Species	Genome size (Mb)	Nombre approximatif de gènes
Tetrahymena thermophila (protozoaire)	125	2700
Arabidopsis thaliana (plante à fleur)	120	26500
Oryza sativa (riz)	430	~45000
Zea mays (maïs)	2200	> 45000
Mus musculus (souris)	2600	22000
Homo sapiens	3200	22000
Paramécie	72	40000

Taille des génomes

On va distinguer dans les génomes de bactéries et d'archaea :

- le core génome (génome cœur) : ensemble des gènes communs à toutes les souches d'une même espèce (génes orthologues).

- le génome accessoire : ensemble des gènes présents uniquement dans la souche étudiée ainsi que ceux présents dans deux ou plusieurs souches

- le pan-génome : ensemble des gènes du génome cœur et des génomes accessoires de toutes les souches de l'espèce.

Alignement de génomes

Utilisé pour comparer la structure et l'organisation des génomes

Problèmes :

- la recombinaison peut conduire à des réarrangements des génomes.
- les HGT introduisent de nouvelles séquences dans certains génomes bactériens.
- des délétions peuvent conduire à la perte de régions dans certains génomes.

Chaque génome est constitué d'une mosaïque de :

- segments uniques à la lignée
- segments conservés dans un sous ensemble de génomes
- segments conservés dans tous les génomes

Quand on compare plusieurs génomes, l'ordre linéaire de ces différents fragments peut être remanié d'un génome à l'autre.

Alignement de génomes

Le programme doit aligner rapidement de longues séquences génomiques

Méthode la plus souvent appliquée en première étape :

Recherche de régions alignées servant de points d'ancrage

Les génomes sont connus pour posséder des régions répétées comme les opérons ribosomiques et les prophages.
Donc quand on recherche des ancres entre plusieurs génomes, des problèmes vont apparaître si un motif répété particulier est rencontré plusieurs fois dans chaque séquence. Il devient difficile de déterminer quelle combinaison de régions il faut aligner.
En effet, si l'élément répété est trouvé r fois dans chacun des G génomes, on a r^G combinaisons possibles et seulement r vont correspondre à l'alignement des régions orthologues.

Alignement de génomes : MAUVE

(Darling *et al.* (2004), Genome Research 14, 1394-1403)

MAUVE évite ce problème en utilisant des « Multiple Maximal Unique Matches » (multi-MUMs) de longueur minimum k comme ancres, c'est-à-dire des régions qui sont trouvées identiques entre deux ou plusieurs génomes mais présentes au plus une fois par génome et qui sont bordées de chaque côté par des résidus qui ne sont plus conservés entre les différentes séquences.

Cette approche réduit la sensibilité de MAUVE dans les régions conservées répétées et dans les régions qui présentent beaucoup d'évènements de substitutions et/ou d'indels.

→ Utilisation d'une stratégie d'ancrage récursive qui progressivement réduit la taille de k pour trouver des ancres de plus petites tailles dans les régions restant non alignées.

Alignement de génomes : MAUVE

Recherche des multi-MUMs :

- pour chaque génome *g*, construction d'une liste triée de *k*-mers.
- les listes ordonnées sont ensuite parcourues pour identifier les *k*-mers qui sont trouvés dans deux ou plusieurs séquences mais qui apparaissent au plus une fois dans chacun des génomes.
- si le multi-MUM qui inclut le *k*-mer détecté n'a pas encore été découvert, alors on procède à une extension dans chaque génome jusqu'à ce les positions alignées restent identiques. L'extension s'arrête dès qu'une substitution est rencontrée.

Chaque multi-MUM est définit comme un tuple ($L, S_1, \ldots, S_j \ldots, S_G$) avec L sa longueur et S_j sa position gauche dans le génome j.
L'ensemble résultant de multi-MUMs trouvé est $\mathrm{M} = \{\mathrm{M}_1, \ldots, \mathrm{M}_i, \ldots, \mathrm{M}_N\}$
La $i^{ème}$ multi-MUM est appelé M_i. Sa longueur est notée $M_i \,.\, L$ et sa position dans le génome j $M_i \,.\, S_j$.
Si le multi-MUM est trouvé dans le génome j dans une région en orientation inverse, le signe de $M_i \,.\, S_j$ est négatif et si le multi-MUM n'existe pas dans le génome j, $M_i \,.\, S_j$ vaut 0.

Alignement de génomes : MAUVE

Sélection des ensembles d'ancres :

Parmi les multi-MUMs sélectionnés, certains peuvent être des contaminants, c'est-à-dire des segments dont la similarité est due au hasard.
Filtre de ces contaminants et construction des blocs localement colinéaires (LCB).

Définition d'un LCB :

Un LCB est une séquence de multi-MUMs $lcb \subseteq \mathrm{M}$, $\mathrm{lcb} = \{\mathrm{M}_1, \mathrm{M}_2, \ldots, \mathrm{M}_{\mathrm{lcb}}\}$ qui satisfait une propriété d'ordre total tel que $M_i \,.\, S_j \leq M_{i+1} \,.\, S_j$ pour tout i, $1 \leq i \leq |lcb|$ et pour tout j, $1 \leq j \leq G$, (G nombre total de génomes comparés).

Recherche des LCB : partitionnement de M en sous ensembles colinéaires -> implémentation d'un algorithme de recherche des points de cassure (breakpoints).

Alignement de génomes : MAUVE

Recherche des LCB :

1. MAUVE ordonne les multi-MUMs de M sur le génome de référence $|M_i \,.\, S_0|$
2. un label est augmenté de façon monotone de 1 à |M| et assigné à chaque MUM correspondant à son index sur $|M_i \,.\, S_0|$. Le label du ith multi-MUM est donné par $M_i.label$.
3. Le groupe de multi-MUMs est réordonné de façon répétée suivant les positions dans chaque génome autre que celui de référence (donc par rapport au $|M_i \,.\, S_j|$ pour $j = 2$ à G.
4. après chaque ré-ordonnancement, recherche des points de cassure :

   Entre M_i et M_{i+1} si :

   $M_i.label + 1 \neq M_{i+1}.label$ (M_i et M_{i+1} brin direct)

   $M_i.label - 1 \neq M_{i+1}.label$ (M_i et M_{i+1} brin complémentaire)

   signe $(M_i \,.\, S_j) \neq M_{i+1} \,.S_j$ (M_i et M_{i+1} pas sur le même brin)

5. Les LCBs sont les sous-séquences de longueurs maximales de multi-MUMs $M_i \ldots M_{i+1}$ qui ne possèdent pas de points de cassure identifiés entre eux.

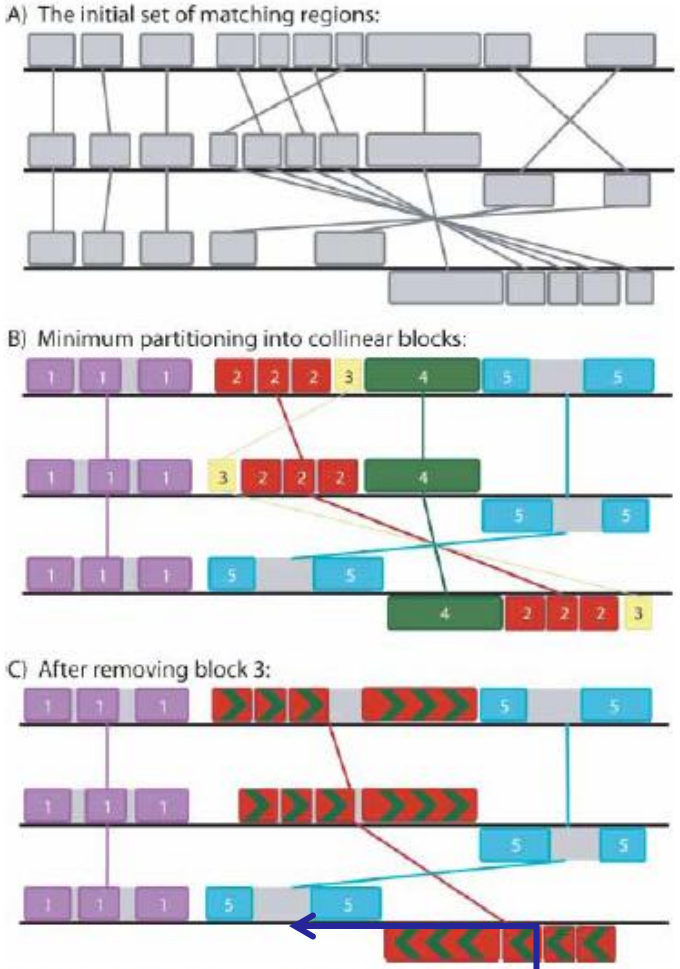


Figure 1 A pictorial representation of greedy breakpoint elimination in three genomes. (*A*) The algorithm begins with the initial set of matching regions (multi-MUMs) represented as connected blocks. Blocks below a genome's center line are inverted relative to the reference sequence. (*B*) The matches are partitioned into a minimum set of collinear blocks. Each sequence of identically colored blocks represents a collinear set of matching regions. One connecting line is drawn per collinear block. Block 3 (yellow) has a low weight relative to other collinear blocks. (*C*) As low-weight collinear blocks are removed, adjacent collinear blocks coalesce into a single block, potentially eliminating one or more breakpoints. Gray regions within collinear blocks are targeted by recursive anchoring.

Multi-MUMs

LCBs

LCB 3 éliminé
Retour en 1, nouveaux LCBs déterminés

Région étiquetée par l'algorithme pour recherche de nouvelles ancres.

Algorithme glouton d'élimination des points de cassure dus à la présence de LCB court pouvant résulter de la détection de multi-MUMs contaminants. Nécessite un critère de poids minimum $MinimumWeight \geq 0$
A partir de l'étape 3, répétition des étapes jusqu'à ce que tous les blocs colinéaires de M satisfassent la condition du *MinimumWeight*

1. Déterminer le partitionnement de **M** en blocs colinéaires **CB**
2. Calculer le poids $w(cb_i)$ de chaque bloc colinéaire $cb_i \in \mathbf{CB}$. Avec :
$$w(cb_i) = \sum_{M_i \in CB} M_i.L$$
3. Soit $z = min_{cb \in \mathbf{CB}} w(cb)$
4. Stop si $z \geq MinimumWeight$
5. identifier les sous ensembles $\mathbf{MinCB} \subseteq \mathbf{CB}$ qui satisfont $w(cb_i) = z$
6. Pour chaque $cb \in \mathbf{MinCB}$ enlever chaque multi-MUM $M \in cb$ de **M**
7. Retourner en 1.

Défaut : $MinimumWeight = 3k$

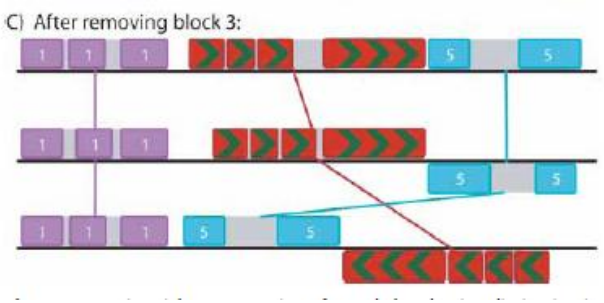
Extrait de Genome Research (2004), 14:1394-1403

Alignement de génomes : MAUVE

Recherche récursive d'ancres:

Recherche initiale d'ancres peut na pas être assez sensible pour détecter des ancres dans les régions répétées et les régions ayant un taux de substitution plus élevé.

➡ Utilisation des ancrages précédents comme guide : deux recherches récursives :



• régions à l'extérieur des LCBs sont examinées pour étendre les frontières des LCBs existants (régions blanches sur figure).
• régions sans ancre à l'intérieur des LCBs sont examinées pour rechercher de nouveaux multi-MUMs (région grise sur figure).

régions à l'extérieur des LCBs : taille de k plus petite. Un k-mer non unique quand on regarde tout le génome peut devenir unique quand on regarde les régions entre LCBs. On peut soit étendre les LCBs existants soit trouver de nouveaux LCBs.
régions à l'intérieur des LCBs : permet de trouver de nouvelles ancres dans des régions moins bien conservées ou répétées (car pas répétées à l'intérieur du LCB).

Alignement de génomes : MAUVE

Alignement des séquences entre les ancres:

Utilisation d'un programme « classique » d'alignement multiple progressif comme ClustalW. Cependant pour des séquences de longueur supérieure à 10 Kb, le temps de calcul devient prohibitif, d'où la nécessité de trouver le maximum d'ancres.

Ces méthodes d'alignement progressif utilisent un arbre guide pour ajouter de façon itérative les séquences dans l'alignement en construction.

MAUVE utilise les multi-MUMs détectés avant la sélection des ancres pour construire cet arbre guide car peut utiliser des multi-MUM pas présents dans tous les génomes analysés.
Le rapport du nombre de bases communes entre deux génomes sur la moyenne de leur longueur fournit une estimation de la similarité de séquence. Cette estimation de la similarité est convertie en distance pour construire la matrice des distances utilisée ensuite par la Neighbor Joining Method ($d = 1-s$).
Si des multi-MUMs se chevauchent, le chevauchement est résolu car chaque résidu aligné ne doit compter qu'une fois.

ClustalW est exécuté pour chaque paires d'ancres adjacentes de chaque LCB. Les répétitions en tandem < 10 Kb sont alignées pendant cette phase. Les régions ≥ 10 Kb sans une ancre sont ignorées.

Alignement de génomes : MAUVE

La procédure d'alignement de MAUVE fournit un alignement global de chaque bloc localement colinéaire dont les éléments de séquences sont conservés entre tous les génomes étudiés. Les nucléotides alignés sont considérés comme orthologues.
MAUVE n'essaie pas d'aligner les régions paralogues.
Les régions restant non alignées peuvent donc être :

- des séquences lignée spécifique
- des régions réarrangées
- des régions paralogues

Alignement de génomes : MAUVE

Résultats de MAUVE : tester à partir de 9 génomes d'entérobactéries qui ont été modifiés par un simulateur d'évolution de génomes :

Table 1. The Published Genome Sequences of These Nine Enterobacteria Are a Target for the Alignment System Presented Here

Species	Genome size	Reference
E. coli K12 MG1655	4,639,221	Blattner et al. 1997
E. coli O157:H7 EDL933	5,524,971	Perna et al. 2001
E. coli O157:H7 VT-2 Sakai	5,498,450	Hayashi et al. 2001
E. coli CFT073	5,231,428	Welch et al. 2002
S. flexneri 2A 2457T	4,599,354	Wei et al. 2003
S. flexneri 2A	4,607,203	Jin et al. 2002
S. enterica Typhimurium LT2	4,857,432	McClelland et al. 2001
S. enterica Typhi CT18	4,809,037	Parkhill et al. 2001
S. enterica Typhi Ty2	4,791,961	Deng et al. 2003

Numerous large-scale evolutionary events such as horizontal transfer and rearrangement are scattered throughout their genomes.

Extrait de Genome Research (2004), 14:1394-1403

Plusieurs expériences et comparaison des résultats avec Multi-LAGAN pour l'alignement de séquences colinéaires ayant subies un taux croissant de substitutions et indels.
Comparaison avec shuffle-LAGAN pour l'alignement de séquences ayant subi un taux croissant d'inversion et de substitutions de nucléotides

Alignement de génomes : MAUVE

Multi-LAGAN qui implémente une approche différente de recherche d'ancres peut aligner des génomes bien plus divergent que MAUVE et est donc mieux adapté à la comparaison inter-espèce.
Shuffle-LAGAN fournit de meilleurs alignements de séquences quand le taux d'inversion est faible et le taux de substitution élevé (77,8% de précision avec ~500 000 substitutions et 400 inversions entre deux génomes) (problème lié à l'identification d'ancre dans le cas de présence de séquences répétées)

MAUVE donne de très bons résultats quand le taux de substitution ou d'inversion est faible, atteignant des score de plus de 90% quand l'un ou l'autre de ces taux est faible. Bonne performance quand les séquences alignées sont réarrangées mais il faut qu'elles présentent un faible taux de substitution. Pour les taux de substitutions et d'indel reportés dans le jeu de données des entérobactéries, MAUVE aligne les génomes simulés avec un très haut degré de précision.

Alignement des 9 génomes d'entérobactéries :

Inversion reportée dans le génome de *E. coli* O157:H7 EDL 933 identifiée

Les génomes de *Shigella* et *Salmonella* sont plus remaniés que ceux *de E. coli*, notamment une grande inversion.

Figure 6 Locally collinear blocks identified among the nine enterobacterial genomes listed in Table 1. Each contiguously colored region is a locally collinear block, a region without rearrangement of homologous backbone sequence. LCBs below a genome's center line are in the reverse complement orientation relative to the reference genome. Lines between genomes trace each orthologous LCB through every genome. Large gray regions within an LCB signify the presence of lineage-specific sequence at that site. Each of the 45 blocks has a minimum weight of 69. The *Shigella* and *Salmonella* genomes have undergone more genome rearrangements than the *E. coli*, possibly because of the presence of specific mobile genetic elements. The computation consumed ~3 h on a 2.4-GHz workstation with 1 GB of memory. The figure was generated by the Mauve rearrangement viewer.

Extrait de Genome Research (2004), 14:1394-1403

Exemple de l'alignement de 3 génomes (alignment viewer) extrait du guide utilisateur de Mauve

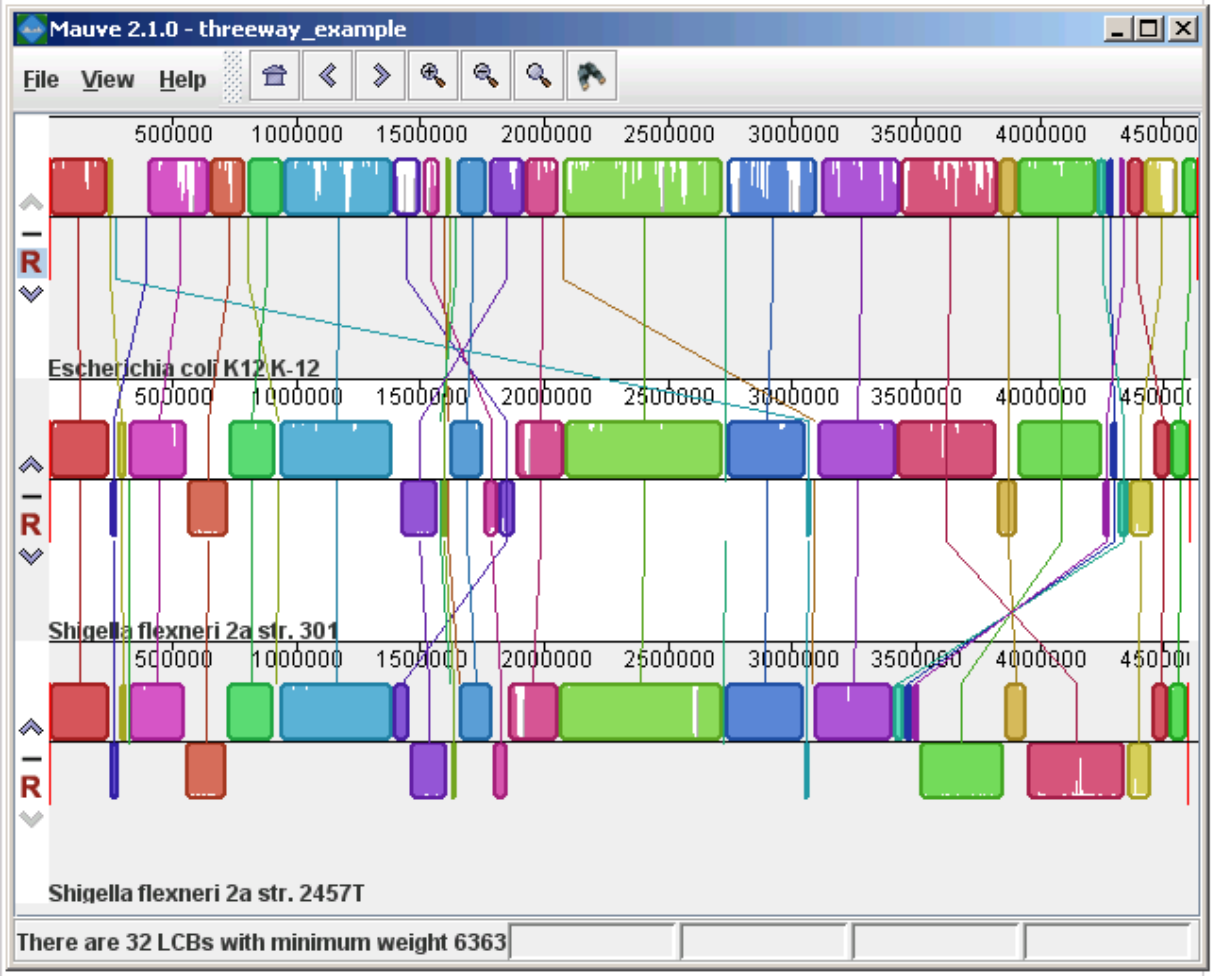


Figure 1 (above) shows an alignment of *E. coli* K12 MG1655, *S. flexneri* 2a 301, and *S. flexneri* 2457T. Notice how inverted regions in the *S. flexneri* are clearly depicted as blocks below a genome's center line. These three genomes were taken from the NCBI FTP site and aligned with Progressive Mauve using default parameters, as described in the previous section.

In Figure 1, colored blocks in the first genome are connected by lines to similarly colored blocks in the second and third genomes. These lines indicate which regions in each genome are homologous. Notice the crossing "X" pattern of lines, which happen to occur in the vicinity of the predicted origin and terminus of replication in these organisms. When viewing genomes with complex rearrangements, the LCB connecting lines can be confusing and they can be hidden (or made visible again) by typing **Shift+L** (pressing shift and L simultaneously) or using the "View" menu.

In the standard color scheme, the region of sequence covered by a colored block is entirely collinear and homologous among the genomes. The boundaries of colored blocks usually indicate the breakpoints of genome rearrangement, unless sequence has been gained or lost in the breakpoint region.

Alignement de génomes : MAUVE

Evolution de la méthode pour pallier les faiblesses de MAUVE : progressiveMAUVE (Darling *et al.* (2010) Plos One volume 5 Issue 6 e11147.

- peut correctement aligner des régions communes à certains mais pas à tous les génomes
- notion de multi-Mums remplacée par des Local Multiple Alignements (LMA) qui permet d'inclure des alignements avec des substitutions.