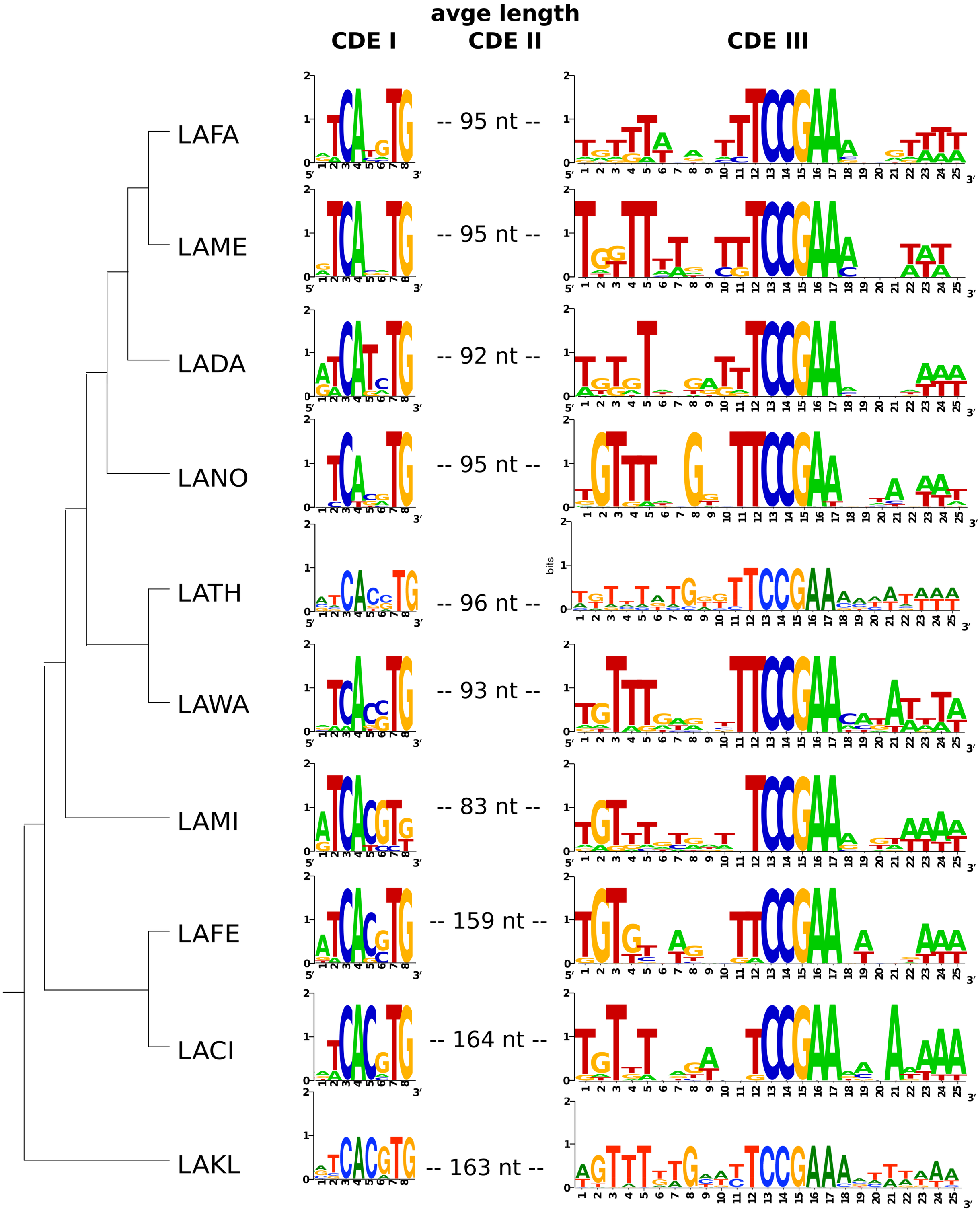
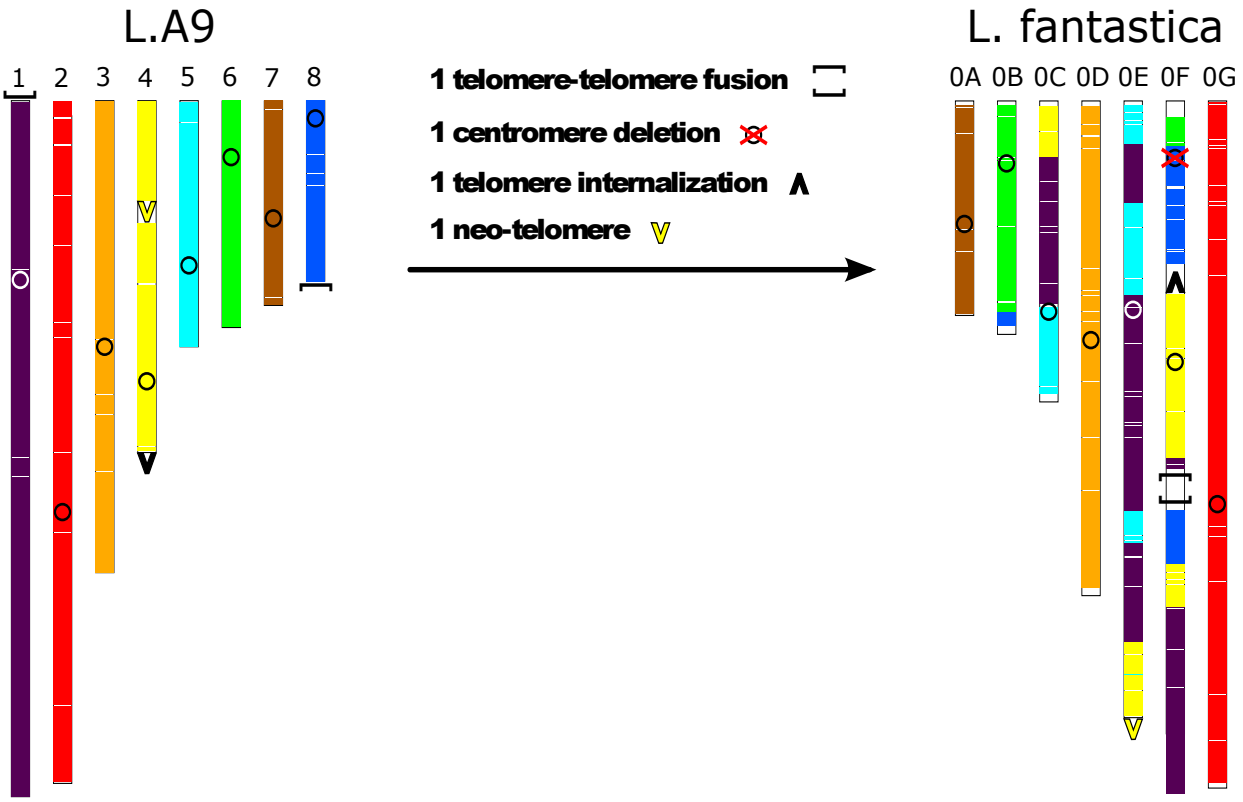


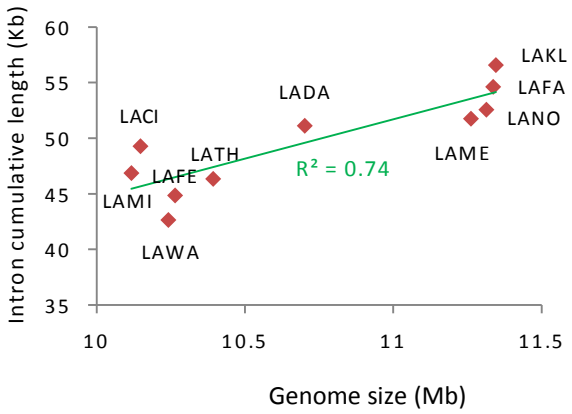
Supplemental Figure S1

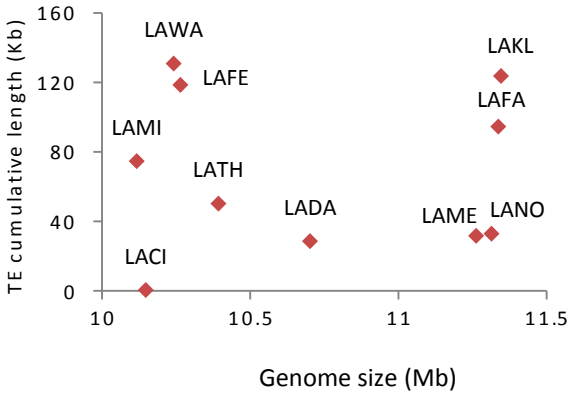
Supplemental Figure S1: Centromere organization in *Lachancea.* Sequence comparison of the CDEI and CDE III elements in the different species. Each logo represents the consensus sequence deduced from the alignment of all the CDEI and CDEIII sequences. The overall height of each stack position indicates the sequence conservation at that position (measured in bits), whereas the height of letters within the stack reflects the relative frequency of the corresponding nucleotide at that position. The logos for *LATH* and *LAKL* come from Souciet *et al.* (2009). In *S. cerevisiae* the AT-rich CDEII element is about 80 bp, while it is about 160 bp in *L. cidri*, *L. fermentati* and *L. kluyveri*, and 95 bp in the seven other species. Abbreviations: LACI=*L. cidri,* LAFE=*L. fermentati,* LAME=*L. meyersii,* LADA=*L. dasiensis,* LAMI=*L. mirantina*, LANO=*L. nothofagi*, LAFA=*L. fantastica,* LATH=*L. thermotolerans*, LAKL=*L. kluyveri* and LAWA=*L. waltii*.

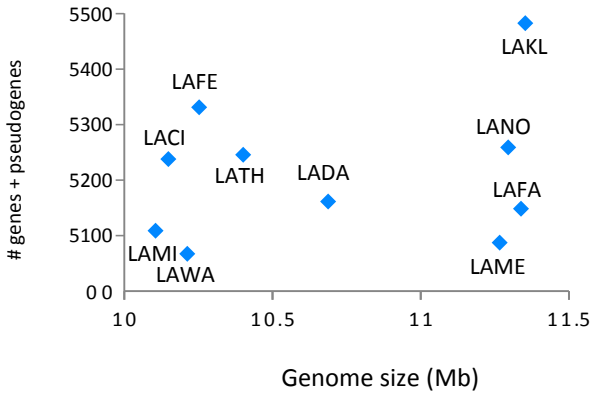


Supplemental Figure S2

Supplemental Figure S2: Chromosome fusion in *L. fantastica.* Chromosomes are colored relatively to the *L.A9* genome. Centromeres are represented as circles. The telomere-telomere fusion between chromosomes 1 and 8 was accompanied by the deletion of the centromere of chromosome 8 from *L.A9,* the telomere Internalization of one end of chromosome 4 from *L.A9* and 1 neo-telomere formation in chromosome 0E of *L. fantastica*.



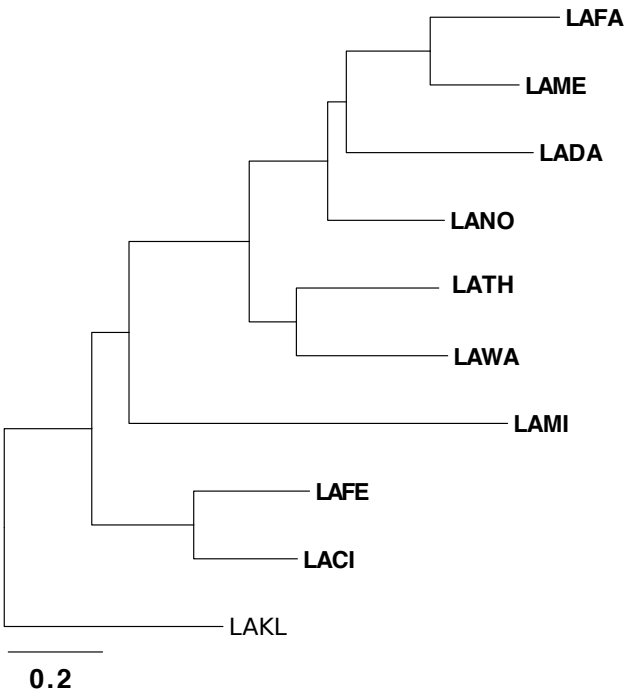




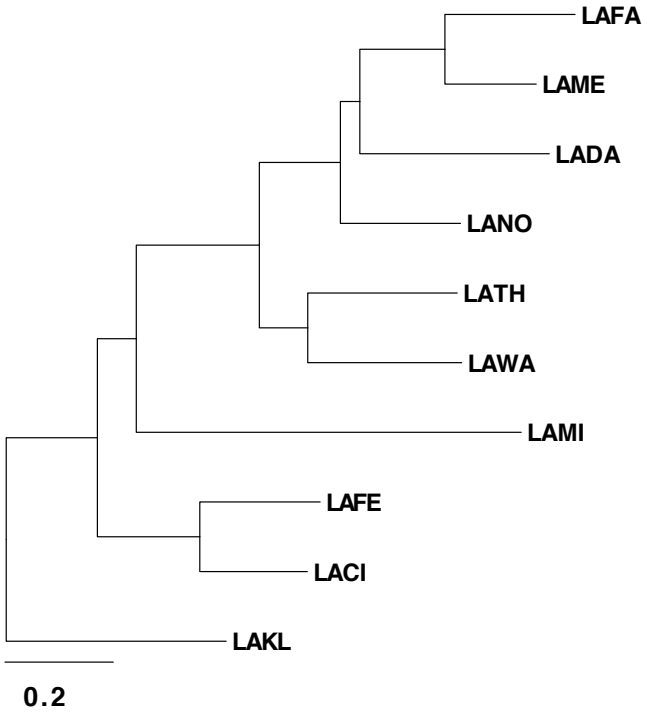
Supplemental Figure S3

Supplemental Figure S3: Genome size in *Lachancea.* Genome size in *Lachancea* positively correlates with cumulated intron length (top) but not with cumulative TE length (middle) or with the number of genes and pseudogenes (bottom). Abbreviations: LACI=*L. cidri,* LAFE=*L. fermentati,* LAME=*L. meyersii,* LADA=*L. dasiensis,* LAMI=*L. mirantina*, LANO=*L. nothofagi*, LAFA=*L. fantastica,* LATH=*L. thermotolerans*, LAKL=*L. kluyveri* and LAWA=*L. waltii*.

472 genes sharing the prevalent topology

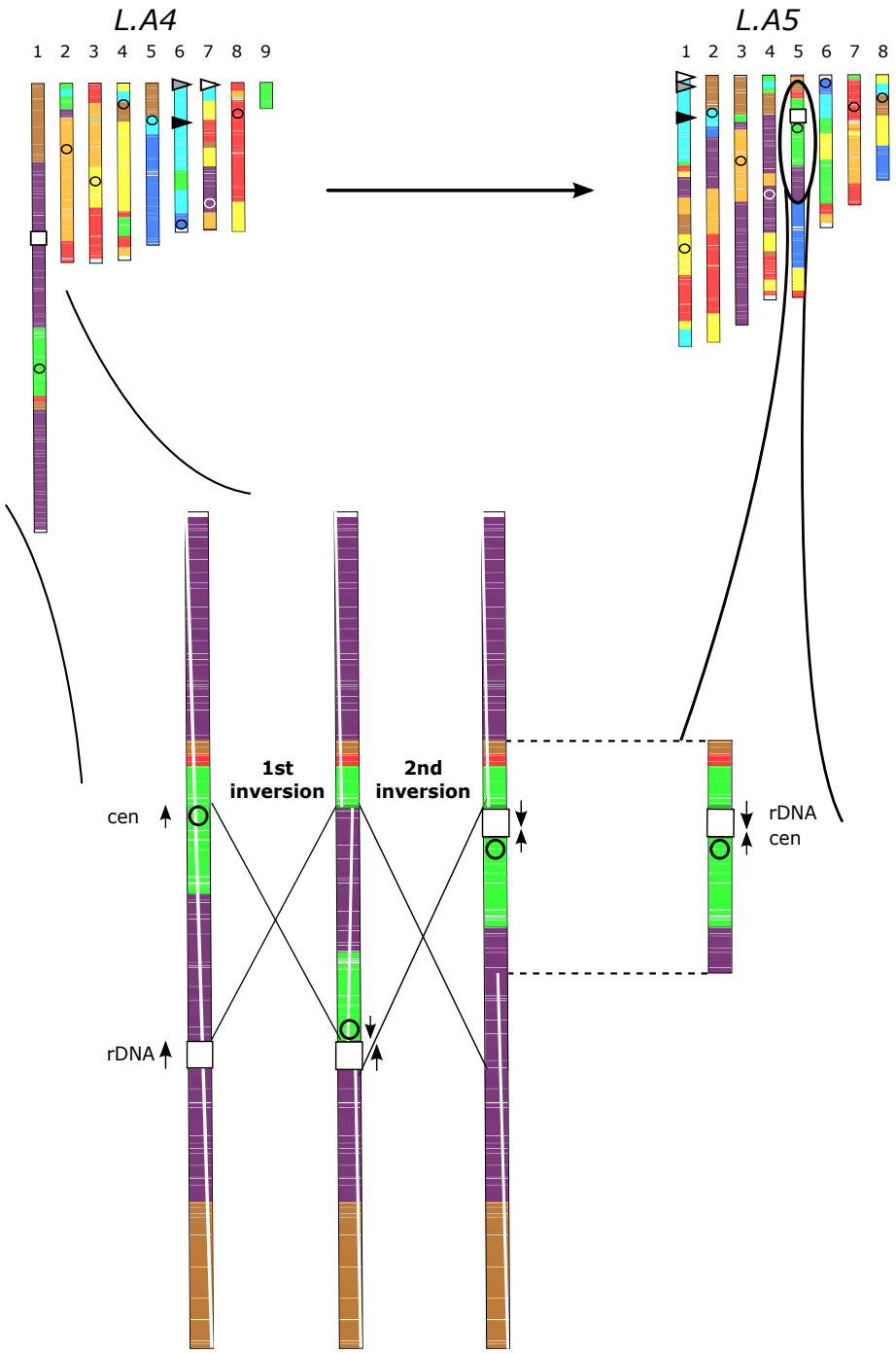


3,598 orthologous genes



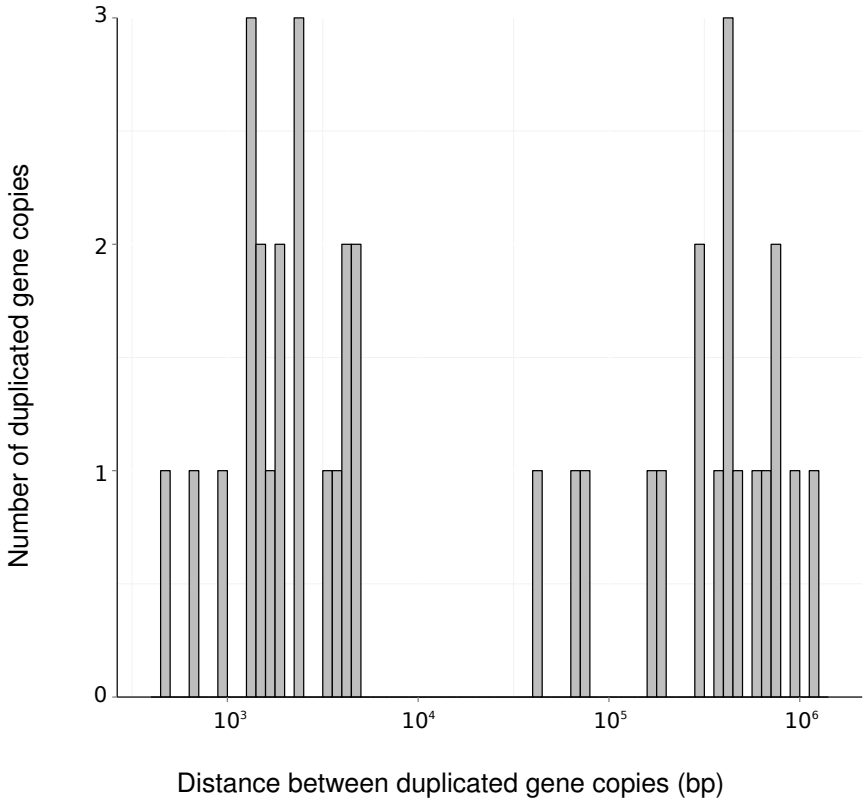
Supplemental Figure S4

Supplemental Figure S4: *Lachancea* species tree. PhyML reconstructions were performed from the concatenation of the multiple alignments of either the 472 orthologous groups whose individual trees have the eMRC topology (387,091 aligned positions) or all 3,598 orthologous genes (1,983,702 aligned positions) using the LG model and a gamma-law distribution with 4 categories of evolution rates. In all cases, 500 bootstrap replicates were performed and resulted in a 100% support at all nodes. Abbreviations: LACI=*L. cidri,* LAFE=*L. fermentati,* LAME=*L. meyersii,* LADA=*L. dasiensis,* LAMI=*L. mirantina*, LANO=*L. nothofagi*, LAFA=*L. fantastica,* LATH=*L. thermotolerans*, LAKL=*L. kluyveri* and LAWA=*L. waltii*.



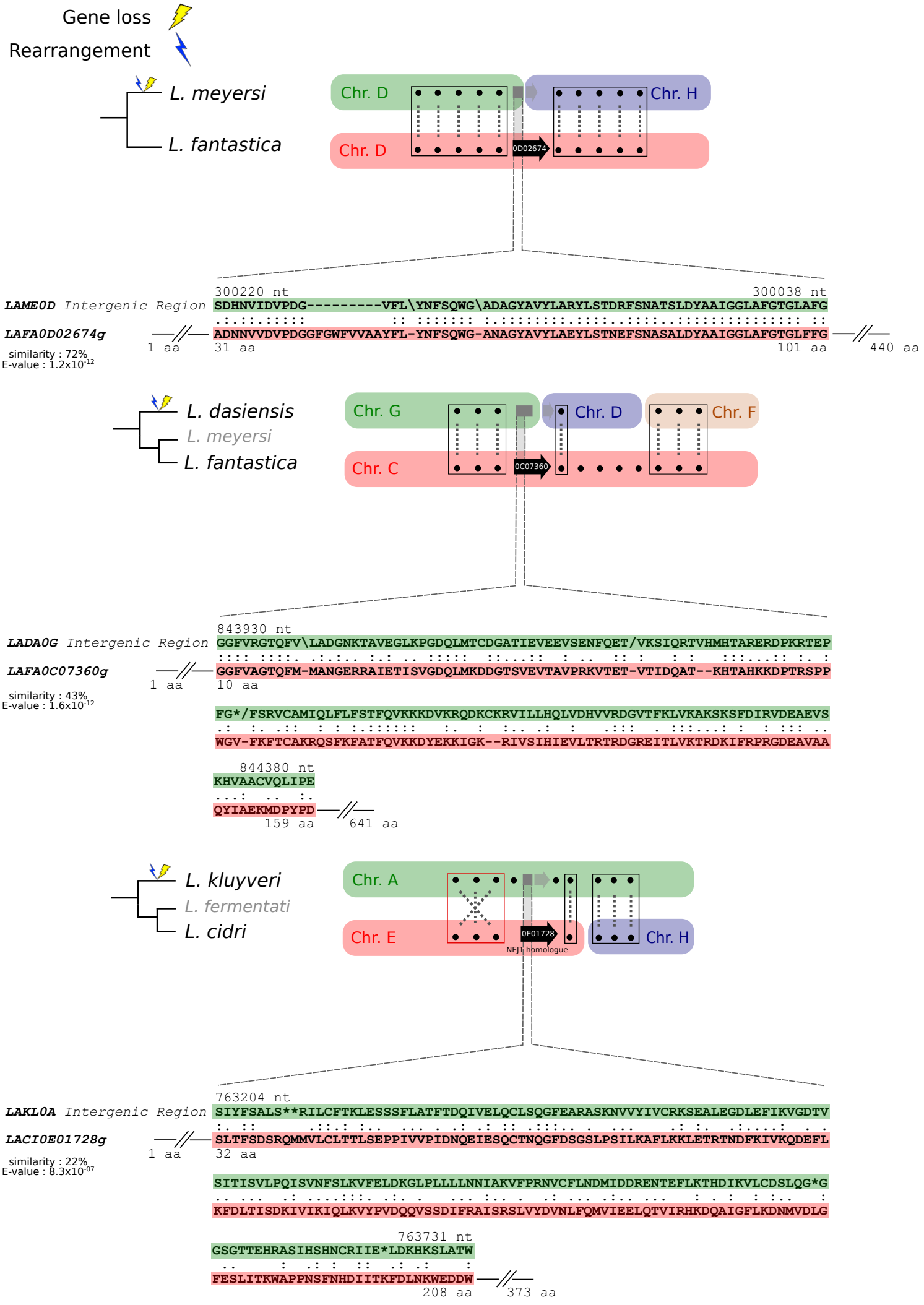
Supplemental Figure S5

Supplemental Figure S5: Relocation of the rDNA locus in the *L.A5* ancestor. Chromosomes are colored relatively to the *L.A1* genome, as in Figure 2. Centromeres are represented as circles and the rDNA locus is symbolized by the white rectangle. The bottom of the figure is a zoom in the structure of chromosome 1 of *L.A4* and chromosome 5 in *L.A5.* The black arrows indicate the transcriptional orientation of the 35S gene and the relative order of the centromeric elements CDEI to CDEIII. The relocation of the rDNA locus next to the centromere was possibly due to two inversions involving one breakpoint reuse.



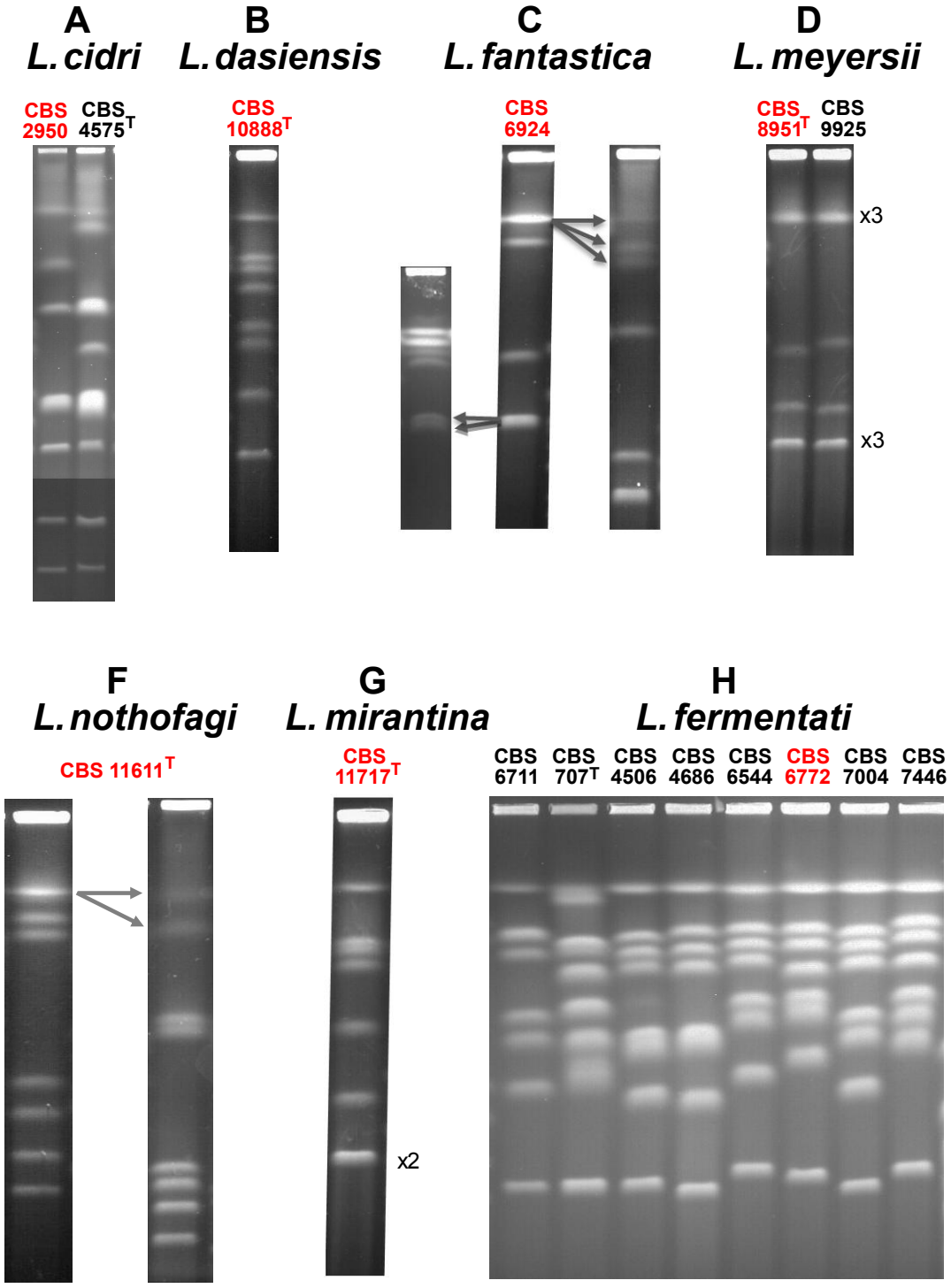
Supplemental Figure S6

Supplemental Figure S6: Distance separating intra-chromosomally duplicated gene copies. There are 20 duplication events separated by less than 3 kb and 18 cases separated by more than 30 kb. No intermediate distance was found out of the total 38 intra-chromosomal duplications.



Supplemental Figure S7

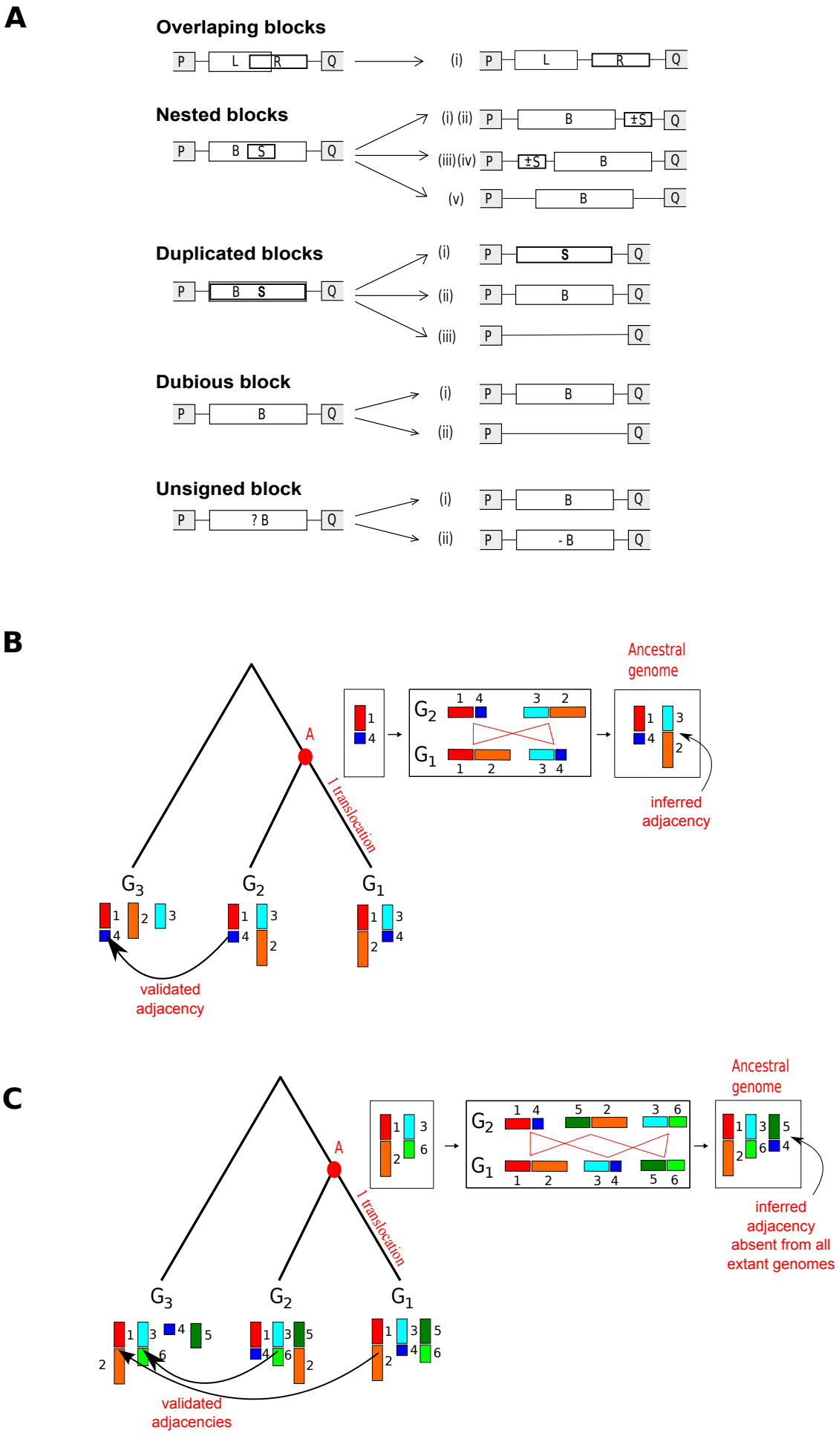
Supplemental Figure S7: Disrupted gene relics at rearrangement breakpoints. The colored rectangles represent the synteny blocks between two species. (Top) One breakpoint intergenic region in *L. meyersii* shows clear traces of homolgy with an intact gene in *L. fantastica*. (Middle) the same situation is found between an intergene in *L. dasiensis* and a gene in *L. fantastica*. Bottom : The *S. cerevisiae* NEJ1 homologue is missing in *L. kluyveri* and its putative intergenic loss position is inferred one gene downstream of an inversion (not immediately adjacent as in the two above cases). Similarity was found between the NEJ1 homolog in *L. cidri* and the intergenic region in *L. kluyveri*.



Supplemental Figure S8

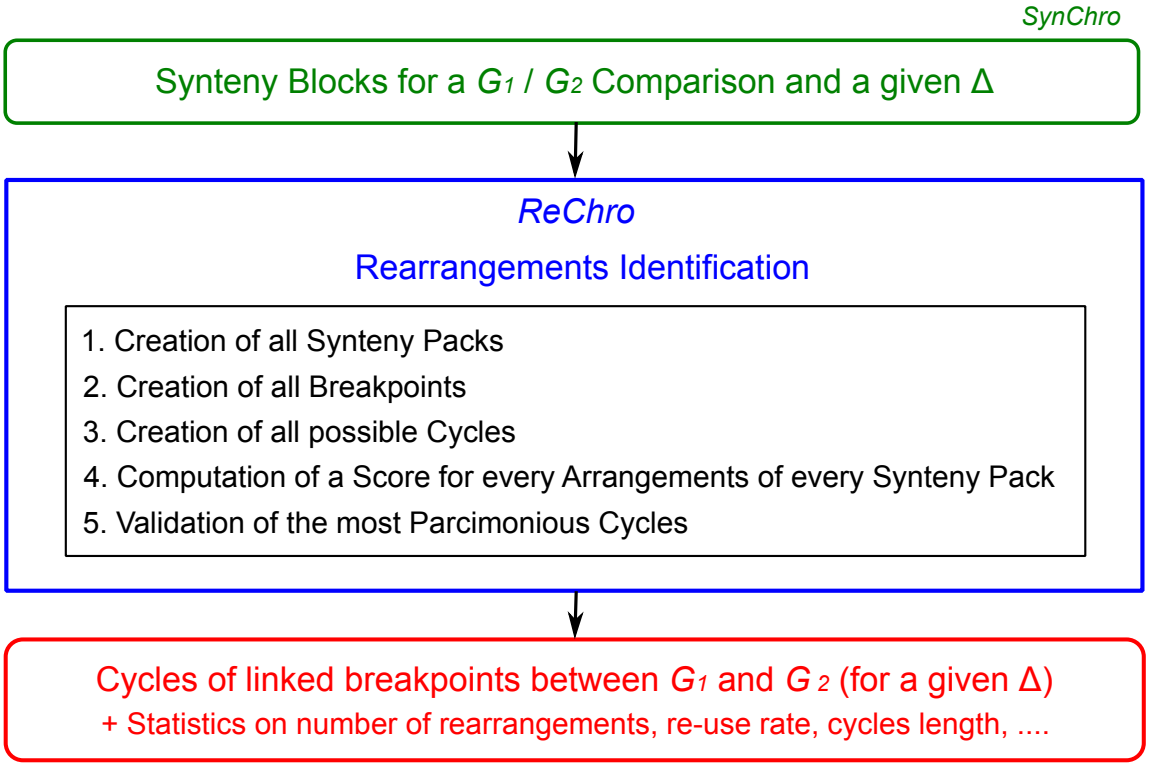
Supplemental Figure S8: Electrpohoretic karyotyping of *Lachancea* strains. The strains chosen for genome sequencing are in red. (A) Karyotypes were established for 4 strains known as *L. cidri*. Two of them showed highly divergent karyotypes (data not shown). PCR amplification and sequencing of their D1D2 region revealed CBS2951 and CBS5666 to be *C. sorbosivorans* and *K. marxianus*, respectively. Karyotyping of CBS 4575T showed that this strain is either aneuploid or diploid with translocations between chromosome pairs (programs 2 and 3, see below). Thus, strain CBS 2950 with 8 chromosomes was preferred for sequencing. (B) *L. dasiensis* strain CBS 10888T has clearly 8 chromosomes (program 1, see below). (C) For *L. fantastica* strain CBS 6924, three programs were required to separate the 3 largest and the 2 smallest chromosomes (programs 1, 2 and 3, see below). Finally, only 7 chromosomes were observed. (D) *L. meyersii* strains showed 4 bands due to the comigration of 3 chromosomes of about 800 kb and 3 chromosomes larger than 2 Mb (program 1, see below). (E) The karyotype of *L. nothofagi* CBS 11611T showed 8 bands corresponding to 8 chromosomes (programs 1 and 2, see below). (F) The chromosome band at ~800 kb corresponds to a doublet that could not be separated in *L. mirantina* CBS 11717T (program 1, see below). (G) Karyotypes were established for 8 strains of *L. fermentati* (program 4, see below). The type strain CBS 707T, which is diploid, showed a complex profile which denotes probable size differences between homologous chromosomes. The haploid strain CBS 6772 was preferred for genome sequencing. Pulsed field gel electrophoresis was carried out on a CHEF-DRII apparatus (Bio-Rad) at 13°C in a 1% agarose gel in TBE 0.5X with the following programs: Program 1: 360 s pulse time, 100 V (24h), ramp between initial pulse time 70 s and final pulse time 90 s, 200 V (21 h); Program 2: 420 s pulse time, 100 V (18h 12 min), 300 s pulse time 130 V (24h), 90 s pulse time, 200 V (12 h); Program 3:

ramp between initial pulse time 40 s and final pulse time 100s, 200 V (28 h); Program 4: 360 s pulse time, 100 V (23h), ramp between initial pulse time 70 s and final pulse time 90 s, 200 V (20 h)



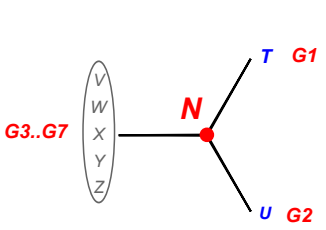
Supplemental Figure S9

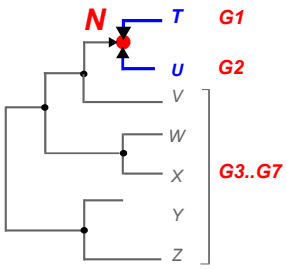
Supplemental Figure S9: (A) Schematic representation of problematic adjacencies resolved by ReChro as 'synteny packs'. For overlapping blocks, a single combination is tested (i). For nested blocks, the two possible orientations of the included block "S" are tested on both sides of the including block "B" (i to iv) as well as the deletion of the included block "S" (v). For duplicated blocks where a single block in one genome corresponds to 2 different blocks in the other genome, both combinations of one block (i and ii) as well as the deletion of the duplicated block (iii) are tested. A block is considered as dubious in one genome when its counterpart in the other genome is included in another block or when the 2 blocks are telomeric in the 2 genomes. In this case the two combinations with (i) and without (ii) the dubious block are tested. A block is considered as unsigned when the local order and orientation of the anchors are different in the 2 genomes. In this case, the two possible orientations of the unsigned block are tested (i and ii). (B) Simplified principle of ancestral genome reconstruction at node *A* from a *G1/G2* comparison and a genome *G3* used as reference. The $(1,4)_{G2;G1}$ block adjacency in *G2* is conserved in the reference genome. The other adjacency $(3,2)_{G2;G1}$ can be inferred as being ancestral because the two blocks 3 and 2 are linked in the same cycle as the blocks 1 and 4. (C) The adjacencies $(1,2)_{G1;G2}$ and $(3,6)_{G2;G1}$ are conserved in G3 and therefore validated as ancestral but the (5,4) adjacency can also be inferred as being ancestral, even if it is absent from all extant genomes, because the two blocks are linked in a single cycle comprising the 2 other adjacencies.

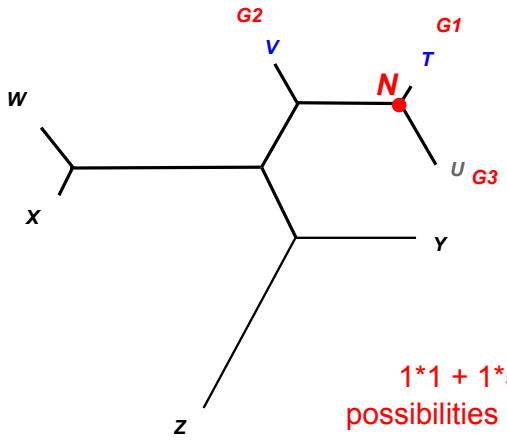


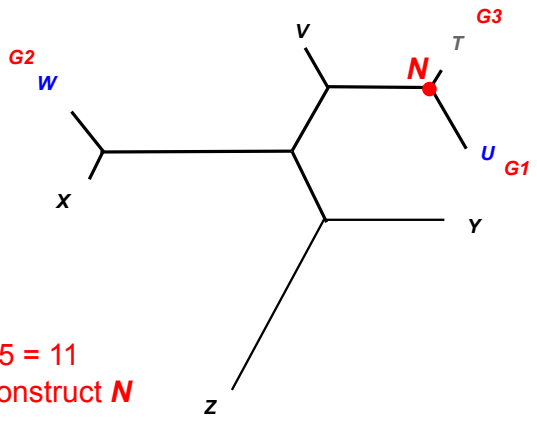
Supplemental Figure S10

Supplemental Figure S10: Flowchart of the *ReChro* program. Input files are represented in green, the program is in blue and the outputs are in red.





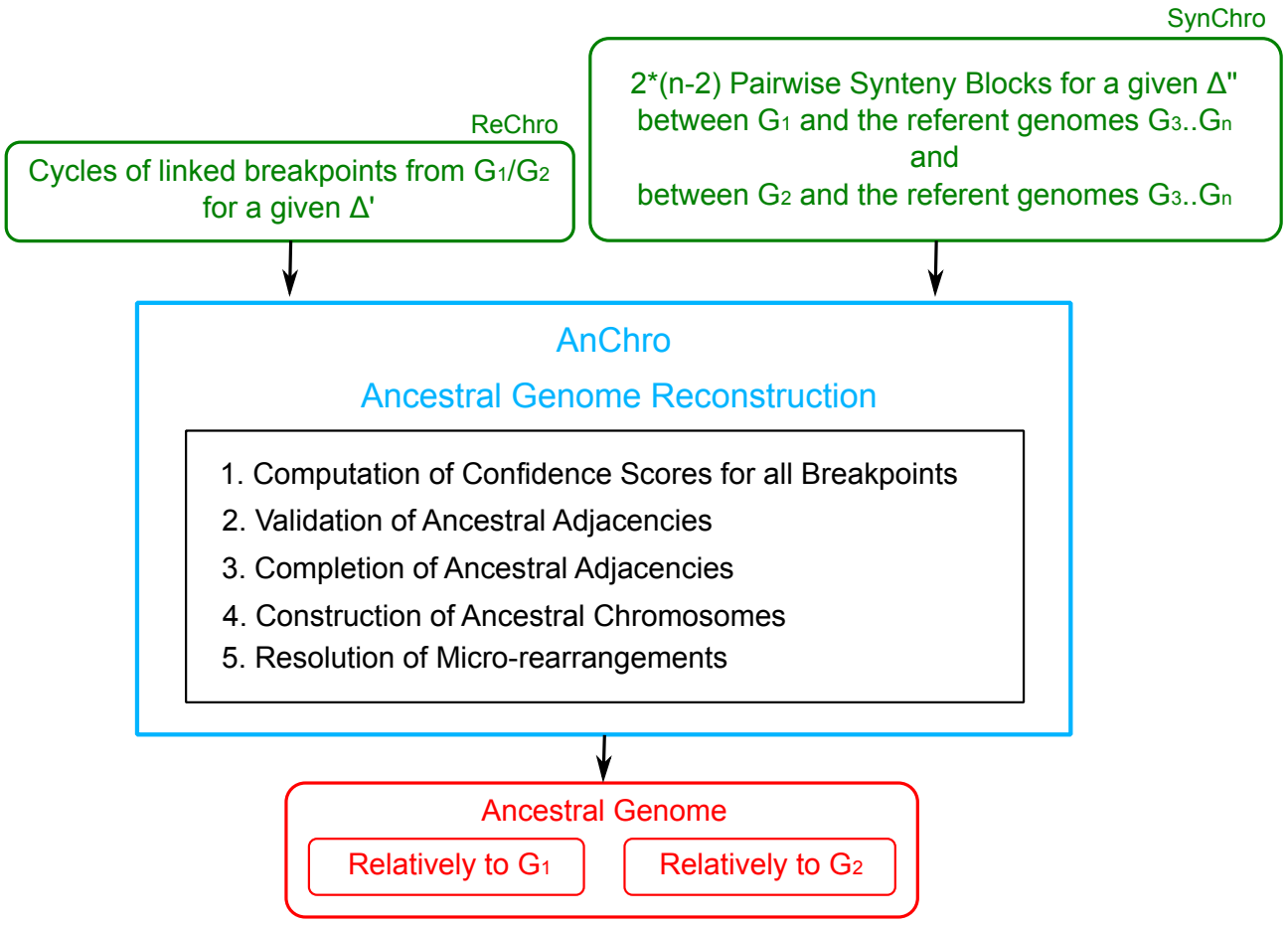




1*1 + 1*5 + 1*5 = 11
possibilities to reconstruct ***N***

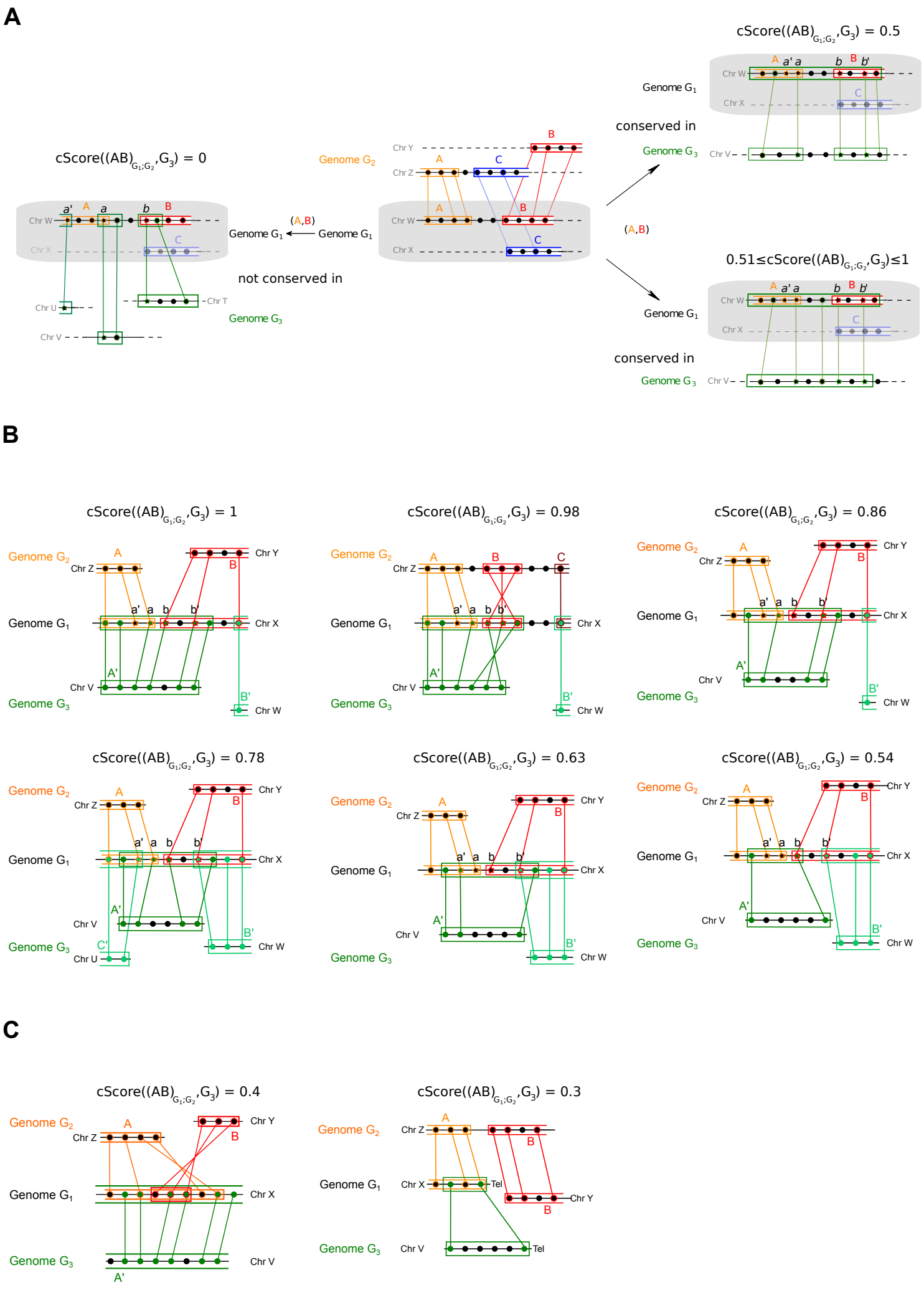
Supplemental Figure S11

Supplemental Figure S11: The choice of the *G1*, *G2* and (*G3..Gn*) genomes to reconstruct the ancestor at node *N.* Top: *G1* and *G2* are chosen on two different branches of the phylogenetic tree such that their paths to join node *N* don't cross. All other genomes which paths to join node *N* do not cross the *N*-*G1* and the *N*-*G2* path are chosen as reference genomes *(G3..Gn).* Bottom: Two different possibilities to choose *G1, G2* and the reference genome(s) to reconstruct the ancestor at node *N*. The total number of all possible reconstructions of the ancestor at node *N* depending on the choice of *G1* and *G2* are calculated.



Supplemental Figure S12

Supplemental Figure S12: Flowchart of the *AnChro* program. Input files are represented in green, the program is in blue and the outputs are in red.



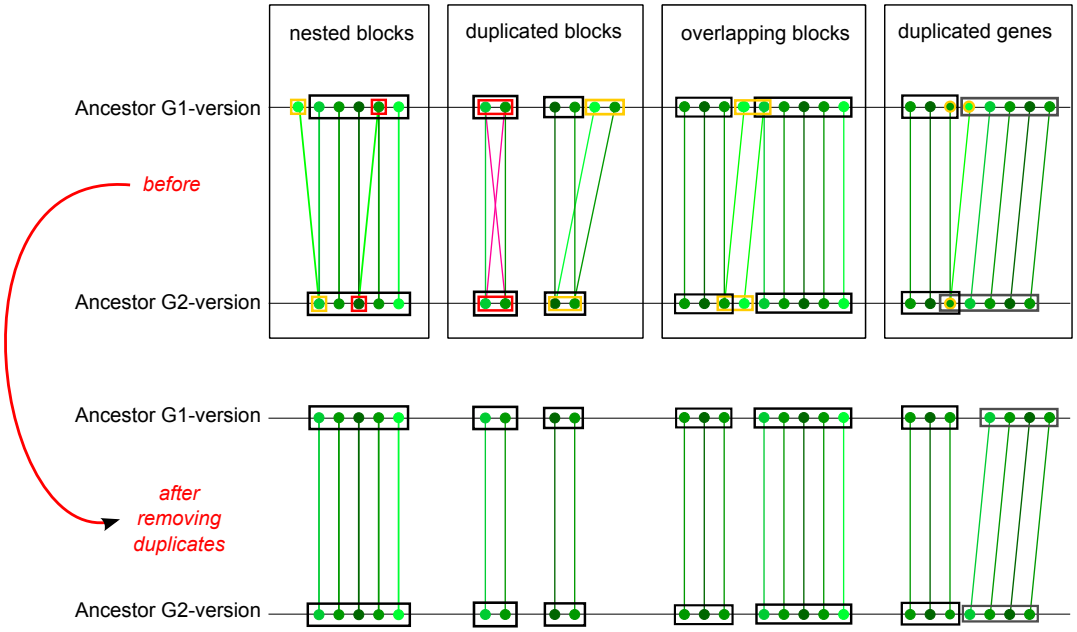
Supplemental Figure S13

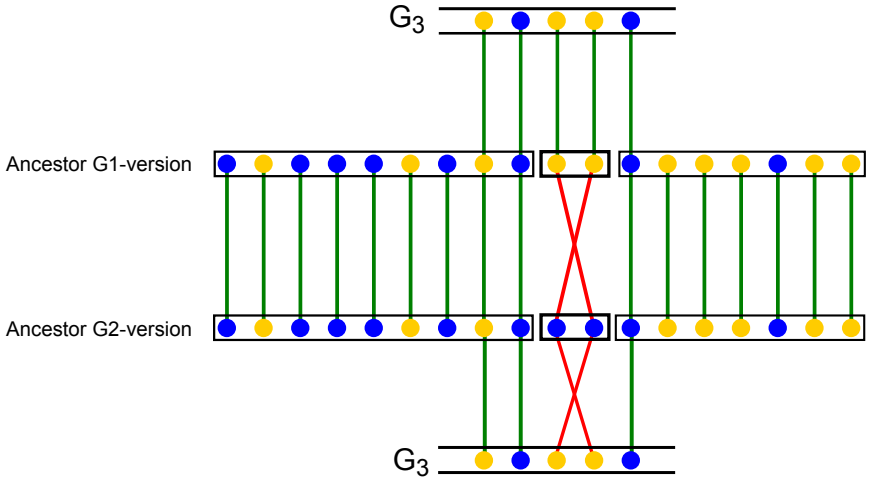
Supplemental Figure S13: Conservation of block adjacency in reference genome. Genes are represented by filled black circles, homology relationships are represented by straight lines and synteny blocks are represented by rectangles. The $cScore\left((AB)_{G_1;G_2}, G_3\right)$ determines the level of confidence (between 0 and 1) of having the $(AB)_{G_1;G_2}$ adjacency conserved in *G₃*. This score depends on the existence and the position of 4 anchor genes called *a, a′, b* and *b′* that are defined as follows: *a* and *a′* are the 2 rightmost anchors of block *A* in *G₁* which also belong to a synteny blocks in the *G₁G₃* comparison; *b* and *b′* are the two leftmost anchors of *B* in *G₁* which also belong to a synteny blocks in *G₁G₃*. (**A**) The $(AB)_{G_1;G_2}$ adjacency is represented in the middle of the figure. On the left, the adjacency is not conserved in *G₃* resulting in $cScore\left((AB)_{G_1;G_2}, G_3\right) = 0$. On the right, two examples of conservation of the $(AB)_{G_1;G_2}$ adjacency in genome *G₃*. Top: the homologs of *a* and *b* in *G₃* are located in 2 neighboring synteny blocks in the *G₁G₃* comparison resulting in $cScore\left((AB)_{G_1;G_2}, G_3\right) = 0.5$. Bottom: the homologs of *a* and *b* in *G₃* belong to a single synteny block in the *G₁G₃* comparison resulting in $0.51 \leq cScore\left((AB)_{G_1;G_2}, G_3\right) \leq 1$. (**B**) When $0.51 \leq cScore\left((AB)_{G_1;G_2}, G_3\right) \leq 1$, the $cScore\left((AB)_{G_1;G_2}, G_3\right)$ can take 50 discrete values depending on a distance called *d0* and 4 additional scores called left(*a*), left(*a′*) right(*b*) and right(*b′*):

$$cScore\left((AB)_{G_1;G_2}, G_3\right) = \begin{cases} 1 - dO & if\ \text{left}(a), \text{left}(a'), \text{right}(b) \text{and right}(b') \geq 2 \\ 0.9 - dO & if\ \text{left}(a), \text{left}(a'), \text{right}(b) \geq 2 \text{ and right}(b') = 1 \\ & or\ if\ \text{left}(a), \text{right}(b), \text{right}(b') \geq 2 \text{ and left}(a') = 1 \\ 0.8 - dO & if\ \text{left}(a), \text{right}(b) \geq 2 \text{ and left}(a'), \text{right}(b') = 1 \\ 0.7 - dO & if\ \text{left}(a) = 1 \text{ and right}(b) \geq 2 \\ & or\ \text{left}(a) \geq 2 \text{ and right}(b) = 1 \\ 0.6 - dO & if\ \text{left}(a), \text{right}(b) = 1 \end{cases}$$

The 4 scores quantify how far *a, a′, b* and *b′* lie from the ends of block *A′* in *G₃*. They correspond to the number of anchors in *A′* to the left of *a* and *a′* and to the right of *b* and *b′*. The distance

dO is based on N_{G_1} and and N_{G_3}, the numbers of intervening genes in *G₁* and in *G₃*, respectively, between the two closest anchors to the $(AB)_{G_1;G_2}$ adjacency in *A'*. *d0* is comprised between 0 and 0.09 and is calculated as follows: $dO = \min(0.09, (abs(N_{G_1} + N_{G_3}) - 2) * 0.01)$. Only 6 different $cScore$ values are exemplified. (**C**) There are two particular cases where the definition of *a, a', b* and *b'* is not applicable. Left: block *B* is included into block *A* in the G_1G_2 comparison: if the genes from block *B* are anchors of *A'* then $cScore((AB)_{G_1;G_2}, G_3) = 0.4$, if not $cScore((AB)_{G_1;G_2}, G_3) = 0$. Right: block *A* is a telomere proximal block in G_1. If at least one gene from block *A* belongs to a telomere proximal block in *G₃* then $cScore((AB)_{G_1;G_2}, G_3) = 0.3$, otherwise $cScore((AB)_{G_1;G_2}, G_3) = 0$.

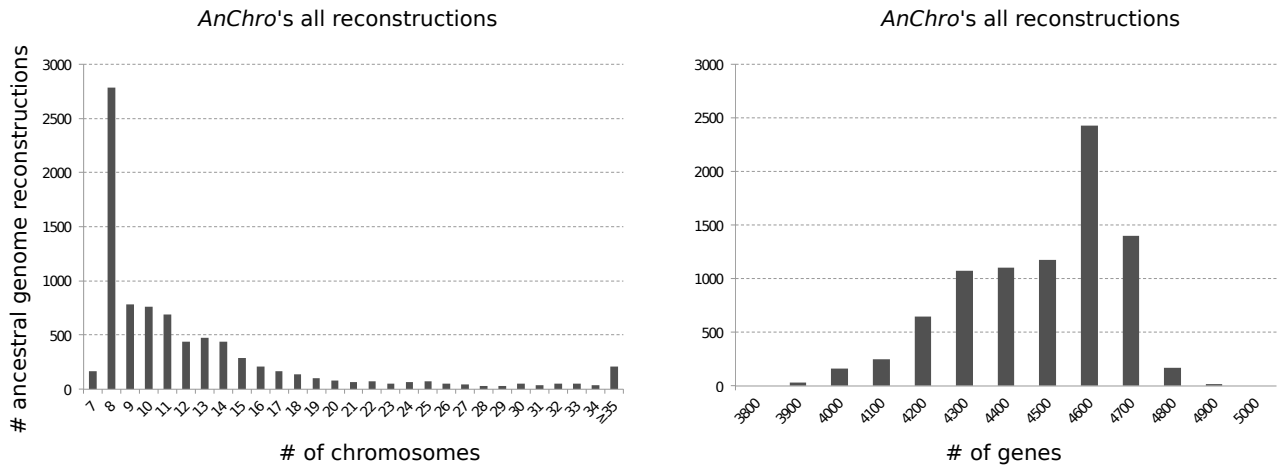




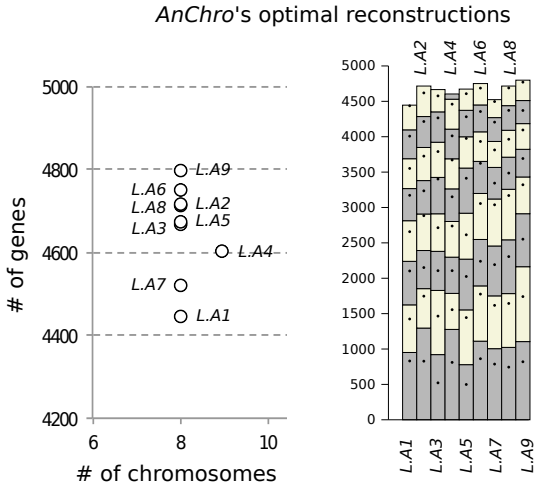
Supplemental Figure S14

Supplemental Figure S14: Resolution of micro-rearrangements. Top: Ancestral chromosomes are assembled as two lists of ordered synteny blocks. They comprise the syntenic homologues corresponding to the anchors of the blocks, one version according to *G1* and the other to *G2.* All duplicated blocks and genes are removed to produce a one-to-one gene relationship between the 2 versions of each ancestral chromosome. Bottom: Micro-synteny blocks are then computed between these 2 versions and when the orientation of a micro-synteny block is different between the *G1* and *G2* versions, the correct orientation is chosen according to the gene orientation found in the *G3* reference genome. Here the *G1* version is chosen as ancestral because it has the same gene orientation as in *G3.*

A



B



Supplemental Figure S15

Supplemental Figure S15: Reconstructions of *Lachancea ancestral genomes with AnChro.* All synteny blocks were computed with *SynChro*. **A.** The default reconstruction for each of the nine ancestral genomes (*L.A1* to *L.A9*) was determined by (i) choosing the pair of genomes G_1, G_2 that minimizes the number of synteny blocks and (ii) setting both synteny block stringency parameters Δ' and Δ'' to 3 (see Supplemental information). **B.** Distribution of the number of chromosomes (left) and the number of genes (right) obtained in the 8,532 reconstructions of the 9 *Lachancea* ancestors. A majority of reconstructions ends up in ancestral genomes containing 8 chromosomes and about 4,600 genes. **C.** The nine optimal reconstructions were chosen among all 8,532 ancestral reconstructions (see Supplemental information). The left panel shows the number of ancestral genes and chromosomes reconstructed for each ancestral genome. The right panel shows a schematic representation of ancestral chromosomes. Each column represents the ancestral chromosomes of a given ancestor as an alternation of grey and beige boxes with size being proportional to the number of reconstructed ancestral genes. The small black circles indicate the centromere position.