Recombination drives the evolution of GC-content in the human genome.

Julien Meunier & Laurent Duret

Supplementary Information

Retrieving chimpanzee and baboon sequences. To construct genomic human / chimpanzee / baboon alignments, we retrieved large ($\geq$ 20 Kb) chimpanzee and baboon (i.e. Pan and Papio species) DNA sequences (respectively 291 and 233) from GenBank (Rel. 133, February 2003). We proceeded in three main steps to derive orthologous triple alignments, mainly to avoid the pitfalls of repetitive elements, and to deal with large-scale chromosomal rearrangements such as duplications, inversions, insertions and deletions.

Computing a rough mapping. Repeated elements in the chimpanzee and baboon sequences were masked with RepeatMasker (Smit A. F. A. and Green P.), using the Repbase Update reference library (Jurka, 2000). RepeatMasker is available at
http://ftp.genome.washington.edu/RM/RepeatMasker.html).
The command line was as follows:

```
prompt$ RepeatMasker seq.fasta
```

which produced a `seq.fasta.masked` file with the masked sequence. We then conducted a similarity search against human chromosomes (Ensembl, release 8.3) using Megablast to roughly map chimpanzee and baboon sequences on their orthologous loci:

```
prompt$ megablast -D 1 -F F -d ENSEMBL.hum.rel8.3 -i
        seq.fasta.masked -o seq.blast.masked -p 95 (90)
        -s 400 (300)
```

with -D 1: an option for the output format; -d ENSEMBL.hum.rel8.3: the bank on which the sequence are blasted; -i seq.fasta.masked: the chimpanzee or baboon sequence to be blasted, which has been previously masked; -o seq.blast.masked: the name of the output file; -p 95 (90): the minimal percentage of identity of a similarity area (95% for chimpanzee and 90% fo baboon). -s 400 (300): the minimal length of a similarity area (400 for chimpanzee and 300 for baboon). At this stage, for each chimpanzee and baboon sequence, we have a set of similarity blocks which are subsequently referred to as HSPs (Fig. 1).

From the set of similarity blocks (HSPs), we derived a rough mapping in three steps. First, we eliminated from each HSP all segments that were involved in more than one HSP (see Fig 1a). This could lead to a reduction in size of the HSPs. The second steps consisted in the elimination of all HSPs which size was below 200 base pair (Fig 1b). Finally, sets of three or more HSPs that were consistently orientated were identified (Fig 1c), and the outer boundaries of the most-left and most-right HSPs defined a first mapping of the chimpanzee or baboon sequence on the human genome (Fig 1d).

Improving the mapping. We then used human/chimpanzee and human/baboon pairwise alignments computed by MGA (Holn, Kurtz and Ohlebusch, 2002) to generate an accurate mapping with non-masked sequences, which enabled us to identify potential triple alignments, and to discard large non-homologous, inverted or deleted areas:

```
prompt$ mkvtree -dna -lcp -suf -tis -indexname ind -db
        seq.fasta hum.fasta
prompt$ mga.32seqs -l 100 33 (13) -gl 200 -always ind
        > align.mga
```

with `seq.fasta`: a non-human primate sequence area that roughly mapped on a human genomic area (`human.fasta`). This command lines parameterized MGA so as to reject any area in the chimpanzee (baboon) sequence larger than 200 base pair that did not present an exact match of 33 (13) base pair, respectively (Fig. 2a). Simulations shows that this criterion guarantees the rejection of areas with a similarity below 96% (92%) for chimpanzee and baboon respectively. At this step, we could derived a set of areas in the sequence that accurately mapped on the human genome (Fig. 2b).

Computing the triple alignments We then identified regions in common to at least one chimpanzee and one baboon sequence (Fig. 3). When several chimpanzee or baboon sequences mapped on a same area of the human genome, we arbitrarily selected in such cases the largest sequence (see Fig. 3a). Finally, triple alignments were generated using ClustalW (see Fig. 3b).

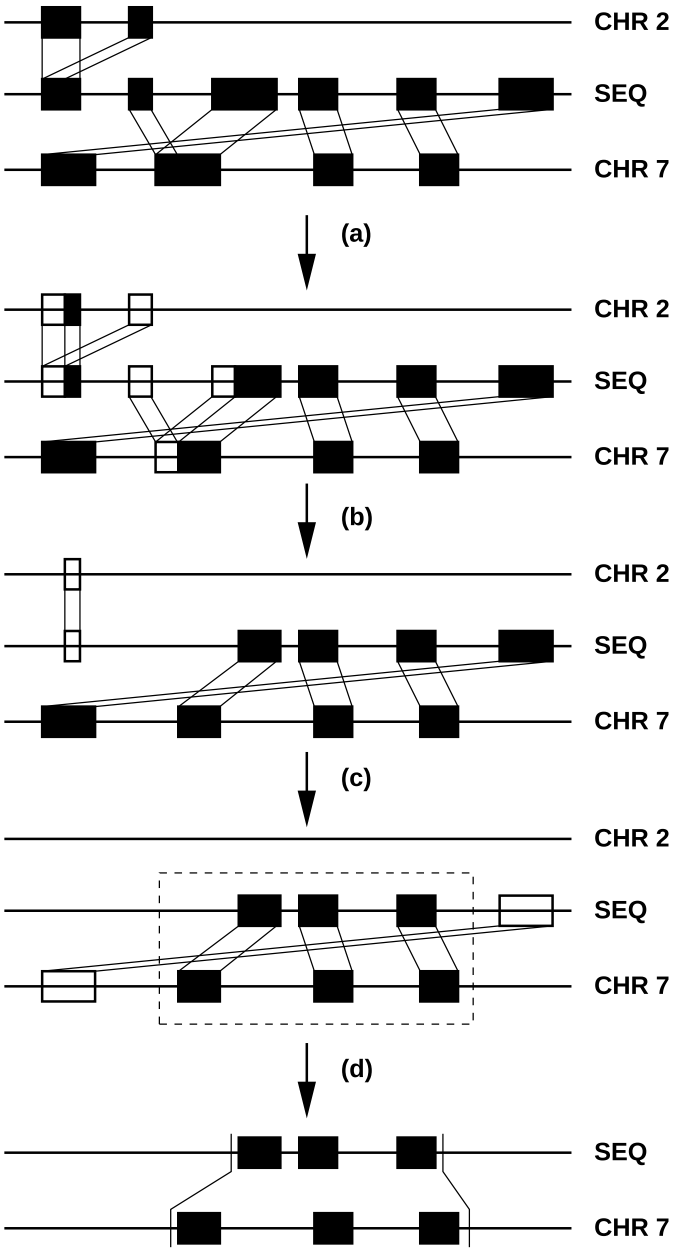

Figure 1: **From similarity blocks to a rough mapping.** The main steps to compute a rough mapping of a chimpanzee or baboon sequence on a human chromosome are presented. SEQ: a non-human primate sequence; CHR n: human chromosome number n; black rectangle: HSP; open rectangle: eliminated area in a HSP.

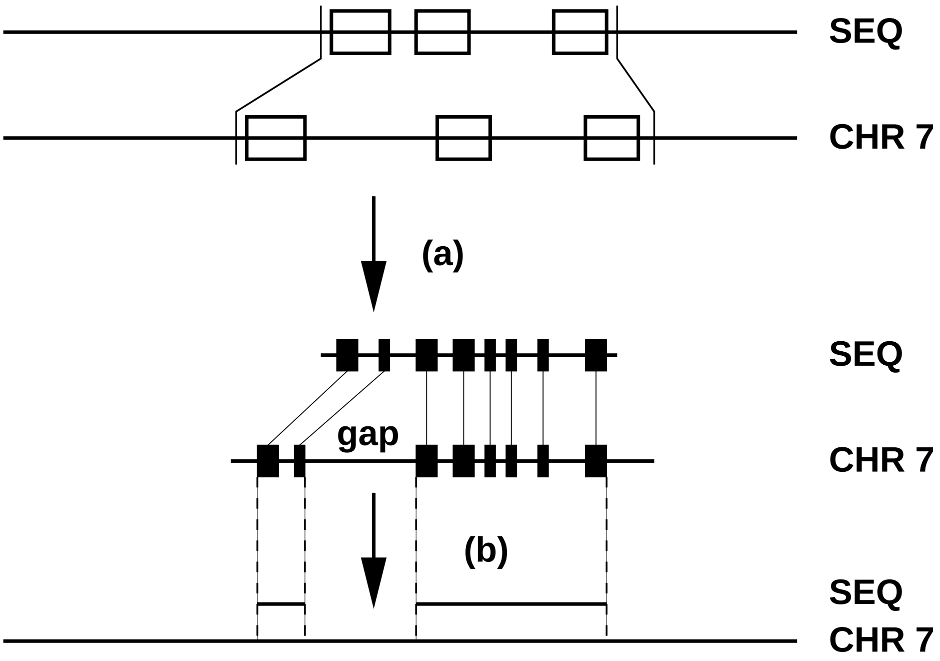

Figure 2: **Computing an accurate mapping.** The use of MGA to compute an accurate mapping of non-human primate sequences on a human chromosome is presented. SEQ: a non-human primate sequence; CHR 7: human chromosome number 7; open rectangle: HSP; large black rectangle: large exact match; small black rectangle: small exact match. From a rough mapping, MGA computed alignments in which all regions with more than 200 base pair that did not present an exact match were tagged as a unaligned gap (a) and eliminated. The outer boundaries of the left-most and right-most exact matches in a remaining region constituted a precise mapping on the human cromosome (b).

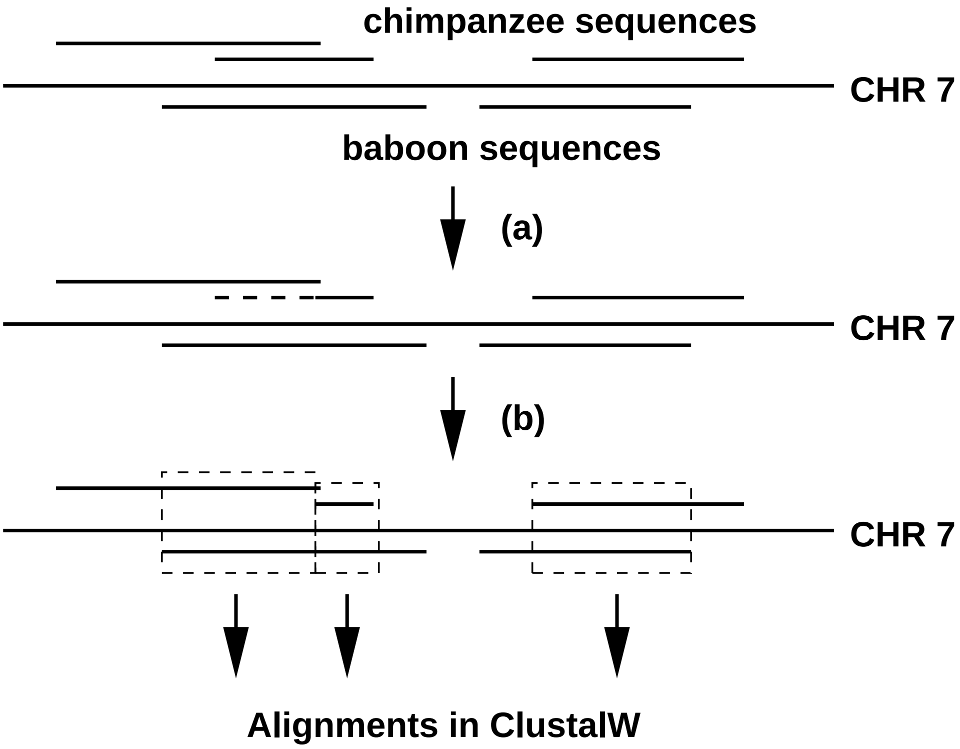

Figure 3: **Generating triples alignments.** The triple alignments were computed in two steps. First, when several chimpanzee or baboon sequence areas mapped on the same human genomic region, we chose the area belonging to the longest sequence and eliminated the others (a). Second, the areas with potential triple alignments were identified and used as input for ClustalW (b). SEQ: a non-human primate sequence; CHR 7: human chromosome number 7.

References

Holn, M., S. Kurtz, and E. Ohlebusch. 2002. Efficient Multiple Genome Alignment. Bioinformatics **S1**:S312–S320.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends. Genet. **16**:418–420.