*Proc. Assoc. Advmt. Anim. Breed. Genet. Vol 14*

# COMPUTATIONAL COMPARATIVE MAPPING BETWEEN MAMMALIAN SPECIES

**A. Zadissa[1], K.G. Dodds[2], J.C. McEwan[2]**

[1]AgResearch Molecular Biology Unit, Dept of Biochemistry, University of Otago, P.O. Box 56, Dunedin, New Zealand
[2]AgResearch, Invermay Agricultural Research Centre, PB 50034, Mosgiel, New Zealand

## SUMMARY
Merging genetic maps is an important tool in comparative genomics. Transferring locus positional information from a well-mapped genome to a map of lower density allows for more rapid genetic analysis in the poorly mapped species. In this report a methodology for merging genetic maps from different species is presented. While this can be used for any type of genetic map, the primary intention for its use is to extrapolate positional information from human genomic sequence to ruminant species via framework Type I loci and ruminant ESTs using *in silico* mapping.
**Keywords**: Comparative mapping, translation table, homology, DNA.

## INTRODUCTION
Cross-species comparison of mammalian gene maps is a powerful tool for identifying and studying genomic segments conserved between species. The basis of comparative gene mapping is that genes closely linked in one species tend to be closely linked in other species, whereas loosely linked genes in one species tend to be unlinked in related species. Although some rearrangements have been observed among ruminants, the number of the rearrangements is low (Crawford *et al.* 1995; Gellin *et al.* 2000). Hence, comparative maps are a potentially important source of information for QTL localisation studies. A direct comparison between gene maps of divergent species requires the existence of a set of orthologous sequences that can serve as landmarks for alignment of conserved segments across the species. The best-suited markers for this purpose would be Type I markers, i.e. expressed genes. The recently completed human genomic map is a valuable resource for the development of comparative maps (Larsen *et al.* 1999). Efforts in livestock species are not as intense, but it has been determined that gene order and chromosome organisation between mammalian species is highly conserved (O'Brien *et al.* 1999). This information can be used to predict the map positions for markers unique to ruminants via translation tables. Currently there are computational techniques available for a number of comparative mapping tasks, but most of these are concerned with predicting the number of syntenic regions, identifying these regions, and predicting syntenic block membership (Goldberg *et al.* 2000; Nadeau *et al.* 1998), or display of aligned maps (Hu *et al.* 2001). Although an aligned map may give a good visual impression of the predicted location of a locus, based on its position in one species, it does not go as far as predicting that location. In this work we propose a methodology for predicting the location of an orthologous locus in another species.

## METHODOLOGY
**Construction of a comparative map.** A translation table is essential for merging maps from different species. A framework marker, usually a Type I marker, is a locus common to maps being compared. A non-framework marker is present on only one map. Framework markers are used for

calculating the position of non-framework markers in the comparative species. There are three different conditions for computing the position of a marker, depending on the arrangement of adjacent framework markers.

**Interpolation.** The interpolation estimation is appropriate when the chromosome regions being compared are syntenic both in order and relative position, i.e. there are no known breakpoints or rearrangements in the region, and when there are two orthologous flanking markers. It will usually be sufficiently accurate to assume a linear relationship between the two regions being compared, in which case the estimated position of the locus is found by simple interpolation, as shown in Equation 1 and Figure 1a.

$$C' \approx \frac{B' - A'}{B - A}(C - A) + A' \qquad (1)$$

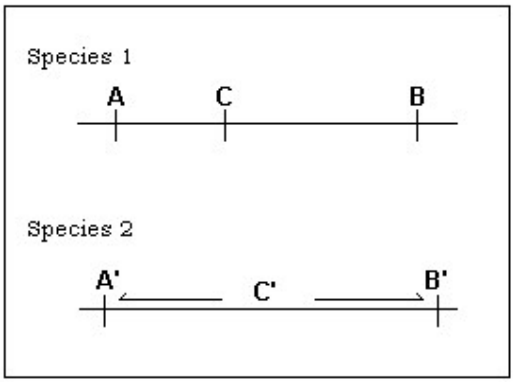



**Figure 1a. The markers A', B' and C' are orthologous to A, B and C. A and B are flanking markers surrounding locus C. The estimated position of C' is calculated using interpolation.**

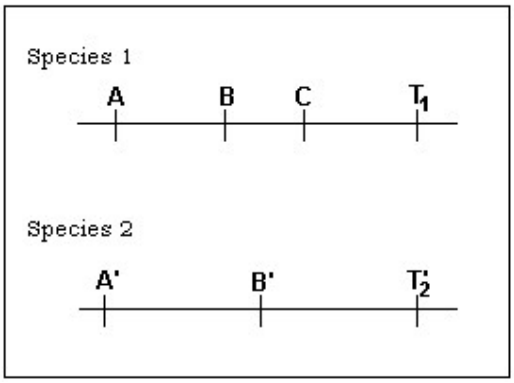
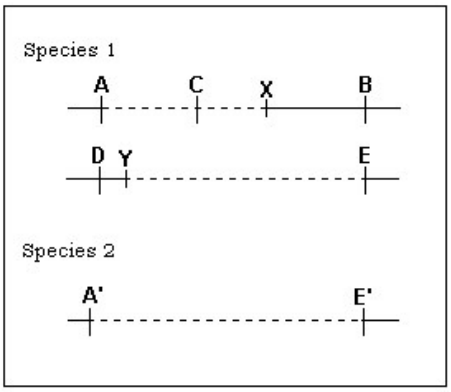



**Figure 1b. Here the marker being mapped, C, is outside the A-B interval. A and B are orthologous to A' and B'. The position of C' is estimated by extrapolating from A and B. Non-orthologous but telomeric markers $T_1$ and $T'_2$ may provide further information.**

**Simple extrapolation.** This method is used for mapping in syntenic regions when there are no orthologous flanking markers. The estimated position of *C'* can be calculated by extrapolation using Equation 1. The absence of orthologous flanking markers introduces greater uncertainty. It may be beneficial to use a long common syntenic baseline (i.e. marker *A* would not be the closest orthologous marker to *B*) for the extrapolation of *C'* to reduce the influence of positional errors of *A*, *A'*, *B* and *B'*, see Figure 1b. Alternatively, we may have additional information provided by distal markers $T_1$ and $T'_2$ which give some indication of the likely distance to the telomeric end of the chromosome in both species. The expected distance of such markers from the telomeres is a function of the total number of markers on the map, assuming the markers have been developed and positioned randomly. Thus if *C'* is mapped to a position outside the likely terminal end of the syntenic chromosome, an alternative is to use the estimated positions of the respective telomeres and translate the problem into the first case and use modified interpolation. In practice, the option chosen

will depend on the number of mapped markers and the number of mapped syntenic markers in both species. The alternative method is particularly suited to situations with few common syntenic markers but large numbers of mapped markers in both species.

**Complex extrapolation.** The third case is where the marker to be positioned is within a breakpoint region (Figure 2). Initially we assume that *C* is positioned between *A* and the breakpoint *X*. Adjacent orthologous markers are used to determine that the segments homologous to *A'-E'* are *A-X* joined with *Y-E*. A linear relationship between these segments is given in Equation 2.




$$E' - A' \approx l\left((X - A) + (E - Y)\right), \quad l \text{ = constant} \quad (2)$$

**Figure 2. Two breakpoints, X and Y, have occurred in the divergence of species 1 and 2 and their exact position is not known. The markers A, B, D and E, in species 1 and A' and E' in species 2 have known positions. The dashed line indicates the homologous segments of interest.**

Assuming this linear relationship holds across the regions of interest, we can interpolate to find:

$$C' = A' + l\,(C - A) \quad (3)$$

A simple estimate for the constant $l$ is the ratio of orthologous syntenic regions, either on the chromosome(s) in question, or in the total genome, between species 2 and species 1. A lower bound on $l$ is found by letting $X=B$ and $Y=D$ in Equation (2). More refined bounds can be found by incorporating information from comparatively mapping the other sides of these breakpoints.

Now that we have estimated the position of *C'*, assuming that *C* is positioned between *A* and *X*, we would like to know how valid that assumption was. This involves estimating the positions of *X* and *Y*, and the distribution of these estimates. A simple method is to place *X* and *Y* so that contributions from each breakpoint region are proportionally equal, i.e.

$$\frac{X - A}{B - A} = \frac{E - Y}{E - D} \quad (4)$$

Substitutions and simplifications using Equation 2 and Equation 4 result in

$$X = \frac{(E' - A')(B - A)}{l\,(E - D) + (B - A)} + A \qquad \text{and} \qquad Y = (X - A) + E - l^{-1}(E' - A'),$$

and we can check whether $C<X$.
A more refined method is to assume that $X$ is uniformly distributed between its bounds:

$$X_{\min} = \begin{cases} D + A - E + l^{-1}(E' - A') & \text{if } Y \le D \text{ at } X = A \\ A & \text{if } Y > D \text{ at } X = A \end{cases}$$

$$X_{\max} = \begin{cases} B & \text{if } Y \le E \text{ at } X = B \\ A + l^{-1}(E' - A') & \text{if } Y > E \text{ at } X = B \end{cases}$$

If P($C<X$) is low, it suggests that *C'* is syntenic to *B'* rather than to *A'*. Further refinements on the bounds of $X$ are possible by incorporating information from other regions, as suggested above. Inevitably there will be times when the method is found to suggest a region incorrectly, but such information would be of value in locating and refining breakpoint positions.

**DISCUSSION**

The human genome project is currently complete with draft human genome sequence and gene transcript maps available (Venter *et al.* 2001). A large amount of information in the form of ruminant ESTs and gene sequences that can be employed in comparative genetics has also been generated. The challenge is to integrate this information using type I markers from genetic maps for agriculturally important ruminants, such as cattle and sheep. Genetic maps for ruminants are now in their second and third generation (Barendse *et al.* 1997; de Gortari *et al.* 1998; Schibler *et al.* 1998). Nevertheless, their density is very low compared to the human genome. Homology of the human DNA coding sequence with cattle and sheep average 84% (Zadissa 2000). In addition, this group found that 62% of bovine EST contigs aligned *in silico* with a putative human ortholog in the human draft sequence. This level of homology is at the lower edge of reliable Southern blotting using heterologous probes and below the threshold where PCR primers designed in one species can be used in other species. The methodology outlined in this paper will allow these matching EST contigs to be approximately positioned on the existing bovine map. Sequences mapped *in silico* to regions of likely economic importance, e.g. under QTL peaks, can then be confirmed experimentally.

**REFERENCES**

Barendse, W., Vaiman, D., Kemp, S.J., *et al.* (1997) *Mamm Genome* **8**: 21.
Crawford, A.M., Dodds, K.G., Ede, A.J., *et al.* (1995) *Genetics* **140**: 703.
Gellin, J., Brown, S., Graves, J.A.M., *et al.* (2000) *Mamm Genome* **11**: 140.
Goldberg, S.D., McCouch, S. and Kleinberg, J. (2000) In "Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment and the Evolution of Gene Families", editors D. Sankoff and J H. Nadeau, Dordrecht, Kluwer Academic Press.
de Gortari, M.J., Freking, B.A., Cuthburtson, R.P., *et al.* (1998) *Mamm Genome* **9**: 204.
Hu, J., Mungall, C., Law, A., *et al.* (2001) *Nucl. Acids Res.* **29**: 106.
Larsen, N.J., Hayes, H., Bishop, M., *et al.* (1999) *Mamm Genome* **10**: 482.
Nadeau, J.H. and Sankoff, D. (1998) *Trends in Genetics* **14**: 495.
O'Brien, J.S., Eisenberg, F.J., Miyamoto, M., *et al.* (1999) *Science* **286**: 463.
Schibler, L., Vaiman, D., Oustry, A., *et al.* (1998) *Genome Res* **8**: 901.
Venter, J.C., Adams, M.D., Myers, E.W., *et al.* (2001) *Science* **291**: 1304.
Zadissa, A. (2000) MSc Thesis, Uppsala, Sweden.