Modelling RH maps uncertainty and application to the validation of whole genome assemblies

B. Servin[*], S. de Givry[†] and T.Faraut[*]

Introduction

Genome maps are imperative to address the genetic basis of an organism machinery. While a growing number of genomes are being completely sequenced, providing therefore ultimate genome maps, the process of constructing intermediate maps (genetic maps, RH maps, ...) remains an inescapable time-consuming task. Radiation hybrid (RH) maps for example have played an important role in facilitating the process of whole genome sequencing and assembly for human, mouse rat and dog (Cheung et al., 2001; Hitte et al., 2005; Kwitek et al., 2004; Rowe et al., 2003) to name only a few. They have provided in addition an independent source of information for the validation of genome assemblies because the comparison of maps produced by independent protocols (genetic, RH, sequence-based) give clues about map accuracies(Lewin et al., 2009). One of the most sensitive aspect of comparing maps however - for example the comparison of a map to a genome assembly - lies in the interpretation of inconsistencies and more particularly in the way to resolve them. Indeed, as the result of the limited nature of experimental data used in the different mapping construction process, the resulting maps are not exempt of errors but drastically lack associated uncertainty measures that could enable to quantify the amount and identify these errors. The usual output of a mapping experiment consists in a single map representing the optimal solution of the optimization problem associated with the mapping experiment, e.g minimum obligate breaks or maximum likelihood order in genetic or RH mapping. This single solution does not come with information concerning the data support for a particular order or for the order of a particular subset of markers in a predefined region.

The aforementionned difficulties are particularly relevant when addressing the quality of genome assembly. Indeed, having in hand a single map resulting from a mapping process and a single genome assembly, there is no straightforward rules that enable to select assembly regions, inconsistent with the map, that deserve further investigations.

Here, we propose to address this difficulty with new methods that provide means of describing precisely the uncertainty in the RH ordering of markers, therefore providing measures of confidence or support we have for a particular map order. This has the concrete effect of enabling to rank the inconsistencies with respect to map uncertainty measures and help in the validation of whole genome assemblies. We provide an example of application of our methods on a dog dataset.

[*]INRA Laboratoire de Génétique Cellulaire, 31326 Castanet Tolosan, France
[†]INRA Unité de Biométrie et Intelligence Artificielle, 31326 Castanet Tolosan, France

Material and methods

Our approach is built on the comparative mapping model of Faraut et al. (2007) which consists in incorporating *a priori* information from a reference order, π_{ref} (a closely related species or an imperfect assembly) into the marker ordering approach by modelling the posterior distribution of the order π:

$$P(\pi|X,\theta) \propto L(\pi;X,\theta)p(\pi|\pi_{ref}) \tag{1}$$

where X are the RH genotypes, θ is a vector of model parameters, namely the retention fraction of the radiated hybrid panel and the breakage probabilities between pairs of markers. We assume these are known (or estimated elsewhere) and concentrate on the inference of the order (π). In Faraut et al. (2007) the proposed application of the comparative model is to find a marker map π^* that maximizes (1). To extend the application of this model to evaluate uncertainty in the marker ordering, we developped a Markov Chain Monte Carlo algorithm to estimate the posterior distribution of orders (1). The description of the algorithm is lengthy and will be presented elsewhere, we will just mention that the main issue to tackle is to explore efficiently the space of orders when dealing with large number of markers such as provided by high throughput genotyping platforms.

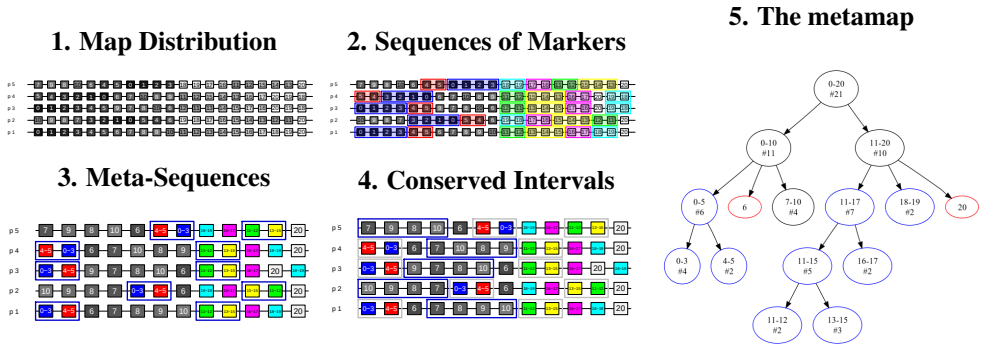

Figure 1: A toy example of the construction of a metamap (**5**) from a map distribution (**1**) by successively identifying sequences of markers **(2)**, meta-sequences (**3**) and conserved intervals (**4**). The metamap (**5**) is an inclusion tree. Individual markers are red nodes (not all are shown), sequences are the blue nodes at the tips of the tree, metasequences are the blue nodes inside the tree and conserved intervals are black nodes. The uncertainty in the marker ordering lies in the possible orderings of each node's children.

Our MCMC algorithm provides us with a map distribution: a set of marker orders with associated posterior probabilities. This set is typically very large (in the order of thousants) so that in order to fully exploit the information, we propose a concise representation of the map distribution: a *metamap*. The principle (illustrated on figure 1) is to construct an inclusion tree

that allows to partition the uncertainty into groups of markers that can be examined independently (as shown on figure 1.5). This metamap summarizes most of the information on the uncertainty in marker ordering.

One important application of the metamap is to extract a *robust map* from the map distribution. A robust map is a map composed of groups of markers which order is invariant in the map distribution. This can be constructed easily by going down the tree from the root to the tips and stopping whenever the ordering of a node's children is uncertain. In that case all the markers lying below the node are gathered in a bin that is included as a single unit on the robust map. Working with a robust map allows to eliminate all the uncertainty in the marker ordering and facilitates the comparison with other maps, such as whole genome assemblies.

Application to the validation of whole genome assemblies

We tested our algorithm on an *a posteriori* validation of the dog genome assembly CanFam1, using RH data on 423 gene-based markers located on dog chromosome 2, typed on the RHDF9000 dog RH panel (Hitte et al., 2005). The human genome order was used as a reference order. The map distribution for these data contains more than ten thousand maps. The robust map is composed of 353 units out of the 423 genes.

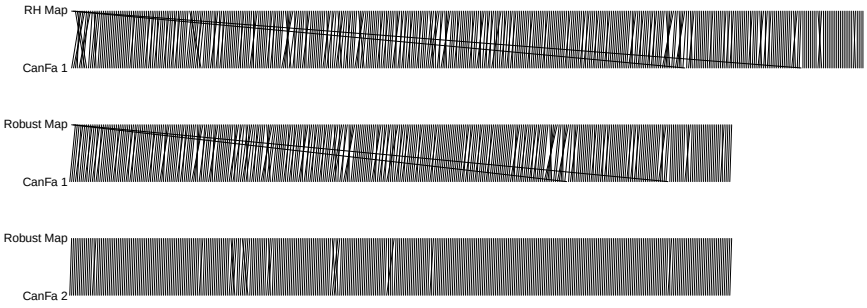

Figure 2: Comparison of the ordering of (i) top : the comparative RH map with the first draft of the dog genome CanFam1 (ii) middle : the robust map derived from the map distribution and CanFam 1 and (iii) bottom : the robust map and the second draft of the dog genome CanFam2

A great number of discrepencies were identified between CanFam1 and the best map as obtained by the TSP approach (Faraut et al. (2007) ; figure 2 top) but also with the robust map constructed using our new approach (figure 2 middle). Most of the discrepencies with the robust map were absent when compared to a more recent version of assembly (figure 2, bottom), indicating that the robust map could have been of great interest to validate and suggest ameliorations to CanFam1.

Aknowledgements

We thank Christophe Hitte for giving us access to the dog RH dataset.

Conclusion

We showed here how RH mapping is a powerful tool to help in the validation and improvement of whole genome assemblies. The methods we developped are applicable to large marker datasets such as provided by 60K SNP chips. With high resolution panels, these datasets provide a sufficient resolutive power to validate genome assemblies. The MCMC algorithm used to obtain the map distribution is implemented within Carthagene (de Givry et al., 2005). The methods to construct and exploit the metamap will be made available as an accompanying software.

References

Cheung, V. G., Nowak, N., Jang, W., et al. (2001). Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*, 409(6822):953–8.

de Givry, S., Bouchez, M., Chabrier, P., Milan, D., and Schiex, T. (2005). CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics*, 21(8):1703–4.

Faraut, T., Givry, S. d., Chabrier, P., Derrien, T., Galibert, F., Hitte, C., and Schiex, T. (2007). A comparative genome approach to marker ordering. *Bioinformatics*, 23(2):50–56.

Hitte, C., Madeoy, J., Kirkness, E. F., et al. (2005). Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat Rev Genet*, 6(8):643–8.

Kwitek, A. E., Gullings-Handley, J., Yu, J., et al. (2004). High-density rat radiation hybrid maps containing over 24,000 SSLPs, genes, and ESTs provide a direct link to the rat genome sequence. *Genome Res*, 14(4):750–7.

Lewin, H. A., Larkin, D. M., Pontius, J., and O'Brien, S. J. (2009). Every genome sequence needs a good map. *Genome Res*, 19(11):1925–8.

Rowe, L. B., Barter, M. E., Kelmenson, J. A., and Eppig, J. T. (2003). The comprehensive mouse radiation hybrid map densely cross-referenced to the recombination map: a tool to support the sequence assemblies. *Genome Res*, 13(1):122–33.