BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 24 2010, pages 3035–3042
doi:10.1093/bioinformatics/btq598

Genome analysis Advance Access publication November 13, 2010

Statistical confidence measures for genome maps: application to the validation of genome assemblies

Bertrand Servin[1,*], Simon de Givry[2] and Thomas Faraut[1]

[1]INRA Toulouse – Laboratoire de Génétique Cellulaire and [2]INRA Toulouse–Unité de Biométrie et Intelligence Artificielle, Castanet-Tolosan, France

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Genome maps are imperative to address the genetic basis of the biology of an organism. While a growing number of genomes are being sequenced providing the ultimate genome maps—this being done at an even faster pace now using new generation sequencers—the process of constructing intermediate maps to build and validate a genome assembly remains an important component for producing complete genome sequences. However, current mapping approach lack statistical confidence measures necessary to identify precisely relevant inconsistencies between a genome map and an assembly.

Results: We propose new methods to derive statistical measures of confidence on genome maps using a comparative model for radiation hybrid data. We describe algorithms allowing to (i) sample from a distribution of maps and (ii) exploit this distribution to construct robust maps. We provide an example of application of these methods on a dog dataset that demonstrates the interest of our approach.

Availability: Methods are implemented in two freely available softwares: Carthagene (http://www.inra.fr/mia/T/CarthaGene/) and a companion software (metamap, available at: http://snp.toulouse.inra.fr/~servin/index.cgi/Metamap)

Contact: Bertrand.Servin@toulouse.inra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 11, 2010; revised on September 28, 2010; accepted on October 18, 2010

1 INTRODUCTION

Genome maps are imperative to address the genetic basis of an organism biology. While a growing number of genomes are being completely sequenced providing the ultimate genome maps, the process of constructing intermediate maps—genetic maps, RH maps to name only a few—remains certainly an important task (Lewin *et al.*, 2009). Indeed, the current whole-genome shotgun approach for genome sequencing produce a large number of contigs and scaffolds of limited length. The N50 scaffold size of the recent panda genome assembly, for example, reaches 1.3 Mb (Li *et al.*, 2010) compared with the 50 Mb of the smallest human chromosome, providing a limited picture of the genome architecture for this species. Dense chromosomal maps remain invaluable for organizing the scaffolds along the chromosomes and can be of great help for checking the order of markers within assemblies (Lewin *et al.*, 2009). Radiation hybrid (RH) maps, for example, have played an important role in facilitating the process of whole-genome sequencing and assembly (Hitte *et al.*, 2005). They provide an independent source of information for the validation of genome assemblies because the comparison of maps produced by independent protocols (genetic, RH, sequence based) gives clues about map accuracies. 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. Indeed, as a result of the limited nature of experimental data used in the mapping construction process, the resulting maps are not exempt of errors. Perhaps more importantly, because of the difference in marker informativeness and also in the density of markers along a map, the experimental data support for the local order of markers at different locations of the map may vary considerably. The usual output of a mapping experiment consists, however, 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. While LOD scores between pairs of markers in genetic or RH maps measure the degree of linkage between adjacent markers, genome maps lack statistical confidence measure to reflect middle to long range order accuracy in contrast, for example, to the long-standing practice of support values in phylogenetic analysis, i.e. the bootstrap values or the Bayesian posterior probabilities for the internal nodes of phylogenetic trees (Felsenstein, 2004).

The aforementioned 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 by constructing statistical confidence measures for maps that reflect locally the confidence or support we have for a particular map order. This enables to rank the inconsistencies with respect to these measures and help in the validation of whole-genome assemblies.

Previous work has been done on the modeling of map uncertainty using Bayesian models, for RH data (Heath, 1997; Lange and Boehnke, 1992) and genetic data (George, 2005). They were, however, limited to a small number of markers. In this article, we propose a new model that (i) exploit the availability of prior information on marker ordering (on a closely related species or on a draft assembly) and (ii) aims at analyzing large datasets such as provided by high-throughput genotyping platforms (SNP chips).

*To whom correspondence should be addressed.

Our approach is based on the comparative mapping approach of Faraut *et al.* (2007), that introduced a model to combine information from reference orders and RH data. Here, we extend the application of this model to be used in the evaluation of map uncertainty in a coherent probabilistic framework. This is done by estimating a posterior distribution of marker ordering (i.e. a set of orders and associated posterior probabilities) using a Markov Chain Monte Carlo (MCMC) approach. The exploitation of this distribution to characterize the uncertainty is an issue that needs to be tackled. To this end, we developed a method to exploit the order distribution and in particular to find a subset of markers which order is strongly supported by the data. We first describe briefly the methods developed and then illustrate the properties of the model on simulated data. Finally, we present an analysis of RH data for the dog genome that exemplifies how the inference provided by our approach can help improve whole-genome assemblies. A detailed description of the methods is given in the Appendices provided in Supplementary Material.

2 METHODS

2.1 Estimation of the distribution of orders

Our approach is based on the comparative mapping model of Faraut *et al.* (2007), which consists in incorporating *a priori* information from a reference order, π_{ref} (closely related species or a draft assembly) into the marker ordering approach by modeling 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 (see Appendix 2 provided in Supplementary Material for a description of the statistical model and associated parameters). In the following, we will assume that these are known (or estimated elsewhere) and concentrate on the inference of the map (π).

In Faraut *et al.* (2007), the proposed application of the comparative model is to find a marker map π^* that maximizes the probability (1) with respect to π. This is done by converting the maximization problem into a symmetric Traveling Salesman Problem (TSP) for which efficient solving algorithms exist (Applegate *et al.*, 2007). Here, we extend the application of this model to evaluate the uncertainty in the marker ordering by using a MCMC algorithm that estimates the posterior distribution of orders (Equation 1). The main issue to tackle here is to explore the space of orders with high posterior probabilities. Indeed, computing $L(\pi;X,\theta)P(\pi|\pi_{ref})$ is reasonably fast (less than 0.1 s on the dog data presented below); however, the number of possible maps π is very large ($n!/2$ for n markers).

Outline of the MCMC algorithm: we combine two sampling strategies:

(1) Sampling of inversions via a Metropolis Hastings step.

(2) Sampling of markers placements via Gibbs sampling.

In practice in each iteration of the MCMC, we perform step 1 one time and step 2 n times, where n is the number of markers.

Step 1: sampling of inversions Given a current map π, we wish to propose a new map π' that differs from π by one inversion of a subsection of the map. In order to propose inversions with a reasonable acceptance probability, we build our proposal distribution in the following way. For each possible inversion on map π of the subsection with extremities $i \in [1 \ldots n-1]$ and $j \in [i+1 \ldots n]$, resulting in the map $I(\pi,i,j)$, we compute the right member of Equation (1). We denote this quantity $Q(\pi,i,j)$. We then propose a new map $\pi' = I(\pi,i',j')$ by sampling an inversion from the probability distribution:

$$P(\pi'|\pi) = \frac{Q(\pi,i',j')}{\sum_{(i,j)} Q(\pi,i,j)} = \frac{L_\theta(\pi';X)P(\pi'|\pi_{ref})}{C(\pi)}$$

with $C(\pi) = \sum_{(i,j)} Q(\pi,i,j)$.

This move is then accepted with probability:

$$\min\left(1, \frac{P(\pi'|X,\theta)P(\pi|\pi')}{P(\pi|X,\theta)P(\pi'|\pi)}\right)$$

which simplifies into:

$$\min\left(1, \frac{C(\pi)}{C(\pi')}\right)$$

Step 2: sampling of individual marker positions The second move used in the MCMC algorithm consists in resampling the position of a marker on the map, conserving the order of all the other markers. This can be done by Gibbs sampling of the new position i' of marker i. Let $T(\pi,i,i')$ be the map such that marker at position i in π is at position i' in $T(\pi,i,i')$ and let π_{-i} be the map π with marker i removed. We wish to sample a new position j for marker i from:

$$\begin{aligned} P(i'|\pi_{-i},X) &= P(T(\pi,i,i')|X)/P(\pi_{-i}|X) \\ &= P(T(\pi,i,i')|X)/\sum_k P(T(\pi,i,k)|X) \\ &= \frac{P(X|T(\pi,i,i'))P(T(\pi,i,i'))}{\sum_k P(X|T(\pi,i,k))P(T(\pi,i,k))} \end{aligned} \tag{2}$$

A more detailed description of the MCMC algorithm is given in Appendix 1 provided in Supplementary Material.

Starting order In principle, we could start the MCMC algorithm from any random ordering of markers. We expect the algorithm to converge to a neighborhood of the best map after enough iterations (called *burning iterations*). We found out, however, that this strategy was not efficient because the rate of convergence is extremely slow. We use a different strategy by first finding π^* using the methods mentioned above, and then starting the MCMC algorithm from this best ordering. This way we obtain consistent inferences across multiple MCMC runs.

2.2 Description and representation of the order distribution

Our MCMC algorithm provides us with a map distribution: a set of marker orders with associated posterior probabilities. Although this set is typically very large (in the order of thousands), the distribution exhibits structures that are shared by a large number of maps. The first and simplest common structure, called a *sequence*, is characterized by markers that are organized consecutively in all the maps. The second type of structure is an extension of the previous one and is characterized by consecutive sequences, called *meta-sequences*. Finally, we observe group of markers that are always grouped together but can be rearranged within the group: the *common intervals*. This last notion has been extensively studied in the context of computational approaches to genome rearrangements (Bergeron *et al.*, 2008). The information included in the map distribution can be summarized in an inclusion tree, called a *metamap*, built using the three structures described above. Each node of the tree is associated to a set of markers or group of markers and the corresponding probabilities of their respective orders in the distribution. The principle (illustrated on Fig. 1) is described in details in Appendix 3 provided in Supplementary Material. Our inclusion tree very much resembles a PQ-tree (Landau *et al.*, 2005) where the Q nodes correspond to the nodes associated to sequences (or meta-sequences) and the P nodes to the ones associated to common intervals. In contrast to PQ-trees, our inclusion tree stores at each node the probabilities of the concurrent orientations, for Q nodes, and the probabilities of the concurrent orders, for the P nodes. This additional information is crucial for the construction of the robust map (see below).

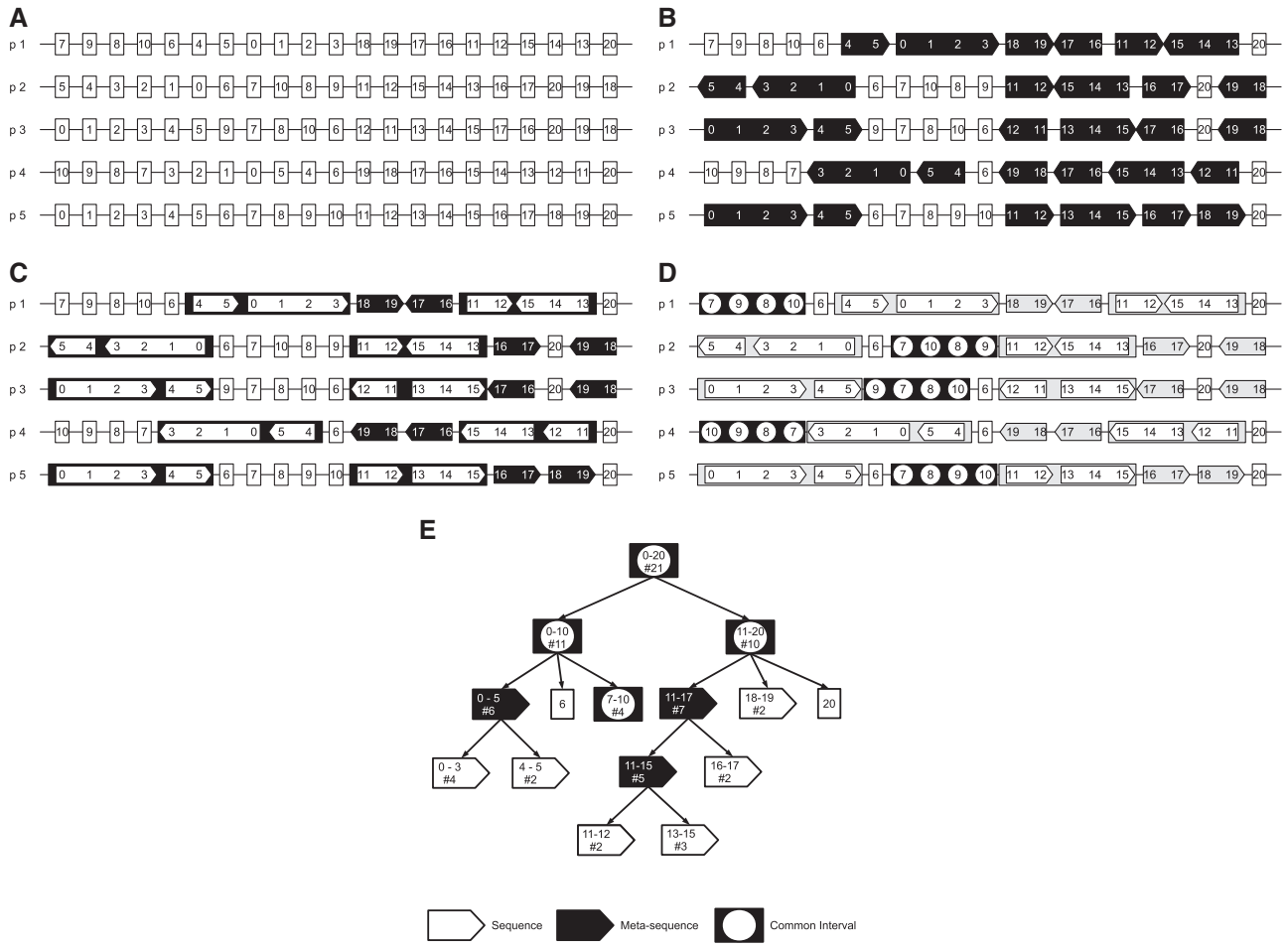

Fig. 1. A toy example of the construction of a metamap (**E**) from a map distribution (**A**) by successively identifying sequences of markers **(B)**, meta-sequences (**C**) and common intervals (**D**). The metamap (**5**) is an inclusion tree. The uncertainty in the marker ordering lies in the possible orderings of each node's children. Each node is labeled with the starting and ending marker and the number of markers.

2.3 Constructing robust maps

To pinpoint the relevant inconsistencies between a map and a whole- genome assembly, our approach is to first identify a *robust map*, i.e. a marker ordering which is invariant in the map distribution. We now show how a robust map can be constructed from the metamap. Starting from the root node of the metamap, we perform the following steps:

(1) If the current node is a sequence with a single possible orientation, add all the markers to the robust map.

(2) If the order of the children nodes is certain, apply the algorithm to each of the children nodes.

(3) Otherwise, in the set of markers that lie below the current node, we identify the longest sequence with a single possible orientation by going down the sub-tree below the current node. We then add markers of this longest sequence to the robust map and place all other markers lying below the current node in a bin. The longest sequence is thus representative of this bin.

2.4 Simulations

We used computer simulations to evaluate the properties of the map posterior distribution obtained via our MCMC algorithm. We studied different parameters that can affect the uncertainty in the mapping process. We tried to adapt our simulations to mimic the amount and characteristics of data that is obtained from high-density genotyping projects. More precisely and for comparison, the dog map presented below includes 423 markers for a length of 50 Rays and an estimated retention fraction of 22%. In a pig dataset (mentioned below) on the 60K SNP chip, the map of the largest chromosome includes around 5000 markers with sizes of 300 and 500 Rays and retention fractions of 29 and 34%, respectively, for two different RH panels (Servin,B. unpublished data). The simulation parameters we considered are a map size of 20 Rays, 200 markers, 100 hybrids, a retention fraction of 30 and 10% missing data. So our simulated dataset have a similar density (about 10 markers per ray) and retention fraction than real datasets. In addition, we simulated datasets with 0, 5 and 10% of genotyping errors.

For each set of parameters studied, we simulated 1000 RH datasets. For each dataset, starting with the identity permutation as the true order, a reference order was generated by applying eight random rearrangements (inversions and transpositions), leading to an average of 11 breakpoints between the reference order and the map of interest. We then performed an analysis consisting in (i) constructing the comparative RH map using methods of Faraut *et al.* (2007), (ii) running 5000 MCMC iterations, discarding the first 1000 as burning iterations and (iii) analyzing the posterior distribution.

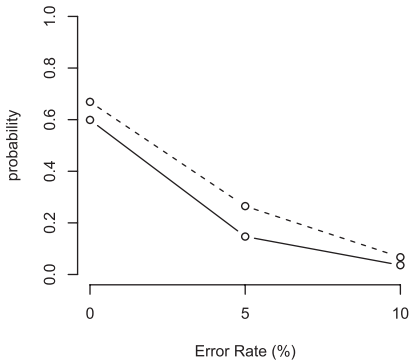

Fig. 2. Probability of recovering the true map as a function of the error rate simulated, using the two-point likelihood (solid line) or the multipoint likelihood (dashed line).

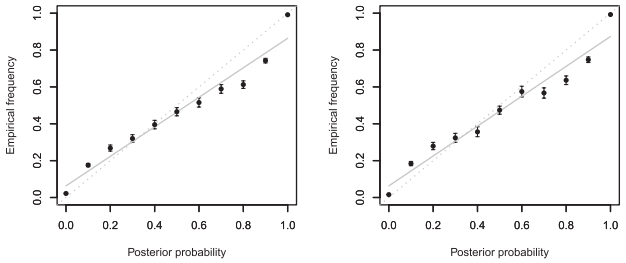

Fig. 3. Frequency of adjacent segments as a function of their estimated posterior probability, for probabilities estimated using the two point likelihood (left) and the multipoint likelihood (right). The results were pooled across datasets of different error rates as the differences were not large (data not shown).

3 RESULTS

3.1 Simulated datasets

For the simulated datasets, Figure 2 shows the probability that the map of highest posterior probability is the true underlying map, as a function of the error rate simulated. We can see that when no errors were simulated, this probability is not negligible (around 0.6). It does not reach a higher probability because of the 10% missing data included in our simulations. Simulations performed with no missing data showed that the probability of finding the true map is much closer to 1.0 (data not shown). This being said, the key information from Figure 2 is that the probability of recovering the true map decreases rapidly with increased error rate. Note also that the map that maximizes the posterior probability computed using the multipoint likelihood is more likely to be the true map than the best map found using the two-point approximation (see Appendix 2 and 1 provided in Supplementary Material). This is a first advantage of sampling from the map distribution using MCMC: we can examine many maps and have a criterion (the posterior probability) to rank them coherently. It is important to note that, for datasets with errors, the probability of finding the true map is quite low. Furthermore, as the panel size, and hence its resolution, is typically fixed and cannot be increased, in contrast to genetic mapping where gathering more data potentially increases the number of informative meiosis, we would expect the probability of finding the true map to be even lower when considering datasets with more markers. This shows that focusing on a single best map for the inference may be too simplistic. We show in the following how to build a more reliable inference by exploiting the map distribution.

Performance of the MCMC sampler Our MCMC algorithm provides us with a distribution of maps that, if the algorithm has converged, should be a sample from the map posterior distribution. Although on small datasets (<10 markers), we found that the posterior probabilities of the maps were well estimated (data not shown), it is not sufficient to insure that this is true for large datasets. As large datasets are typically what interests us here, we had to rely on other methods to evaluate the quality of the sample produced. Indeed, even for datasets of modest size, the very large number of possible orders prevents us from comparing our map sample to the true map distribution. We compared instead inferences based on the sample to the underlying true ordering of markers, known in our simulations. Specifically, we used the adjacency probabilities of markers, derived from the map distribution. If the map distribution is correctly estimated by our algorithm, we should have, among the pairs of markers that have been assigned a posterior probability of adjacency of p, a proportion of p truly adjacent markers. Considering all pairs of markers in all datasets, we grouped them into bins according to their posterior probability of adjacency. Then, within each bin, we computed the frequency of truly adjacent segments. Figure 3 illustrates how the posterior probability of adjacency assigned to pairs of markers correlates with the frequency of adjacent markers, for inferences based on the two-point posterior probabilities of the map or the multipoint probabilities. There is a very high correlation between the estimated probability of adjacency and the empirical frequency (gray line on Fig. 3), suggesting that our MCMC algorithm provides a good mean to estimate the map distribution. The small bias observed tends to be negative for high probabilities and positive for small probabilities. However, for very extreme probabilities (>0.95 and <0.05), the probabilities are very well calibrated.

Robust maps Figure 4 presents the characteristics of the robust maps obtained on our simulated datasets. The figure is divided in three plots, one for each of the error rate simulated. The top part of a plot shows the size of the longest increasing sub-sequence (LIS) of the robust map compared with the true order as a function of the number of markers on the robust map. Recall that the true order is the identity permutation so that the size of the LIS provides a measure of the distance of a map to the true ordering of markers. The open circles on the top plot corresponds to robust maps with a perfect ordering of the markers. The bottom part shows the distribution of the number of markers on the robust maps across the datasets. For datasets with no errors, the vast majority of the robust maps include more than 190 markers out of 200, and all of them are perfect or nearly perfect maps. In only 7 simulations out of 1000, the uncertainty in the data is such that the robust maps include less than 150 markers. For datasets with 5% errors, the uncertainty in the marker ordering is higher and robust maps tend to include less markers. However, they are usually still good maps (i.e. with LIS close to its maximum value). In a very few cases, the LIS is far from its ideal value, denoting an incorrect ordering of markers. For datasets with 10% errors, the number of markers on the robust maps is vastly decreased with a lot of cases where the robust map contain less than 10 markers,

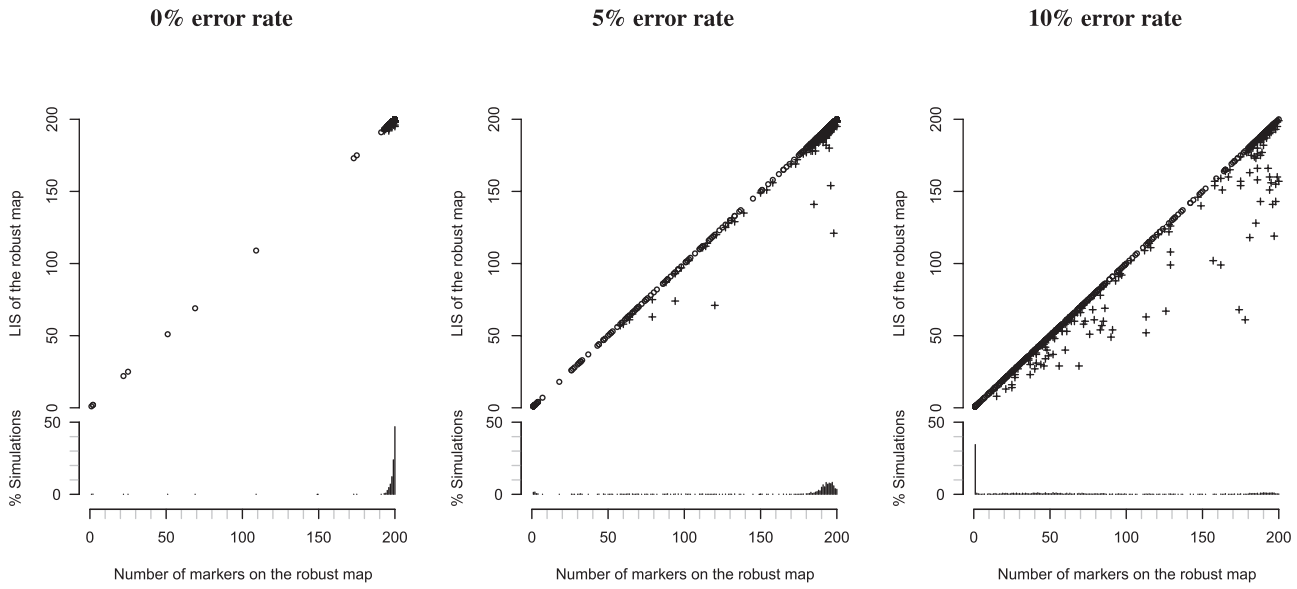

Fig. 4. Quality of the robust maps produced, for 0, 5 and 10% error rates. On each plot, the top part shows the longest increasing sub-sequences of the robust map as a function of the number of markers on the robust map. The open circles correspond to robust maps with a perfect ordering of markers and crosses to robust maps with at least one error. The bottom histogram shows the distribution of the number of markers across all simulations.

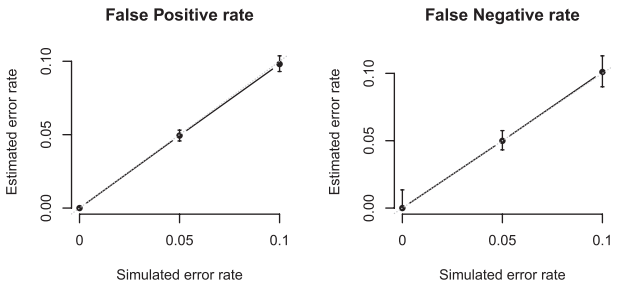

Fig. 5. Estimation of the error rates in the simulations. The error model is estimated conditional on the map of highest posterior probability. The dots represent the median of the estimates across simulations and the bars the 0.05 to 0.95 quantiles.

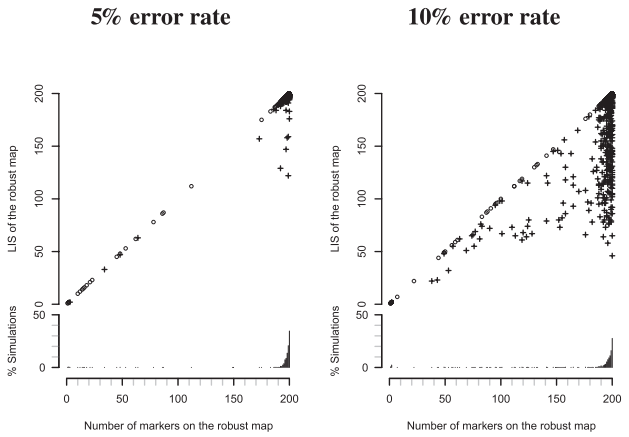

Fig. 6. Robust maps constructed after genotype imputation.

meaning essentially a failure to construct useful robust maps. The number of simulations where the robust map has a very bad ordering of markers is also increased but not dramatically (recall that there are 1000 points on the plot and only a few tens can be identified as being clearly problematic).

Modeling genotyping errors As we observed that the ability to produce useful robust maps is impaired when the error rate in the data increases, we tried to improve our inferences by incorporating an error model (see Appendix 2 provided in Supplementary Material for a detailed description of the model). One issue of the error model is that it has to be fitted conditional on a marker ordering, which is precisely unknown in our map construction process. Our experiments show that an approximate order, the maximum likelihood order for example, enables to approximate error rates satisfactorily (Fig. 5 and Supplementary Fig. S1). The error model provides a posterior probability of error for each genotype (presence or absence) and therefore a mean to correct genotypes. Here again the order is crucial for an appropriate imputation. We observed that using the expected posterior probability of error with respect to the map distribution was an efficient approach to improve the estimation of the map distribution and the corresponding robust map (see Appendix 2 provided in Supplementary Material for the details of the imputation protocol).

A new map distribution and associated robust map is subsequently computed using the imputed data. The results are shown on Figure 6. For datasets with a 5% error rate, the results show a clear improvement, with an increased number of markers in the robust maps and large LIS values. For datasets with a 10% error rate, although many robust maps contain a large number of markers, many of them have very low LIS suggesting that the noise introduced by the high error rate prevents from recovering sufficient signal for ordering the markers.

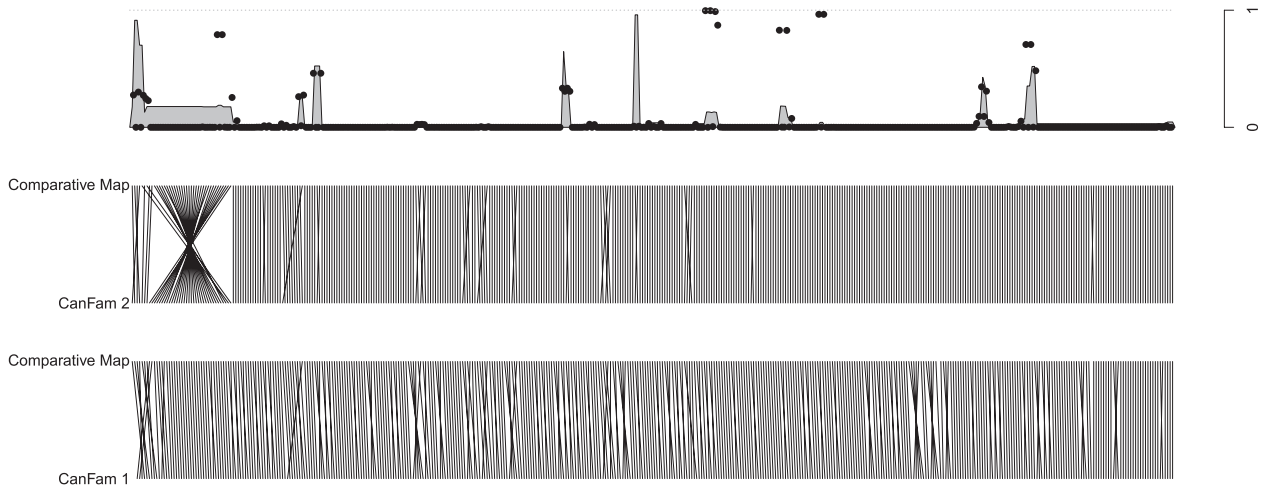

Fig. 7. Comparison of the marker ordering of the comparative RH map with the first draft of the dog genome (bottom figure) and the second draft (top figure). For each marker, a segment is plotted between its respective positions on the orders considered. The top plot shows two measures of uncertainty on the comparative map: the probability that two consecutive markers are not adjacent (points) and the probability that the map is inverted relative to the best map (gray area). These probabilities range from 0 to 1 as indicated by the scale on the right hand side of the plot.

3.2 Application to the validation of the assembly of dog chromosome 2

We tested our algorithm in an *a posteriori* validation of the dog genome assembly using RH data on 423 gene-based markers located on dog chromosome 2, typed on the RHDF9000 dog RH panel (Hitte *et al.*, 2005), and used the human genome as a reference order. We compared our inference both to a first version of the dog genome assembly (CanFam1) and a later version (CanFam2). The Canfam2 assembly is an improvement of the Canfam1 7.5X assembly that is essentially the result of algorithmic improvements made to the ARACHNE assembler leading to the interim ARACHNE2+ assembler [see Supplementary Material 2 of Lindblad-Toh *et al.* (2005)].

The map distribution for these data contains more than 10 000 maps. The false positive error rate estimate is quite low (0.004) and the false negative rate estimate is 0.09. Figure 7 compares the map maximizing the posterior probability using the multipoint likelihood to CanFam1 (bottom) and CanFam2 (top). To illustrate the uncertainty present in the data, we added the posterior probability that two markers are *not* adjacent for each interval between successive markers on the best map (points on the figure) and the posterior probability that a segment of the map is inverted as compared with the most probable map (gray area on the figure). The metamap representation of the uncertainty is provided on the `metamap` software web page.

Comparison between the RH map and the first version of the whole-genome assembly was initially performed by Faraut *et al.* (2007) which showed that the comparative mapping approach provided a map that is more colinear to the genome assembly than regular RH mapping. Looking now at the later version of the dog genome (top), it is striking to see that a lot of discrepancies with the first version of the genome assembly do not exist anymore. And perhaps even more importantly, most of the discrepancies remaining correspond to regions where other possible orderings are seen in the map distribution. This is true in particular at the left end of the chromosome. This can be exemplified further by looking at the comparison of CanFam1 and CanFam2 with the robust map constructed from the map distribution. The robust map is composed of 353 markers out of the original 423. A great number of discrepancies can be identified between CanFam1 and the robust map constructed using our new approach (Fig. 8, top), while most of them have disappeared when compared with a more recent version of assembly (Fig. 8, bottom). Only nine discrepancies remain between the robust map and CanFam2. Two reasons can explain these discrepancies (i) an incorrect assembly ordering of these genes or (ii) an error in the robust map. We found that for all nine discrepancies remaining, the robust map ordering is the same as the Human order, indicating that the RH data do not support a rearrangement between human and dog in these regions, although it is an hypothesis that could be tested by other means. In conclusion, the results of this *a posteriori* analysis show how our methods to construct a robust map could have been of great interest to validate and suggest amelioration to CanFam1.

4 DISCUSSION

Radiation hybrid mapping is a powerful tool to facilitate the localization of genes of interest on animal genomes. Integrating comparative genomics into RH mapping allows to further improve the mapping process by exploiting the important colinearity (conserved synteny) between genomes of (closely) related species. In this work, we present methods that exploit this model to help in the process of producing robust whole-genome assemblies. We exploit a comparative model which has the key property that only when the RH data is informative enough with respect to a different ordering than that of the reference order will an alternative order be accepted. As a consequence, our approach allows to pinpoint regions where the assembly order disagree with an RH map order that is strongly supported by the RH data.

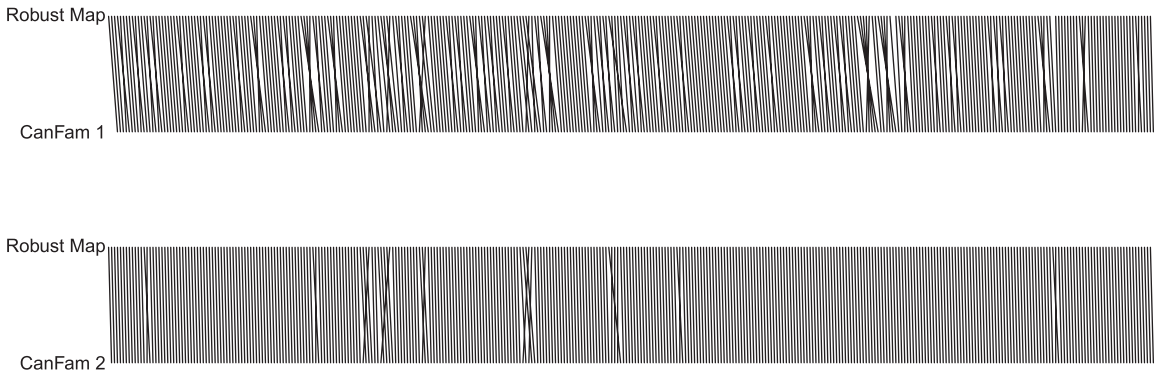

Fig. 8. Comparison of the ordering of the robust map derived from the map distribution with CanFam 1 (top) and CanFam2 (bottom).

As was previously explained by Lunetta *et al.* (1995), we found that the main impact of the presence of genotyping errors was to increase the uncertainty in the marker ordering. This can be explained by the resulting inflation of markers distances, corresponding to higher estimates of breakage probability. The main consequence is that robust maps tend to have a smaller number of markers. However, the quality of the robust maps, which we measured by the LIS criterion, is only marginally impacted. This shows how estimating the map uncertainty allows to take into account and to some extent overcome the problems created by genotyping errors. To further reduce the uncertainty in the maps when errors are present, we propose to combine an error model with the map distribution. We observed, however, that for high error rates (10% in our simulations), using imputations was not enough to produce good-quality robust maps. Indeed, it tended to produce robust maps with a higher number of markers but low LIS, i.e. a biased inference. Our explanation for these results is that the initial map distribution, which we use for imputation, is too far from the true distribution to provide accurate imputation of the genotypes, and consequently good robust maps. In particular, we observed that some of the markers tend to have a very large number of imputations, most of them being erroneous (see Supplementary Fig. S3). This is in great contrast with imputation on datasets with a 5% error rate and explains the differences in performance in these two cases. More generally, our results stress the fact that the estimation of the map distribution is greatly impacted by markers with a large number of errors. For real data, care as to be taken for including only markers for which genotyping is *a priori* of good quality. The criteria to identify good-quality marker will depend on the genotyping method used. For SNP chips, our experience is that genotyping controls (markers and non-irradiated genome) and using conservative thresholds for genotype calling from the raw data are quite important. Considering that today the density of markers provided by SNP chips is larger than the one that the resolution of RH panels allows to separate, working with a subset of high-quality markers is not problematic. This being said, it is reassuring to note that even large error rates can be estimated confidently with the error model (Supplementary Fig. S1). This has the practical consequence of providing a mean to quantify the level of confidence that can be placed in the data to produce good RH maps.

The robust maps presented here correspond to a subset of markers with an order common to all the maps in the distribution. We can note that a search for the longest common subsequence of all the maps would lead to the same subset of markers. The notion of robust maps, however, can easily be extended, using the inclusion tree as a guideline, to a subset of markers with preserved order in a controlled proportion of the distribution (e.g. 99.9%). This last notion is very close to the notion of framework maps well known in the context of RH mapping (Agarwala *et al.*, 2000; Ben-Dor and Chor, 1997; Schaffer *et al.*, 2007). The metamap representation of the distribution of maps could therefore eventually lead to new approaches for the construction of framework maps.

A classical problem when using MCMC is to make sure that the chain has converged and provides samples from the target distribution. Our main argument to show that the MCMC chains have converged in our simulation is the good calibration of the posterior distributions of the probability of adjacency. We obtained this result while performing a relatively modest number of iterations but in the same time using proposal distributions that tend to propose moves with high probability of acceptance (see details in Appendix 1 provided in Supplementary Material). On the dog data, we checked that our results were consistent across several MCMC runs, indicating that we have performed a sufficient number of iterations. In previous Bayesian approaches to RH mapping, the authors were dealing with much smaller datasets, and incorporated the estimation of breakage probabilities and retention fraction into the model. With large datasets, this becomes computationally impractical and so we relied on simple points estimates for these quantities. Although trying to incorporate an update of these parameters without increasing too much of the computation time could be interesting, our results show that good robust maps can be obtained with our more simple approach.

An important aspect of our model is the incorporation of a reference order as a prior. For the sampling of maps via the MCMC algorithm, we found that this was particularly true. When running the algorithm on the RH data without a reference order (which corresponds to specifying a uniform prior on all possible orders), the number of maps visited becomes huge after very few iterations. Using comparative information allows to reduce the space of orders visited to those that are compatible with the mechanisms of chromosome evolution known to preserve large conserved segments between genomes of related species. In our example of application in the dog, we used the human genome as a reference. This was possible in large part because the markers used were gene coding sequences

that are relatively easily mapped on a reference genome, due to their high level of conservation between species. When the markers used are SNPs such as those available on chips, using a reference genome as a prior order is more complicated. Indeed, sequences that define SNPs on the chips are short and not necessarily within conserved regions between genomes. In this case, we can use an assembly draft, for which position of many SNPs are available, as a reference order. This approach is currently used for the validation and improvement of the pig genome assembly (Swine Genome Sequencing Consortium, manuscript in preparation). On this dataset, we were able to apply our methods for the construction of maps with as much as 5000 markers on one chromosome, using data on 180 radiation hybrids.

The MCMC algorithm used to obtain the map distribution is implemented within Carthagene (de Givry *et al.*, 2005). The methods to construct the metamap and produce robust maps from the map distribution are included in the `metamap` software.

ACKNOWLEDGEMENTS

We thank three anonymous referees for their insightful comments. The authors are grateful to Christophe Hitte (CNRS-UMR6061) for kindly providing the dog RH dataset as a case study. The simulations were performed on the cluster of the Toulouse Bioinformatics Platform.

Conflict of Interest: none declared.

REFERENCES

Agarwala,R. *et al.* (2000) A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res.*, **10**, 350–364.

Applegate,D.L. *et al.* (2007) *The Traveling Salesman Problem: A Computational Study*. Princeton University Press.

Ben-Dor,A. and Chor,B. (1997) On constructing radiation hybrid maps. *J. Comp. Biol.*, **4**, 517–533.

Bergeron,A. *et al.* (2008) Computing common intervals of k permutations, with applications to modular decomposition of graphs. *SIAM J. Dis. Math.*, **22**, 1022–1039.

de Givry,S. *et al.* (2005) CARHTA GENE: multipopulation integrated genetic and radiation hybrid mapping. *Bioinformatics*, **21**, 1703–1704.

Faraut,T. *et al.* (2007) A comparative genome approach to marker ordering. *Bioinformatics*, **23**, 50–56.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer.

George,A.W. (2005) A novel Markov chain monte carlo approach for constructing accurate meiotic maps. *Genetics*, **171**, 791–801.

Heath,S.C. (1997) Markov chain Monte Carlo methods for radiation hybrid mapping. *J. Comput. Biol.*, **4**, 505–515.

Hitte,C. *et al.* (2005) Facilitating genome navigation: survey sequencing and dense radiation-hybrid gene mapping. *Nat. Rev. Genet.*, **6**, 643–648.

Landau,G.M. *et al.* (2005) Gene proximity analysis across whole genomes via PQ trees. *J. Comput. Biol.*, **12**, 1289–1306.

Lange,K. and Boehnke,M. (1992) Bayesian methods and optimal experimental design for gene mapping by radiation hybrids. *Ann. Hum. Genet.*, **56**(Pt 2), 119–144.

Lewin,H.A. *et al.* (2009) Every genome sequence needs a good map. *Genome Res.*, **19**, 1925–1928.

Lindblad-Toh,K. *et al.* (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, **438**, 803–819.

Li,R. *et al.* (2010) The sequence and de novo assembly of the giant panda genome. *Nature*, **463**, 311–317.

Lunetta,K.L. *et al.* (1995) Experimental design and error detection for polyploid radiation hybrid mapping. *Genome Res.*, **5**, 151–163.

Schaffer,A.A. *et al.* (2007) rh_tsp_map 3.0: end-to-end radiation hybrid mapping with improved speed and quality control. *Bioinformatics*, **23**, 1156–1158.