USING PIG BACTERIAL ARTIFICIAL CHROMOSOMES TO VALIDATE THE CONSTRUCTION OF THE VIRTUAL SHEEP GENOME

A. Ratnakumar and B.P. Dalrymple

CSIRO Livestock Industries, St Lucia, QLD 4067, Australia, International Sheep Genomics Consortium (http://www.sheephapmap.org)

SUMMARY
The virtual sheep genome (VSG) was created using comparative genomics to position and orient sheep bacterial artificial chromosomes (BACs) on the framework of the human genome. Segments of the human genome defined by overlapping sheep BACs were then reordered and reoriented into their predicted order in the sheep genome to create the VSG. The VSG allows the prediction of the gene order across the sheep genome and the gene content of almost half the BACs in the CHORI-243 BAC library. However, the same absence of a sequenced sheep genome that drove the construction of the VSG also means that it is difficult to know whether the BAC assignments are in fact correct. This prompted us to test our so called “MegaBAC” method, which requires only BAC-end sequence (BES) data, using an independent dataset. This paper describes the use of pig BAC sequences to compare the locations of pig BACs positioned on the human genome using the MegaBAC method, to the positions of the pig BACs determined using the partial or full sequence of the same BAC. Of the 1059 BACs positioned by both methods, we initially identified 33 putative false position BACs. However, after detailed analysis the false position BACs proved to be BACs which either had had sequences misidentified during the sequencing process or with GenBank entries that appeared to contain sequence derived from two different BACs. In conclusion, we did not find a single genuine discrepancy, thus confirming the high level of accuracy of the MegaBAC method.

INTRODUCTION
The VSG (http://www.livestockgenomics.csiro.au/vsheep/) was created to provide the sheep research community with the best possible surrogate for a real sheep genome sequence. The construction of the VSG used a novel bioinformatics method to maximise the utilisation of the information contained in the sheep BES from the CHORI-243 sheep BAC library and the sequenced genomes of the dog, human and cow (Dalrymple et al., 2007). The MegaBAC approach enabled 84,624 sheep BACs to be positioned on the human genome with their BESs in a tail-to-tail structure, i.e. the way in which they were originally located in the sheep genome. The BESs of a BAC are determined such that the left end sequence is oriented 5' to 3', and the right end sequence is oriented 3' to 5', relative to the inserted DNA fragment in the BAC. The 3' ends of the sequence reads are designated the tails and the 5' ends the heads of the BESs. In a tail-to-tail BAC the 3' ends of the two BESs are internal to the BAC (Figure 1). A BAC positioned in the human genome with BESs tail-to-tail is assumed to have high synteny to the equivalent region of the human genome, i.e. contains the orthologous genes and only small rearrangements. Overlapping tail-to-tail BACs identify regions of extended synteny between the genomes. The more BACs that are positioned tail-to-tail the larger the regions of synteny, and the smaller number of regions to be reordered. One advantage of the MegaBAC method is that it allows the positioning of many more BACs than conventional methods, which significantly aided the construction of the VSG. The false position rate of the BACs positioned by the MegaBAC
method estimated using randomisation of the positions of the BESs was found to be 0.056%. The absence of a sequenced sheep genome means that it is very difficult to validate this low false positive rate estimate in practice. However, the availability of BAC-end sequenced and partially or fully sequenced BACs from a pig BAC library provided an opportunity to validate this method with an independent dataset.

MATERIALS AND METHODS
The BES and other sequencing data for the CHORI-242 pig BAC library were downloaded from NCBI on the 31st of January, 2007. There were 1962 BACs either fully sequenced, or with sequencing underway, with corresponding BESs. The set of 1655 BACs with at least 2 BESs associated with them were used for the analysis described below. The BESs were aligned with the human (hg17), dog (canfam2) and cow (Btau2.0) genome sequences, and the human genome was used as the assembly framework for the MegaBAC analysis as previously described (Dalrymple et al., 2007). The set of partially and fully sequenced pig BACs were positioned on the human genome with BLAT using the default parameters. The human chromosome and location of the top hit was taken as the position of the BAC. At least one base overlap between the positions determined using BLAT and the MegaBAC method was used to discriminate between congruent and non-congruent positions. BACs with non-congruent positions were flagged as BACs with potential false positions. MegaBLAST (-D 3 -W 32) was used to align the full set of BESs (315,982) against the set of pig BACs. Neither set of sequences was repeat masked and only hits with an e-value of 0.0 were retained. A map of the pig genome based on the restriction enzyme fingerprint patterns of ~270k CHORI-242 pig BACs is also available (http://www.sanger.ac.uk/Projects/S_scrofa/mapping.shtml). This map provided us with another set of position assignments for the BACs, and their relationship to overlapping and adjacent BACs, that is largely independent of the BESs.

RESULTS AND DISCUSSION
MegaBAC analysis using pig BACs. 1059 (64%) of the 1655 pig BACs analysed were positioned by both the MegaBAC and BLAT approaches (Figure 1A). Of these 33 BACs were flagged as potentially false position BACs, i.e. the BAC location based on the BESs did not overlap with the BAC location based on the internal sequence (Figure 1B). The false position rate of 3.1% is much higher than the calculated rate of 0.056% for sheep BACs (Dalrymple et al 2007). Each of the false position BACs was then investigated to determine if the MegaBAC approach was at fault or if there were other explanations. There are several possible reasons why BACs could appear to have false positions from the MegaBAC analysis: the BES-based position is incorrect (initially thought to be the most likely case); the BLAT-based position is incorrect (probably very unlikely, see Figure 1C); the BAC is chimeric; the identity of the BAC is not the same in the end-sequencing and full sequencing projects (Figure 1G); and more than one BAC is represented in the sequences deposited in GenBank with the same identifier (Figure 1H).

Procedure for identifying underlying causes of the putative false positions. Upon close inspection of the set of putative false position BACs one pattern became apparent very quickly. This was that the genomic location at which the MegaBAC analysis had placed a BAC seemed to correspond quite well to the genomic location that a different BAC in the putative false position set had been positioned using BLAT (five pairs of BACs). In most of these cases a reciprocal relationship between two such BACs was observed (Figure 1G). Alignment of the BAC end sequences and the putative
reciprocally swapped BACs confirmed that the most likely explanation for the relationship was that BACs had been swapped at some point in the sequencing process. The positions of the apparently swapped BACs in the fingerprint map also supported the reciprocal swaps.

Figure 1. Diagrammatic representation of the observed and potential relationships between pig BACs and the human genome. BESs are represented by short horizontal arrows flanking the longer inserted sequence. A tail-to-tail BAC is shown diagrammatically as -> -------- <-. Solid and dashed lines represent matches to different regions of the human genome. Matches from BACs flanking the putative false position BAC in the fingerprint map are also shown in parts A-F.

Such swaps may have occurred between BACs in our set and other BACs, including those being sequenced, but not positioned in the MegaBAC analysis and BACs not currently being sequenced. The full set of BESs was then aligned to the full set of BACs with partial or complete sequences. An additional set of nineteen BACs potentially swapped with BACs other than the putative false positive BACs were identified (Figure 1G). For the remaining BACs, the correct BESs were clearly present in the partially or fully sequenced BACs. Analysis of the matches of these BACs to the human genome identified that they all contained highly significant matches to two distinct regions, the second best of which was consistent with the position predicted by the MegaBAC analysis. These BACs could
represent junction points of rearrangements between the human and pig genomes, although in theory they should not be mapped by the MegaBAC method. In this case the flanking BACs should map to the two different regions of the human genome identified in the sequence searches (Figure 1D). But in all cases the flanking BACs mapped to the same region of the human genome. These BACs could be potential chimeric BACs, although again they should not be mapped by the MegaBAC method as they would not have fingerprint patterns consistent with other BACs in the region (Figure 1E). There is no evidence that any of the putative false position BACs are chimeric. Further possibilities are that these BACs contain a segment of the pig genome, that in the pig matches the human genome in a different region from the flanking sequences and regions (Figure 1F), or that during the sequencing data from two or more BACs has been accidentally combined (Figure 1H). Comparison of the length of the deposited sequence with the length of the BACs determined from restriction enzyme fingerprinting indicated that without exception the GenBank accessions for these BACs contained substantially more sequence than was consistent with the physical length of the BACs. Overall the weight of evidence supports the presence of DNA sequences from more than one different BAC in the GenBank accession that was used as the source of sequence for this analysis. Thus all of the putative false positive BACs have now been ruled out as true false position BACs (Table 1).

Table 1. Number of putative false position BACs in the groups

<table>
<thead>
<tr>
<th>Groups</th>
<th>Number of BACs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swapped BACs, mainly reciprocal swaps</td>
<td>19</td>
</tr>
<tr>
<td>Multiple BACs, correct BAC present</td>
<td>10</td>
</tr>
<tr>
<td>Multiple BACs, possibly also swapped</td>
<td>4</td>
</tr>
</tbody>
</table>

CONCLUSIONS
The false position rate for the pig BACs using the MegaBAC approach is less than 0.1%, which is close to the theoretical rate calculated from the randomisation of the sheep BAC-ends. Although the pig BACs may represent a somewhat biased sample, these analyses suggest that the MegaBAC method of assignment of BACs to the framework of the human genome has a very low false position rate. This analysis also identifies problems that we will need to be aware of when sequencing the sheep genome based on the assignment of BACs using the MegaBAC method. The issues identified here have been communicated to the sequencing centre concerned.

ACKNOWLEDGMENTS
Sean McWilliam and Wes Barris for their contributions to the construction of the VSG. This work was supported by SheepGenomics which is an initiative of Australian Wool innovation Limited and Meat & Livestock Australia together with 11 leading research organisations in Australia and New Zealand. This is Sheep Genomics publication number 79.

REFERENCES