Supporting Information

Reference-Assisted Chromosome Assembly

Jaebum Kim, Denis M. Larkin, Qingle Cai, Asan, Yongfen Zhang, Ri-Li Ge, Loretta Auvil, Boris Capitanu, Guojie Zhang, Harris A. Lewin, and Jian Ma

Supplementary Methods

Posterior probability of a SF adjacency

The posterior probability of a SF adjacency was calculated by examining the evolution of a SF configuration along a given phylogenetic tree. Here, we developed a probabilistic framework for computing the posterior probabilities of SF adjacencies in a descendant's genome given the SF adjacencies in genomes of other related species. Suppose we want to compute the posterior probabilities of SF adjacencies in the target genome T with R as a reference genome, and O_1 and O_2 as outgroup genomes (Fig. S4A). Then, we reroot the tree by creating a new root A_0 on the branch between A_1 and T, and compute the posterior probabilities in A_0 (Fig. S4B). By setting the branch length $t(A_0T) = t(A_1R)$, and $t(A_0A_1) = 0$, we can incorporate the adjacency information both from the target genome T and the subtree rooted by the ancestral genome A_1.

Specifically, if there exists a SF b_i in a rerooted target genome T, the predecessor $p_T(i)$ and the successor $s_T(i)$ are defined as the *signed* index of a SF that immediately precedes and succeeds b_i on the same chromosome, respectively. When b_i appears first (or last) on a chromosome, we set $p_T(i) = \$$ (or $s_T(i) = \$$). When $p_T(j) = i$ and $s_T(i) = j$, we call b_i and b_j are adjacent in the genome T, i.e. $A_T(i, j) = 1$. We use the following approximation to calculate the posterior probability of adjacency (i,j) as:

$$Prob\ (i,j) = P(A_T(i,j) = 1|D_T) = P(p_T(j) = i|D_T)P(s_T(i) = j|D_T) \quad \text{(S1)}$$

where $P(p_T(j) = i|D_T)$ and $P(s_T(i) = j|D_T)$ are posterior probabilities that i precedes j, and j succeeds i, respectively. D_T is all the observed data in all leaves of the subtree rooted by the genome T. Using the Bayes' theorem, the posterior probability that i precedes j is defined as follows by assuming equal prior probabilities $P(p_T(j) = i)$.

$$P(p_T(j) = i|D_T) = \frac{P(D_T|p_T(j) = i)}{\sum_k P(D_T|p_T(j) = k)} \quad \text{(S2)}$$

If there are two child genomes *L* and *R* of the parent genome *T*, then the likelihood $P(D_T|p_T(j) = i)$ can be calculated recursively as:

$$P(D_T|p_T(j) = i) = P(D_L|p_T(j) = i)P(D_R|p_T(j) = i) \quad \text{(S3)}$$

$$= \sum_k P(D_L|p_L(j) = k)P(p_L(j) = k|p_T(j) = i) \quad \text{(S4)}$$

$$\times \sum_k P(D_R|p_R(j) = k)P(p_R(j) = k|p_T(j) = i)$$

where $P(p_L(j) = k|p_T(j) = i)$ is the probability that the preceding SF of *j* has changed from *i* to *k* in the course of evolution from *T* to *L*, which can be represented by the extended Jukes-Cantor model for breakpoints (1).

In a given phylogenetic tree, suppose *T* is a parent and *L* and *R* are the left and right child nodes of *T*. Based on the extended Jukes-Cantor model, the probability that the preceding SF of *j* has changed from *i* to *k* in the course of evolution from *T* to *L* can be represented as:

$$P(p_L(j) = i|p_T(j) = i) = \frac{1}{2n-1} + \frac{2n-2}{2n-1}e^{-(2n-1)\mu t_{TL}} \quad \text{(S5)}$$

$$P(p_L(j) = k|p_T(j) = i) = \frac{1}{2n-1} - \frac{1}{2n-1}e^{-(2n-1)\mu t_{TL}} \quad \text{(S6)}$$

where *n* is the total number of SFs, μ is the rate parameter, and t_{TL} is the branch length between *T* and *L*. We assumed that the rate μ is the same across all tree branches and the branch lengths are given. At a leaf genome *Z*, the likelihood $P(D_Z|p_Z(j) = \text{i})$ is defined as:

$$P(D_Z|p_Z(j)=i)=\begin{cases}1 & \text{if } p_Z(j)=i \text{ in } Z\\ 0 & \text{otherwise}\end{cases} \quad \text{(S7)}$$

Then, the probability that the preceding SFs of a SF j are the same in the nodes L and R is:

$$\begin{aligned}
P(p_L(j)=\ p_R(j)=i) &= \sum_{k\neq j,-j} P(p_L(j)=i|p_T(j)=k)\,P(p_R(j)=i|p_T(j)=k)\\
&= P(p_L(j)=i|p_T(j)=i)P(p_R(j)=i|p_T(j)=i)\\
&\quad +(2n-2)P(p_L(j)=i|p_T(j)=k)P(p_R(j)=i|p_T(j)=k)\\
&= \left(\frac{1}{2n-1}+\frac{2n-2}{2n-1}e^{-(2n-1)\mu t_{TL}}\right)\left(\frac{1}{2n-1}+\frac{2n-2}{2n-1}e^{-(2n-1)\mu t_{TR}}\right)\\
&\quad +(2n-2)\left(\frac{1}{2n-1}-\frac{1}{2n-1}e^{-(2n-1)\mu t_{TL}}\right)\left(\frac{1}{2n-1}-\frac{1}{2n-1}e^{-(2n-1)\mu t_{TR}}\right)\\
&= \frac{1}{2n-1}+\frac{2n-2}{2n-1}e^{-(2n-1)\mu(t_{TL}+t_{TR})}
\end{aligned} \quad \text{(S8)}$$

Since,

$$P\big(p_L(j)\neq\ p_R(j)\big)=1-P\big(p_L(j)=\ p_R(j)\big)=\frac{2n-2}{2n-1}\big(1-e^{-(2n-1)\mu(t_{TL}+t_{TR})}\big) \quad \text{(S9)}$$

Therefore,

$$\begin{aligned}
\mu &= -\frac{1}{(2n-1)(t_{TL}+t_{TR})}\ln\left(1-\frac{2n-1}{2n-2}P\big(p_L(j)\neq\ p_R(j)\big)\right)\\
&\approx -\frac{1}{(2n-1)(t_{TL}+t_{TR})}\ln\left(1-\frac{2n-1}{2n-2}\frac{d(L,R)}{n}\right)
\end{aligned} \quad \text{(S10)}$$

where $d(L,R)$ is a breakpoint distance between L and R. Here, we assumed that the probability $P\big(p_L(j)\neq\ p_R(j)\big)$ can be approximated by $d(L,R)/n$, which is the number of breakpoints per synteny blocks.

Computation of $N_{ir}(i,j)$

To compute the number of paired-end reads that link two SFs b_i and b_j from two different scaffolds with the orientation indicated by the signs of i and j in (i, j), we retained those paired-end reads that the mapping of two end reads was within a certain range (length of insert library size plus two standard deviations). The score $N_{ir}(i, j)$ is for two SFs from two different scaffolds. Therefore, each end read is mapped to a different scaffold, and the mapping distance between them is defined as the sum of the distances from the mapped positions to one of two ends of that scaffold that is determined by the orientation of the mapped end read (Fig. S3B-C).

Computation of $P_{ia}(i, j)$

This score is for two SFs from the same scaffold that could represent a potential breakpoint in that scaffold. For each position of a scaffold, the number of paired-end reads that span that position is computed by only considering the paired-end reads that the mapping distance between two end reads is within +/- two standard deviations of an insert library size. Then, for each region between two SFs including L_f up and down flanking regions, the average coverage for each window of length L_w that overlaps $L_w/2$ and covers the entire region is obtained. In our case, we used 50 Kbp and 1 Kbp as the values of L_f and L_w, respectively. The ratio of the coverage for each window with respect to the average coverage across the entire set of scaffolds is subsequently computed, denoted as p_a. The score $P_{ia}(i, j)$ is the minimum of p_a of all the windows. To include only reliable joins of SFs, a cutoff score was used, which corresponds to the bottom 5% of a background value distribution that was estimated by the p_a values from across all scaffolds. The 50 Kbp up and down flanking regions were included because the precise breakpoint region may not exist in the region between two SFs due to alignment errors.

Link score of a SF adjacency

To compute the link score $Link(i, j)$ of the adjacency of two SFs b_i and b_j, the generic function is defined to consider both $N_{ir}(i, j)$ for two SFs from two different scaffolds and $P_{ia}(i, j)$ for two SFs

from the same scaffold. First, $P_{ir}(i,j)$ is computed, which is the percentage of $N_{ir}(i,j)$ with respect to the average across all possible edges $e(i',j')$. Then, the percentage score $P(i,j)$ is defined as:

$$P(i,j) = \begin{cases} P_{ir}(i,j) & sf(i) \neq sf(j) \\ P_{ia}(i,j) & sf(i) = sf(j) \end{cases} \quad \text{(S11)}$$

where $sf(i)$ is the scaffold to which the SF b_i belongs. The link score $Link(i,j)$ in the range from 0 to 1 is a min-max normalized version of $P(i,j)$ defined as:

$$Link(i,j) = \frac{P(i,j) - \min\left(P(i',j'), \forall i', j'\right)}{\max(P(i',j'), \forall i', j') - \min(P(i',j'), \forall i', j')} \quad \text{(S12)}$$

Parameter estimation for α

The parameter α controls the relative contribution of the posterior probability of an adjacency and the support from the paired-end read mapping. The assessment of those two scores requires reliable benchmarking datasets. To this end, we collected the Mammalian Gene Collection (MGC) genes from the UCSC genome browser (2) that have orthologous genes in both human and cattle, and identified adjacencies of SFs (A_R) whose boundary is spanned by the genes. By assuming that those adjacencies are highly reliable and therefore have maximum posterior probabilities and link scores, we computed the sum of squared errors (SSE) of each score and used those SSEs to adjust the parameter α (more weight to the score with less SSE) as follows:

$$\alpha = \frac{SSE_L}{SSE_P + SSE_L} \quad \text{(S13)}$$

where SSE_P and SSE_L are SSEs of the posterior probability of a SF adjacency and the link score, respectively, which are defined as:

$$SSE_L = \sum_{(i,j)\in A_R} (1 - Link(i,j))^2 \qquad SSE_P = \sum_{(i,j)\in A_R} (1 - Prob(i,j))^2 \tag{S14}$$

Comparison with existing reference-based methods

As many organisms have been sequenced and assembled, we now have a large volume of genome assemblies of species that represent major phylogenetic clades. In this sense, utilizing a reference genome as a guide is a promising approach to address the problems of the assembly algorithms. To this end, Pop et al. (3) proposed a method that aligns reads to a single reference genome and uses the mapping information to assemble the reads into contigs. Gnerre et al. (4) developed an algorithm that improves a *de novo* genome assembly by aligning reads to reference genome sequences, which resulted in both the join of *de novo* scaffolds and the detection of potential mis-assembly regions. In addition, tools such as ABACAS (5), CONTIGuator (6), OSLay (7), and Projector 2 (8) have been developed to extend contigs into longer scaffolds by mapping to a reference genome. Recently, Husemann and Stoye (9) proposed an algorithm that computes the adjacency score of sequence contigs by taking advantage of several genomes of related species and their pairwise relationships in a phylogenetic tree. However, these methods are based on the mapping to only one reference genome, or they use several related genomes with only pairwise comparisons. More recently, Gao et al. (10) proposed a method for an optimal scaffolding problem by utilizing paired-end sequences. However all previous methods are based on the reads mapping to only one reference genome, or they use several related genomes with only pairwise comparisons. Our RACA algorithm is significantly different from these previous methods: (i) we proposed a novel framework to specifically consider the tree topology and branch lengths of the phylogeny when computing the posterior probability of adjacencies in the target genome in the context of genome evolution; and (ii) we proposed a new model to consider both

phylogenetic comparative information and paired-end read mapping. The method can also be used to detect and correct mis-assembled scaffolds.

Construction of simulated genome assemblies

To create simulated genome assemblies, we used the Evolver program (11) that simulates the evolution of genome sequences by creating inter-chromosomal mutations, such as chromosome fission, fusion, and segmental moves and copies, as well as intra-chromosomal mutations, such as substitutions, insertions, deletions, moves, and copies of sequences, in a given length of time (Fig. 1A). We first prepared the sequences of the human chromosome 21 and 22 (NCBI36/hg18 assembly; total length 69 Mbp) and their annotations for sequence elements, such as genes and conserved non-gene elements, as an input of the Evolver program. We chose to base the simulation only on 69 Mbp of the original human sequence in part because the simulation of whole human genome by the program Evolver is impractical due to computational constraints, and in part because 69 Mbp are enough to represent real genome data in terms of the difficulty of the chromosome reconstruction task. We then simulated a reference dataset *R* and 11 target datasets from *D0* to *D10* by using ten different evolutionary divergence times and event rates from the reference dataset *R* (Fig. 1A), based on parameter settings provided by the authors of the Evolver program. The order of the indexes of the datasets represents the relative divergence from the dataset *R*. For example, the dataset *D4* is more divergent than *D2* from *R*. The simulated chromosome sequences in each dataset were then fragmented into multiple short sequence fragments by following the length distribution of Tibetan antelope scaffolds, whose lengths were scaled down (90 %) to allow a reasonable number of fragments in a small number of chromosomes. We generated a total five different sets of sequence fragments for each dataset by repeating this fragmentation, and created a predefined number (6% estimated from Tibetan antelope scaffolds) of chimeric fragments by combining two randomly chosen fragments.

Evaluation by using simulated genome assemblies

We used the sequences of the dataset *R* as reference genome sequences and predicted the order and orientation of the sequence fragments of the target datasets *D0* - *D9* (the remaining dataset *D10* was used only as an outgroup genome) by first building SFs and next applying RACA. For the size limit of SFs, we used 5 Kbp to produce a similar number of breakpoints as compared to those from five real genome assemblies (chimpanzee: panTro2, orangutan: ponAbe2, rhesus: rheMac2, mouse: mm9, and cattle: UMD3.0; downloaded from the UCSC Genome Browser (2)) by using human (NCBI36/hg18 assembly) as a reference with a minimum SF size 150 Kbp (Table S2). To compute the number of breakpoints, we constructed SFs by comparing the reference and target genomes, and counted the number of cases where two SFs are adjacent in the reference genome but not in the target genome. The predicted order and orientation of sequence fragments in each dataset was compared with the true order and orientation that were inferred from the sequences of those datasets (we know which fragment comes from which part of the sequence). To measure the effect of outgroup species, we varied the outgroup species. Specifically, for each target dataset, we used more divergent datasets, for example *D2* - *D10* for the target dataset *D1*, as the genome sequences of outgroup species. We note that there is no paired-end read mapping data and therefore RACA only uses the posterior probabilities of SF adjacencies. As evaluation measures, we used (i) recall, which is the fraction of the true order and orientation of sequence fragments that was found in the predicted sequence fragments, and (ii) precision, which is the fraction of the predicted order and orientation of sequence fragments that agree with the true order and orientation. For each dataset, the evaluation measures were computed for each different set of fragmented sequences (total five), and only averages across those five fragmentations were reported.

Evaluation of RACA using the GAGE data sets

We excluded the ABySS assembly from the GAGE dataset because its scaffold N50 was too small (2.1 Kbp). The GAGE website has three different versions of assemblies that are generated by using original and two error corrected paired-end reads by the ALLPATHS-LG and Quake (12) programs, respectively. Among the three versions, we selected the best assembly and corresponding paired-end reads for each genome assembler based on the reported results in the supplementary material of the GAGE paper (13).

The phylogenetic trees for the two settings we used were ((ponAbe2:0.0183, human:0.0187):0.1331, umd3:0.2195) and ((mm9:0.3526, human:0.1312):0.0207, umd3:0.2195), respectively. The branch lengths were estimated based on the substitution rate, which was obtained from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/multiz46way/46way.corrected.nh.

SF ordering algorithm

We developed a greedy algorithm as an approximate solution to solve the SF ordering problem.

1: **Begin**
2: *C: initially empty set of connected components*
3: **For each** *edge e in descending order of weight ≥ 0.1:*
4: **If** *e is not inconsistent with any previously used edges*
5: **AND** *does not introduce a cycle*
6: **If** *e can be added to any connected component in C*
7: *Add e to that connected component*
8: **Else**
9: *Create a new connected component with e and add it to C*
10: **End**

Whenever a new edge is under consideration, it should be checked whether the order and orientation of two SFs that are implied by the edge are consistent with those already processed (line 4). For example, if the edge (i, j) that connects b_i^h and b_j^h is in consideration, then the head of the SF b_i

and the head of the SF b_j should not be connected to any other SFs previously. Otherwise, the edge (i, j) is ignored (line 4).

Time and memory complexity

The time complexity of the RACA algorithm itself is $O(N^2C)$, where N is the number of SFs and C is the final number of connected components. This is because, for each pair of SFs (the time complexity is $O(N^2)$), the consistency check only takes a constant time due to an efficient implementation by using a hash table, and the cycle check needs to consider all connected components that are constructed so far (the number is less than C) in a constant time also due to an efficient implementation by using a hash table ($C << N$). The memory complexity is $O(N^2 + C)$ because the algorithm needs to store the weights for $O(N^2)$ pairs of SFs and $O(C)$ connected components. When we applied this algorithm to the Tibetan antelope assembly, it took just a few minutes with a very small amount of memory. However, the processing of the paired-end read mapping data and the estimation of the threshold parameters took roughly a day or so on Intel Xeon 2.80 GHz machine with parallel execution of 10 threads. In the case of simulated genome assemblies, it took just a few minutes with a very small amount of memory because we could not use paired-end read mapping.

Whole-genome alignment

We used the UCSC whole-genome alignment pipeline to generate alignments used in this study. For each pair of genomes (e.g. reference genome vs. scaffolds from a *de novo* assembly), we first used LASTZ (14) to perform whole-genome pairwise alignment. Specifically, we used these parameters in LASTZ: M=254, O=600, E=150, K=4500, Y=15000, and T=2. Then we used the UCSC Chains/Nets pipeline (15) to generate pairwise chains and nets to be used as input for RACA to produce SFs. The net alignment can be generally regarded as orthologous regions between two genomes (15).

Construction of syntenic fragments (SFs) between two genome assemblies

Given two genome assemblies (one is a reference and the other is a target), they were first aligned using LASTZ (14). Alignment nets, which are the putative orthologous regions, were then created using the tools downloaded from the UCSC Genome Browser (2). Syntenic fragments (SFs) were constructed by merging co-linear alignments (16), and discarding SFs of size less than a given threshold. Two other programs, SatsumaSynteny (17) and SyntenyTracker (18) were used for aligning the genome sequences and building SFs respectively. The two methods produced highly consistent results but only results obtained with LASTZ and the UCSC tools are reported in the present work.

Reconstruction of the chromosome fragments of a target genome

Based on the predicted order and orientation of SFs, the chromosome fragments of a target genome are reconstructed by concatenating the target scaffolds that the SFs are contained. Gaps of length 100 bp are inserted between two adjacent target scaffolds unless there is a physical overlap ($\geq$ 5bp) between them. If there is more than one SF in the same target scaffold and they are not placed at the same reconstructed chromosome fragment, the target scaffold is split at the middle of the boundaries of two adjacent SFs.

Tibetan antelope assembly

The *de novo* assembly of the Tibetan antelope (*Pantholops hodgsonii*) genome was shotgun sequenced by the Illumina GAII with paired-end libraries up to 20 Kbp insert size and assembled using SOAPdenovo (19) by the Beijing Genomics Institute (http://www.genomics.cn; see Supplementary Table S14 for the details of the insert libraries; the short read archive accession number is SRA052275 (20)).

Paired-end reads mapping for Tibetan antelope

The mapping of paired-end reads to Tibetan antelope scaffold sequences was done by SOAP2 (21) with the maximum of five mismatches (-v 5), -m and –x according to the insert library size, and every other default parameter ("-M 4 -l 256 -n 5 -r 1 -s 255 -g 0 -e 5"). From the mapping of paired-end reads, we only collected uniquely mapped paired-end reads (96.41 % of all mappings; Supplementary Table S14) and used them to reconstruct Tibetan antelope chromosomes.

Reconstruction of Tibetan antelope predicted chromosome fragments

The phylogenetic tree that was used to compute the posterior probability of a SF adjacency is ((umd3:0.0832, panHod2:0.0832):0.0832, hg18:0.2163). The branch lengths were estimated based on the neutral substitution rate between human and cattle (0.3828) (22), and the relative numbers of SFs between cattle and human in comparison with Tibetan antelope. The parameter α was estimated as 0.5 because both the $Prob$ and $Link$ scores performed equally well with the common Mammalian Gene Collection (MGC (23)) genes regions between human and cattle. The estimated value of the rate parameter μ of the extended Jukes-Cantor model was 0.000165, which was obtained by averaging the estimated μ values separately from two pairs of species: (Tibetan antelope, cattle) and (Tibetan antelope, human).

Overlapping Tibetan antelope scaffolds

When reconstructing the Tibetan antelope predicted chromosome fragments (PCFs), we found that there are adjacent Tibetan antelope scaffolds that physically overlap. For example, two Tibetan antelope scaffolds 475 and 693 that were predicted as adjacent by our method have the same sequence of length 37 bp at the ends toward each other. We investigated the physical overlap of Tibetan antelope scaffolds in conjunction with their alignment distances on cattle chromosomes. There were 216 (15%) adjacent pairs of Tibetan antelope scaffolds whose sequences physically overlap (Table S15). This fraction was increased as the two adjacent scaffolds became closer on a cattle chromosome, and 166

(36%) pairs of adjacent Tibetan antelope scaffolds overlapped when the adjacency distance on a cattle chromosome was 0. The physical overlap between two Tibetan antelope scaffolds may be artifacts introduced by an assembly algorithm, and it could be used as a strong indicator of the true scaffold adjacency. In this analysis, we used 5 bp as a minimum overlap size resulting in 0.005 false positive rate in the dataset of randomly chosen pairs (not adjacent in PCFs) of Tibetan antelope scaffolds. We performed PCR to validate the overlapping Tibetan antelope scaffolds. All except one of the PCR primer pairs designed from the overlapping scaffolds produced PCR products of expected sizes (Fig. S5A). The only case when the product was missed is a 12 bp overlap between the scaffold 1177 and scaffold 2280. We generated the second primer pair (non-overlapping with the first pair) for this adjacency that also failed to produce PCR product. Further investigation showed that the overlapping sequence between these two scaffolds contained 1 bp mismatch in the overlapping sequence. Scaffold 2280 starts with "C" (scaffold 1177 has "A" in the corresponding position) followed by 12 bases of an exact match with the end sequence of scaffold 1177. This might indicate that the end sequences of the scaffolds are not really overlapping but duplicated sequences. In total we verified 9 out of 10 (90%) overlaps between predicted adjacent scaffolds in the Tibetan antelope genome.

Chimeric Tibetan antelope scaffolds

Among the 1,434 Tibetan antelope scaffolds that were aligned to the cattle genome (Table S9), 130 (9%) were split into more than one SF. As an example, Fig. S2 shows that Tibetan antelope scaffold 63 was partitioned into two fragments, one of which mapped to the reconstructed PCF 21c_27. These 130 scaffolds may be chimeric or contain authentic Tibetan antelope-specific EBRs (Table 1). Our reconstruction of Tibetan antelope PCFs predicted that 84 out of 130 Tibetan antelope scaffolds are chimeric (6% of the total scaffolds aligned to cattle genome).

Validating predicted adjacencies and mis-assemblies by PCR

We conducted PCR validation for the adjacent but not overlapping Tibetan antelope scaffolds. Out of 14 PCR primer pairs, three pairs (scaffolds 131 and 2887, scaffolds 1041 and 1560, scaffolds 1153 and 1701) produced multiple PCR products of similar intensity and were thus excluded from the further analysis because of lack of specificity (Fig. S5B). The remaining 11 primer pairs (11/14=79%) produced single products, of which four (33.3%) were of the size expected from the gap distance between adjacent Tibetan antelope scaffolds aligned to the cattle genome.

Using PCR we tested two adjacencies within Tibetan antelope scaffolds (63 and 358) that had SFs mapping to two different cattle chromosomes and therefore could contain inter-chromosomal EBRs between the Tibetan antelope and cattle genomes. The selection of these two scaffolds out of 83 was made on the basis of the distance between SFs within Tibetan antelope scaffolds that could be spanned by a PCR product. In both cases we failed to produce PCR products that would confirm adjacency of scaffolds 63 and 358 in the Tibetan antelope genome. Moreover, we generated a PCR product that connects a part of scaffold 63 with another Tibetan antelope scaffold, 321. This adjacency additionally confirms that scaffold 63 is chimeric.

The reason we presented just two examples of how RACA can be used to identify mis-assemblies was the difficulty of producing PCR products that would span large genomic intervals. In such cases the absence of a PCR product is not convincing evidence because the reaction could fail to amplify a long DNA fragment that spans a true adjacency. Therefore, we focused on those examples where we could produce a PCR product of <1000 bp from the Tibetan antelope genome and the primers could be found in non-repetitive sequences. Because our resolution threshold was ≥150 Kb for the SF sizes, and there is high repetitive content within mis-joins or EBRs, there were only two potentially mis-joined regions that could be selected for validation. In both selected examples, we did show that RACA allows for detection and correction of assembly errors.

PCR primer selection

We designed PCR primers for 10 physically overlapping scaffolds, 20 adjacent scaffolds with <200 bp distance in the cattle genome (UMD3.0) and two pairs of adjacent SFs that represent putative inter-chromosomal rearrangements between Tibetan antelope and cattle genomes. Repeats within Tibetan antelope sequences were masked with the RepeatMasker -species cow option to reduce chances that primers will be selected within ruminant repetitive sequences. Primer selection has been performed using Primer3 software (24) with the target primer size of 23 bp and product size of 100 –600 bp.

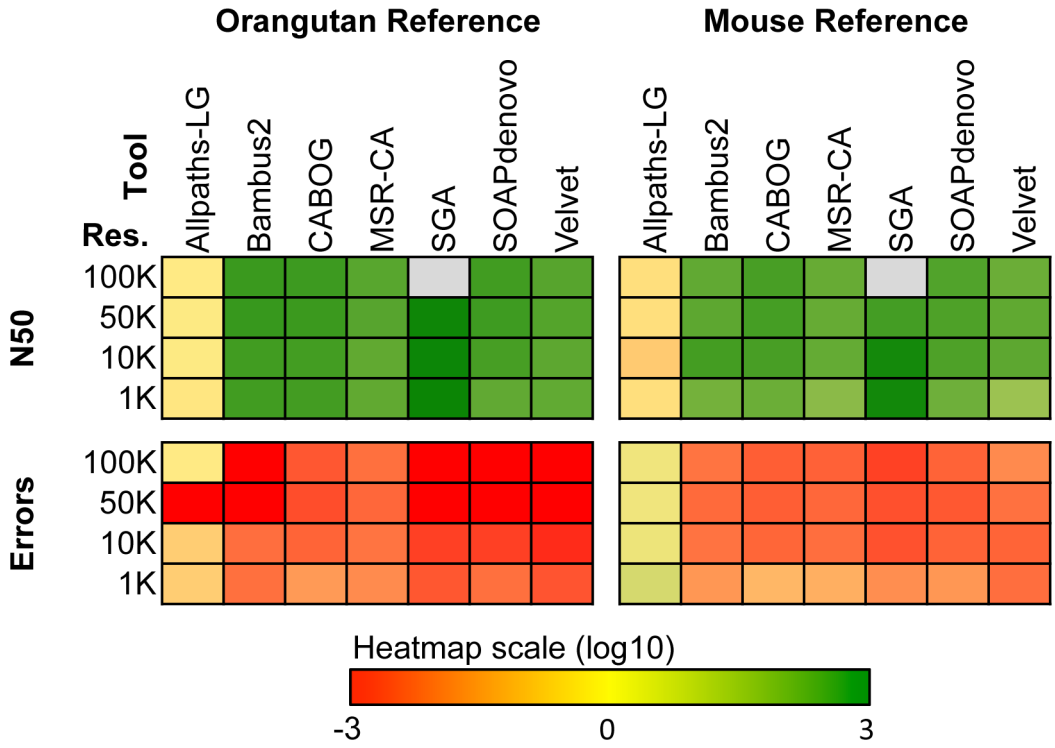
Polymerase chain reaction was performed in 50 μl volume containing 1.25 U TakaRa Ex Taq (TakaRa, China), 100 ng template genomic Tibetan antelope DNA. The initial denaturation step at 94°C for 2 min was followed by 30 cycles at 94°C for 30 s, 55-60°C for 30 s, and at 72°C for 30 s finalized by a 5-min extension at 72°C. For the PCR reactions resulting in multiple products we decreased the touchdown temperature from 68°C to 60°C in the 4th cycle.

References

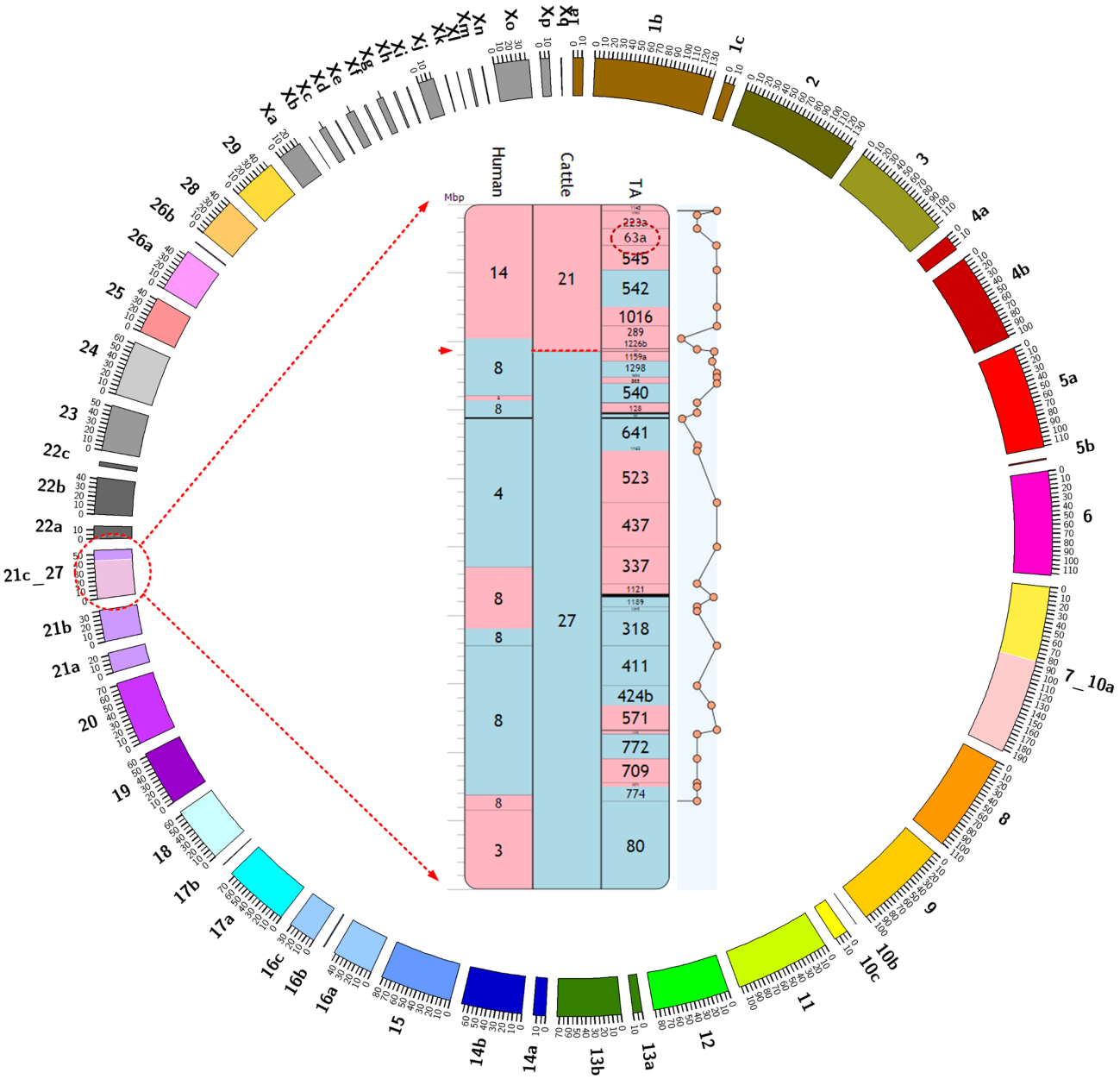
1. Sankoff D & Blanchette M (1999) Probability models for genome rearrangements and linear invariants for phylogenetic inference. *3rd International Conference on Computational Molecular Biology (RECOMB99)*, pp 302-309.
2. Kent WJ*, et al.* (2002) The human genome browser at UCSC. *Genome Res* 12(6):996-1006.
3. Pop M, Phillippy A, Delcher AL, & Salzberg SL (2004) Comparative genome assembly. *Briefings in bioinformatics* 5(3):237-248.
4. Gnerre S, Lander ES, Lindblad-Toh K, & Jaffe DB (2009) Assisted assembly: how to improve a de novo genome assembly by using related species. *Genome Biol* 10(8):R88.
5. Assefa S, Keane TM, Otto TD, Newbold C, & Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* 25(15):1968-1969.
6. Galardini M, Biondi EG, Bazzicalupo M, & Mengoni A (2011) CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine* 6:11.
7. Richter DC, Schuster SC, & Huson DH (2007) OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics* 23(13):1573-1579.
8. van Hijum SA, Zomer AL, Kuipers OP, & Kok J (2005) Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic acids research* 33(Web Server issue):W560-566.
9. Husemann P & Stoye J (2010) Phylogenetic comparative assembly. *Algorithms for molecular biology : AMB* 5:3.
10. Gao S, Sung WK, & Nagarajan N (2011) Opera: reconstructing optimal genomic scaffolds with high-throughput paired-end sequences. *J Comput Biol* 18(11):1681-1691.
11. Edgar RC, Asimenos G, Batzoglou S, & Sidow A (2010) Evolver: a whole-genome sequence evolution simulator.
12. Kelley DR, Schatz MC, & Salzberg SL (2010) Quake: quality-aware detection and correction of sequencing errors. *Genome Biol* 11(11):R116.
13. Salzberg SL & Yorke JA (2005) Beware of mis-assembled genomes. *Bioinformatics* 21(24):4320-4321.
14. Harris RS (2007) Improved pairwise alignment of genomic DNA. Ph.D. (The Pennsylvania State University).

15. Kent WJ, Baertsch R, Hinrichs A, Miller W, & Haussler D (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc Natl Acad Sci U S A* 100(20):11484-11489.
16. Ma J*, et al.* (2006) Reconstructing contiguous regions of an ancestral genome. *Genome Res* 16(12):1557-1565.
17. Grabherr MG*, et al.* (2010) Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* 26(9):1145-1151.
18. Donthu R, Lewin HA, & Larkin DM (2009) SyntenyTracker: a tool for defining homologous synteny blocks using radiation hybrid maps and whole-genome sequence. *BMC Res Notes* 2:148.
19. Li R*, et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265-272.
20. Ge R-L*, et al.* (2012) Genome data from the Tibetan Antelope (Pantholops hodgsonii). *GigaScience*.
21. Li R*, et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25(15):1966-1967.
22. Pollard KS, Hubisz MJ, Rosenbloom KR, & Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20(1):110-121.
23. Temple G*, et al.* (2009) The completion of the Mammalian Gene Collection (MGC). *Genome Res* 19(12):2324-2333.
24. Rozen S & Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132:365-386.
25. Krzywinski M*, et al.* (2009) Circos: an information aesthetic for comparative genomics. *Genome Res* 19(9):1639-1645.
26. Elsik CG*, et al.* (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* 324(5926):522-528.

Supplementary Figures

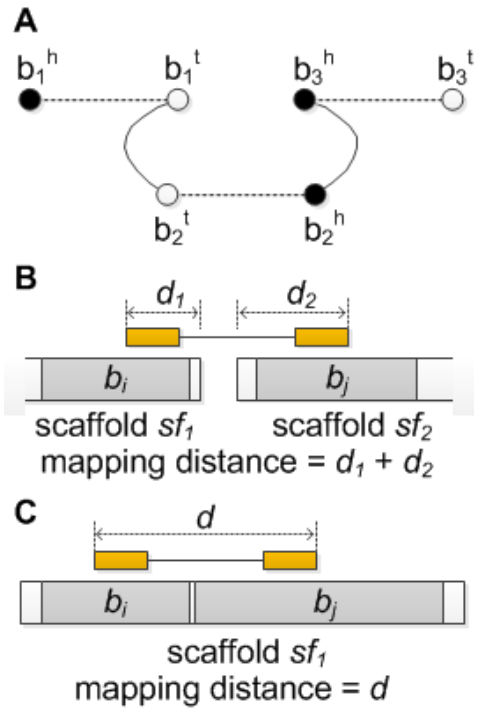


Supplementary Fig. S1. Evaluating RACA improvement of the GAGE assemblies without using paired-end read information. RACA improved the original assemblies created by seven genome assemblers in the GAGE data sets. The final RACA assemblies were compared with the original assemblies in terms of N50 and the number of adjacency errors. The heat maps show the log ratio of RACA N50 to the N50 of the original assembly (top horizontal block), and the log ratio of RACA adjacency errors to the errors of the original assembly (lower horizontal block), with the orangutan genome as a reference (vertical block on the left) and mouse genome as a reference (vertical block on the right). Four different resolutions of SF size were used, 100, 50, 10, and 1 Kbp; gray blocks in the top and bottom horizontal blocks represent the results for which there was no N50 data due to low coverage at certain resolutions and where the number of errors is zero in the RACA assemblies, respectively. For the complete data set see SI Tables S7-8.

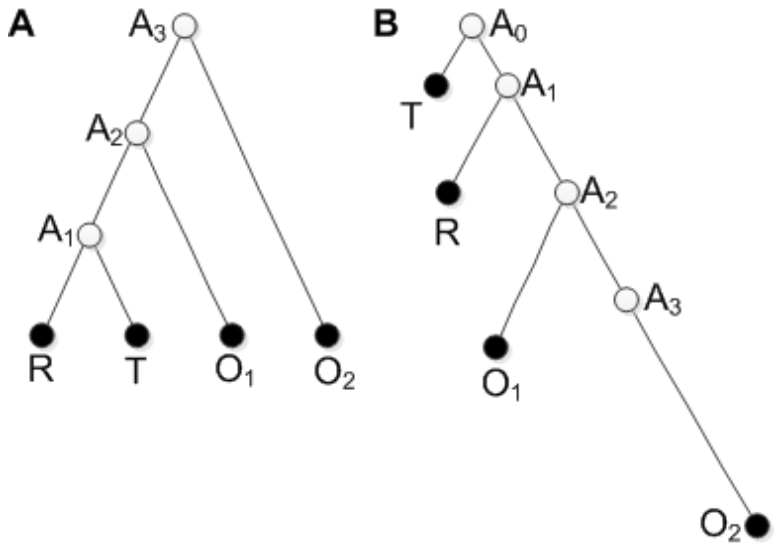


Supplementary Fig. S2. Circos plot (25) of reconstructed Tibetan antelope (TA) predicted chromosome fragments (PCFs) and an example of Evolution Highway tracks. The circle shows the mapping between TA PCFs and cattle chromosomes with different colors for each cattle chromosome. The outer labels are the names of TA PCFs, which were numbered to correspond to the mapped cattle chromosome. Using this naming system, joins representing more than one cattle chromosome within TA PCFs are shown by concatenating corresponding cattle chromosome names

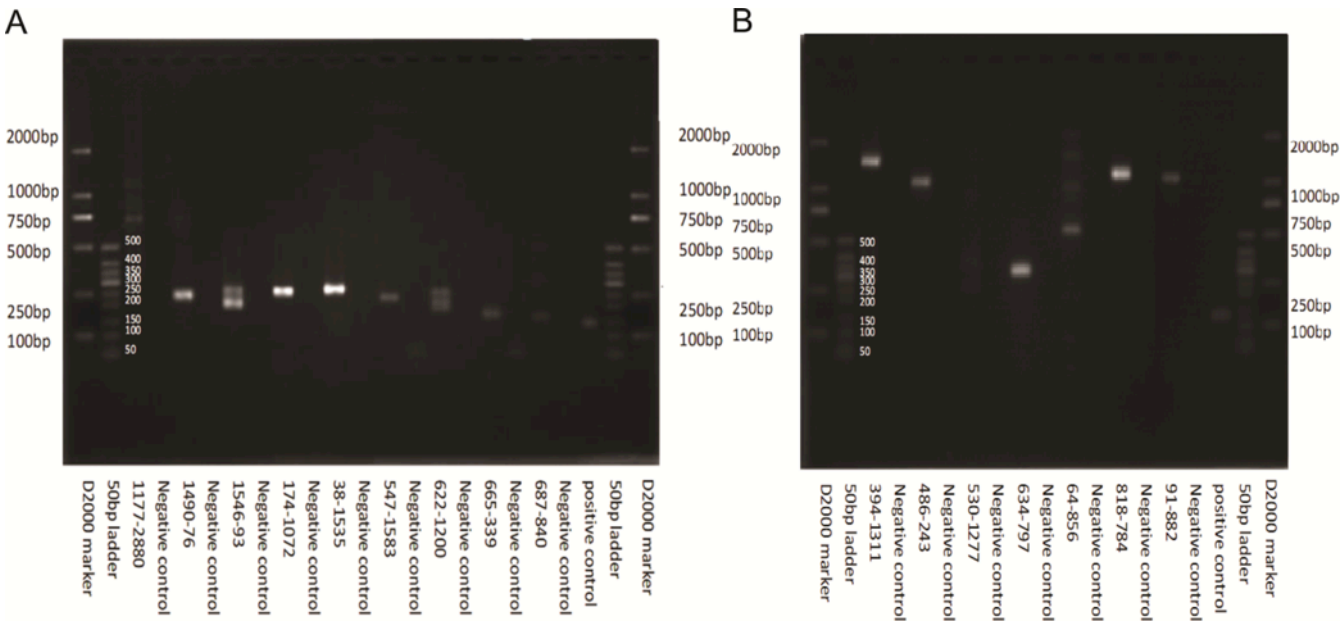
with "_". In the Evolution Highway tracks (center panel), cattle (middle), human (left), and TA (right) genome sequences are mapped to the TA PCF 21c_27. The red arrow indicates a cattle-specific evolutionary breakpoint region (EBR) that splits the ancestral chromosome corresponding to the TA PCF 21c_27 into the two cattle chromosomes, 21 and 27. The adjacency scores of the TA scaffolds are shown as a sparkline with orange circles along the tracks, and the blue and pink shading represents positive and negative orientation of the mapped blocks, respectively (Supplementary Data S1-5).



Supplementary Fig. S3. An example of the SF graph and a SF configuration. (A) Closed and open circles represent heads (b_i^h) and tails (b_i^t) of SFs respectively. The head and tail vertices from the same SF are always connected (dashed edge), and those which belong to different SFs are connected only when the edge weight is greater than 0 (solid edge; edge weights are not shown). The order and orientation of SFs can be inferred by traversing this graph starting from one of the two ends. The order and orientation of three SFs, b_1, b_2, and b_3 by this graph is (b_1, $-b_2$, b_3) or ($-b_3$, b_2, $-b_1$). (B) Two potentially adjacent SFs, b_i and b_j, from two different scaffold sf_1 and sf_2 respectively. In this case, the mapping distance is the sum of two distances from the end of each read to the end of the scaffold that the end read is mapped. (C) Two SFs, b_i and b_j, that belong to the same scaffold sf_1, representing a breakpoint between the two SFs or an assembly error. In this case, the mapping distance is simply the distance between the two end reads.



Supplementary Fig. S4. An example of rerooting of a phylogenetic tree. (A) An original phylogenetic tree with a reference genome *R*, target genome *T*, and two outgroup genomes O_1 and O_2. (B) A rerooted phylogenetic tree by adding a new root A_0 on the branch between A_1 and *T*. In the rerooted tree, branch length $t(A_0T) = t(A_1R)$, and $t(A_0A_1) = 0$.



Supplementary Fig. S5. PCR analysis. (a) PCR amplification of the DNA regions present in nine scaffolds with detected overlapping sequences. The results show that only one pair of overlapping scaffolds (scaffold1117 & scaffold2880) failed producing a product of an expected size (see Supplementary Data S6). (b) PCR amplification of predicted adjacent scaffolds. Out of seven cases, one (scaffold530 & scaffold1277) showed a very weak single PCR product. Other six pairs produced strong single amplification products of various sizes often different from the sizes expected from the distance between the scaffolds in the cattle genome, suggesting that antelope-specific sequences might be present in these Tibetan antelope genomic intervals (Supplementary Data S6). The images also show a negative control for each primer pair, positive control and 100 bp and 2000 bp ladders.

Supplementary Tables

Supplementary Table S1. Statistics of simulated datasets.

Dataset	D0	D1	D2	D3	D4
Total length	69,018,842	69,005,334	69,038,264	68,977,146	69,060,491
GC level	44.03	44.03	44.03	44.03	44.03
Repeats frac.	44.06	44.05	44.05	44.05	43.99
Avg. Frag. Coverage	0.9862	0.9849	0.981	0.981	0.9773

Dataset	D5	D6	D7	D8	D9	D10
Total length	68,985,069	68,991,281	68,994,533	68,933,700	68,968,144	68,966,587
GC level	44.03	44.03	44.03	44.03	44.02	44.03
Repeats frac.	44.04	44.04	43.99	44.06	44.03	44.01
Avg. Frag. Coverage	0.9777	0.9764	0.9756	0.9726	0.9702	0.9684

Supplementary Table S2. The number of breakpoints in real and simulated genome assemblies.

Real genome assembly[1]					Simulated genome assembly									
Chimp (panTro2)	Orangutan (ponAbe2)	Rhesus (rheMac2)	Mouse (mm9)	Cattle (UMD3.0)	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
112	89	218	418	563	18	43	81	138	172	194	215	275	295	332

[1]Human (NCBI36/hg18 assembly) was used as a reference with a minimum SF size of 150 Kbp.

[2]The dataset *R* was used as a reference with a minimum SF size of 5 Kbp.

Supplementary Table S3. Statistics of the original assemblies in the GAGE data sets and RACA assemblies using orangutan (ponAbe2 assembly) as a reference genome.

Resolution (Kbp)	Tool	Assembly[1]	Total Scaffolds	N50[2] (bp)	Misjoin Errors	Unjoin Errors	Total Errors[3]	Coverage[4]
100	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	85,369,765	0	1	1	0.98
	Bambus2	org	186	324,289	26	174	200	0.87
		raca	1	67,596,306	0	0	0	0.87
	CABOG	org	206	392,605	25	195	220	0.87
		raca	1	75,284,819	3	0	3	0.87
	MSR-CA	org	109	893,428	158	59	217	0.92
		raca	2	52,689,077	8	0	8	0.92
	SGA	org	216	-	0	215	215	0.40
		raca	1	-	0	0	0	0.40
	SOAPdenovo	org	211	453,540	3	203	206	0.75
		raca	1	78,803,340	0	0	0	0.75
	Velvet	org	143	1,190,421	208	83	291	0.81
		raca	1	114,571,702	0	0	0	0.81
50	ALLPATHS-LG	org	4	81,646,936	0	3	3	0.99
		raca	1	86,776,395	0	0	0	0.99
	Bambus2	org	238	324,289	31	227	258	0.92
		raca	1	72,104,657	0	0	0	0.92
	CABOG	org	285	392,605	38	268	306	0.94
		raca	1	81,393,373	3	0	3	0.94
	MSR-CA	org	135	893,428	206	71	277	0.94
		raca	3	28,099,128	9	0	9	0.94
	SGA	org	499	81,968	0	498	498	0.61
		raca	1	57,458,484	0	0	0	0.61
	SOAPdenovo	org	281	453,540	3	273	276	0.80
		raca	1	84,423,938	0	0	0	0.80
	Velvet	org	177	1,190,421	396	64	460	0.86
		raca	1	123,014,014	0	0	0	0.86
10	ALLPATHS-LG	org	8	81,646,936	8	4	12	0.99
		raca	1	86,849,014	5	0	5	0.99
	Bambus2	org	345	324,289	113	305	418	0.97
		raca	2	55,445,675	10	1	11	0.97
	CABOG	org	384	392,605	58	358	416	0.98
		raca	3	64,369,337	7	1	8	0.98
	MSR-CA	org	266	893,428	309	165	474	0.98
		raca	2	87,194,931	14	0	14	0.98
	SGA	org	1202	81,968	2	1201	1203	0.80

		raca	1	77,355,660	7	0	7	0.80
	SOAPdenovo	org	428	453,540	41	407	448	0.83
		raca	2	67,580,633	2	1	3	0.83
	Velvet	org	258	1,190,421	1845	22	1867	0.90
		raca	2	98,322,257	6	1	7	0.90
1	ALLPATHS-LG	org	41	81,646,936	72	10	82	1.00
		raca	9	32,141,672	30	3	33	1.00
	Bambus2	org	410	324,289	1463	324	1787	0.98
		raca	3	55,999,946	48	1	49	0.98
	CABOG	org	428	392,605	65	403	468	0.99
		raca	6	64,501,768	40	2	42	0.99
	MSR-CA	org	345	893,428	680	195	875	0.98
		raca	5	87,324,992	52	0	52	0.98
	SGA	org	2387	82,400	813	2041	2854	0.84
		raca	4	80,320,208	35	0	35	0.84
	SOAPdenovo	org	1914	453,540	877	1501	2378	0.88
		raca	8	32,520,040	58	2	60	0.88
	Velvet	org	555	1,190,421	8347	45	8392	0.92
		raca	9	83,737,280	103	1	104	0.92

[1]org and raca represent the original and RACA assembly respectively.

[2]N50 values were calculated by using the same genome size (88,289,540 bp: ungapped size of human chromosome 14).

[3]Sum of misjoin and unjoin errors. The misjoin errors occur when two adjacent contigs in the predicted assembly are not actually adjacent in the human genome assembly. The unjoin errors occur when two contigs are actually adjacent in the human genome assembly, but they are in separate scaffolds in the predicted assembly.

[4]Coverage value of each tool was calculated against the total size of the original assembly that was produced by the tool.

Supplementary Table S4. Statistics of the original assemblies in the GAGE data sets and RACA assembly using mouse (mm9 assembly) as a reference genome.

Resolution (Kbp)	Tool	Assembly[1]	Total Scaffolds	N50[2] (bp)	Misjoin Errors	Unjoin Errors	Total Errors[3]	Coverage[4]
100	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	58,307,776	3	0	3	0.98
	Bambus2	org	150	319,249	20	141	161	0.79
		raca	2	23,083,727	5	0	5	0.79
	CABOG	org	177	387,988	18	167	185	0.82
		raca	2	55,875,679	3	0	3	0.82
	MSR-CA	org	99	893,428	106	61	167	0.90
		raca	4	23,801,373	6	0	6	0.90
	SGA	org	145	-	0	144	144	0.31
		raca	2	-	0	1	1	0.31
	SOAPdenovo	org	168	446,090	2	159	161	0.69
		raca	2	49,005,779	3	0	3	0.69
	Velvet	org	86	1,104,606	53	65	118	0.59
		raca	2	60,591,686	7	0	7	0.59
50	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	58,307,776	3	0	3	0.98
	Bambus2	org	194	319,249	33	183	216	0.86
		raca	2	26,240,107	5	0	5	0.86
	CABOG	org	233	392,605	32	217	249	0.89
		raca	2	60,026,227	4	0	4	0.89
	MSR-CA	org	122	893,428	162	68	230	0.93
		raca	4	55,267,092	7	0	7	0.93
	SGA	org	373	72,866	0	372	372	0.52
		raca	3	11,253,325	4	0	4	0.52
	SOAPdenovo	org	225	446,095	4	214	218	0.75
		raca	2	53,718,374	3	0	3	0.75
	Velvet	org	146	1,190,421	192	84	276	0.81
		raca	2	88,600,139	8	0	8	0.81
10	ALLPATHS-LG	org	4	81,646,936	0	3	3	0.98
		raca	2	71,824,886	3	1	4	0.98
	Bambus2	org	291	324,289	111	254	365	0.94
		raca	3	52,943,436	7	0	7	0.94
	CABOG	org	319	392,605	42	298	340	0.95
		raca	1	82,131,575	5	0	5	0.95
	MSR-CA	org	195	893,428	245	104	349	0.96
		raca	2	85,466,930	7	0	7	0.96
	SGA	org	884	81,590	2	882	884	0.73

		raca	3	47,964,203	10	0	10	0.73
	SOAPdenovo	org	322	446,095	10	310	320	0.80
		raca	1	84,839,717	3	0	3	0.80
	Velvet	org	213	1,190,421	1000	44	1044	0.88
		raca	4	97,517,636	20	0	20	0.88
1	ALLPATHS-LG	org	17	81,646,936	18	7	25	0.99
		raca	8	50,434,774	70	1	71	0.99
	Bambus2	org	342	324,289	636	275	911	0.97
		raca	11	10,574,731	68	3	71	0.97
	CABOG	org	390	392,605	61	361	422	0.98
		raca	11	26,721,154	77	3	80	0.98
	MSR-CA	org	286	893,428	374	169	543	0.98
		raca	9	57,348,185	87	2	89	0.98
	SGA	org	1358	82,822	71	1330	1401	0.80
		raca	11	51,580,655	90	2	92	0.80
	SOAPdenovo	org	838	453,540	308	705	1013	0.84
		raca	12	25,265,899	78	2	80	0.84
	Velvet	org	390	1,190,421	4780	32	4812	0.91
		raca	25	17,622,254	126	4	130	0.91

[1]org and raca represent the original and RACA assembly respectively.

[2]N50 values were calculated by using the same genome size (88,289,540 bp: ungapped size of human chromosome 14).

[3]Sum of misjoin and unjoin errors. The misjoin errors occur when two adjacent contigs in the predicted assembly are not actually adjacent in the human genome assembly. The unjoin errors occur when two contigs are actually adjacent in the human genome assembly, but they are in separate scaffolds in the predicted assembly.

[4]Coverage value of each tool was calculated against the total size of the original assembly that was produced by the tool.

Supplementary Table S5. Effect of using an outgroup species (cattle) on reducing adjacency errors. Results show RACA using orangutan (ponAbe2 assembly) as a reference species.

Resolution (Kbp)	Tool	Effect of outgroup		
		with	Without	Reduced Errors
100	ALLPATHS-LG	1	0	-1
	Bambus2	0	0	0
	CABOG	3	3	0
	MSR-CA	8	8	0
	SGA	0	0	0
	SOAPdenovo	0	0	0
	Velvet	0	0	0
50	ALLPATHS-LG	0	0	0
	Bambus2	0	0	0
	CABOG	3	3	0
	MSR-CA	9	9	0
	SGA	0	0	0
	SOAPdenovo	0	0	0
	Velvet	0	0	0
10	ALLPATHS-LG	5	10	5
	Bambus2	11	12	1
	CABOG	8	8	0
	MSR-CA	14	20	6
	SGA	7	10	3
	SOAPdenovo	3	7	4
	Velvet	7	12	5
1	ALLPATHS-LG	33	53	20
	Bambus2	49	76	27
	CABOG	42	59	17
	MSR-CA	52	72	20
	SGA	35	50	15
	SOAPdenovo	60	76	16
	Velvet	104	132	28

Supplementary Table S6. Effect of using an outgroup species (cattle) on reducing adjacency errors. Results show RACA using mouse (mm9 assembly) as a reference species.

Resolution (Kbp)	Tool	Effect of outgroup		
		with	without	Reduced Errors
100	ALLPATHS-LG	3	5	2
	Bambus2	5	5	0
	CABOG	3	4	1
	MSR-CA	6	8	2
	SGA	1	4	3
	SOAPdenovo	3	5	2
	Velvet	7	6	-1
50	ALLPATHS-LG	3	5	2
	Bambus2	5	5	0
	CABOG	4	5	1
	MSR-CA	7	7	0
	SGA	4	4	0
	SOAPdenovo	3	5	2
	Velvet	8	7	-1
10	ALLPATHS-LG	4	12	8
	Bambus2	7	10	3
	CABOG	5	10	5
	MSR-CA	7	11	4
	SGA	10	12	2
	SOAPdenovo	3	11	8
	Velvet	20	17	-3
1	ALLPATHS-LG	71	175	104
	Bambus2	71	150	79
	CABOG	80	178	98
	MSR-CA	89	184	95
	SGA	92	168	76
	SOAPdenovo	80	166	86
	Velvet	130	184	54

Supplementary Table S7. Statistics of the original assemblies in the GAGE data sets and RACA assemblies generated without using paired-end read information (orangutan ponAbe2 assembly) used as a reference).

Resolution (Kbp)	Tool	Assembly[1]	Total Scaffolds	N50[2] (bp)	Misjoin Errors	Unjoin Errors	Total Errors[3]	Coverage[4]
100	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	3	81,646,936	0	2	2	0.98
	Bambus2	org	186	324,289	26	174	200	0.87
		raca	1	67,596,306	0	0	0	0.87
	CABOG	org	206	392,605	25	195	220	0.87
		raca	1	75,284,819	3	0	3	0.87
	MSR-CA	org	109	893,428	160	59	219	0.92
		raca	1	81,820,160	6	0	6	0.92
	SGA	org	216	-	0	215	215	0.40
		raca	1	-	0	0	0	0.40
	SOAPdenovo	org	211	453,540	3	203	206	0.75
		raca	1	78,803,340	0	0	0	0.75
	Velvet	org	143	1,190,421	208	83	291	0.81
		raca	1	114,571,702	0	0	0	0.81
50	ALLPATHS-LG	org	4	81,646,936	0	3	3	0.99
		raca	1	86,776,395	0	0	0	0.99
	Bambus2	org	238	324,289	31	227	258	0.92
		raca	1	72,104,657	0	0	0	0.92
	CABOG	org	285	392,605	38	268	306	0.94
		raca	1	81,393,373	3	0	3	0.94
	MSR-CA	org	135	893,428	208	71	279	0.94
		raca	1	83,701,987	6	0	6	0.94
	SGA	org	499	81,968	0	498	498	0.61
		raca	1	57,458,484	0	0	0	0.61
	SOAPdenovo	org	281	453,540	3	273	276	0.80
		raca	1	84,423,938	0	0	0	0.80
	Velvet	org	177	1,190,421	396	64	460	0.86
		raca	1	123,014,014	0	0	0	0.86
10	ALLPATHS-LG	org	8	81,646,936	8	4	12	0.99
		raca	1	86,849,014	5	0	5	0.99
	Bambus2	org	345	324,289	113	305	418	0.97
		raca	2	55,445,675	10	1	11	0.97
	CABOG	org	384	392,605	58	358	416	0.98
		raca	3	64,369,337	7	1	8	0.98
	MSR-CA	org	266	893,428	310	164	474	0.98
		raca	3	66,103,700	14	1	15	0.98
	SGA	org	1202	81,968	2	1201	1203	0.80

		raca	2	59,079,960	7	1	8	0.80
	SOAPdenovo	org	428	453,540	41	407	448	0.83
		raca	2	67,580,633	2	1	3	0.83
	Velvet	org	258	1,190,421	1845	22	1867	0.90
		raca	2	98,322,577	6	1	7	0.90
1	ALLPATHS-LG	org	41	81,646,936	72	10	82	1.00
		raca	6	72,176,652	34	1	35	1.00
	Bambus2	org	410	324,289	1463	324	1787	0.98
		raca	3	55,999,946	48	1	49	0.98
	CABOG	org	428	392,605	67	402	469	0.99
		raca	4	64,989,108	43	1	44	0.99
	MSR-CA	org	345	893,428	681	194	875	0.98
		raca	6	66,044,405	52	1	53	0.98
	SGA	org	2387	82,400	813	2041	2854	0.84
		raca	4	62,050,568	36	1	37	0.84
	SOAPdenovo	org	1914	453,540	877	1501	2378	0.88
		raca	9	32,517,937	64	1	65	0.88
	Velvet	org	555	1,190,421	8350	45	8395	0.92
		raca	9	83,737,280	103	1	104	0.92

[1]org and raca represent the original and RACA assembly respectively.

[2]N50 values were calculated by using the same genome size (88,289,540 bp: ungapped size of human chromosome 14).

[3]Sum of misjoin and unjoin errors. The misjoin errors occur when two adjacent contigs in the predicted assembly are not actually adjacent in the human genome assembly. The unjoin errors occur when two contigs are actually adjacent in the human genome assembly, but they are in separate scaffolds in the predicted assembly.

[4]Coverage value of each tool was calculated against the total size of the original assembly that was produced by the tool.

Supplementary Table S8. Statistics of the original assemblies in the GAGE data sets and RACA assembly generated without using paired-end read information (mouse mm9 assembly used as a reference).

Resolution (Kbp)	Tool	Assembly[1]	Total Scaffolds	N50[2] (bp)	Misjoin Errors	Unjoin Errors	Total Errors[3]	Coverage[4]
100	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	58,307,776	3	0	3	0.98
	Bambus2	org	150	319,249	20	141	161	0.79
		raca	2	23,083,727	5	0	5	0.79
	CABOG	org	177	387,988	18	167	185	0.82
		raca	2	55,875,679	3	0	3	0.82
	MSR-CA	org	99	893,428	107	60	167	0.90
		raca	2	56,129,580	3	0	3	0.90
	SGA	org	145	-	0	144	144	0.31
		raca	2	-	0	1	1	0.31
	SOAPdenovo	org	168	446,090	2	159	161	0.69
		raca	2	49,005,779	3	0	3	0.69
	Velvet	org	86	1,104,606	53	65	118	0.59
		raca	2	60,591,686	7	0	7	0.59
50	ALLPATHS-LG	org	3	81,646,936	0	2	2	0.98
		raca	2	58,307,776	3	0	3	0.98
	Bambus2	org	194	319,249	33	183	216	0.86
		raca	2	26,240,107	5	0	5	0.86
	CABOG	org	233	392,605	32	217	249	0.89
		raca	2	60,026,227	4	0	4	0.89
	MSR-CA	org	122	893,428	162	68	230	0.93
		raca	3	56,686,002	5	0	5	0.93
	SGA	org	373	72,866	0	372	372	0.52
		raca	3	11,253,325	4	0	4	0.52
	SOAPdenovo	org	225	446,095	4	214	218	0.75
		raca	2	53,718,374	3	0	3	0.75
	Velvet	org	146	1,190,421	192	84	276	0.81
		raca	2	88,600,139	8	0	8	0.81
10	ALLPATHS-LG	org	4	81,646,936	0	3	3	0.98
		raca	3	32,558,124	3	2	5	0.98
	Bambus2	org	291	324,289	111	254	365	0.94
		raca	4	52,750,795	11	0	11	0.94
	CABOG	org	319	392,605	42	298	340	0.95
		raca	2	56,141,028	7	0	7	0.95
	MSR-CA	org	195	893,428	245	104	349	0.96
		raca	3	58,203,700	9	0	9	0.96

	SGA	org	884	81,590	2	882	884	0.73
		raca	3	47,964,203	10	0	10	0.73
	SOAPdenovo	org	322	446,095	10	310	320	0.80
		raca	2	53,817,611	6	0	6	0.80
	Velvet	org	213	1,190,421	1000	44	1044	0.88
		raca	4	97,517,636	20	0	20	0.88
1	ALLPATHS-LG	org	17	81,646,936	18	7	25	0.99
		raca	9	57,710,117	78	2	80	0.99
	Bambus2	org	342	324,289	636	275	911	0.97
		raca	12	14,377,178	74	4	78	0.97
	CABOG	org	390	392,605	61	361	422	0.98
		raca	13	21,361,010	88	4	92	0.98
	MSR-CA	org	286	893,428	374	169	543	0.98
		raca	11	20,579,869	92	3	95	0.98
	SGA	org	1358	82,822	71	1330	1401	0.80
		raca	13	51,580,655	91	4	95	0.80
	SOAPdenovo	org	838	453,540	310	703	1013	0.84
		raca	14	22,638,976	83	4	87	0.84
	Velvet	org	390	1,190,421	4783	32	4815	0.91
		raca	23	17,773,919	125	4	129	0.91

[1]org and raca represent the original and RACA assembly respectively.

[2]N50 values were calculated by using the same genome size (88,289,540 bp: ungapped size of human chromosome 14).

[3]Sum of misjoin and unjoin errors. The misjoin errors occur when two adjacent contigs in the predicted assembly are not actually adjacent in the human genome assembly. The unjoin errors occur when two contigs are actually adjacent in the human genome assembly, but they are in separate scaffolds in the predicted assembly.

[4]Coverage value of each tool was calculated against the total size of the original assembly that was produced by the tool.

Supplementary Table S9. Syntenic fragments (SFs) shared between the cattle and Tibetan antelope genomes.

No. aligned cattle chromosomes (total length)	30 (2,661 Gbp)
No. aligned TA scaffolds (total length)	1,434 (2,601 Gbp, 96%[1])
No. SFs	1,597
Total length of SFs in cattle	2.596 Gbp (98%[2])
Total length of SFs in Tibetan antelope	2.571 Gbp (95%[1])

[1]Coverage compared to the total length of the Tibetan antelope assembly (2.699 Gbp).

[2]Coverage compared to the total length of the 30 cattle chromosome sequences (2.661 Gbp).

Supplementary Table S10. Characterization of predicted adjacencies between two syntenic fragments of Tibetan antelope.

Support from other genomes	Both cattle and human	1,396 (90.8%)
	Only cattle	68 (4.4%)
	Only human	63 (4.1%)
	None	10 (0.7%)
	Total	1,537
Support from paired-end reads	With paired-end reads support	1,056 (68.7%)
	Without paired-end reads support	481 (31.3%)
	Total	1,537
Support from other genomes and paired-end reads	Both cattle and human with paired-end reads	963 (62.7%)
	Only cattle with paired-end reads	30 (2.0%)
	Only human with paired-end reads	53 (3.4%)
	Either paired-end reads only or comparative genomic information	491 (31.9%)
	Total	1,537

Supplementary Table S11. Statistics of Tibetan antelope predicted chromosome fragments (PCFs) by using adjacencies with both comparative genome and paired-end reads support.

No. PCFs	512
No. PCFs that correspond to complete cattle chromosomes	0
Total length of PCFs	2.601 Gbp
Max. length of PCFs	37 Mbp
Min. length of PCFs	164 Kbp
PCF N50	9.5 Mbp
Max. no. Tibetan antelope scaffolds in PCFs	22
Min. no. Tibetan antelope scaffolds in PCFs	1
No. cattle EBRs	57
No. other EBRs	385
No. Tibetan antelope scaffolds that have more than one SF	130 (9%[1])
No. Tibetan antelope scaffolds predicted as chimeric[2]	66 (5%[1])

[1]Percentage of the total number of aligned Tibetan antelope scaffolds.

[2]Among 66 scaffolds, 5 were mapped to three different PCFs, 60 were mapped to two different PCFs, and the remaining 1 was mapped to the same PCF at different and non-adjacent locations.

Supplementary Table S12. Ratio of the number of syntenic fragments (SFs) to the number of sequence fragments.

Dataset	D0	D1	D2	D3	D4	D5	D6	D7	D8	D9
Ratio	1.0873	1.1500	1.2319	1.3446	1.4191	1.5235	1.5757	1.7312	1.7049	1.7983

SFs were constructed by using the dataset *R* as a reference with the minimum SF size of 5 Kbp.

For each dataset, average across five different sets of sequence fragments is reported.

Supplementary Table S13. Known evolutionary breakpoint regions (26).

Chromosome	Start	End	EBR Type	Size	Spanned by TA scaffold
chr1	140637389	140713496	cattle	76108	Yes
chr1	153097418	153236312	cattle	138895	No
chr10	4863074	5085594	cattle	222521	No
chr10	11062121	11859768	cattle	797648	Yes
chr10	20534282	20582030	cattle	47749	Yes
chr10	65609815	65760700	cattle	150886	No
chr11	9330539	9514305	cattle	183767	Yes
chr11	14065354	14174278	cattle	108925	Yes
chr11	43766273	44081098	cattle	314826	Yes
chr11	45980519	46106594	cattle	126076	Yes
chr11	68458408	68602338	cattle	143931	Yes
chr11	92075075	92217401	cattle	142327	Yes
chr12	36821348	37215305	cattle	393958	No
chr13	9645734	11520700	cattle	1874967	No
chr13	37338235	37571107	cattle	232873	No
chr13	43203199	43304808	cattle	101610	Yes
chr13	47290805	47485694	cattle	194890	No
chr13	51210519	51368258	cattle	157740	Yes
chr13	53751336	54156303	cattle	404968	No
chr13	60155298	60235042	cattle	79745	Yes
chr14	20438925	20760131	cattle	321207	No
chr14	46850889	46999999	cattle	149111	Yes
chr14	83283856	83308985	cattle	25130	Yes
chr15	16232407	16335808	cattle	103402	No
chr15	34911095	34932579	cattle	21485	Yes
chr15	57462840	57582400	cattle	119561	Yes
chr16	36704676	36740286	cattle	35611	Yes
chr16	52804836	52887301	cattle	82466	Yes
chr16	55793954	56366044	cattle	572091	Yes
chr17	36598033	36687755	cattle	89723	Yes
chr17	44905980	45042468	cattle	136489	Yes
chr18	14761834	15044093	cattle	282260	Yes
chr18	40289451	40439853	cattle	150403	Yes
chr19	35992121	36090948	cattle	98828	Yes
chr2	5391490	5500507	cattle	109018	Yes
chr2	79582982	79695537	cattle	112556	No
chr21	25517465	25658342	cattle	140878	No
chr21	27660782	27906521	cattle	245740	No
chr21	35022024	35200027	cattle	178004	No

chr21	55507507	55542387	cattle	34881	Yes
chr21	56496660	56624459	cattle	127800	Yes
chr22	16544591	16725742	cattle	181152	Yes
chr22	54896835	54944928	cattle	48094	Yes
chr23	6922424	7113859	cattle	191436	Yes
chr23	25148171	25372693	cattle	224523	Yes
chr24	12754657	12916848	cattle	162192	Yes
chr24	43961913	44156654	cattle	194742	No
chr28	12815222	12977543	cattle	162322	No
chr29	33404945	33604545	cattle	199601	No
chr29	37554336	37673007	cattle	118672	Yes
chr3	112883146	113005398	cattle	122253	Yes
chr4	72495447	72536027	cattle	40581	Yes
chr4	75131543	75247024	cattle	115482	Yes
chr5	25226164	25519589	cattle	293426	Yes
chr5	60292857	60312327	cattle	19471	Yes
chr5	76291290	76502027	cattle	210738	No
chr5	109832859	109911358	cattle	78500	Yes
chr6	2764216	3357658	cattle	593443	Yes
chr6	38302693	38636309	cattle	333617	No
chr6	104307383	104503691	cattle	196309	Yes
chr7	17190737	17284255	cattle	93519	No
chr7	39168049	39197714	cattle	29666	No
chr7	41121138	42033445	cattle	912308	No
chr7	82805598	82861662	cattle	56065	Yes
chr8	11068091	11138570	cattle	70480	Yes
chr8	65775439	66678745	cattle	903307	Yes
chr8	75976901	76051647	cattle	74747	Yes
chr8	86969431	86995322	cattle	25892	No
chr9	23965228	24092472	cattle	127245	Yes
chr9	66545893	66670543	cattle	124651	No
chrX	38178576	42203401	cattle	4024826	No
chrX	90844040	99879172	cattle	9035133	No
chrX	97922988	103010533	cattle	5087546	No
chr1	144809649	145469012	cetartyodactyl	659364	No
chr10	2301145	2370416	cetartyodactyl	69272	Yes
chr10	58021703	59195218	cetartyodactyl	1173516	No
chr10	58214589	59444054	cetartyodactyl	1229466	No
chr13	18277163	18442984	cetartyodactyl	165822	Yes
chr13	28002739	28143903	cetartyodactyl	141165	Yes
chr17	57624873	57720839	cetartyodactyl	95967	Yes
chr18	62055889	62228890	cetartyodactyl	173002	No

chr21	5321045	5385847	cetartyodactyl	64803	Yes
chr22	56481590	56504504	cetartyodactyl	22915	No
chr22	57778940	58158578	cetartyodactyl	379639	No
chr25	27729542	27891174	cetartyodactyl	161633	No
chr4	32368679	32439913	cetartyodactyl	71235	No
chr7	3340783	3600508	cetartyodactyl	259726	No
chr7	14270266	14330866	cetartyodactyl	60601	No
chr8	59365919	59663699	cetartyodactyl	297781	No
chr8	77458294	77515575	cetartyodactyl	57282	Yes

Supplementary Table S14. Statistics of insert libraries.

Library ID	Insert size (bp)	Avg read length (bp)	GC%	Lanes	Usable reads[1] (Mb)	No. mapped reads	No. uniquely mapped reads	Frac. of uniquely mapped reads[2]
PHOlcpDAFDFAAPE	350	65.26	40.49	4	179.34	80,954,532	78,743,568	97%
PHOlcpDAFDFBAPE	353	61.29	41.78	4	162.77	72,891,310	71,115,590	98%
PHOlcpDAADEAAPE	358	60.38	46.06	10	315.44	137,879,528	134,601,756	98%
PHOlcpDAADEBAPE	364	57	48.27	10	287.55	125,212,662	121,119,303	97%
PHOlcpDAADECAPE	386	52.28	46.1	10	323.54	142,570,584	140,077,523	98%
PHOlcpDAADEDAPE	393	50	47.91	1	25.66	11,200,639	11,013,347	98%
PHOlcpDAADIAAPE	497	63.14	43.64	3	60.07	25,490,978	25,309,616	99%
PHOlcpDAADIBAPE	542	66.28	42	5	154.84	66,456,669	66,070,100	99%
PHOlcpDAFDJAAPE	559	63.37	40.22	6	270.93	119,175,926	118,439,185	99%
PHOlcpDAFDKAAPE	565	60.01	40.36	4	173.15	76,734,022	75,953,810	99%
PHOlcpDABDWABPE	2350	44	52.11	11	295.45	96,178,272	89,077,002	93%
PHOlcpDABDWAAPE	2730	44	47.33	11	378.23	120,779,668	115,654,617	96%
PHOlcpDADDLBAPE	5000	44	47.64	3	97.99	34,794,073	33,082,911	95%
PHOlcpDACDLAAPE	5340	44	46.69	8	267.97	85,894,780	83,445,531	97%
PHOlcpDACDLACPE	9000	44	47.35	3	90.19	31,540,535	30,550,479	97%
PHOlcpDBADTAAPE	9370	44	45.08	4	115.55	42,611,116	39,281,355	92%
PHOlcpDAADTAAPE	10000	44	46.02	1	25.52	9,159,349	8,964,911	98%
PHOlcpDAADUAAPE	20000	44	44.31	3	147.87	56,317,816	45,447,125	81%

[1]Amount of reads after filtering out uninformative reads that (i) have an 'N' over 10% of its length, (ii) have more than 40% of bases with low quality, (iii) are more than 10 bp from the adapter sequence (allowing ≤2 bp mismatches), (iv) are small insert size paired-end reads that overlap ≥10 bp between two ends, and (v) have completely identical reads (both ends) in another paired-end reads (hence considered to be the products of PCR duplication). SRA accession no. SRA052275.

[2] Out of total mapped paired-end reads, 96.41% were mapped uniquely.

Supplementary Table S15. Adjacent Tibetan antelope scaffolds on cattle chromosomes and their overlaps.

Distance on cattle chromosome (Kbp)	No. predicted adjacencies	No. adjacencies between overlapping scaffolds (%)	Avg. overlapping size in bp (stdev)
0	463	166 (36)	32 (9.7)
< 1	421	19 (5)	28.6 (12.2)
< 10	305	15 (5)	31.3 (11.1)
< 150	195	12 (6)	30.6 (11.9)
>= 150	77	4 (5)	32 (5.8)
Total	1461	216 (15)	

Minimum overlap size = 5bp (false positive rate = 0.005 in a dataset of random adjacencies)

Supplementary Data (as attachments)

Supplementary Data S1: TA PCFs

Supplementary Data S2: Mapping between TA PCFs and Cattle genome

Supplementary Data S3: Mapping between TA PCFs and TA scaffolds

Supplementary Data S4: Mapping between TA PCFs and Human genome

Supplementary Data S5: Predicted adjacency scores in TA PCFs

Supplementary Data S6: Selected primer pairs and PCR analysis results