LODE: A Locus Ordering Procedure Based On Linkage Disequilibrium Applied To Bovine Assembly.

M.S. Khatkar[*], M. Hobbs[*], M. Neuditschko[†] J. Sölkner[†], F. W. Nicholas[*] and H. W. Raadsma[*]

Introduction

A number of species are being sequenced with rapid advances in low-cost sequencing technologies (Benson *et al.* 2008). Next generation sequencing (NGS) generally procedures sequencing data in short reads which pose challenges in the creation and ordering of contigs and scaffolds in the absence of a mature reference genome. A key output of NGS is discovery of abundance of Single Nucleotide Polymorphisms (SNP) (Eck *et al.* 2009), and indeed many such genome sequencing efforts may be coupled with high-throughput SNP-genotytping platforms to undertake population diversity characterization. Linkage disequilibrium (LD) analysis of such data can provide information for ordering such markers (Miller *et al.* 2006; Sölkner *et al.* 2008). We have developed a locus ordering procedure based on linkage disequilibrium (LODE) which provides an alternative means for verification of marker order and position as well as to estimate chromosomal positions of unaligned SNPs and scaffolds. First, we investigated the efficiency of using genome-wide LD information by using mapped SNPs as a test set from bovine data. Next, we applied the procedure to assign positions for 4688 orphan SNPs from three high-density SNP panels on Btau4.0, which were either un-assigned or assigned with ambiguity based on BLAST against Btau4.0,. We also suggest the chromosomal locations of un-ordered scaffolds. Finally, we used the LODE procedure to confirm the order of mapped SNPs across the genome as a means to check the quality of genome assembly.

Material and methods

Genotypic Data: Data from three bovine genotyping SNP arrays, namely 15k (Khatkar *et al.* 2007), 25k (Affymetrix; http://www.affymetrix.com) and 54k (Illumina; http://www.illumina.com/), were used for genotyping 1536, 441 and 377 Australian Holstein-Friesian (HF) bulls, respectively. These data were combined into a single dataset for the current analyses. The final combined dataset represented 73,569 unique SNPs and 1,943 bulls with an average of 628 bulls genotyped per SNP. The location of these SNPs in the bovine genome was assessed from BLAST alignment of SNP flanking sequences with the Btau4.0 assembly (ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Btaurus/fasta/Btau20070913-freeze/), which includes a considerable amount of sequence data organised as scaffolds (not assigned to a chromosome and referred here as 'Un'). SNP positions on Btau4.0 were categorised as follows: i) 'mapped' (single assignment to a chromosome); ii) 'ambiguous'

[*] The Univsrsity of Sydney, NSW, Australia

[†] University of Natural Resources and Applied Life Sciences, Vienna, A-1180, Austria

(more than one assignment in the genome); iii) 'Un' (single assignment to 'Un' sequences only); iv) 'unassigned' (no assignments in the genome). Collectively, the last three categories (ambiguous, Un and unassigned) are here called 'unpositioned'. We used the Batu4.0 assembly to demonstrate utility of the LODE procedure since the assembly contained a number of SNPs not assigned to chromosomes. Comparison of LODE positions were also made against another bovine assembly build UMD3 (ftp://ftp.cbcb.umd.edu/pub/data/assembly/Bos_taurus/Bos_taurus_UMD_3.0/) recently became available.

LODE procedure: The location of each unpositioned SNP was estimated on the basis of its LD (estimated as r^2) (Abecasis and Cookson 2000) with all mapped SNPs on the assembly. From these estimates of r^2 we computed the maximum r^2 (r^2_{max}, as an indicator of the strength of LD) and number of mapped SNPs with $r^2 > 0.1$ ($n_{0.1}$, only $n_{0.1} > 2$ was considered useful) for each chromosome. A chromosome with highest ranking for both the parameters was identified as the candidate chromosome for that unpositioned SNP. SNPs with MAF <0.05 required an additional check to improve accuracy of placement. In addition to the above strategy, if the r^2_{max} of the second best chromosome exceeded 2/3 r^2_{max} of the candidate chromosome, the SNP was not assigned to any chromosome. SNPs which didn't meet these criteria were left unpositioned. For each unpositioned SNP that could be assigned to a chromosome, its location on that chromosome was allocated the same position as that of the mapped SNP with which the unpositioned SNP has r^2_{max} . After testing the LODE procedure with three test sets, the same procedure was applied to unpositioned bovine SNPs.

Results and discussion

Validation of LODE procedure: The LODE procedure was first tested for three test sets by masking the locations of the SNPs whose location was actually known. Each set comprised SNPs sampled randomly within a different MAF class, namely 0.001<MAF<0.01 (300 SNPs, 10 from each chromosome); 0.01<MAF<0.05 (300 SNPs, 10 from each chromosome); >0.05 (900 SNPs, 30 from each chromosome). The efficiency of placing SNPs (the proportion of masked SNPs that were assigned a location) in these three test sets was 2.0%, 30.6% and 96.7%, with accuracy (the proportion of masked SNPs that were assigned correctly) of 33.3, 98.9% and 99.9, respectively. The average precision for placement of the SNPs taken as the difference between allocated position and actual position in the genome was 914, 3137 and 6853 Kb, for the three test sets, respectively. Overall, the LODE procedure can position SNPs with MAF>0.01 with high accuracy. Rare SNPs (0.001<MAF<0.01) could not be positioned using the LODE.

Application of LODE to unpositioned bovine SNPs and scaffolds: In the Btau4.0 assembly, there are 6470 'unpositioned' SNPs from the three bovine chips. Of these, 5314 SNPs have MAF>0.01, making them suitable for LODE mapping. Using the LODE strategy, 4468 of the 'unpositioned' SNPs were positioned. Of these, 2060 had ambiguous positions, 1499 were aligned to 494 unique 'Un' sequences, and 1129 were unaligned on Btau4.0. Assignment of 1499 SNPs belonging to 'Un' sequences to definite chromosomes suggests the assignment and positions of respective "Un" scaffolds to the same chromosome as well.

Out of these, 210 “Un” scaffolds had two or more SNPs (mean=5.04) with all the SNPs aligned to one chromosome. These 210 “Un” scaffolds with multiple SNPs could be assigned and some of them could be oriented on the chromosome, based on the SNP position estimates by the LODE procedure.

Checking the integrity bovine genome assembly with LODE: The position of each SNP was recomputed using LODE procedure based on the LD information of the remaining SNPs in the genome. The chromosomal assignments and positions estimated by LODE were compared with original positions on Btau4.0 and also with UMD3. The comparison of the overall agreement between LODE and SNPs positioned on Btau4.0 is shown in Figure 1a by the way of an Oxford grid. Most of the SNPs (99.9 %) were given the same chromosomal assignments which indicates in general a high level of integrity of Btau4.0. However, there were 81 SNPs mapped to different chromosomes by LODE. Of these 6 SNPs form a contiguous segment on chromosome 3 (77488731-77859105) and were all assigned to chromosome 2 (126296914-121859331) by the LODE procedure. Similarly chromosomal segments containing multiple SNPs from chromosome 3, 10, 10, X-chromosome on Btau4.0 were assigned to chromosomes 22, 1, 4, and 8, respectively by the LODE procedure. These LODE reassignments were also supported when compared with UMD3. These blocks suggest problem areas within the Btau4.0 bovine assembly.

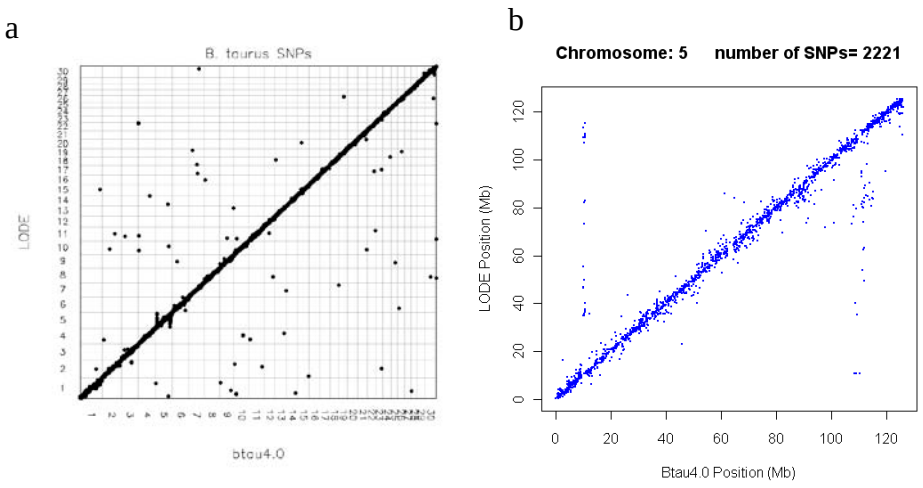

Figure 1(a): Comparison of repositioned locations of SNP by LODE with original location on Btau4.0; (b). The comparison of chromosomal assignments of 1776 SNPs repositioned by LODE procedure with original positions on Btau4.0 on chromosome 5. Two potential problem regions (10-11 Mb and 90-120 Mb) on Btau4.0 on this chromosome can be noted.

In this study we reported and validated a procedure to accurately and efficiently map SNPs based on LD information. The LODE procedure offers particular advantages in the positioning of problem SNPs for which no unambiguous assignment on a draft genome

assembly could be made, as well as a means for positioning of unordered scaffolds containing SNPs. The LODE procedure can be used for checking the integrity of assembly by sampling and reassigning the positions of SNPs as shown in the above section. Where there are many discrepancies in the predicted position by LODE and the proposed SNP position from the assembly, this, would indicate a general problem in assembly as shown in case of BTA5 in Figure 1b.

The LODE procedure presented here can also be very helpful in refining sequence and genetic maps for species where comparative genome assemblies are used to build a virtual assembly for the species of interest, such as has recently been done for sheep (Dalrymple *et al.* 2007). In fact we are applying the LODE procedure to sheep hapmap data (International Sheep Genomics Consortium; http://www.sheephapmap.org/), initially using two breeds separately we could identify chromosomal locations for 70 orphan SNPs out of 110 (Khatkar et al. unpublished on behalf of the ISGC). We also identified 101 SNPs which were given a different chromosomal assignment by LODE as compared to their proposed position on the virtual sheep genome. Comparison of intra chromosomal SNP positions by LODE and original map positions, indicated a number of problem areas in the ovine genome which required further analysis particularly overall assembly of the X-chromosome.

Acknowlegement

The research was funded by the Co-operative Research Centre for Innovative Dairy Products (CRC-IDP). We thank Genetics Australia for semen samples for genotyping, the Australian Dairy Herd Improvement Scheme for pedigree data and Drs, Julie Cavanagh, Natasha Ellis, Kyall Zenger, Bruce Tier and Mrs Gina Attard for providing input at various stages of development of the core data resource.

References

Abecasis G.R. and Cookson W.O. (2000). *Bioinformatics* 16: 182-3.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., *et al.* (2008). *Nucleic Acids Res* 36: D25-30.

Dalrymple B.P., Kirkness E.F., Nefedov M., *et al.* (2007). *Genome Biol* 8: R152.

Eck S.H., Benet-Pages A., Flisikowski K., *et al.* (2009). *Genome Biol* 10: R82.

Khatkar M.S., Zenger K.R., Hobbs M., *et al.* (2007). *Genetics* 176: 763-72.

Miller S.P., Hayes B.J. and Goddard M.E. (2006). *8th World Congress on Genetics Applied to Livestock Production, August 13-18, 2006, Belo Horizonte, MG, Brasil.*

Sölkner J., Neuditschko M., Khatkar M.S., *et al.* (2008). *59th Annual Conference of the European Association for Animal Production. Wageningen Publishers, Vilnius, Lithuania.*