Fine-Scale Study of Meiotic Recombination in Cattle

C. Oget-Ebrad[1*], G.C.M. Moreira[1], L. Karim[2], W. Coppieters[2], C. Charlier[1], M. Georges[1] and T. Druet[1]

[1] Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège (B34), 4000 Liège, Belgium; [2] Genomics Platform, GIGA, University of Liège (B34), 4000 Liège, Belgium; [*] claire.oget-ebrad@inrae.fr

Abstract

We herein characterized fine-scale meiotic recombination rate (RR) variation in a whole-genome sequenced cattle pedigree including 264 individuals. With a family-based approach, we identified 6,644 cross-over (CO) events (median length: 3.9kb) whereas with a population-based method, we observed the presence of 12,776 recombination hotspots (median length: 1.5kb). Although these hotspots represented only 0.97% of the genome, 38.3% of the CO overlapped with these hotspots (40.6% in males, 34.8% in females), suggesting that they might still be in use in the current population. Subsequently, we determined that RR was higher in open chromatin regions, in CpG islands, and in predicted PRDM9 motifs. Conversely, lower RR were estimated close to genes except in the transcription start sites. Finally, higher RR were estimated around H3K4me3 histone mark in testis, further supporting the implication of the PRDM9 zinc finger protein in meiotic recombination in cattle.

Introduction

Meiotic recombination is an important biological process as it ensures the proper segregation of homologues during meiosis and it shapes the patterns of linkage disequilibrium (LD) by creating new combinations of alleles (see for instance Coop and Przeworski, 2007, Baudat *et al.*, 2010, and Peñalba and Wolf, 2020, for reviews). Two approaches can be used to study meiotic recombination from genotypes (Peñalba and Wolf, 2020). First, family-based approaches allow to identify sex-specific cross-over (CO) events and provide current recombination rates (RR). Second, population (or LD)-based approaches provide historical sex-averaged RR. These latter maps represent relative estimates as they are influenced by other parameters such as past effective population size or selection intensity. Two mechanisms directing recombination are currently known (Baudat *et al.*, 2010). The first one is PRDM9-independent, where recombination is concentrated at TSS and CpG islands, and where we observe stable patterns of recombination across generations, as seen in yeasts, plants, and vertebrate species such as birds. In contrast, in most mammals such as in human, recombination is PRDM9-directed, that is, concentrated at the positions where PRDM9 protein is binding, away from genes, and where we observe a fast evolution of PRDM9 binding-motifs and a divergence between close species (Baudat *et al.*, 2010). Here, we take advantage of a unique sequenced cattle pedigree to characterize the patterns of RR in the bovine genome and to study the relationship between recombination and genomic features.

Materials & Methods

Sequence data. Whole-genome sequence data was available for a small pedigree containing 264 Holstein-Friesian individuals. Based on filtering and processing steps described in Oget-Ebrad *et al.* (2021), we selected 8.4 million SNPs for further analysis.

CO events identification. We used the family-based approach implemented in LINKPHASE3 (Druet and Georges, 2015) to identify CO events in the sequence data. We removed double

CO occurring within 1Mb, and multiple identical CO within a family to remove putative errors. CO intervals were then refined using ShapeIT4.1 (Delaneau *et al.*, 2019).

LD-based RR and hotspots. We used LDhelmet (Chan *et al.*, 2012) to estimate population-based RR and sequenceLDhot (Fearnhead, 2006) to identify recombination hotspots.

PRDM9-related features. We predicted the PRDM9 binding-motifs for each allele of the BTA1 paralog previously characterized for the Holstein breed (Zhou *et al.*, 2018), using the approach from Persikov and Singh (2014). We scanned the genome for potential binding locations of each of these predicted motifs with FIMO (Grant *et al.*, 2011).

Other genomic features. We used available public database to locate genomic features such as the transcription start sites (TSS) (Goszczynski *et al.*, 2021), and CpG islands (https://hgdownload.soe.ucsc.edu/goldenPath/bosTau9/database/). Functional annotation was available from experiments performed in testis, including an Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq), and epigenetic data for histone marks (H3K27me3, H3K4me3, H3K4me1, H3K27ac) and for one DNA-binding protein (CTCF).

Results

Broad-scale variation of RR. We identified 6,644 CO events (3,895 for males and 2,749 for females; median length: 3.9kb) with the family-based approach and characterized 12,776 LD-based hotspot regions (median length: 1.5kb; RR=4.1/kb in the hotspots *vs.* 0.18/kb as whole-genome average), the latter representing 0.97% of the genome. Thirty-eight percent of the CO overlapped with LD-based hotspots (40.6% in males, 34.8% in females) and more than 50% were within 5kb from the hotspots (**Figure 1**). This proportion was highly significant compared to random sets of CO events (p-value < 0.001; 2.4% overlap on average with 1000 simulations). We also observed that both CO and hotspots density were higher at telomeres.

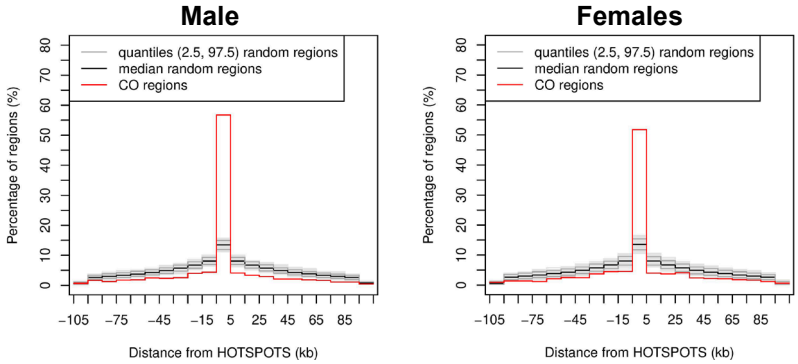

Figure 1. Percentage of CO events (red) identified with the family-based approach in males and females as a function of the distance (in kb) from the closest hotspot from the population-based approach, compared to 1000 simulated random sets of CO matching in number and size (grey for quantiles, black for median).

Fine-scale variation of RR. In our population, we observed the five alleles of the BTA1 *PRDM9* paralog previously described by Zhou *et al.* (2018) at following frequencies: 12.3% for allele 1, 51.1% for alleles 2 and 3, 23.1% for allele 4, and 13.4% for allele 5. A total of 149,620 PRDM9 predicted motifs were identified in the genome. Although we observed increased RR around these motifs (**Figure 2a**), only 11.3% of the LD-based hotspots contained at least one motif. We observed reduced RR close to TSS (**Figure 2b**) and

increased RR in CpG islands (**Figure 2d**). RR were higher in accessible DNA regions (ATAC-seq peaks, **Figure 2e**) and in CTCF signals. Methylation marks from active promoters (H3K4me1 and H3K27ac) presented lower RR. We observed a strong increase of RR in the H3K4me3 methylation mark (**Figure 2f**). This was specific to assays performed in testis and not observed with ATAC-seq data from other tissues (e.g., liver). Some of these associations were also observed with the set of identified CO (**Figure 2c**), with no striking difference between males and females, although the patterns were less clear than with the population-based approach.

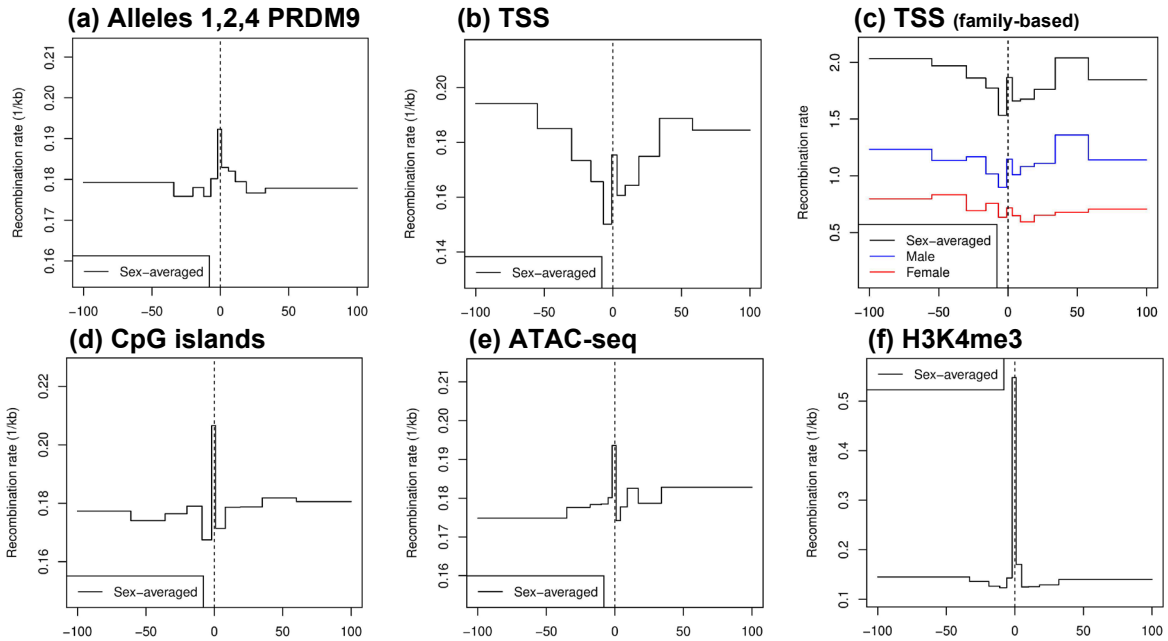

Figure 2. Recombination rates (RR) as a function of the distance (in kb) from the closest (a) BTA1 PRDM9 predicted binding-motif for alleles 1, 2 & 4, (b-c) transcription start site (TSS), (d) CpG island, (e) ATAC-seq peak, and (f) H3K4me3 signal. RR were estimates from the population-based approach, except in (c).

Discussion

In cattle, association between PRDM9 and individual RR and CO events location (at relatively broad scale) has been previously reported, as in human and mice. Meiotic recombination has been indeed previously studied in cattle with sparser marker maps (Ma *et al.*, 2015). Based on whole-genome sequence data, our findings further support the role of PRDM9 in this species. First, we observed the presence of recombination hotspots (median length: 1.5kb) in cattle, as in human and mice (Baudat *et al.*, 2010). These recombination hotspots represent 0.97% of the genome (1.7-1.9% in human (Kong *et al.*, 2010)). They were significantly overlapping with identified CO, suggesting that the historical hotspots are still in use in the present population. Then, we observed increased RR around BTA1 PRDM9 binding-motifs, as well as a small proportion of hotspots (11.3%) containing at least one PRDM9 binding-motif. This relatively low proportion (10.1% expected with random location) might result from inaccuracies in hotspots identification or in binding-motif predictions. The hotspots might also be associated with other *PRDM9* alleles that we did not identify, for instance alleles associated with the *PRDM9* paralogue on BTAX or alleles that are no longer present in the population. Overall, when multiple *PRDM9* alleles contribute to recombination hotspots, motifs enrichment analysis might fail to identify specific *PRDM9* alleles. Finally, a

strong increase of RR in the H3K4me3 methylation mark was observed. This mark is known to be deposited by PRDM9 after binding DNA and is mandatory for the initiation of the recombination process (Diagouraga *et al.*, 2018), further supporting the implication of PRDM9 in meiotic recombination in cattle.

More generally, we found that RR increased in CpG islands and in open chromatin regions as seen in human (Halldorsson *et al.*, 2019), and decreased in regions associated to active promoters. CO would thus be kept away from genes and motifs should evolve rapidly; the recombination landscape could thus differ among breeds or between closely related species.

References

Baudat F., Buard J., Grey C., Fledel-Alon A., Ober C., *et al.* (2010) Science 327(5967):836-840. https://doi.org/10.1126/science.1183439

Chan A.H., Jenkins P.A., and Song Y.S. (2012) PLoS Genet. 8(12):e1003090. https://doi.org/10.1371/journal.pgen.1003090

Coop G. and Przeworski M. (2007) Nat. Rev. Genet. 8(1):23-34. https://doi.org/10.1038/nrg1947

Delaneau O., Zagury J.-F., Robinson M.R., Marchini J.L., and Dermitzakis E.T. (2019) Nat Commun 10(1):5436. https://doi.org/10.1038/s41467-019-13225-y

Diagouraga B., Clément J.A.J., Duret L., Kadlec J., de Massy B., *et al.* (2018) Molecular Cell 69(5):853-865.e6. https://doi.org/10.1016/j.molcel.2018.01.033

Druet T. and Georges M. (2015) Bioinformatics 31(10):1677-1679. https://doi.org/10.1093/bioinformatics/btu859

Fearnhead P. (2006) Bioinformatics 22(24):3061-3066. https://doi.org/10.1093/bioinformatics/btl540

Goszczynski D.E., Halstead M.M., Islas-Trejo A.D., Zhou H., and Ross P.J. (2021) Genome Res. 31(4):732-744. https://doi.org/10.1101/gr.267336.120

Grant C.E., Bailey T.L., and Noble W.S. (2011) Bioinformatics 27(7):1017-1018. https://doi.org/10.1093/bioinformatics/btr064

Halldorsson B.V., Palsson G., Stefansson O.A., Jonsson H., Hardarson M.T., *et al.* (2019) Science 363(6425). https://doi.org/10.1126/science.aau1043

Kong A., Thorleifsson G., Gudbjartsson D.F., Masson G., Sigurdsson A., *et al.* (2010) Nature 467(7319):1099-1103. https://doi.org/10.1038/nature09525

Ma L., O'Connell J.R., VanRaden P.M., Shen B., Padhi A., *et al.* (2015) PLoS Genet. 11(11):e1005387. https://doi.org/10.1371/journal.pgen.1005387

Oget-Ebrad C., Kadri N.K., Moreira G.C.M., Karim L., Coppieters W., *et al.* (2021) bioRxiv:2021.10.27.466052. https://doi.org/10.1101/2021.10.27.466052

Peñalba J.V. and Wolf J.B.W. (2020) Nat. Rev. Genet. 21(1):476-492. https://doi.org/10.1038/s41576-020-0240-1

Persikov A.V. and Singh M. (2014) Nucleic Acids Res. 42(1):97-108. https://doi.org/10.1093/nar/gkt890

Zhou Y., Shen B., Jiang J., Padhi A., Park K.-E., *et al.* (2018) DNA Res. 25(2):183-194. https://doi.org/10.1093/dnares/dsx048