Author
Fardokhtsadat Mohammadi

Submission
Institute of Biophysics

Thesis Supervisor
Assoc. Prof. Dr. Irene Tiemann-Boege

Assistant Thesis Supervisor
Dr. Philipp Hermann

November 2020

FINE-SCALE RECOMBINATION MAPS OF THE CATTLE GENOME INFERRED BY LINKAGE DISEQUILIBRIUM

Bachelor's Thesis

to confer the academic degree of

Bachelor of Science

in the Bachelor's Program

Bioinformatics

JOHANNES KEPLER UNIVERSITY LINZ
Altenberger Straße 69
4040 Linz, Austria
jku.at

Bibliographical Detail

Mohammadi, F., 2020: Fine-Scale Recombination Maps of the Cattle Genome Inferred by Linkage Disequilibrium. Bachelor Thesis, in English. - 35 p., Institute for Biophysics, Johannes Kepler University, Linz, Austria

Annotation

Recombination is a genetic event that occurs during meiosis and leads to the exchange of genetic material between paternal and maternal homologous chromosomes. The intensity of recombination is shown to vary across genomes between and within species, yet the determinants of recombination patterns among populations of the same species are not fully understood. In this thesis, we estimated fine-scale, breed-specific recombination maps of a subset of chromosome 25 of Braunvieh and Fleckvieh cattle for different populations with respect to inbreeding coefficients using the R package *LDJump* under two assumptions, neutrality and demography. Moreover, we studied the association between recombination rates and genomic features such as SNP count, GC content, and the density and nature of genes. We observed a statistically-significant, weak negative correlation between recombination rates and SNP count, where low recombination rates are accompanied by higher SNP count, and vice versa. More complex demographic scenarios as well as the level of inbreeding should be incorporated in further research using *LDJump* to address this possible association between SNP count and recombination rates. On the contrary, we observed no such relationship between recombination rates and GC content. We detected a substantial difference in gene density between the lowest and highest SNP-count regions of chromosome 25.

Declaration

I hereby declare that I have worked on my Bachelor's thesis independently and used only the sources listed in the bibliography.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my Bachelor's thesis, which is kept in full form in the Faculty of Science archive and in electronic form in the publicly accessible part of the STAG database operated by University of South Bohemia in České Budějovice accessible through its web pages.

Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defense in accordance with the aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagiarism detection system.

....
Place, Date

....
Fardokhtsadat MOHAMMADI

Acknowledgments

I would like to thank my supervisors, Assoc. Prof. Dr. Irene Tiemann-Boege and Dr. Philipp Hermann, for their consistent support and guidance. Furthermore, I would like to thank Prof. Dr. Hubert Pausch and his team for providing some of the datasets used in this thesis. Finally, I would like to express my gratitude and appreciation for my family, whose guidance, support and encouragement has been invaluable throughout my life.

Contents

List of Figures

List of Tables

1 Introduction

Recombination is an evolutionary important biological process in eukaryotes that leads to the shuffling of genetic material and the creation of new traits in the offspring (Jensen-Seaman et al., 2004). Several studies have shown that recombination does not occur randomly across genomes, but is concentrated in specific regions known as recombination hotspots (Thomsen et al., 2001; Paigen & Petkov, 2010).

Genomic sequence features such as distance from the centromere (Jensen-Seaman et al., 2004), GC content (Weng et al., 2014; Galtier et al., 2001; Jensen-Seaman et al., 2004), the presence of repeats, and the density and nature of genes (Kong et al., 2002) can affect the frequency of recombination (Majewski & Ott, 2000).

The intensity of recombination events can be either species specific or sex specific. For example, humans show higher recombination rates in comparison to rats and mice (Jensen-Seaman et al., 2004). In many species such as the human (Kong et al., 2002; Otto & Lenormand, 2002), mouse (Dietrich et al., 1996), and dog (Neff et al., 1999), females usually show higher recombination rates compared to males (Shen et al., 2018). In cattle, the recombination activity in males is shown to be higher (Shen et al., 2018) or equal (Paigen & Petkov, 2010) to that of females.

Recombination events can be studied using several approaches differing in genome-wide coverage and resolution (Hermann, Heissl, et al., 2019) such as i) sperm typing (Li et al., 1988), which leads to high-resolution events in regions of a few hundred base pairs (Arnheim et al., 2007); ii) pedigree analysis (Sobel & Lange, 1996), which provides resolution in the order of tens of kilobases (Arnheim et al., 2003); and iii) the analysis of patterns of linkage disequilibrium, which presents the accumulation of genome-wide historical recombination events (Tapper et al., 2005).

High levels of recombination have shown to decrease the accuracy of phasing and genotype imputation and, therefore, a better understanding of variability in recombination rates across a genomic region may help to improve the accuracy of haplotype phasing and genotype imputation (Weng et al., 2014). Moreover, producing genome-wide recombination maps may facilitate breeding strategies implemented to decrease inbreeding levels and increase effective population size (Shen et al., 2018; Thomsen et al., 2001).

Thus far, several studies have analyzed species-specific and sex-specific recombination rates, but only a few have studied breed-specific recombination in cattle while taking the coefficient of inbreeding into account (Thomsen et al., 2001; Sandor et al., 2012; Ma et al., 2015; Kadri et al., 2016; Shen et al., 2018). Moreover, to our knowledge, no study has investigated the correlation between recombination rates and the SNP count of a genomic region in cattle.

In the present study, we extended and used an R package entitled *LDJump* (Hermann, Futschik, & Mohammadi, 2019) to infer fine-scale recombination maps based on patterns of linkage disequilibrium (Hermann, Heissl, et al., 2019) in two Swiss breed cattle populations, Braunvieh and Fleckvieh. Our aims were to i) identify the highest and lowest SNP-count regions along chromosome 25 for both populations, ii) split each population into three subsets based on the levels of inbreeding among the individuals to allow the detection of inbreeding patterns, iii) compute the recombination rates of the aforementioned genomic regions under neutrality and demography, iv) detect breed-specific and species-specific recombination patterns, and v) investigate the correlation between local recombination rates and several factors such as SNP count, GC content, and the number and nature of genes.

2 Background

2.1 Single Nucleotide Polymorphism

Recombination shuffles genetic material and produces new variants in the given genome (Jensen-Seaman et al., 2004). One such variation is known as *single nucleotide polymorphism* (Gu et al., 1998), abbreviated to SNP. SNPs are single nucleotide variations in specific positions on the genome. A variation can be considered a SNP if the less frequent allele is present in more than 1% of the general population (Brookes, 1999).

2.2 Inbreeding, hybridization, homozygosity, heterozygosity, and fitness

The mating of closely related individuals within a population is known as inbreeding (Pekkala et al., 2014). Inbreeding increases the likelihood of deleterious traits in a population, leads to the loss of genetic diversity, and increases homozygosity (Stachowicz et al., 2011; Fenster & Galloway, 2000; Pekkala et al., 2014). Homozygosity refers to the possession of two identical alleles of a particular gene, whereas heterozygosity is a condition wherein two different alleles of a particular gene are present (Ayala, 1978). In contrast to inbreeding, hybridization among different lines or populations potentially reduces the effect of inbreeding by increasing heterozygosity and producing offspring which are fitter than their ancestors (Fenster & Galloway, 2000; Pekkala et al., 2014). Fitness, in population genetics, is a term that describes reproductive success and the adaption of the individual to its environment (Orr, 2009).

2.3 Demography and neutrality

Demography is the study of populations and the processes through which populations and their size change (Tarsi & Tuff, 2012). Genetic bottlenecks and population growth are examples that take the demographic history of a population into account leading to a reduction and an increase in the population size, respectively. Neutrality refers to the neutral theory of molecular evolution which holds that most changes at the molecular level of cells are caused by random genetic drift and are not due to natural selection (Kimura, 1979).

2.4 Variant Call Format

Files in the variant call format (VCF) are used to store genetic variation data, such as insertions, deletions, and SNPs (Danecek et al., 2011). The VCF format allows the storage of multi-sample sequence variation, meaning that the genetic information of multiple individuals of a population can be stored. A VCF file consists of two sections: a header section and a data section. The header section stores meta-information with a standard description of the data. The data section comprises several columns that describe the sequence variations. Each variant is described by the chromosome (CHROM), the position (POS), a unique identifier (ID), the reference allele (REF), the alternative non-reference allele (ALT), a phred-scaled quality score (QUAL), site-filtering information (FILTER), and user-extensible annotation (INFO). Each row in the data section represents one variant for all individuals in the dataset specifying the zygosity of the individual. In diploid organisms, an individual can be either homozygous or heterozygous; where homozygosity is denoted as "0|0" or "1|1", and heterozygosity is denoted as "1|0" or "0|1". The value of 0 refers to the reference allele, 1 refers to the alternative allele. If the variant is either not present or information about it is missing, the genotype for the individual is denoted as ".|.". The separator can be of two types: "|" or "/", indicating whether the genotype is phased or unphased, respectively. Phased data indicates whether a variant is inherited from the father or the mother, whereas unphased data does not determine which one of the pair of chromosomes holds the variant.

3 Materials

The data analysis of this study is based on genome datasets from two Swiss cattle breeds known as “Braunvieh” and “Fleckvieh”. The Braunvieh dataset comprises 91 individuals, the Fleckvieh 161 individuals. The genotyped data is provided in VCF format, together with the reference genome (ARS-UCD1.2) in FASTA format (Rosen et al., 2020a). In this thesis, we choose chromosome 25 for the data application, which has a length of 42,350,435 base pairs. The total number of SNPs is 338,122 and 428,439 SNPs in the Braunvieh and Fleckvieh dataset, respectively. The dataset and R scripts used in this thesis can be found in the GitHub repository `https://github.com/fardokhtsadat/LDJump-thesis`.

4 Methods

4.1 *LDJump*

4.1.1 Update of *LDJump*

LDJump (Version: 0.2.2) is an R package estimating parsimonious recombination maps of population genetic data provided in FASTA format in a two-step process. First, the DNA sequence under study is divided into segments of user-defined length. For each segment several summary statistics are computed and input into a regression model to estimate the constant recombination rate. Next, *LDJump* estimates the change points in the recombination rate using a segmentation algorithm (Frick et al., 2014). This method allows demography to be taken into account. The newly introduced update of *LDJump* (Version: 0.3.1) enables the analysis using VCF files as input.

4.1.2 *LDJump*’s workflow with VCF files

In order to run *LDJump* on VCF files, two types of files are required: i) a VCF file to be used for the analysis and ii) a reference FASTA file of the same genomic region as the VCF file. The workflow of *LDJump* for both file formats, FASTA and VCF, is shown in Figure 1. We implemented two new functions - *vcf_statistics()* and *vcfR_to_fasta()* - using the reference FASTA file to convert the VCF file into FASTA format. The function *vcf_statistics()* uses *VCFTools* (Adam Auton, 2020) to segment the VCF file according to the segment length defined by the user. Each segmented VCF file is then converted into FASTA format using the *vcfR* package (Knaus & Grünwald, 2017). The newly produced FASTA files serve as input to *LDJump*. Subsequently, *LDJump* computes the recombination rates for each segment. The computation of recombination rates can be sped up through parallelization and the use of several threads.

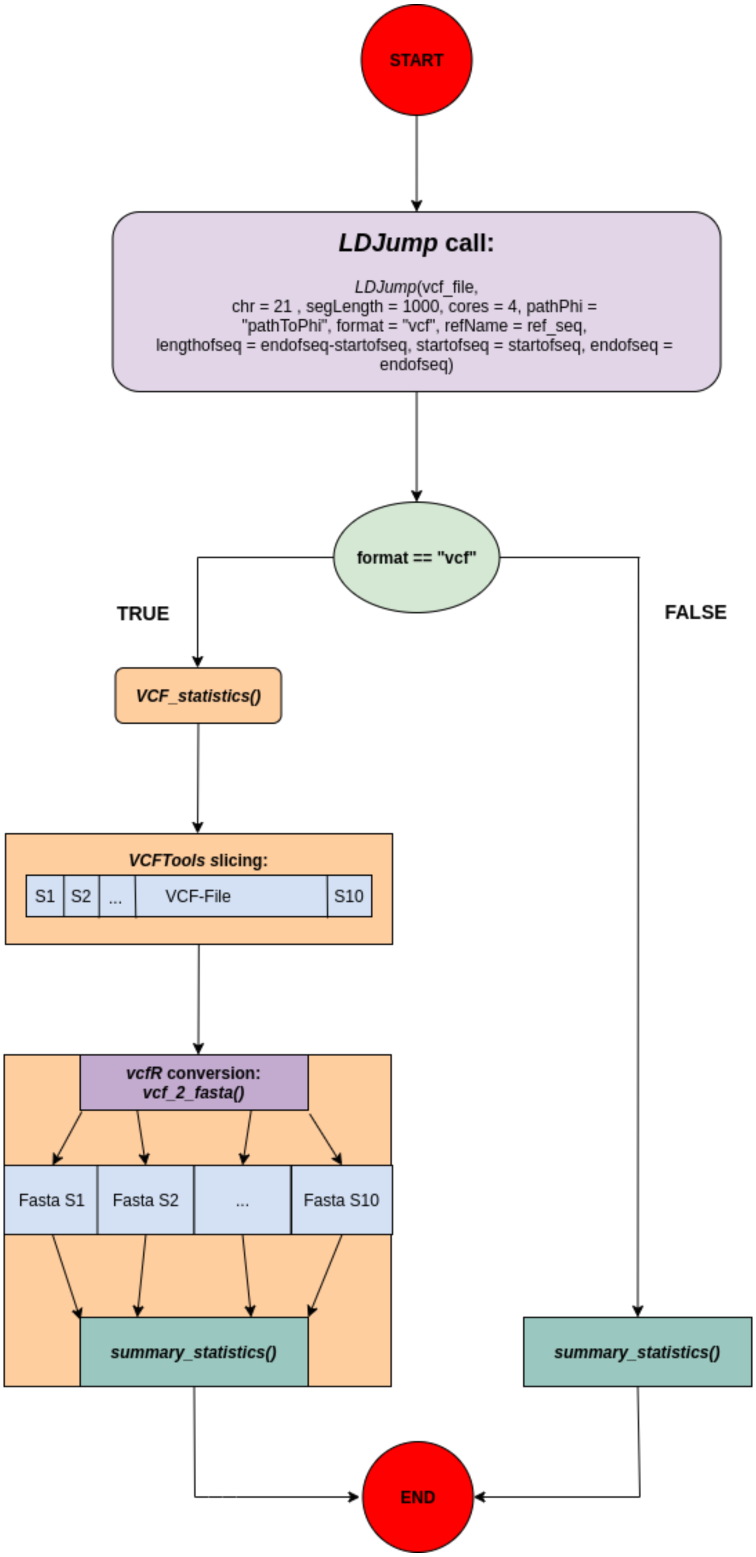

Figure 1: Workflow of *LDJump* is shown for both file formats: VCF (left), FASTA (right). In this example, *LDJump* is applied on a VCF file of chromosome 21. The *segLength* argument is set to 1000, meaning that *LDJump* will divide the VCF file into 1000 base-pair segments. Based on the format of the input file, *LDJump* selects the next applied function. In the case of FASTA files, the summary statistics are calculated for each segment immediately, whereas for VCF files, *VCFTools* is used to segment the file and then convert it to FASTA. Subsequently, the summary statistics are calculated for each segment. To speed up the calculation, four cores are used.

4.1.3 Validation of the update

In order to check the equality of the new update with the existing algorithm we performed a test run where *LDJump* is applied to chromosome 21:41,187,000-41,290,679 (103,679bp) - once using the VCF format (Figure 2A) and once using the equivalent FASTA format (Figure 2B). The population under study comprises 107 human individuals with 3505 SNPs. The recombination rates are estimated per 1000 base pairs for both file formats (FASTA and VCF). The output of *LDJump* for both input files proved to be identical. The recombination maps are shown in Figure 2.

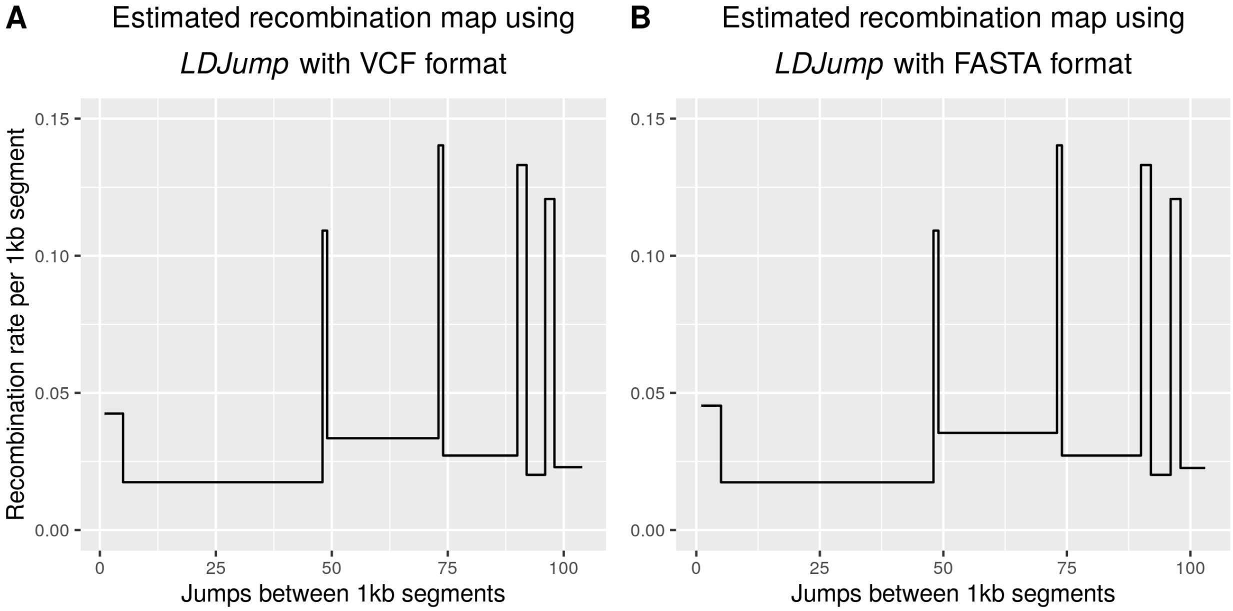

Figure 2: Comparison of the results of applying *LDJump* to the FASTA and VCF formats. Here, we present the recombination maps that resulted from applying *LDJump* to two different file formats (FASTA, VCF) on the same genomic region. The genomic region is 103679 base pairs long and the dataset comprises 107 individuals with a total SNP count of 3505. The x-axis represents the number of segments, the y-axis shows the recombination rate per segment. The segment size in the test run was set to 1000 base pairs.

In this test run, we also measured the run time for each file format. On a standard desktop (Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz, 8 GB RAM) using three threads, the run time of the two applications for the FASTA file and VCF file were 2.17 (FASTA) and 2.28 (VCF) hours, respectively. The different run time arises from the conversion of every segment from VCF to FASTA.

4.2 Estimating the degree of relationship among individuals in each cattle population

The relationship between two individuals can be described with a coefficient of relationship ranging from 0 to 1. Coefficient values close to 1 indicate a higher degree of relationship and higher levels of inbreeding, whereas values close to 0 refer to individuals with a distant common ancestor (Wright, 1922). Table 1 provides an overview of several coefficients of relationship.

Degree of Relationship	Relationship	Coefficient of Relationship
0	identical twins, clones	100%
1	parent-offspring	50%
2	full siblings	50%
2	3/4-sibling or sibling-cousins	37.5%
2	grandparent-grandchild	25%
2	half-siblings	25%
3	aunt/uncle - nephew/niece	25%
3	great grandparent-great grandchild	12.5%
4	first cousins	12.5%
6	quadruple second cousins	12.5%
6	triple second cousins	9.38%
4	half-first cousins	6.25%
5	first cousins	6.25%
6	double second cousins	6.25%
6	second cousins	3.13%
8	third cousins	0.78%
10	fourth cousins	0.20%

Table 1: The coefficient of relationship and the corresponding degrees of relationship are shown.

To account for different levels of inbreeding, the individuals of each cattle population (Braunvieh, Fleckvieh) are grouped into three categories differing in the degree of relationship among the individuals. The first category comprises all individuals - no cut-off is imposed. For the second category individuals with a coefficient value of greater than 0.125 are excluded. The third category only contains individuals with coefficients of lower than 0.0625.

The coefficient of relationship between each pair of individuals is estimated using the *PLINK* software package (Purcell, 2020). *PLINK* is an open-source whole-genome association study (WGAS) tool that allows the efficient manipulation and analysis of large datasets (Purcell et al., 2007). For each population, the VCF files of all chromosomes are merged (Heng Li, 2020) and *PLINK* is applied. All variants with a minor allele frequency below the threshold of 0.01 are filtered and the sex of individuals is ignored.

The output of *PLINK* is a symmetric $n \times n$ square matrix where n denotes the number of individuals. The relationship matrix contains coefficients of relationship for each pair. The complete relationship matrices for both populations can be found in the GitHub repository `https://github.com/fardokhtsadat/LDJump-thesis`.

To obtain the subsets based on the cut-off values of 0.125 and 0.0625 for each individual, we count the pairs in which the relationship coefficient is higher than the cut-off. Then, we repeatedly remove the individual with the highest sum of relationships until no individual exceeds the threshold (0.125, 0.0625). If two individuals have an identical maximum number of relationships, one of them will be removed randomly.

4.3 Detecting the highest and lowest SNP-count regions

In this thesis, we search for the genomic regions that most likely contain information in the form of variation between the two populations. Using *VCFTools* (Adam Auton, 2020) we compute the SNP count along the chromosome 25 per 4000 base-pair segments for both populations (Braunvieh, Fleckvieh). The SNP count is then used to scan the chromosome for the highest and lowest SNP-count region. We define the highest count region (HCR) and the lowest count region (LCR) as genomic regions of a certain length that contain the maximum and minimum number of SNPs along the chromosome, respectively. To obtain the HCR and LCR, a sliding window (Anderson et al., 2019) of two million base pairs in size is passed along chromosome 25 with a step size of 4000 base pairs; the concept of the sliding window algorithm is visualized in Figure 3. Next, the SNP count among all windows is compared and only the regions with highest and lowest SNP count are chosen.

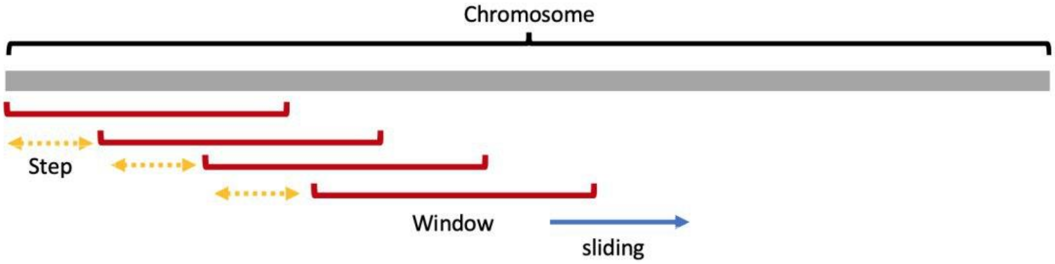

Figure 3: Visualization of a sliding window approach. In the sliding window algorithm, a window of certain length is passed along data allowing to capture different portions of it. This window is passed along the chromosome with a certain step size. In each step, the SNP count of that window is obtained.

5 Results

5.1 SNP-distribution analysis along chromosome 25 of two cattle populations

From the genotyped data of chromosome 25, we identified 385,119 SNPs in the Braunvieh population and 471,754 SNPs in the Fleckvieh population. The distribution of SNP counts per 4000 base pair segments is shown in Figure 4. The distribution is more symmetrical for the Fleckvieh population, whereas the distribution of the Braunvieh population is more skewed to the right and exhibits a strong peak at approximately 20 SNPs per segment. In Table 2, the SNP distribution of both cattle populations is described by means of summary statistics. The minimum SNP count for both populations is 0, meaning that there is at least one segment with 0 SNPs. The maximum number of SNPs found in one segment is 358 for Braunvieh and 310 for Fleckvieh. The mean SNP count is 36.37 SNPs and 44.56 SNPs, whereas the median is 30 SNPs and 39 SNPs for Braunvieh and Fleckvieh, respectively.

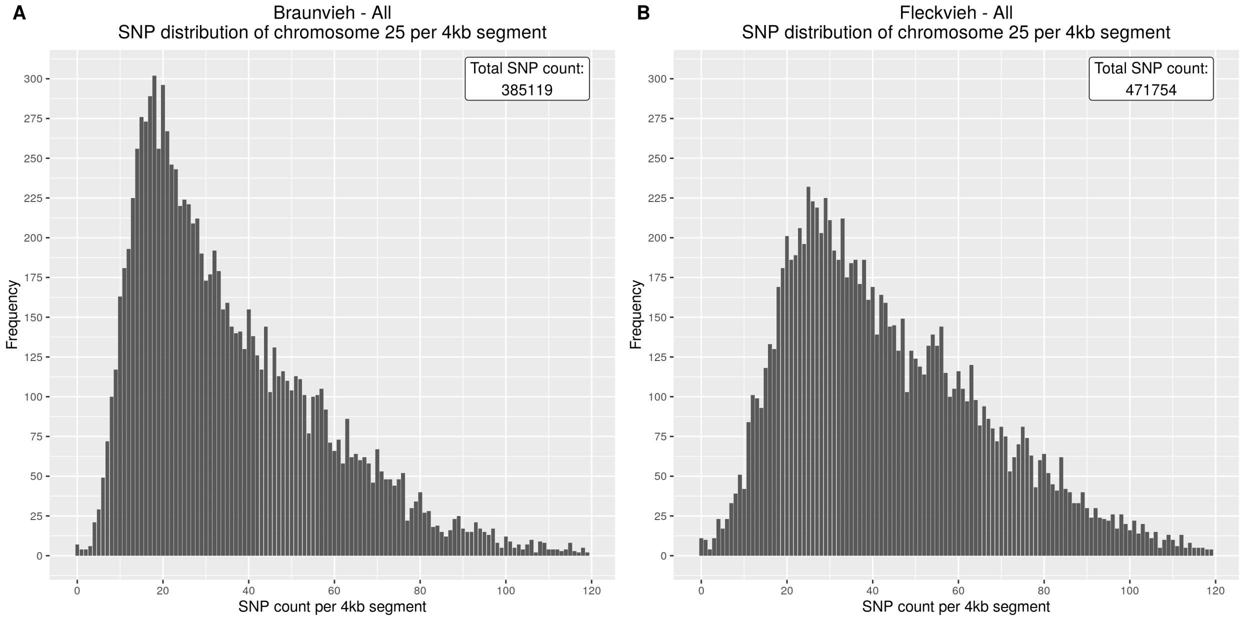

Figure 4: Distribution of SNP count of chromosome 25 per 4000 base-pair segments. A histogram for each cattle population is shown representing the frequency of SNPs per 4000 base-pair segments. The Braunvieh population's SNP distribution (panel A) contains a strong peak at approximately 20 SNPs per 4000 base-pair segments. The histogram of the Fleckvieh population (panel B) shows a more symmetrical SNP distribution. With regard to total SNP counts, the Fleckvieh population has 86635 more SNPs than the Braunvieh population.

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
Braunvieh	0	19	30	36.37	49	358
Fleckvieh	0	26	39	44.56	59	310

Table 2: Description of the SNP distribution of chromosome 25 per 4000 base-pair segments. The SNP distribution of each cattle population (Braunvieh, Fleckvieh) is described by the minimum, 1st quartile, median, mean, 3rd quartile, and maximum per 4000 base pair segment.

5.2 Identification of the highest and lowest SNP-count regions of chromosome 25 for two cattle populations

Using the sliding window algorithm with a window size of two million base pairs, we identify the regions of lowest and highest SNP count along chromosome 25 for each cattle population (Braunvieh, Fleckvieh). The lowest SNP-count region ranges from 36,000 to 2,036,000 base pairs for both populations. The highest SNP-count regions range from 10,728,000 to 12,728,000 and from 10,692,000 to 12,692,000 for Braunvieh and Fleckvieh, respectively. An overview of the region coordinates is given in Table 3.

Population	Region	Start position	End position
Braunvieh	LCR	36000	2036000
Fleckvieh	LCR	36000	2036000
Braunvieh	HCR	10728000	12728000
Fleckvieh	HCR	10692000	12692000

Table 3: Starting and ending positions of the lowest and highest SNP-count regions. "HCR" denotes the highest SNP-count region and "LCR" the lowest SNP-count region. The lowest SNP-count region of both populations (Braunvieh, Fleckvieh) is the same for a window size of two million base pairs, whereas the highest SNP-count region differs by 36,000 base pairs.

A SNP-count map for both cattle populations is shown in Figure 5, which depicts the SNP count of the highest and lowest SNP-count regions. To obtain the SNP-count maps, we calculate the SNP count per 4000 base pair segment. The SNP-count maps show the genomic range starting from 36,000 base pairs to 12,728,000 base pairs, including both the highest and lowest SNP-count regions for each cattle population. Additionally, the average number of SNPs per 4000 base-pair segments is shown with a horizontal solid line, which indicates an average SNP count of ~36 SNPs for Fleckvieh and ~45 SNPs for Braunvieh.

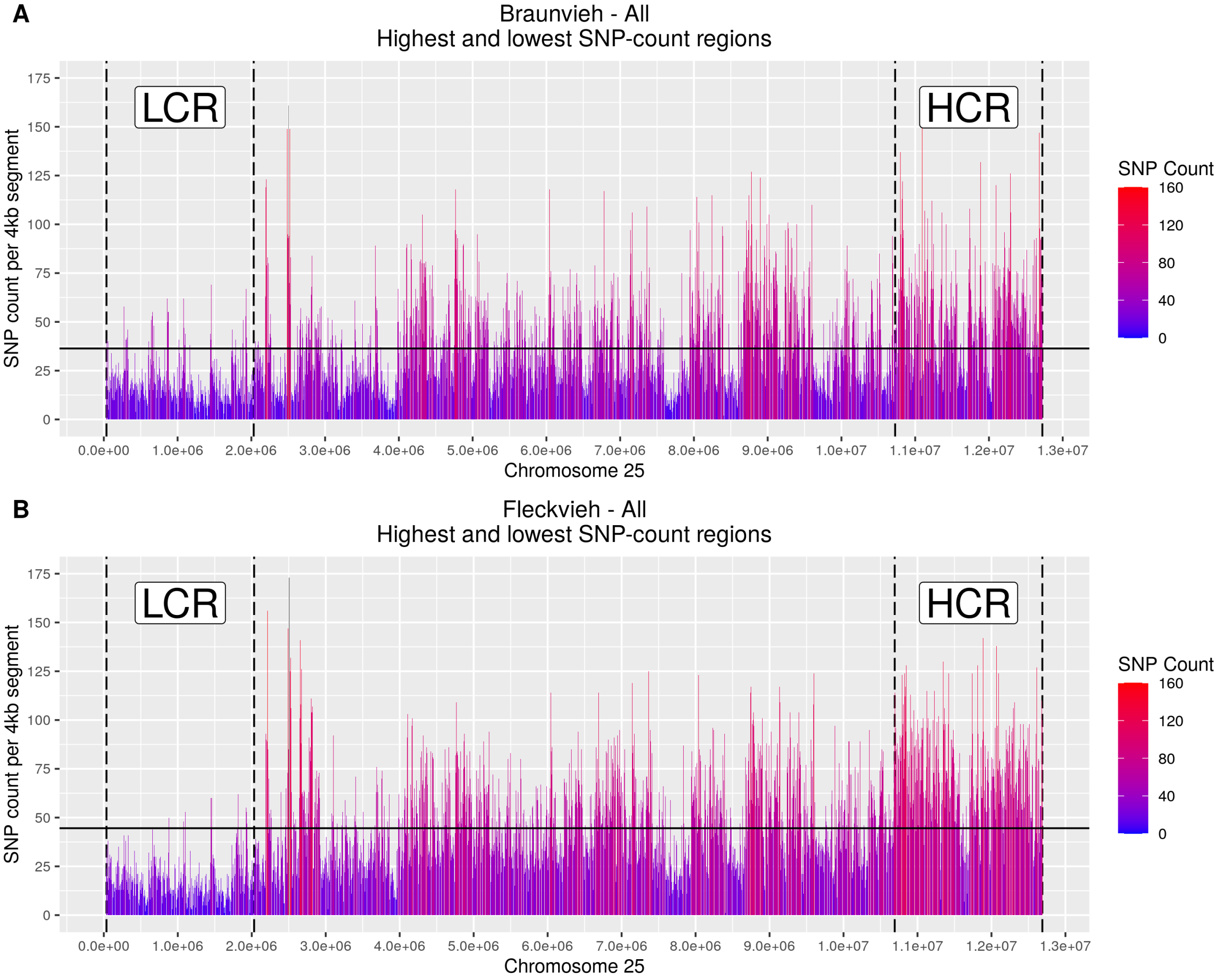

Figure 5: SNP-count plot of the highest and lowest SNP-count region of chromosome 25. The genomic range starts at 36,000 base pairs and ends at 12,728,000 base pairs. The regions of highest and lowest SNP count for both populations are labelled as "HCR" and "LCR", respectively, and their starting and ending positions are marked with vertical long-dashed lines. The solid horizontal line denotes the average SNP count of the population of chromosome 25. The average SNP count per 4000 base-pair segments is 36 SNPs and 45 SNPs for Braunvieh and Fleckvieh, respectively.

5.3 Estimation of recombination rates under neutrality and demography using genotyped cattle data

Using the software *LDJump* (Hermann, Futschik, & Mohammadi, 2019), we estimated fine-scale recombination maps of the highest and lowest SNP-count regions of chromosome 25 in two Swiss cattle breeds, Fleckvieh and Braunvieh. Each cattle population is grouped into three subpopulations according to the degree of relationship among the individuals. The relationship coefficients are estimated using *PLINK* (Purcell, 2020), where a higher value implies a stronger relationship.

The first group represents the whole population, i.e. no cut-off is imposed. In the second group, we impose a cut-off of 0.125, meaning that individuals with a relationship coefficient of higher than 0.125 are removed. The third and smallest group consists of individuals with a relationship coefficient of lower than 0.0625. The Braunvieh population comprises 91 individuals. Setting a cut-off value of 0.125 removes 34 individuals from the population (57 individuals remain); the more stringent value of 0.0625 removes 49 individuals (42 individuals remain). The Fleckvieh population comprises 161 individuals. Imposing a cut-off value of 0.125 removes 84 individuals (77 individuals remain); the more stringent value 0.0625 removes 108 individuals (53 individuals remain). Table 4 provides an overview of the remaining individuals after the cut-off has been applied.

	Number of individuals		
	No cut-off	0.125	0.0625
Braunvieh	91	57	42
Fleckvieh	161	77	53

Table 4: Number of analysed individuals in the analysis for each cattle population. The remaining individuals after data selection are shown for each cattle population. The individuals were selected according to their degree of relationship. Each cattle population is grouped, where: i) no cut-off, ii) a 0.125 cut-off, or iii) a 0.0625 cut-off was applied.

The recombination rates for all subpopulations are computed per 4000 base-pair segments i) under neutrality, and ii) considering the demography of the population. If demography is considered in the calculation, the regression model estimates the recombination rates based on samples from populations under a bottleneck followed by rapid growth (Hermann, Heissl, et al., 2019). We are well aware of the fact that this demography scenario probably might not reflect the true demography of cattle breeds. However, as we are more interested in a first picture of a recombination map for cattle we ignore the discrepancies between the true and assumed demography setup. In any case, valuable information can be drawn from the comparison with the neutrality setup. The recombination maps for all subsets (no cut-off, 0.125, 0.0625) of both populations (Braunvieh, Fleckvieh) in the highest and lowest SNP-count region of chromosome 25 are shown in Figure 6 and 7, respectively.

The recombination maps estimated under demography contain more breakpoints and exhibit higher recombination rates, see Table 5. Moreover, the recombination maps of the Fleckvieh population show a higher number of breakpoints than the Braunvieh population, which may be due to the higher SNP count in Fleckvieh. In the highest SNP-count region, the Fleckvieh population contains 35169 SNPs, whereas the Braunvieh population contains 26613 SNPs. Furthermore, the number of peaks decreases with decreasing sample size in both populations. Due to a limited number of SNPs in the lowest SNP-count regions presented in Figure 7, the estimation of recombination rates based on summary statistics is aggravated (Hermann, Heissl, et al., 2019).

	Braunvieh			Fleckvieh	
Group	Demography	Neutrality	Group	Demography	Neutrality
No cut-off	13	5	No cut-off	22	6
0.125	10	3	0.125	21	3
0.0625	10	4	0.0625	21	2

Table 5: Number of breakpoints introduced for each cattle population under demography or neutrality.

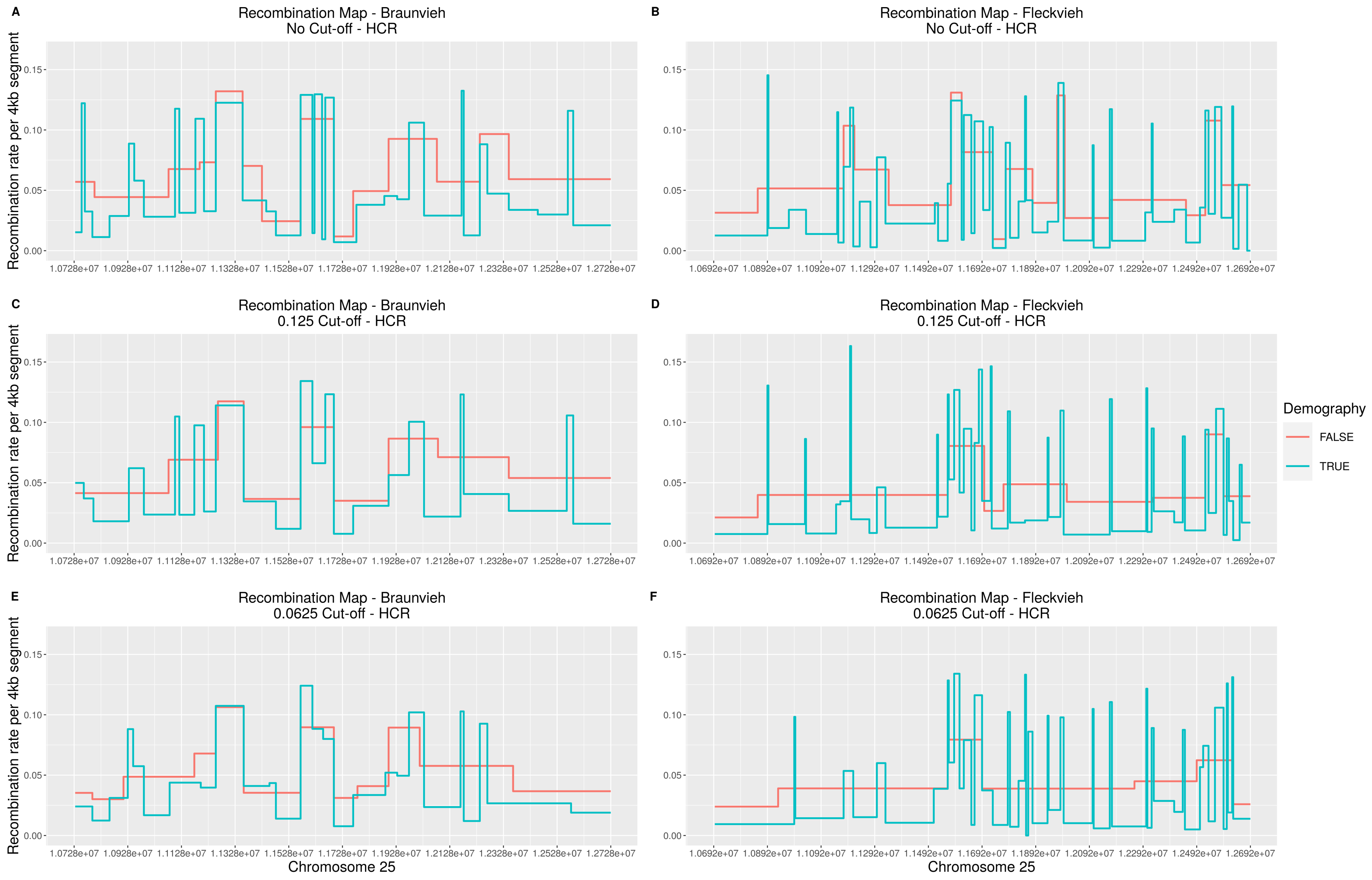

Figure 6: Recombination maps of the highest SNP-count region with different demography settings are shown. The recombination rates are estimated for all subsets (no cut-off - 1. row, 0.125 - 2. row, 0.0625 - 3. row) of both populations (Braunvieh - left panel A, C, E; Fleckvieh - right panel B, D, F) in the highest SNP-count region of chromosome 25. Each recombination map contains the recombination rates computed per 4000 base-pair segments under neutrality (red) or demography (blue).

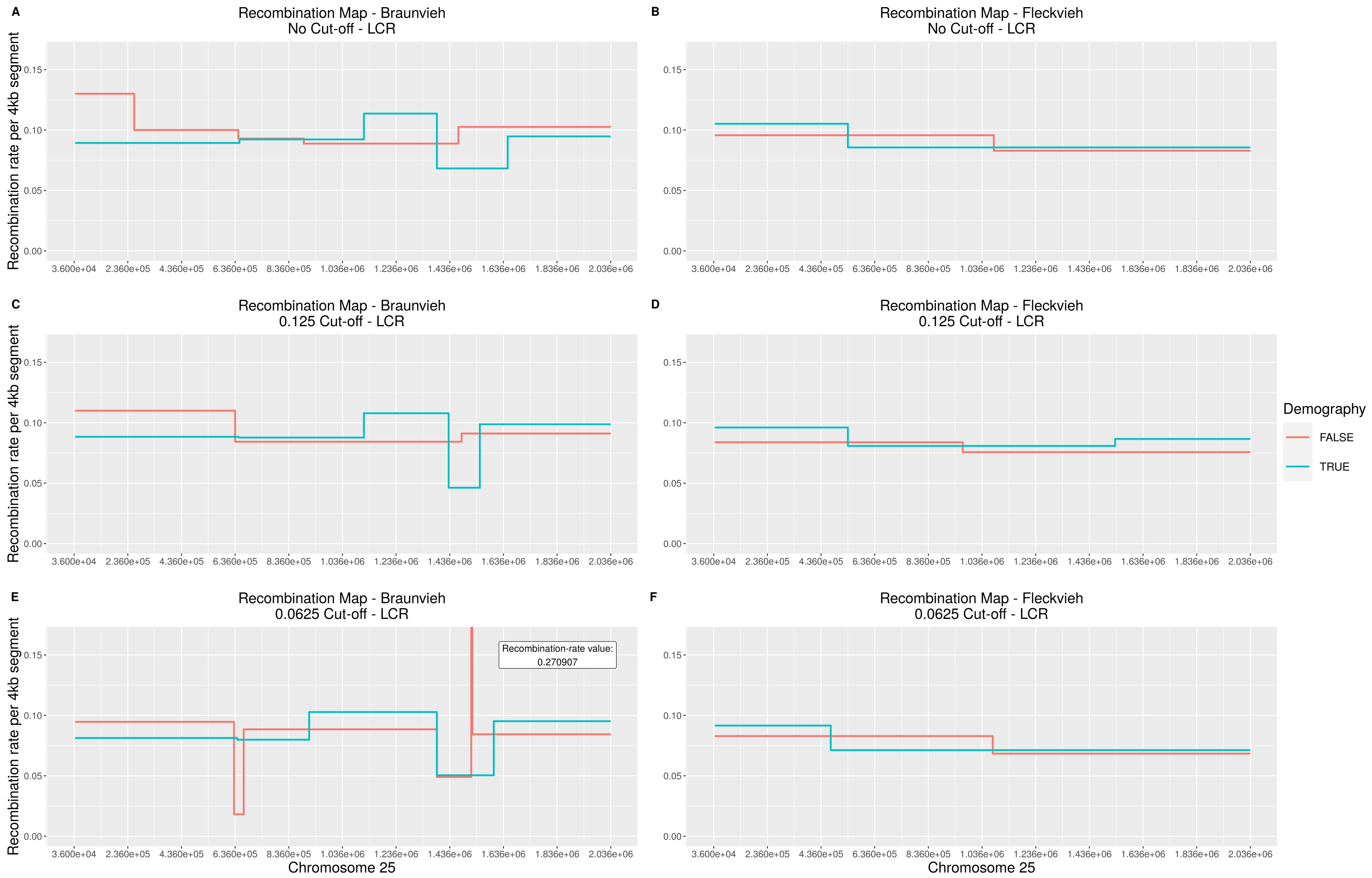

Figure 7: Recombination maps of the lowest SNP-count region with different demography settings are shown. The recombination is estimated for all subsets (no cut-off - 1. row, 0.125 - 2. row, 0.0625 - 3. row) of both populations (Braunvieh - left panel A, C, E; Fleckvieh - right panel B, D, F) in the lowest SNP-count region of chromosome 25. Each recombination map contains the recombination rates computed per 4000 base-pair segments under neutrality (red) or demography (blue).

5.4 Comparison of recombination patterns between two cattle breeds

To identify breed-specific recombination patterns we overlay the recombination maps of the Braunvieh and Fleckvieh population of all subpopulations (no cut-off, 0.125, 0.0625). Figure 8 shows the recombination maps estimated on the highest SNP-count region of chromosome 25 from 10,692,000 to 12,728,000 base pairs. Despite the difference of 36,000 base pairs, the recombination maps are aligned according to the genomic region. All recombination maps are estimated under demography using a segment length of 4000 base pairs.

The recombination background rate is defined as the median of the recombination rates, and recombination hotspots as regions having at least a 5-fold increase in recombination rates compared to the background rate (McVean et al., 2004; Chan et al., 2012; Hermann, Heissl, et al., 2019). To quantify the number of hotspots detected by *LDJump* in each subset we define a hotspot as a region with a minimum 3-fold increase in the background rate. In Figure 8, we present the lower boundary of the hotspot threshold as a dashed and dotted line for the Braunvieh and Fleckvieh population, respectively.

Based on the lower boundary of the hotspot threshold we count hotspots for each subpopulation. If a hotspot of one breed overlaps fully or partially with a hotspot of the other breed, it is considered a shared hotspot. To obtain the breed-specific hotspots, we count all hotspots that are not shared. The Braunvieh and Fleckvieh populations share 8 and 7 hotspots in the categories no cut-off and 0.125, respectively. In the 0.0625 category, Braunvieh shares 4 hotspots with Fleckvieh, whereas Fleckvieh shares 5 hotspots with Braunvieh. The total number of breed-specific and shared hotspots in both breeds (Braunvieh, Fleckvieh) is shown in Table 6.

	Breed-specific hotspots		Shared hotspots
Cutoff	Braunvieh	Fleckvieh	
No cut-off	2	9	8
0.125	1	12	7
0.0625	2	14	4 | 5

Table 6: Total number of breed-specific and shared hotspots in Braunvieh and Fleckvieh. The number of hotspots are listed for each category: i) no cut-off, ii) 0.125, and iii) 0.0625. The first two columns list the total number of breed-specific hotspots for Braunvieh and Fleckvieh, whereas the third and last column lists hotspots that are shared between the two populations.

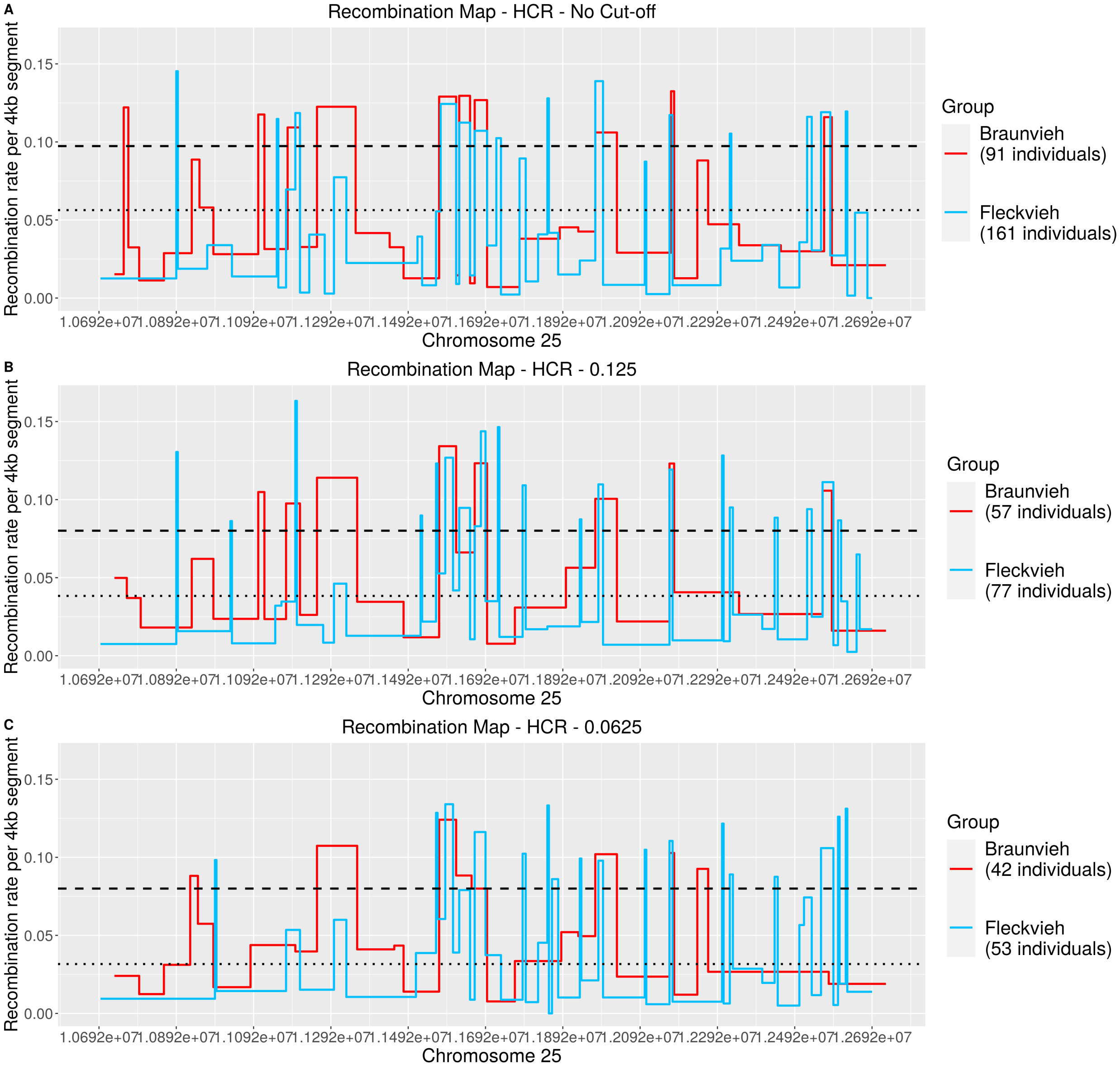

Figure 8: Comparison of recombination maps between the two cattle breeds (Braunvieh, Fleckvieh). The recombination rates between the Fleckvieh and Braunvieh populations are compared. Figure 10A, 10B, and 10C compare the recombination rates for each subset: i) no cut-off (top), ii) 0.125 (middle), iii) 0.0625 (bottom), respectively. The region with the highest count for Braunvieh ranges from 10,728,000 to 12,728,000 base-pairs, whereas the respective region for Fleckvieh ranges from 10,692,000 to 12,692,000 base-pairs. The dashed and the dotted line set the threshold for detecting recombination hotspots in Braunvieh and Fleckvieh, respectively. Recombination rates exceeding the lines are considered to be recombination hotspots.

5.5 Shared recombination patterns between breeds

In addition to the breed-specific comparison of recombination patterns, we constructed two recombination maps by combining the datasets of both populations (Braunvieh, Fleckvieh) with coefficients of relationship of 0.125 and 0.0625. Figure 9 shows these two recombination maps, in addition to the estimated recombination rates of the combined dataset.

In the combined dataset, we identified a total of 17 and 12 hotspots for the subpopulation 0.125 and 0.0625, respectively. The 0.125 subpopulation of the combined dataset shares 7 hotspots with both populations, 0 hotspots specifically with the Braunvieh and 6 hotspots with the Fleckvieh population, and 4 hotspots that are completely distinct from either population. The 0.0625 subpopulation of the combined dataset shares 5 hotspots with both populations, 1 hotspot is specifically shared with Braunvieh, 5 hotspots are shared with the Fleckvieh population, and 1 hotspot is distinct from either population. The total number of species-specific and shared hotspots in the combined dataset is shown in Table 7.

	Total number of HS	Individually shared HS		Overall shared HS	Distinct HS
Cut-off	Combined	Braunvieh	Fleckvieh		
0.125	17	0	6	7	4
0.0625	12	1	5	5	1

Table 7: Total number of specific-specific and shared hotspots in Braunvieh and Fleckvieh. The number of hotspots are listed for each subpopulation: i) 0.125, and ii) 0.0625. The first column lists the total number of hotspots for the combined dataset. The second and third column show the hotspots shared individually between the combined datasets and the two breeds (Braunvieh, Fleckvieh). The fourth and the fifth column describe the number of hotspots that are shared among all groups and the distinct hotspots for the combined dataset, respectively.

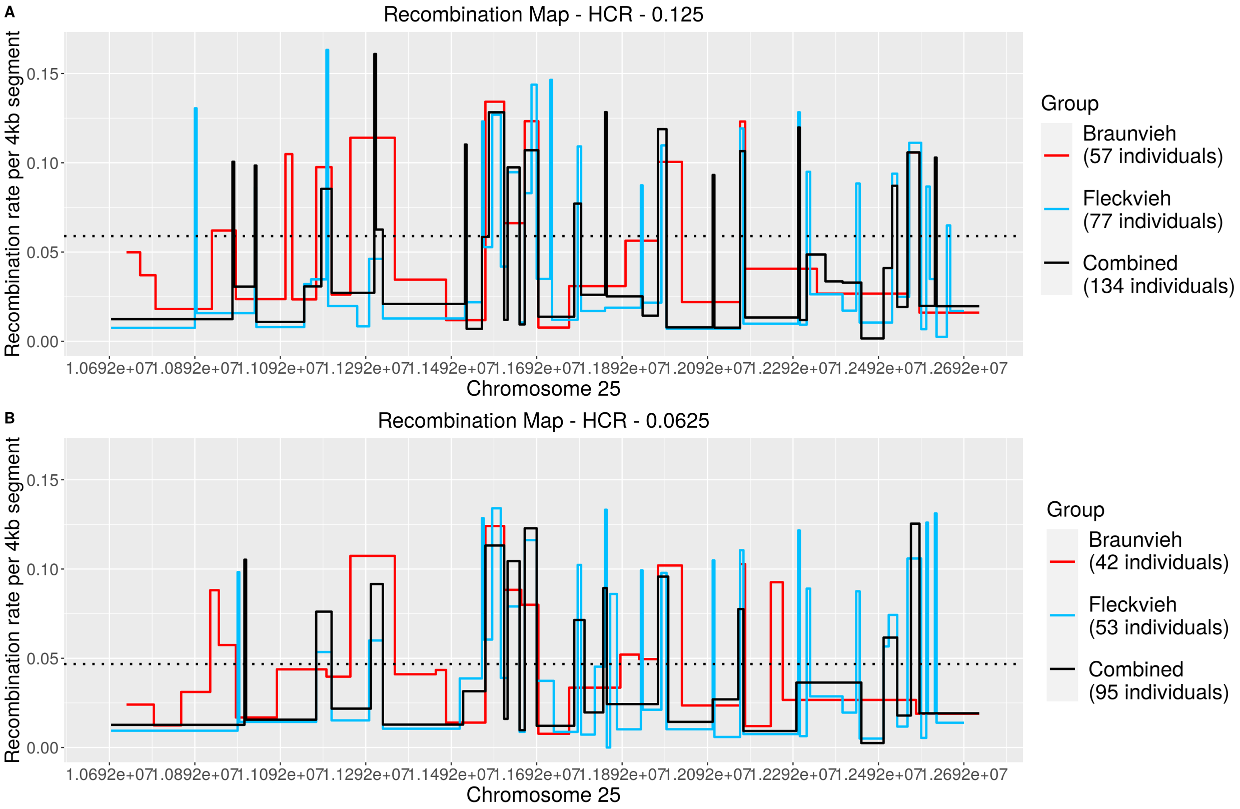

Figure 9: Shared recombination patterns between breeds. The recombination rates of the combined dataset are compared to both cattle breeds (Braunvieh, Fleckvieh) for the two subsets: i) 0.125 (top), ii) 0.0625 (bottom). The genomic region analyzed ranges from 10,692,000 to 12,728,000 base pairs. Figure 12A shows the recombination maps for the 0.125 subset, whereas Figure 12B shows the recombination maps for the 0.0625 subset. The dotted line sets the threshold for recombination hotspots in the combined dataset.

5.6 Comparison of recombination patterns with varying levels of inbreeding and their correlation to SNP count

We analysed the relationship between recombination rates and SNP count by overlapping the recombination maps of all subsets (no cut-off, 0.125, 0.0625) within a population (Braunvieh, Fleckvieh) and aligning the recombination maps to the SNP count of the respective genomic region.

In Figure 10A and 11A, we compare the subpopulation specific recombination maps of the Braunvieh and the Fleckvieh population, respectively. These figures contain the combined information of Figure 8 and Figure 5.

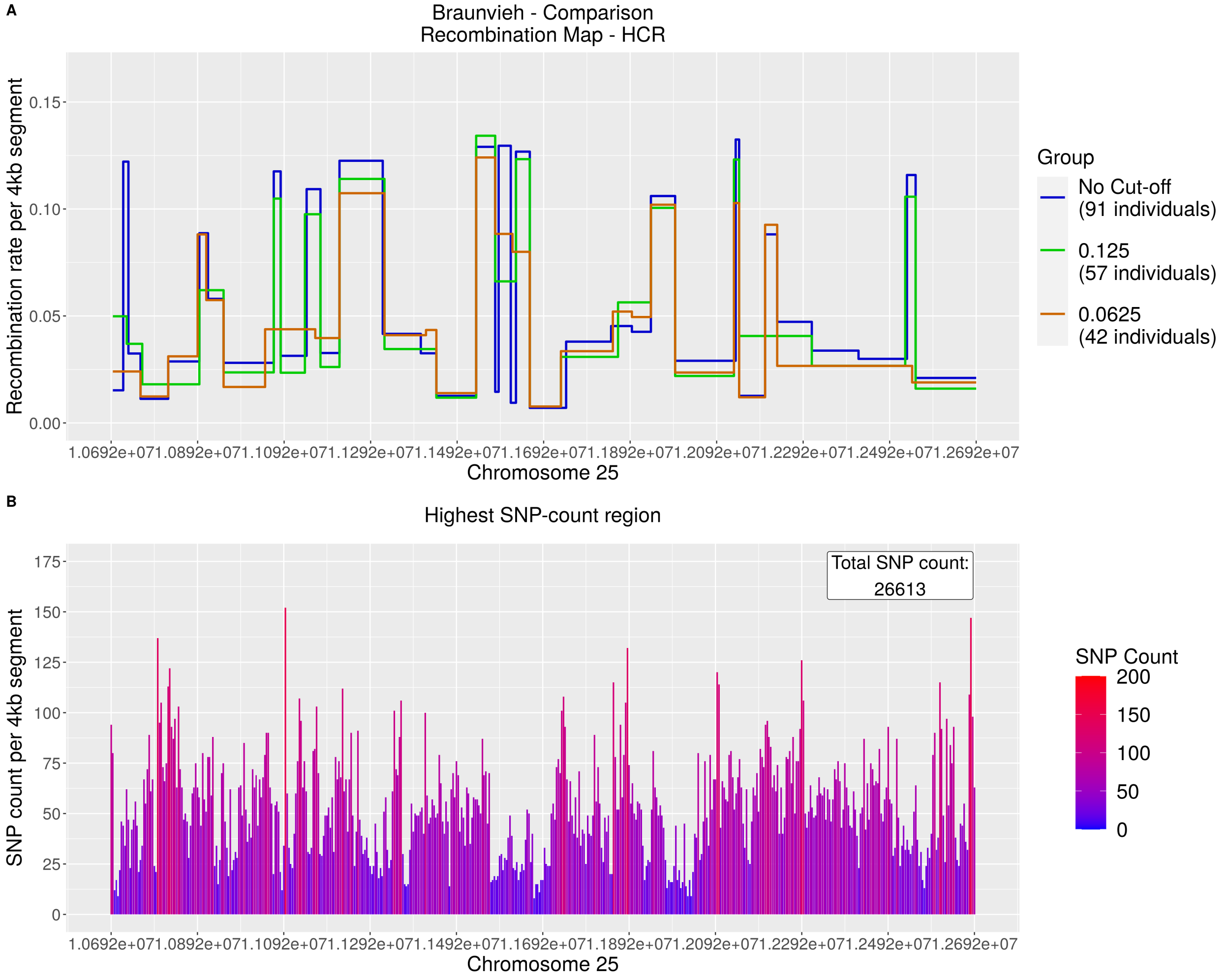

Figure 10: Collapsed recombination maps for Braunvieh. In panel A, the recombination maps for the region 10,692,000-12,692,000 base pairs of chromosome 25 of the Braunvieh subsets (no cut-off - blue, 0.125 - green, 0.0625 - brown) are collapsed. The recombination rates of the highest SNP-count region are estimated under demography per 4000 base-pair segments. Figure 10B shows the SNP count per 4000 base-pair segments of the Braunvieh population, with a total SNP count of 26613 SNPs.

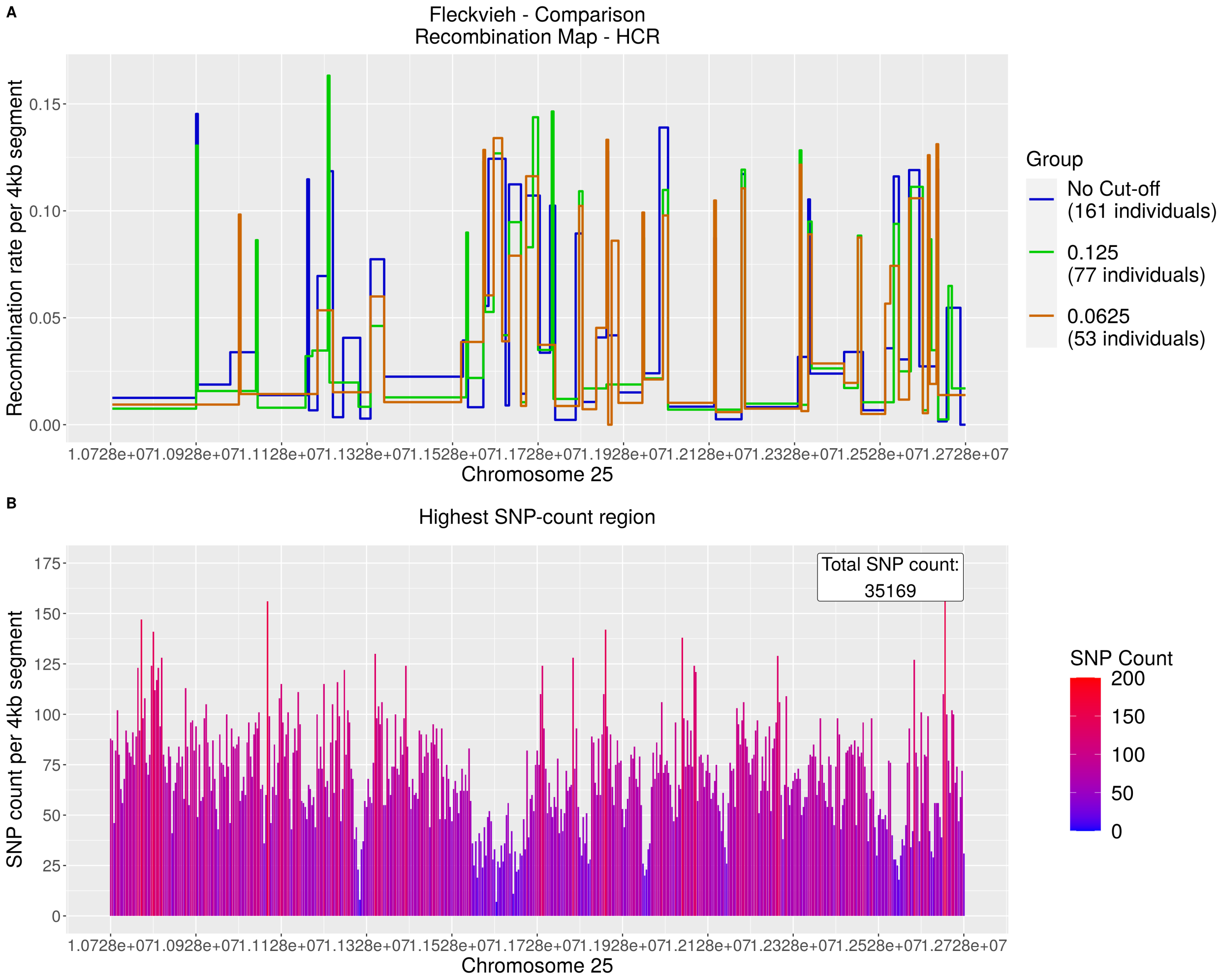

Figure 11: Collapsed recombination maps for Fleckvieh. In panel A, the recombination maps for the region 10,728,000-12,728,000 base pairs of chromosome 25 of the Fleckvieh subsets (no cut-off - blue, 0.125 - green, 0.0625 - brown) are collapsed. The recombination rates of the highest SNP-count region are estimated under demography per 4000 base-pair segments. Figure 11B shows the SNP count per 4000 base-pair segments of the Fleckvieh population, with a total SNP count of 35169 SNPs. For comparative reasons, the spike in panel B is not fully included in the plot; we report the SNP count of this spike to be 156 SNPs.

5.7 Correlation between recombination rate and SNP count

The aligned SNP-count maps in Figure 10 and 11 of both populations (Braunvieh, Fleckvieh) might indicate an inverse relationship between recombination rate and SNP count where high recombination rates are accompanied by low SNP count and vice versa. To further investigate this pattern we tested whether there is a significant correlation between recombination rate and SNP count using the Pearson's product-moment correlation. The weak negative correlation between recombination rate and the SNP count is present in both populations and in all subsets. The correlation is weak but statistically significant in all subsets of both Braunvieh ($p < 0.019$) and Fleckvieh ($p < 0.0054$). Figure 12 plots the SNP count relative to the recombination rate and fits a polynomial surface with a 0.95 confidence interval using local fitting in R. In the current version of *LDJump* (Version: 0.3.1), the level of inbreeding and a more appropriate demographic scenario are not taken into account and could be addressed in further research. Therefore, we wish to highlight that this inverse relationship should be interpreted with caution until further analyses and simulation studies are performed with *LDJump* to account for the level of inbreeding as well as the correct demographic scenario in the population under study.

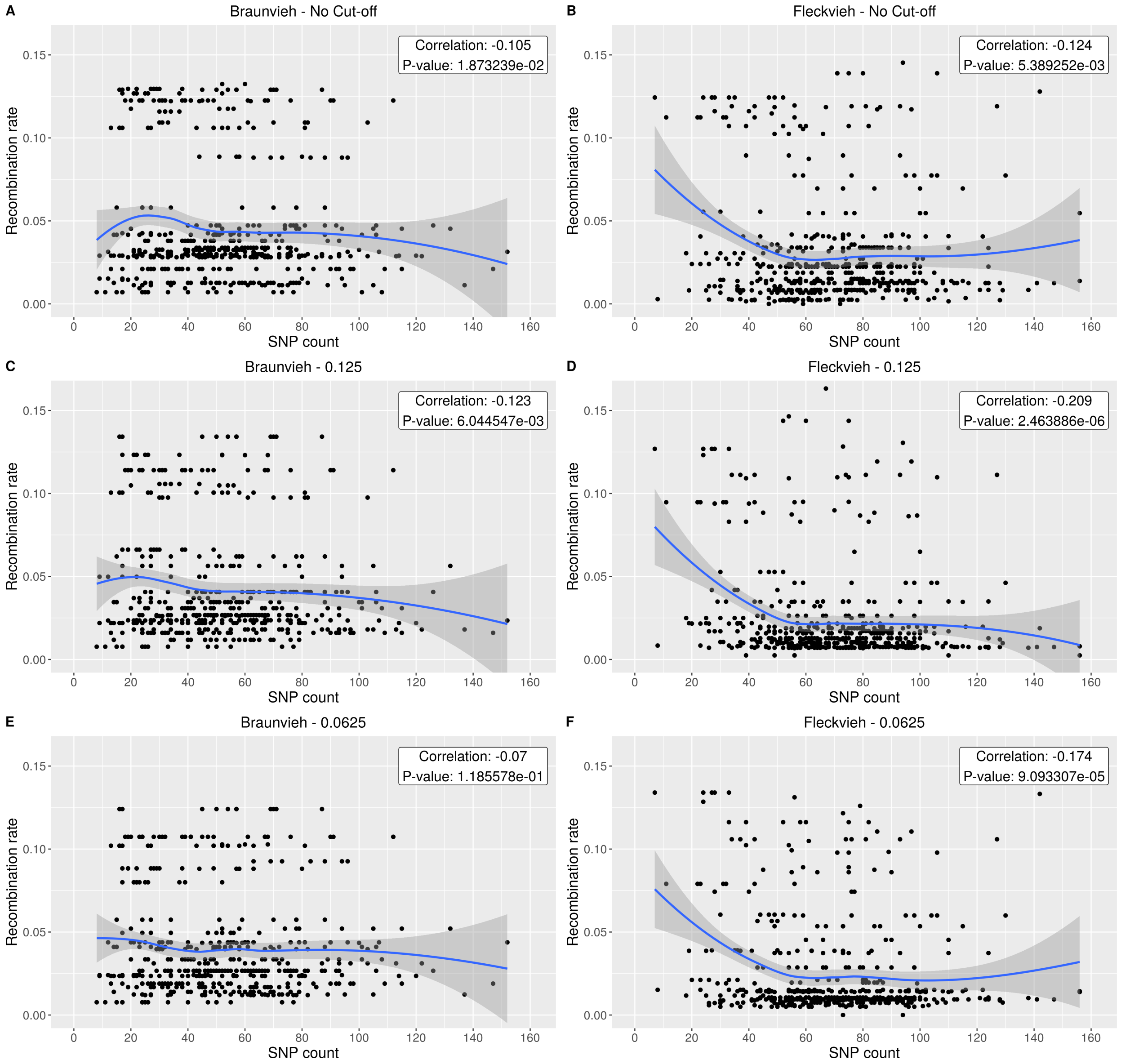

Figure 12: The recombination rate per 4000 base-pair segments is plotted along with the SNP count for the Braunvieh and Fleckvieh populations for all subsets (no cut-off, 0.125, 0.0625). The correlation estimate and the respective p-value are shown for each plot. Figure 12A, 12C, and 12E (left panel) show the correlation of the Braunvieh population for all subsets. Figure 12B, 12D, and 12F (right panel) show the correlation of the Fleckvieh population for all subsets.

5.8 Correlation between recombination rate and GC content

To further investigate the global recombination patterns in the Braunvieh and Fleckvieh population we tested whether there is an association between recombination rate and the GC content. Here, we obtain the GC content per 4000 base-pair segments from the reference sequence (ARS-UCD1.2) of the cattle genome (Rosen et al., 2020b) of the highest SNP-count region. We analysed the correlation between the GC content and recombination rate in both populations, and observed no statistically significant correlation in any of the subsets. Figure 13 plots the GC content relative to the recombination rate and fits a polynomial surface with a 0.95 confidence interval using local fitting in R.

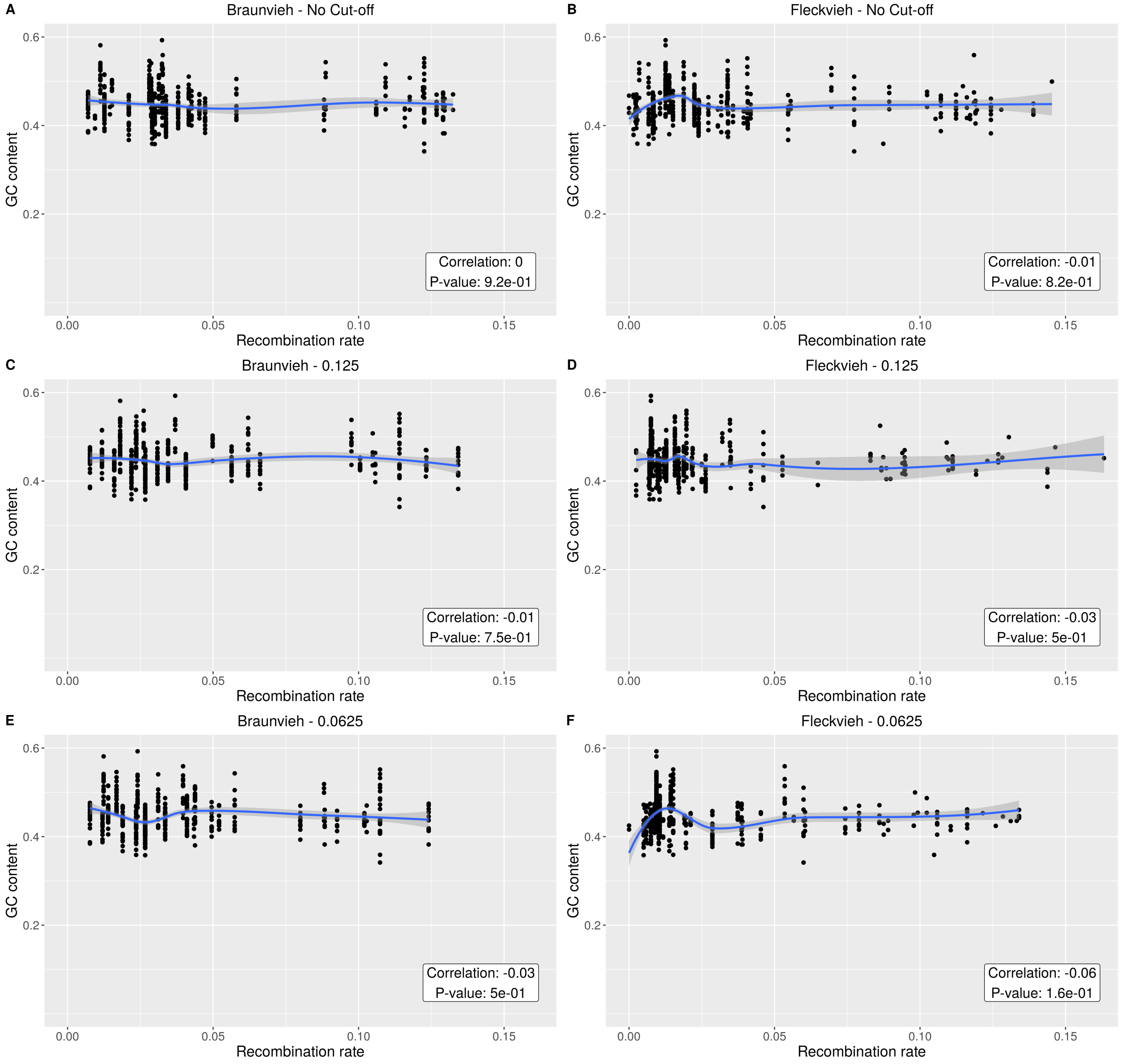

Figure 13: The recombination rate per 4000 base-pair segments is plotted along with the GC content for the Braunvieh and Fleckvieh populations for all subpopulations (no cut-off, 0.125, 0.0625). The correlation estimate and the respective p-value are shown for each plot. Figure 13A, 13C, and 13E (left panel) show the correlation of the Braunvieh population for all subsets. Figure 13B, 13D, and 13F (right panel) show the correlation of the Fleckvieh population for all subsets.

5.9 Annotated genes in the HCR and LCR

Using *NCBI* (NCBI, 1988, 2004) we annotated all genes within the highest and lowest SNP-count region of chromosome 25. The highest SNP-count region contains four genes; three are protein-coding genes and one is a non-coding gene. The lowest SNP-count region contains 139 genes in total; 117 are protein-coding genes, 19 are non-coding genes, and three are pseudogenes. Figure 14 overlays the annotated genes onto the corresponding recombination maps of the two studied genomic regions. Additionally, Table 8 and 9-11 in the supplementary material list the gene ID, name, and starting and ending positions of genes contained in the highest and lowest SNP-count region, respectively.

We observe a striking difference in the gene densities of the lowest and highest SNP-count regions, where the lowest SNP-count region contains 139 genes and the highest SNP-count region contains 4 genes. Intriguingly, the genes in the highest SNP-count region are considerably larger than the genes in the lowest SNP-count region.

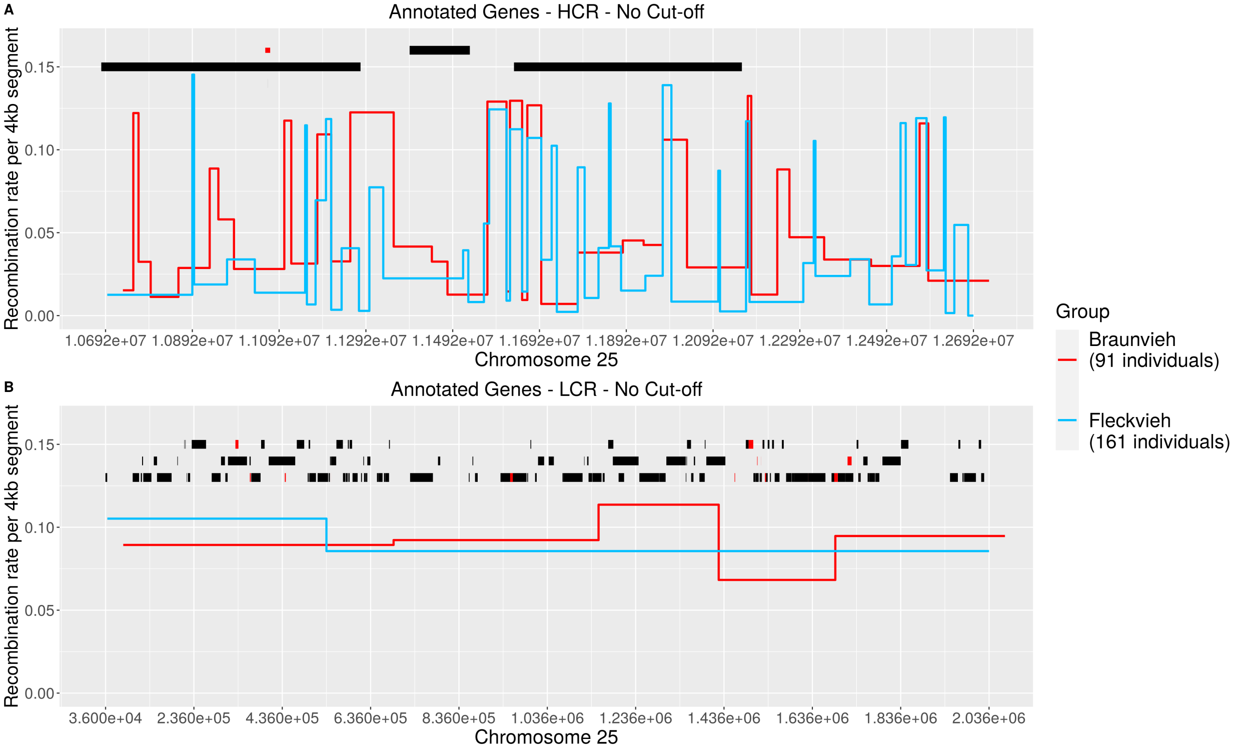

Figure 14: Annotated genes in the HCR and LCR. In this figure, genes present in the highest and lowest SNP-count region of chromosome 25 in the cattle genome are shown in Figure 14A and 14B, respectively. The black lines indicate protein-coding genes, whereas red lines present non-coding genes. The highest and lowest SNP-count regions contain a total of 4 and 139 genes, respectively.

6 Discussion

LDJump estimates recombination rates under neutrality or demography. In this study, we estimated fine-scale recombination maps in two cattle breeds (Braunvieh, Fleckvieh) using both models. The recombination maps estimated in each of the two models remain similar across all subsets. The estimation of recombination rates under demography seems to show a higher resolution as more breakpoints are introduced compared to recombination rates estimated under neutrality.

In agreement with the results of a previous study (Hermann, Heissl, et al., 2019), recombination rate comparisons between the lowest and highest SNP-count region show higher resolution in estimated recombination rates in the highest SNP-count region, suggesting that the lowest SNP-count region does not contain enough SNP information for *LDJump* to estimate informative recombination rates.

The recombination background rate is defined as the median of the recombination rates, and recombination hotspots as regions having at least a 5-fold increase in recombination rates compared to the background rate (McVean et al., 2004; Chan et al., 2012; Hermann, Heissl, et al., 2019). Using this threshold, we obtain an estimate of hotspots for Fleckvieh but not for Braunvieh, which might be due to an overestimation of the background rate in the Braunvieh population. Hence, in our study, we defined hotspots as regions of 3-fold increase to the median of the background rates. The breed-specific recombination maps suggest a higher number of breakpoints in the Fleckvieh population compared to Braunvieh, which might be due to the higher SNP count in Fleckvieh. Intruigingly, the number of hotspots increases with a decreasing relationship coefficient in Fleckvieh, but not in Braunvieh. Adding a model to *LDJump* to deal with inbreeding structures might elucidate why this pattern is observed in one population and not the other.

In addition to the breed-specific comparison of recombination patterns we constructed two recombination maps by combining the datasets of both populations of the categories 0.125 and 0.0625. The results indicate shared recombination patterns between both breeds, which might be specific to cattle as a species. A further study with a possibly larger dataset might explain whether the hotspots detected solely in the combined dataset present regions of high recombination specific to the cattle species, or occur due to missing SNP information in either breed.

We investigated the relationship between recombination rates and SNP count by overlapping the recombination maps of all subsets within a population and aligning the recombination maps to the SNP count of the genomic region with the highest SNP count (HCR). Pearson's product-moment correlation testing suggests a statistically significant, negative, and weak correlation between recombination rate and SNP count. To further investigate this correlation we applied *LDJump* to a randomly selected genomic region on chromosome 25 and also observed a significant, negative correlation. The recombination maps of both populations including a correlation plot are visualized in Figure 15, 16, and 17, respectively. In the current version of *LDJump*, the level of inbreeding as well as a more appropriate scenario of demography are not taken into account and could be addressed in further research. Therefore, we wish to highlight that this inverse relationship should be interpreted with caution unless further analyses and simulation studies are performed with *LDJump* to account for the level of inbreeding and demography in the population under study.

In the genome of warm-blooded vertebrates, GC-rich regions show more recombination events (Bernardi, 1989, 1993, 1995) with higher recombination rates (Duret & Arndt, 2008) compared to GC-poor regions. In our study, we observed no association between local GC content and recombination intensity in cattle. Further investigation is necessary to determine whether the effect of GC content i) is not adequately represented in the reference genome, ii) is chromosome or species specific, or iii) the nonexistent correlation vanishes if a larger genomic region is studied.

Lastly, the analysis of gene density showed a striking difference in the gene densities of the lowest and highest SNP-count regions, the lowest SNP-count region containing 139 genes and the highest SNP-count region containing 4 genes. Intriguingly, the genes in the highest SNP-count region are considerably larger than the genes in the lowest SNP-count region. This finding suggests a greatly larger number of genes in a region of low recombination with a possibly lower chance of change due to recombination.

7 Conclusion

In this thesis, we attempted to identify determinatives of recombination patterns in cattle by analyzing the effect of SNP count, GC content, and the density and nature of genes on recombination. The estimation of recombination rates under different assumptions showed a finer recombination map construction under demography compared to neutrality. Excluding related individuals led to an increase in the estimated number of hotspots in the Fleckvieh population, but to a decrease in the Braunvieh population. Moreover, we estimated recombination rates by combining the datasets of both populations and possibly detected species-specific recombination hotspots.

The analysis of recombination patterns in relation to the respective SNP count suggests an inverse relationship between recombination rates and SNP count. Pearson's product-moment correlation testing indicates a weak, but statistically significant, negative correlation between recombination rates and SNP count, where regions of lower SNP count show higher recombination rates. In contrast, we observed no correlation between local GC content and recombination intensity. Further analysis needs to be performed in order to address this correlation in other genomic regions and under an underlying model in *LDJump* correctly addressing the level of inbreeding as well as demography in a population.

Lastly, we detected a considerable difference in the number of genes between the lowest and highest SNP-count regions, where the lowest SNP-count region contains 139 genes and the highest SNP-count region contains 4 genes.

References

Adam Auton, A. M., Petr Danecek. (2020). Vcftools (0.1.15) [Computer software manual]. Retrieved from `http://vcftools.sourceforge.net`

Anderson, N., Adams, R. H., Demuth, J. P., & Blackmon, H. (2019). evobir: Evolutionary biology in r [Computer software manual]. Retrieved from `https://github.com/coleoguy/evobir` (R package version 1.3)

Arnheim, N., Calabrese, P., & Nordborg, M. (2003). Hot and cold spots of recombination in the human genome: the reason we should find them and how this can be achieved. *The American Journal of Human Genetics*, *73*(1), 5–16.

Arnheim, N., Calabrese, P., & Tiemann-Boege, I. (2007). Mammalian meiotic recombination hot spots. *Annu. Rev. Genet.*, *41*, 369–399.

Ayala, F. J. (1978). The mechanisms of evolution. *Scientific American*, *239*(3), 56–69.

Bernardi, G. (1989). The isochore organization of the human genome. *Annual review of genetics*, *23*(1), 637–659.

Bernardi, G. (1993). The vertebrate genome: isochores and evolution. *Molecular biology and evolution*, *10*(1), 186–204.

Bernardi, G. (1995). The human genome: organization and evolutionary history. *Annual review of genetics*, *29*(1), 445–476.

Brookes, A. J. (1999). The essence of snps. *Gene*, *234*(2), 177–186.

Chan, A. H., Jenkins, P. A., & Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in drosophila melanogaster. *PLoS Genet*, *8*(12), e1003090.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... others (2011). The variant call format and vcftools. *Bioinformatics*, *27*(15), 2156–2158.

Dietrich, W. F., Miller, J., Steen, R., Merchant, M. A., Damron-Boles, D., Husain, Z., ... others (1996). A comprehensive genetic map of the mouse genome. *Nature*, *380*(6570), 149–152.

Duret, L., & Arndt, P. F. (2008). The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, *4*(5), e1000071.

Fenster, C. B., & Galloway, L. F. (2000). Inbreeding and outbreeding depression in natural populations of chamaecrista fasciculata (fabaceae). *Conservation Biology*, *14*(5), 1406–1412.

Frick, K., Munk, A., & Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 495–580.

Galtier, N., Piganeau, G., Mouchiroud, D., & Duret, L. (2001). Gc-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics*, *159*(2), 907–911.

Gu, Z., Hillier, L., & Kwok, P.-Y. (1998). Single nucleotide polymorphism hunting in cyberspace. *Human mutation*, *12*(4), 221–225.

Heng Li, B. H. (2020). Bcftools (1.3.1) [Computer software manual]. Retrieved from `http://github.com/samtools/bcftools`

Hermann, P., Futschik, A., & Mohammadi, F. (2019). Ldjump: Estimating variable recombination rates from population genetic data [Computer software manual]. (R package version 0.3.1)

Hermann, P., Heissl, A., Tiemann-Boege, I., & Futschik, A. (2019). Ldjump: Estimating variable recombination rates from population genetic data. *Molecular ecology resources*, *19*(3), 623–638.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C.-F., ... Jacob, H. J. (2004). Comparative recombination rates in the rat, mouse, and human genomes. *Genome research*, *14*(4), 528–538.

Kadri, N. K., Harland, C., Faux, P., Cambisano, N., Karim, L., Coppieters, W., ... others (2016). Coding and noncoding variants in hfm1, mlh3, msh4, msh5, rnf212, and rnf212b affect recombination rate in cattle. *Genome research*, *26*(10), 1323–1332.

Kimura, M. (1979). The neutral theory of molecular evolution. *Scientific American*, *241*(5), 98–129.

Knaus, B. J., & Grünwald, N. J. (2017). vcfr: a package to manipulate and visualize variant call format data in r. *Molecular ecology resources*, *17*(1), 44–53.

Kong, A., Gudbjartsson, D. F., Sainz, J., Jonsdottir, G. M., Gudjonsson, S. A., Richardsson, B., ... others (2002). A high-resolution recombination map of the human genome. *Nature genetics*, *31*(3), 241–247.

Li, H., Gyllensten, U. B., Cui, X., Saiki, R. K., Erlich, H. A., & Arnheim, N. (1988). Amplification and analysis of dna sequences in single human sperm and diploid cells. *Nature*, *335*(6189), 414–417.

Ma, L., O'Connell, J. R., VanRaden, P. M., Shen, B., Padhi, A., Sun, C., . . . others (2015). Cattle sex-specific recombination and genetic control from a large pedigree analysis. *PLoS Genet*, *11*(11), e1005387.

Majewski, J., & Ott, J. (2000). Gt repeats are associated with recombination on human chromosome 22. *Genome Research*, *10*(8), 1108–1114.

McVean, G. A., Myers, S. R., Hunt, S., Deloukas, P., Bentley, D. R., & Donnelly, P. (2004). The fine-scale structure of recombination rate variation in the human genome. *Science*, *304*(5670), 581–584.

NCBI. (1988). *National center for biotechnology information (ncbi). bethesda (md): National library of medicine (us), national center for biotechnology information;.* Retrieved 2020-10-06, from https://www.ncbi.nlm.nih.gov/

NCBI. (2004). *Gene. bethesda (md): National library of medicine (us), national center for biotechnology information;.* Retrieved 2020-10-06, from https://www.ncbi.nlm.nih.gov/gene/

Neff, M. W., Broman, K. W., Mellersh, C. S., Ray, K., Acland, G. M., Aguirre, G. D., . . . Rine, J. (1999). A second-generation genetic linkage map of the domestic dog, canis familiaris. *Genetics*, *151*(2), 803–820.

Orr, H. A. (2009). Fitness and its role in evolutionary genetics. *Nature Reviews Genetics*, *10*(8), 531–539.

Otto, S. P., & Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, *3*(4), 252–261.

Paigen, K., & Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nature Reviews Genetics*, *11*(3), 221–233.

Pekkala, N., Knott, K. E., Kotiaho, J. S., Nissinen, K., & Puurtinen, M. (2014). The effect of inbreeding rate on fitness, inbreeding depression and heterosis over a range of inbreeding coefficients. *Evolutionary Applications*, *7*(9), 1107–1119.

Purcell, S. (2020). Plink (1.90) [Computer software manual]. Retrieved from http://pngu.mgh.harvard.edu/purcell/plink/

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . others (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, *81*(3), 559–575.

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., . . . others (2020a). De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, *9*(3), giaa021.

Rosen, B. D., Bickhart, D. M., Schnabel, R. D., Koren, S., Elsik, C. G., Tseng, E., . . . Medrano, J. F. (2020b, 03). De novo assembly of the cattle reference genome with single-molecule sequencing. *GigaScience*, *9*(3). Retrieved from https://doi.org/10.1093/gigascience/giaa021 (giaa021) doi: 10.1093/gigascience/giaa021

Sandor, C., Li, W., Coppieters, W., Druet, T., Charlier, C., & Georges, M. (2012). Genetic variants in rec8, rnf212, and prdm9 influence male recombination in cattle. *PLoS Genet*, *8*(7), e1002854.

Shen, B., Jiang, J., Seroussi, E., Liu, G. E., & Ma, L. (2018). Characterization of recombination features and the genetic basis in multiple cattle breeds. *BMC genomics*, *19*(1), 1–10.

Sobel, E., & Lange, K. (1996). Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. *American journal of human genetics*, *58*(6), 1323.

Stachowicz, K., Sargolzaei, M., Miglior, F., & Schenkel, F. (2011). Rates of inbreeding and genetic diversity in canadian holstein and jersey cattle. *Journal of dairy science*, *94*(10), 5160–5175.

Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S., & Morton, N. (2005). A map of the human genome in linkage disequilibrium units. *Proceedings of the National Academy of Sciences*, *102*(33), 11835–11839.

Tarsi, K., & Tuff, T. (2012). Introduction to population demographics. *Nat Educ Knowl*, *3*, 3.

Thomsen, H., Reinsch, N., Xu, N., Bennewitz, J., Looft, C., Grupe, S., . . . others (2001). A whole genome scan for differences in recombination rates among three bos taurus breeds. *Mammalian genome*, *12*(9), 724–728.

Weng, Z.-Q., Saatchi, M., Schnabel, R. D., Taylor, J. F., & Garrick, D. J. (2014). Recombination locations and rates in beef cattle assessed from parent-offspring pairs. *Genetics Selection Evolution*, *46*(1), 34.

Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, *56*(645), 330–338.

Supplementary Material

7.1 Annotated Genes

GeneID	Description	Start	End	Region	Function
112444394	transfer RNA glycine (anticodon CCC)	11065729	11065801	HCR	Non-coding
537938	calcineurin like phosphoesterase domain containing 1	11392701	11531518	HCR	Protein coding
518366	sorting nexin 29	10682180	11279625	HCR	Protein coding
100139490	shisa family member 9	11633245	12158399	HCR	Protein coding

Table 8: Description of detected genes in the highest SNP-count region of chromosome 25 in the cattle genome from *NCBI*. The gene ID, name, and starting and ending positions of genes contained in the highest SNP-count region (HCR) of chromosome 25 are listed. In the HCR, there are a total of four genes listed in *NCBI*.

GeneID	Description	Start	End	Region	Function
104797423	microRNA mir-1842	1742695	1742754	LCR	Non-coding
100313098	microRNA mir-1225	1627688	1627780	LCR	Non-coding
100313089	microRNA mir-940	1790895	1790987	LCR	Non-coding
100313464	microRNA mir-2382	1742942	1743018	LCR	Non-coding
112444449	transfer RNA glycine (anticodon CCC)	546202	546272	LCR	Non-coding
112444391	small nucleolar RNA ACA64	1522006	1522132	LCR	Non-coding
112444384	small nucleolar RNA SNORD60	1693066	1693148	LCR	Non-coding
112444383	small nucleolar RNA SNORA64/SNORA10 family	1519879	1520009	LCR	Non-coding
112444382	small nucleolar RNA SNORA64/SNORA10 family	1520525	1520658	LCR	Non-coding
112444352	uncharacterized LOC112444352	362633	364715	LCR	Non-coding
112444351	uncharacterized LOC112444351	329951	336755	LCR	Non-coding
112444348	uncharacterized LOC112444348	951820	957941	LCR	Non-coding
112444341	uncharacterized LOC112444341	1510969	1511922	LCR	Non-coding
112444337	uncharacterized LOC112444337	1459891	1461129	LCR	Non-coding
104976739	uncharacterized LOC104976739	1528912	1531211	LCR	Non-coding
104975826	uncharacterized LOC104975826	1491712	1502252	LCR	Non-coding
104975822	uncharacterized LOC104975822	441678	444050	LCR	Non-coding
101906195	uncharacterized LOC101906195	1715673	1724829	LCR	Non-coding
100847802	uncharacterized LOC100847802	1685469	1693593	LCR	Non-coding
512439	hemoglobin, alpha 2	216496	217264	LCR	Protein coding
618441	methionine sulfoxide reductase B1	1502594	1507246	LCR	Protein coding
532494	insulin like growth factor binding protein acid labile subunit	1367704	1370515	LCR	Protein coding
508713	N-acetylglucosamine-1-phosphate transferase subunit gamma	1072859	1083887	LCR	Protein coding
282412	calcium voltage-gated channel subunit alpha1 H	930068	988927	LCR	Protein coding
550622	ATPase H+ transporting V0 subunit c	2012685	2018060	LCR	Protein coding
768005	SLC9A3 regulator 2	1576576	1588862	LCR	Protein coding
515573	ubiquitin conjugating enzyme E2 I	1039143	1050827	LCR	Protein coding
513545	chloride voltage-gated channel 7	1138163	1158736	LCR	Protein coding
510343	tektin 4	872389	878381	LCR	Protein coding
281545	tryptase beta 2	997749	999475	LCR	Protein coding
327701	NADH:ubiquinone oxidoreductase subunit B10	1516661	1519322	LCR	Protein coding
618357	mitochondrial ribosomal protein S34	1350950	1352204	LCR	Protein coding
614560	CASK interacting protein 1	1710648	1728751	LCR	Protein coding
352959	ras homolog family member T2	573403	578699	LCR	Protein coding
516233	NADPH oxidase organizer 1	1534932	1538859	LCR	Protein coding
100140149	hemoglobin, alpha 1	219522	220318	LCR	Protein coding
511837	cytosolic iron-sulfur assembly component 3	628542	635399	LCR	Protein coding
504985	TSC complex subunit 2	1595277	1626215	LCR	Protein coding
286867	ribosomal protein S2	1519528	1521705	LCR	Protein coding
535236	MTOR associated protein, LST8 homolog	1736234	1740142	LCR	Protein coding
511692	TNF receptor associated factor 7	1693795	1710429	LCR	Protein coding
618598	chromosome 25 C16orf91 homolog	1119282	1120218	LCR	Protein coding
538173	phosphoglycolate phosphatase	1742545	1745400	LCR	Protein coding
534485	NPR3 like, GATOR1 complex subunit	152452	185421	LCR	Protein coding
616898	tubulin epsilon and delta complex 2	1966625	1970947	LCR	Protein coding
787961	netrin 3	1975804	1978415	LCR	Protein coding
530342	adenine nucleotide translocase lysine methyltransferase	619507	621019	LCR	Protein coding
511836	hydroxyacylglutathione hydrolase like	625429	628537	LCR	Protein coding
618440	ribosomal protein L3 like	1507603	1514414	LCR	Protein coding
515661	splA/ryanodine receptor domain and SOCS box containing 3	1355611	1361300	LCR	Protein coding

Table 9: Description of detected genes in the lowest SNP-count region of chromosome 25 in the cattle genome from *NCBI*. he gene ID, name, and starting and ending positions of genes contained in the lowest SNP-count region (LCR) of chromosome 25 are listed. In the LCR, there are a total of 139 genes listed in *NCBI*.

GeneID	Description	Start	End	Region	Function
515660	nucleotide binding protein 2	1361638	1366861	LCR	Protein coding
513526	small nuclear ribonucleoprotein U11/U12 subunit 25	118663	120818	LCR	Protein coding
509274	hydroxyacylglutathione hydrolase	1380350	1392352	LCR	Protein coding
100336895	rhomboid like 1	579620	582720	LCR	Protein coding
100140603	hemoglobin subunit mu	214283	215064	LCR	Protein coding
100139898	BRICHOS domain containing 5	1740015	1742390	LCR	Protein coding
100139697	intraflagellar transport 140	1184359	1242591	LCR	Protein coding
789464	F-box and leucine rich repeat protein 16	595081	606604	LCR	Protein coding
787784	deoxyribonuclease 1 like 2	1761110	1764865	LCR	Protein coding
767973	ring finger protein 151	1523353	1527393	LCR	Protein coding
618512	Jupiter microtubule associated homolog 2	1289616	1303831	LCR	Protein coding
618487	meiosis specific with OB-fold	1395977	1439102	LCR	Protein coding
618429	transducin beta like 3	1528782	1534840	LCR	Protein coding
618423	growth factor, augmenter of liver regeneration	1539898	1542223	LCR	Protein coding
618415	synaptogyrin 3	1544835	1549036	LCR	Protein coding
618325	hemoglobin, theta 1	222389	227408	LCR	Protein coding
618306	N-methylpurine DNA glycosylase	145190	152508	LCR	Protein coding
618296	RNA polymerase III subunit K	115710	118437	LCR	Protein coding
618294	interleukin 9 receptor	97378	111918	LCR	Protein coding
618053	NHL repeat containing 4	495734	499070	LCR	Protein coding
618031	RAB40C, member RAS oncogene family	516050	539810	LCR	Protein coding
618020	MAPK regulated corepressor interacting protein 2	546336	558052	LCR	Protein coding
615464	transmembrane protein 204	1197395	1209837	LCR	Protein coding
613745	Rho GDP dissociation inhibitor gamma	307142	309575	LCR	Protein coding
540893	heparan sulfate-glucosamine 3-sulfotransferase 6	1485429	1491648	LCR	Protein coding
540233	WFIKKN2	541147	543770	LCR	Protein coding
537598	mitogen-activated protein kinase 8 interacting protein 3	1305496	1349468	LCR	Protein coding
535203	nth like DNA glycosylase 1	1589472	1595206	LCR	Protein coding
535174	enoyl-CoA delta isomerase 1	1765101	1777754	LCR	Protein coding
535131	LUC7 like	231075	263454	LCR	Protein coding
531296	mesothelin like	654193	663061	LCR	Protein coding
530317	SRY-box transcription factor 8	788544	793550	LCR	Protein coding
529167	rhomboid 5 homolog 1	121329	139772	LCR	Protein coding
529002	TBC1 domain family member 24	1979510	2006263	LCR	Protein coding
526097	telomere maintenance 2	1173824	1185499	LCR	Protein coding
525521	RNA pseudouridine synthase domain containing 1	666142	669366	LCR	Protein coding
524646	C1q and TNF related 8	866580	868276	LCR	Protein coding
524063	pentraxin 4	1161366	1165731	LCR	Protein coding
522441	unk like zinc finger	1083088	1115799	LCR	Protein coding
522068	calpain 15	468919	485677	LCR	Protein coding
521401	amidohydrolase domain containing 2	2018207	2024931	LCR	Protein coding

Table 10: Description of detected genes in the lowest SNP-count region of chromosome 25 in the cattle genome from *NCBI*. The gene ID, name, and starting and ending positions of genes contained in the lowest SNP-count region (LCR) of chromosome 25 are listed. In the LCR, there are a total of 139 genes listed in *NCBI*.

GeneID	Description	Start	End	Region	Function
521040	regulator of G protein signaling 11	297326	306152	LCR	Protein coding
520515	cramped chromatin regulator homolog 1	1244372	1289189	LCR	Protein coding
517007	chromosome transmission fidelity factor 18	669617	677704	LCR	Protein coding
517006	G protein subunit gamma 13	677672	679635	LCR	Protein coding
516237	mesothelin	650341	653924	LCR	Protein coding
515997	E4F transcription factor 1	1751303	1761026	LCR	Protein coding
515675	RAB26, member RAS oncogene family	1679439	1685522	LCR	Protein coding
515663	NME/NM23 nucleoside diphosphate kinase 3	1349470	1350659	LCR	Protein coding
515662	essential meiotic structure-specific endonuclease subunit 2	1352279	1358331	LCR	Protein coding
515528	meteorin, glial cell differentiation regulator	614323	616436	LCR	Protein coding
514636	methyltransferase like 26	544097	545876	LCR	Protein coding
511835	coiled-coil domain containing 78	621029	624962	LCR	Protein coding
510344	somatostatin receptor 5	858288	860104	LCR	Protein coding
509273	fumarylacetoacetate hydrolase domain containing 1	1392519	1394143	LCR	Protein coding
508714	TSR3 ribosome maturation factor	1070336	1072788	LCR	Protein coding
508216	mitochondrial ribosomal protein L28	361870	364486	LCR	Protein coding
508215	post-glycosylphosphatidylinositol attachment to proteins 6	365776	384039	LCR	Protein coding
508048	phosphatidylinositol glycan anchor biosynthesis class Q	497149	512238	LCR	Protein coding
507528	WD repeat domain 24	588614	594065	LCR	Protein coding
507493	family with sequence similarity 234 member A	276261	296400	LCR	Protein coding
505787	ATP binding cassette subfamily A member 3	1794636	1836056	LCR	Protein coding
505200	cyclin F	1947654	1965422	LCR	Protein coding
505124	lipase maturation factor 1	725402	777060	LCR	Protein coding
504565	STIP1 homology and U-box containing protein 1	584421	586777	LCR	Protein coding
504506	RNA binding protein with serine rich domain 1	1778569	1787635	LCR	Protein coding
504357	axin 1	313099	356133	LCR	Protein coding
504356	protein disulfide isomerase family A member 2	309730	312277	LCR	Protein coding
101904581	coiled-coil domain containing 154	1129348	1137409	LCR	Protein coding
101902709	WD repeat domain 90	558586	573227	LCR	Protein coding
101902553	proline rich 35	493585	496018	LCR	Protein coding
511647	RAB11 family interacting protein 3	405324	465303	LCR	Protein coding
505086	zinc finger protein 598, E3 ubiquitin ligase	1550721	1561664	LCR	Protein coding
504986	polycystin 1, transient receptor potential channel interacting	1626212	1665628	LCR	Protein coding
112444354	neuropeptide W	1567057	1571224	LCR	Protein coding
100139040	jumonji domain containing 8	585670	588473	LCR	Protein coding
100138582	hemoglobin subunit zeta	198445	199606	LCR	Protein coding
789799	mastin	1008284	1011995	LCR	Protein coding
789324	NME/NM23 nucleoside diphosphate kinase 4	383578	386875	LCR	Protein coding
768256	2,4-dienoyl-CoA reductase 2	388100	396057	LCR	Protein coding
516108	CG2446-like	582898	584292	LCR	Protein coding
104975846	proline and glutamate rich with coiled coil 1	1126167	1129243	LCR	Protein coding
786948	tryptase-2-like	990597	992454	LCR	Protein coding
789192	cyclin-G1	35707	39689	LCR	Protein coding
777692	uncharacterized LOC777692	1836756	1853491	LCR	Protein coding
617663	mastin	1015003	1028789	LCR	Protein coding
100294963	small nuclear ribonucleoprotein polypeptide E pseudogene	225324	225602	LCR	Pseudogenes
787289	heterogeneous nuclear ribonucleoprotein A1 pseudogene	1169962	1171024	LCR	Pseudogenes
100137913	sorting nexin-12 pseudogene	204432	205240	LCR	Pseudogenes

Table 11: Description of detected genes in the lowest SNP-count region of chromosome 25 in the cattle genome from *NCBI*. The gene ID, name, and starting and ending positions of genes contained in the lowest SNP-count region (LCR) of chromosome 25 are listed. In the LCR, there are a total of 139 genes listed in *NCBI*.

7.2 Randomly selected region

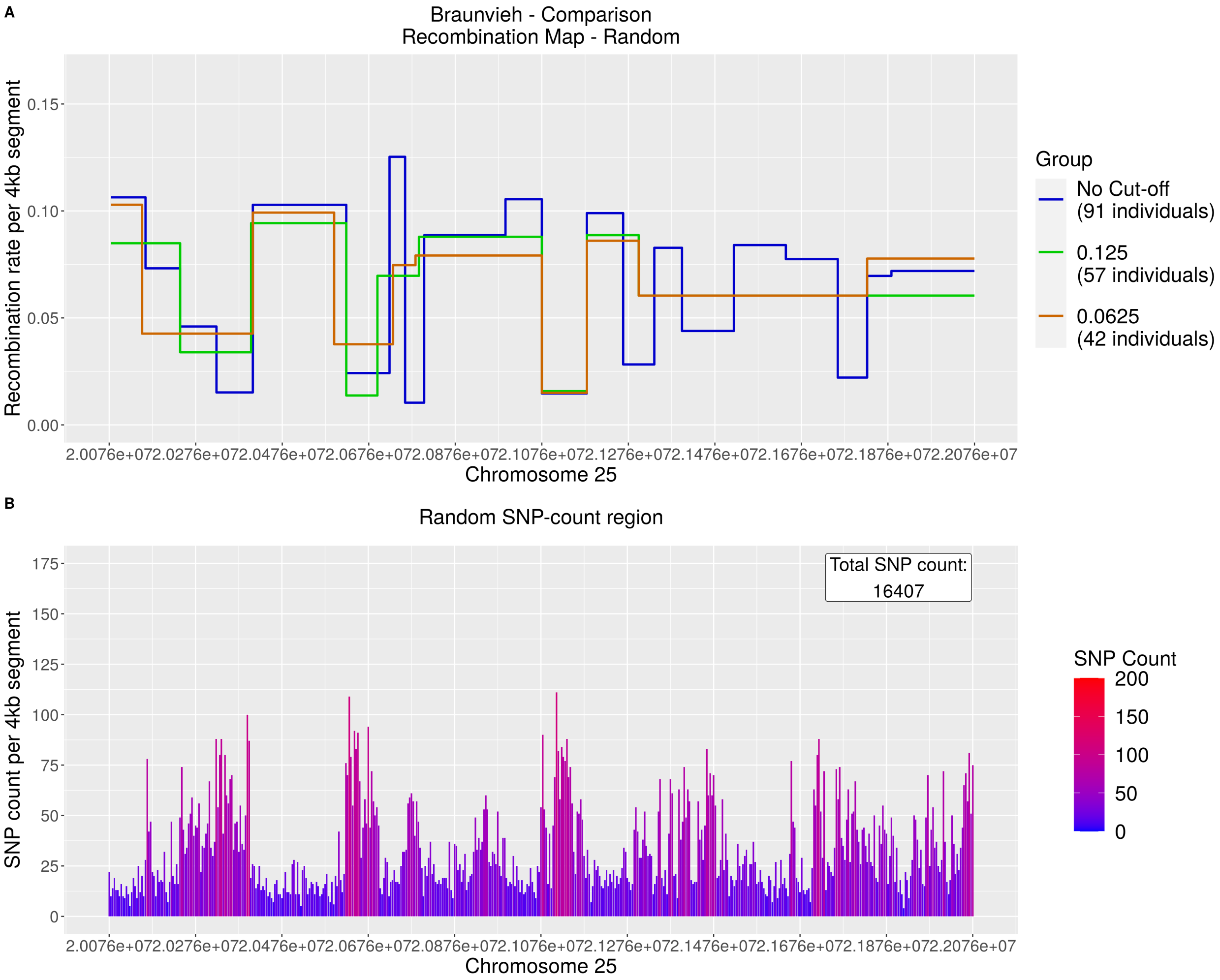

Figure 15: Collapsed recombination maps of a randomly selected region for Braunvieh. In panel A, the recombination map for the region 20,076,000-22,076,000 base-pairs of chromosome 25 of the Braunvieh subsets (no cut-off - blue, 0.125 - green, 0.0625 - brown) are collapsed and shown in Figure 15A. The recombination rates of the randomly selected region are estimated under demography per 4000 base-pair segments. Figure 15B shows the SNP count per 4000 base-pair segments of the Braunvieh population, with a total SNP count of 16407 SNPs.

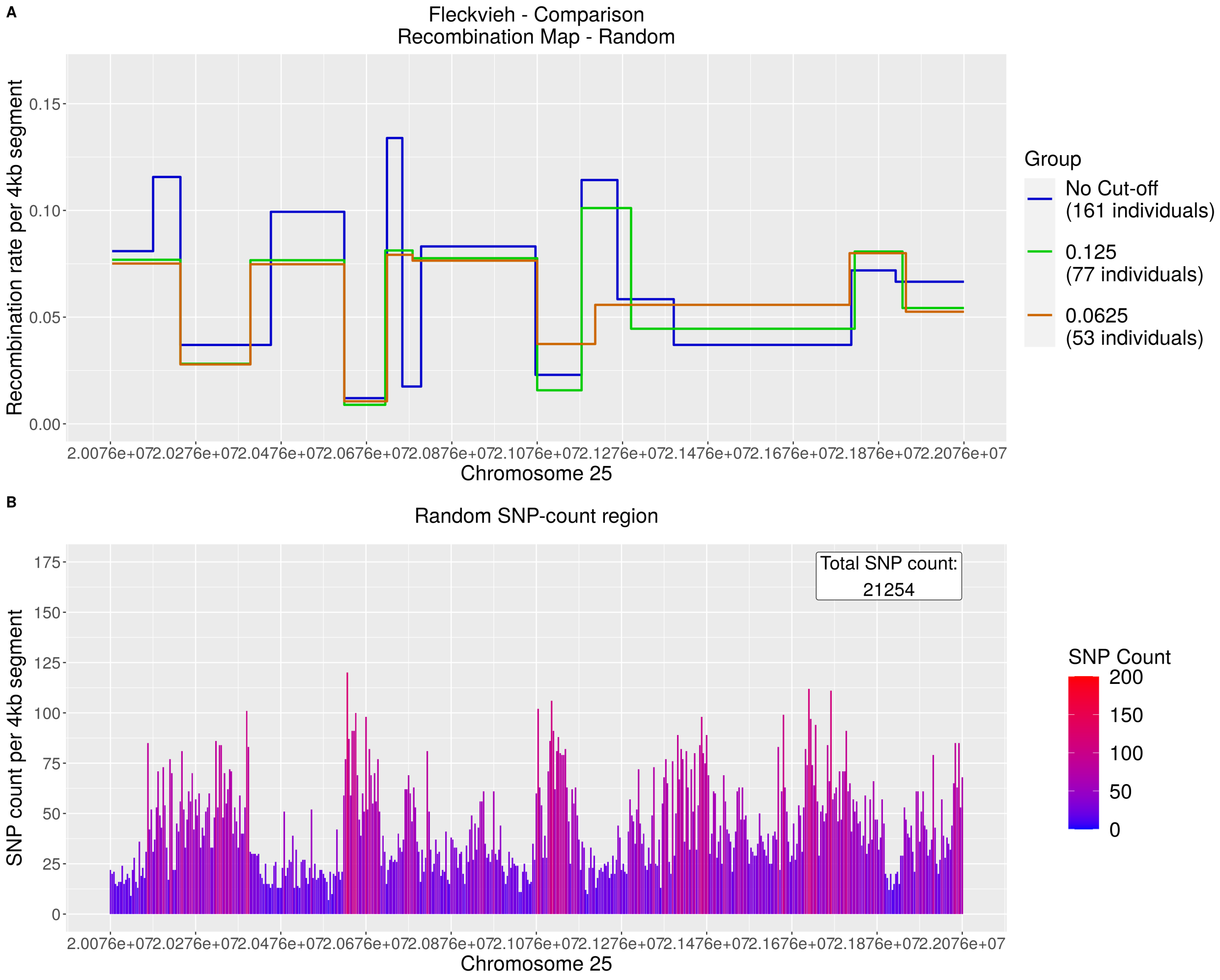

Figure 16: Collapsed recombination maps of a randomly selected region for Fleckvieh. In panel A, the recombination maps for the region 20,076,000-22,076,000 base-pairs of chromosome 25 of the Fleckvieh subsets (no cut-off - blue, 0.125 - green, 0.0625 - brown) are collapsed and shown in Figure 16A. The recombination rates of the randomly selected region are estimated under demography per 4000 base-pair segments. Figure 16B shows the SNP count per 4000 base-pair segments of the Fleckvieh population, with a total SNP count of 21254 SNPs.

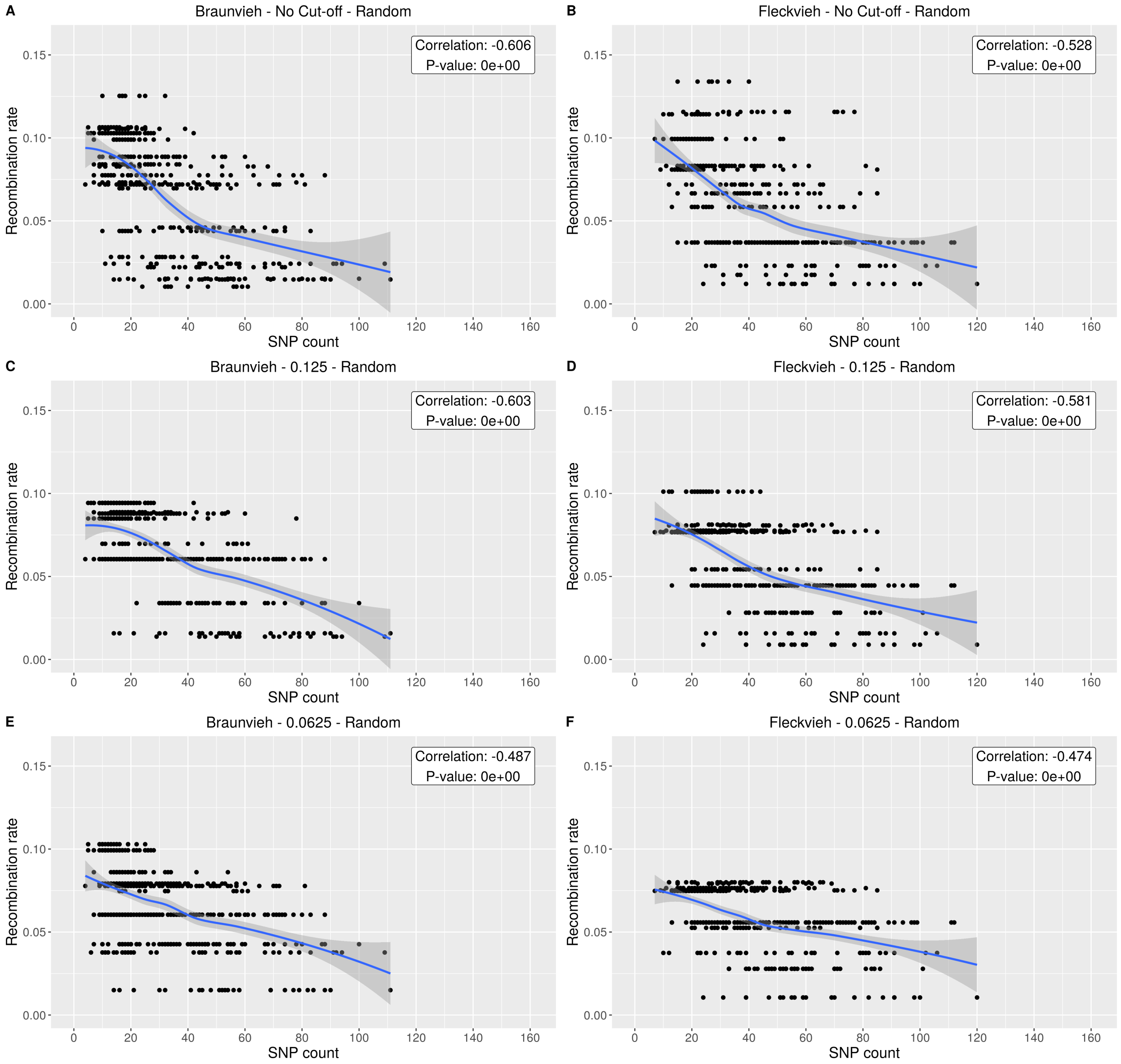

Figure 17: The recombination rate per 4000 base-pair segments is estimated for the region 20,076,000-22,076,000 of chromosome 25 and plotted along with the SNP count for the Braunvieh and Fleckvieh populations for all subsets (no cut-off, 0.125, 0.0625). The correlation estimate and the respective p-value are shown for each plot. Figure 17A, 17C, and 17E (left panel) show the correlation of the Braunvieh population for all subsets. Figure 17B, 17D, and 17F (right panel) show the correlation of the Fleckvieh population for all subsets.