bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BREC: An R package/Shiny app for automatically identifying heterochromatin boundaries and estimating local recombination rates along chromosomes

Yasmine MANSOUR[1,2*], Annie CHATEAU[2], and Anna-Sophie FISTON-LAVIER[1*]

[1]Institute of Evolution Science of Montpellier *(ISEM)*, France
[2]Laboratory of Computer Science, Robotics and Microelectronics of Montpellier *(LIRMM)*, France
[*]Corresponding authors: yasmine.mansour@umontpellier.fr
anna-sophie.fiston-lavier@umontpellier.fr

Abstract

Motivation: Meiotic recombination is a vital biological process playing an essential role in genomes structural and functional dynamics. Genomes exhibit highly various recombination profiles along chromosomes associated with several chromatin states. However, eu-heterochromatin boundaries are not available nor easily provided for non-model organisms, especially for newly sequenced ones. Hence, we miss accurate local recombination rates, necessary to address evolutionary questions.

Results: Here, we propose an automated computational tool, based on the Marey maps method, allowing to identify heterochromatin boundaries along chromosomes and estimating local recombination rates. Our method, called **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates) is non-genome-specific, running even on non-model genomes as long as genetic and physical maps are available. BREC is based on pure statistics and is data-driven, implying that good input data quality remains a strong requirement. Therefore, a data pre-processing module (data quality control and cleaning) is provided. Experiments show that BREC handles different markers density and distribution issues. BREC's heterochromatin boundaries have been validated with cytological equivalents experimentally generated on the fruit fly *Drosophila melanogaster* genome, for which BREC returns congruent equivalents. Also, BREC's recombination rates have been compared with previously reported estimates. Based on the promising results, we believe our tool has the potential to help bring data science into the service of genome biology and evolution. We introduce BREC within an R-package and a Shiny web-based user-friendly application yielding a fast, easy-to-use, and broadly accessible resource.

Availability: BREC R-package is available at the GitHub repository `https://github.com/ymansour21/BREC`.

Key words Genomics, Heterochromatin regions, Centromere position, Recombination rate, non-genome-specific, Graphical user interface.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

Meiotic recombination is a vital biological process which plays an essential role for investigating genome-wide structural as well as functional dynamics. Recombination events are observed in almost all eukaryotic genomes. Crossover, a one-point recombination event, is the exchange of DNA fragments between sister chromatids during meiosis. Recombination is a fundamental process that ensures genotypic and phenotypic diversity. Thereby, it is strongly related to various genomic features such as gene density, repetitive DNA, and DNA methylation (Coop and Przeworski, 2007; Duret and Galtier, 2009; Auton and McVean, 2012).

Recombination rate varies not only between species, but also within species and within chromosomes. Different heterochromatin regions exhibit different profiles of recombination events. Therefore, in order to understand how and why recombination rate varies, it is important to break down the chromosome structure to smaller blocks where several genomic feature besides, recombination rate, are known to also exhibit different profiles. Chromatin boundaries allow to distinguish between two main states of chromatin that can be defined as (1) euchromatin which is lightly compact with a high gene density, and on the contrary (2) heterochromatin which is highly compact with a paucity in genes. The heterochromatin is represented in different chromosome regions: the centromere and the telomeres. Euchromatic and heterochromatic regions exhibit different behaviours in terms of genomic features and dynamics related to their biologic function such as the cell division process that insures the organism viability. Consequently, easily distinguishing chromatin states is necessary for conducting further studies in various research fields and to be able to address questions related to cell processes such as: meiosis, gene expression, epigenetics, DNA methylation, natural selection and evolution, genome architecture and organization among others (Chan *et al.*, 2012; Stapley *et al.*, 2017; Morata *et al.*, 2018). In particular, a profound understanding of centromeres, their complete and precise structure, organization and evolution is currently a hot research area. These repeat-rich heterochromatin regions are currently still either poorly or not assembled at all across eukaryote genomes. Despite the huge advances offered by NGS technologies, centromeres are still considered as enigmas, mostly because they are preventing genome assembly algorithms form reaching their optimal performance in order to achieve more complete whole genome sequences (Muller *et al.*, 2019). In addition, the highly diverse mechanisms of heterochromatin positioning (Vanrobays *et al.*, 2017) and repositioning (Lu and He, 2019) remain a complicated obstacle in face of fully understanding genome organization. Thus, generating high resolution genetic, physical and recombination maps, and locating heterochromatin regions is increasingly interesting the community across a large range of taxa (Schueler *et al.*, 2001; Weinstock *et al.*, 2006; Silva-Junior and Grattapaglia, 2015; Robert L. Nussbaum *et al.*, 2015; Shen *et al.*, 2017; Gui *et al.*, 2018; Rowan *et al.*, 2019).

Numerous methods for estimating recombination rates exist. Population genetic based-methods (Stumpf and McVean, 2003) provide accurate fine-scale estimates. Nevertheless, these methods are very expensive, time-consuming, require a strong expertise and, most of all, are not applicable on all kinds of organisms. Moreover, the sperm-typing method (Jeffreys, 2000), which is also extremely accurate, providing high-density recombination maps, is male-specific and is applicable only on limited genome regions. On the other hand, a purely statistical approach, the Marey Maps (Chakravarti, 1991), could avoid some of the above issues based on other available genomic data: the genetic and physical distances of genomic markers.

The Marey maps approach consists in correlating the physical map with the genetic map representing respectively physical and genetic distances for a set of genetic markers on the same chromosome. Despite the efficiency of this approach and mostly the availability of

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

physical and genetic maps, generating recombination maps rapidly and for any organism is still challenging. Hence, the increasing need of an automatic, portable and easy-to-use solution.

Some Marey map-based tools already exist, two of which are largely used: (1) the MareyMap Online (Rezvoy *et al.*, 2007; Siberchicot *et al.*, 2017) which is applicable on multiple species, however, it does not allow accurate estimate of recombination rates on specific regions like the chromosome extremities, and (2) the *Drosophila melanogaster* Recombination Rate Calculator (RRC) (Fiston-Lavier *et al.*, 2010) which solves the previous issue by adjusting recombination rate estimates on such chromosome regions, yet, as indicated by its name, the RRC is *D. melanogaster*-specific. With the emerging Next Generation Sequencing (NGS) technologies, accessing whole chromosome sequences has become more and more possible on a wide range of species. Therefore, we may expect an exponential increase in markers number which will require more adapted tools to better handle such new scopes of data.

Here, we propose a new Marey map-based method as an automated computational solution that aims to, firstly, identify heterochromatin boundaries along chromosomes, secondly, estimate local recombination rates, and lastly, adjust recombination rates on chromosome along the chromosomal regions marked by the identified boundaries. Our proposed method, called **BREC** (heterochromatin **B**oundaries and **REC**ombination rate estimates), is provided with an R-package (R Core Team, 2018) and a Shiny (RStudio, Inc, 2014) web-based graphical user interface. BREC takes as input the same genomic data, genetic and physical distances, as in previous tools. It follows a workflow that, first, tests the data quality and offers a cleaning option, then, estimates local recombination rates and identify heterochromatin boundaries. Finally, BREC re-adjusts recombination rate estimates along heterochromatin regions, the centromere and telomere(s), in order to keep the estimates as authentic as possible to the biological process (Termolino *et al.*, 2016). Identifying the boundaries delimiting euchromatin and heterochromatin allows investigating recombination rate variations along the whole genome, which will help comparing recombination patterns within and between species. Furthermore, such functionality is fundamental for identifying the position of the centromeric and telomeric regions. Indeed, the position of the centromere on the chromosome has an influence on the chromatin environment and recent studies are interested in investigating how genome architecture may change with centromere organization (Muller *et al.*, 2019).

Our results have been validated with cytological equivalents, experimentally generated on the fruit fly *D. melanogaster* genome (Chan *et al.*, 2012; Langley *et al.*, 2012; Thurmond *et al.*, 2019). Moreover, since BREC is non genome-specific, it could efficiently been run on other model as well as non-model organisms for which both genetic and physical maps are available. Even though it is still an ongoing study, BREC have also been tested with different further species and results are reported.

This paper is organized as follows: in Section 2, BREC is presented in a detailed step by step workflow. Section 3 presents the data and methods involved in BREC and describes how the methods was calibrated and validated. Section 4 introduces the set of results, using both simulated and real data. The results are then discussed in Section 5 and concluding remarks with some perspectives are given in Section 6.

2 New Approach: BREC

BREC is designed following the workflow represented in Figure 1. In order to ensure that the widest range of species could be analyzed by our tool, we designed a pipeline which adapts behaviour with respect to input data. Mostly, each step of the pipeline relies on statistical analysis, adaptive algorithms and decision proposals led by empirical observation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

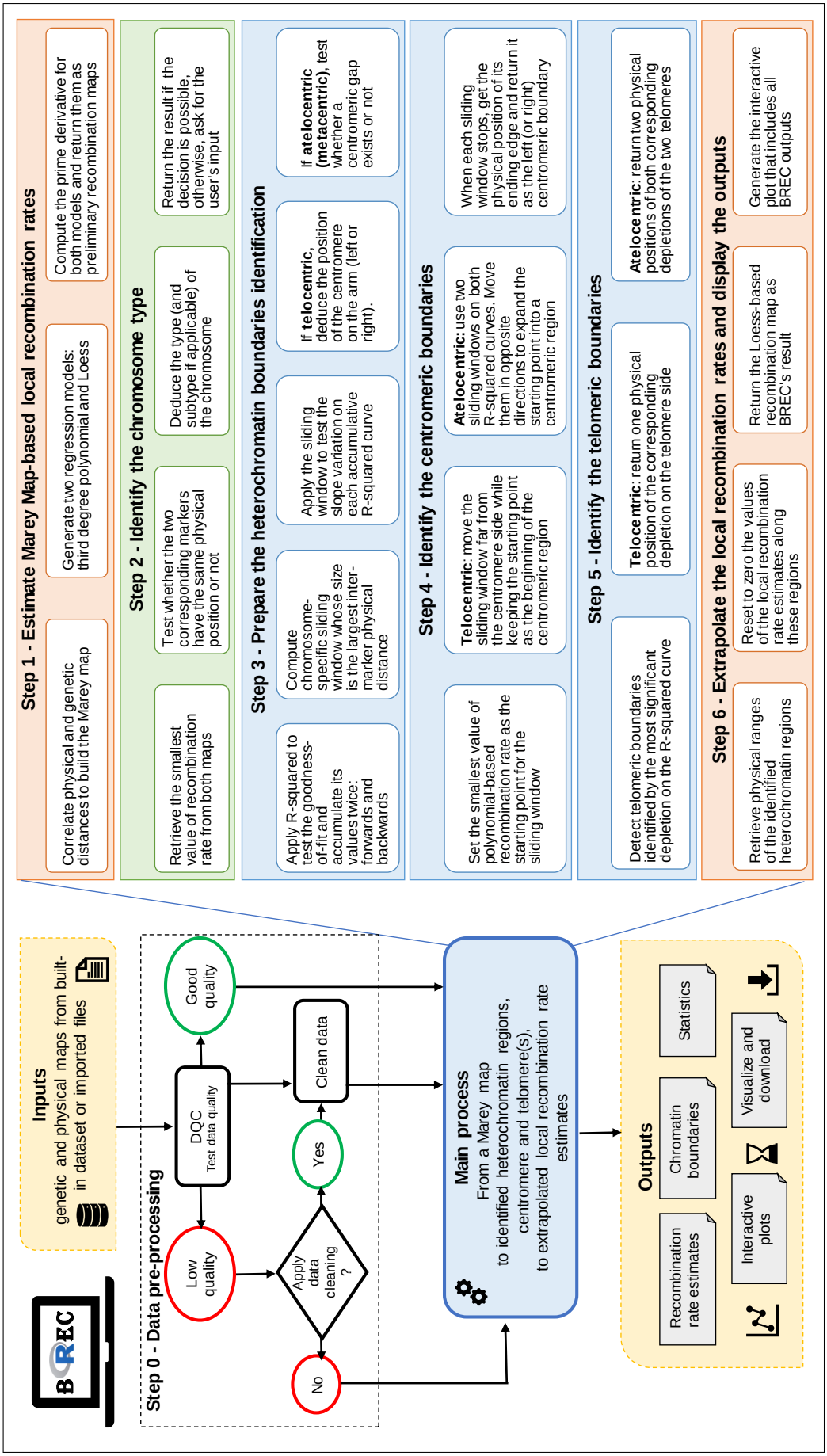


Figure 1: **BREC workflow.** This figure provides an overview of the tool design explaining how the different modules are linked together and how BREC functionalities are implemented. The diagram starting from the top left side shows input data, how are they processed, and what outputs are expected to be returned. Then, the right part of the figure, representing a zoom-in on BREC's main module (estimating recombination rates, identifying chromosome type, identifying heterochromatin boundaries, extrapolating the recombination map and generating the interactive plot), clarifies each step following a more detailed scheme.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The workflow starts with a pre-processing module (called "Step 0") aiming to prepare the data prior to the analysis. Then, it follows six main steps: (1) estimate Marey Map-based local recombination rates, (2) identify chromosome type, (3) prepare the heterochromatin boundaries identification, (4) identify the centromeric boundaries, (5) identify the telomeric boundaries, and (6) extrapolate the local recombination rate map and generate an interactive plot encompassing all BREC outputs (see Figure 1). Each step is detailed hereafter.

2.1 Step 0 - Apply data pre-processing

Since we have noticed that BREC estimates are sensitive to the quality of input data, we propose a pre-processing step to assess data quality and suggest an optional data cleaning for outliers. As such, we could ensure a proper functioning during further steps.

Data quality control The quality of input data is tested regarding two criteria: (1) density of markers and (2) the homogeneity of their distribution on the physical map, along a given chromosome. First, the mean density, defined as the number of markers per physical map length, is computed. This value is compared with the minimum required threshold of 2 markers/Mb. Based on the displayed results, the user gets to decide if data cleaning is required or not. The threshold of 2 markers/Mb is selected based on a simulation process that allowed to test BREC results while decreasing markers density until the observed heterochromatin boundaries estimates seemed to be no longer exploitable (see Materials and Methods in Section 3.2). Second, the distribution of input data is tested *via* a comparison with a simulated uniform distribution of identical markers density and physical map length. This comparison is applied using Pearson's $Chi - squared$ test (Agresti, 2007) which allows to examine how close the observed distribution (input data) is to the expected one (simulated data).

Data cleaning The cleaning step aims to reduce the disruptive impact of noisy data, such as outliers, in order to provide more accurate recombination rate and heterochromatin boundary results. If the input data fails to pass the Data Quality Control (DQC) test, the user has the option to apply or not a cleaning process. This process consists of identifying the extreme outliers and eliminating them upon the user's confirmation. Outliers are detected using the distribution statistics of the genetic map (see Figure S1). More precisely, inter-marker distances (separating each two consecutive points) are computed along the genetic map. Using a boxplot, distribution statistics (quartiles, mean, median) are applied on these inter-marker distances in order to identify outliers, which are chosen as the 5% of the data points with a genetic distance greater than the maximum extreme value, and should be discarded. Thus, the cleaning is targeting markers for which the genetic distance is quite larger than most of the rest. After the first cleaning iteration, DQC is applied again to assess the new density and distribution. The user can also choose to bypass the cleaning step, but in such case, BREC's behaviour is no longer guaranteed.

2.2 Step 1 - Estimate Marey Map-based local recombination rates

Once the data are cleaned, the recombination rate can be estimated based on the Marey map (Chakravarti, 1991) approach by: (1) correlating genetic and physical maps, (2) generating two regression models -third degree polynomial and Loess- that better fits these data, (3) computing the prime derivative for both models which will represent preliminary recombination maps for the chromosome. The main purpose of interpolation here is to provide

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

local recombination rate estimates for any given physical position, instead of only the ones corresponding to available markers.

At this point, both recombination maps are used to identify the chromosome type as well as the approximate position of centromeric and telomeric regions. Yet, as a final output, BREC will return only the Loess-based adjusted map for recombination rates since it provides finer local estimates than the polynomial-based map.

2.3 Step 2 - Identify chromosome type

BREC provides a function to identify the type of a given chromosome, with respect to the position of its centromere. This function is based on the physical position of the smallest value of recombination rate estimates, which primarily indicate where the centromeric region is more likely to be located. Our experimentation allowed to come up with the following scheme (see Figure S2). Two main types are identified: telocentric and atelocentric (Levan *et al.*, 1964). Atelocentric type could be either metacentric (centromere located approximately in the center with almost two equal arms) or not metacentric (centromere located between the center and one telomere of the chromosome). The latter includes the two most known subtypes, submetacentric and acrocentric (recently considered as types rather than subtypes). It is tricky for BREC to correctly distinguish between submetacentric and acrocentric chromosomes because the position of their centromeres varies slightly, and capturing this variation (based on the smallest value of recombination rate on both maps -polynomial and Loess-) could not be achieved, yet. Therefore, we chose to provide this result only if the identification process allowed to automatically identify the subtype. Otherwise, the user gets the statistics on the chromosome and is invited to decide according to further *a priori* knowledge. The two subtypes (metacentric and not metacentric) are distinguished following an intuitive reasoning inspired by their definition found in the literature. First, BREC identifies whether the chromosome is an arm (telocentric) or not (atelocentric). Then, test if the physical position of the smallest value of the estimated recombination rate is located between 40% and 60% interval, the subtype is displayed as *metacentric*, otherwise, it is displayed as *not metacentric*. The recombination rate is estimated using the Loess model ("LOcal regrESSion") (Cleveland and Devlin, 1988; Cleveland and Loader, 1996).

2.4 Step 3 - Prepare the heterochromatin boundaries identification

The heterochromatin boundaries identification is a purely statistical approach relying on the coefficient of determination R^2, which measures how good the generated regression model fits the input data (Zhang, 2017). We chose this approach because the Marey map usually exhibits lower quality of markers (density and distribution) on the heterochromatin regions. Thus, we aim to capture this transition from high to low quality regions (or *vice versa*) as it reflects the transition from euchromatin to heterochromatin regions (or *vice versa*). The coefficient R^2 is defined as the cumulative sum of squares of differences between the interpolation and observed data. R^2 values are accumulated along the chromosome. In order to eliminate the biased effect of accumulation, R^2 is computed twice: $R^2 - forward$ starts the accumulation from the beginning of the chromosome to provide the left centromeric and left telomeric boundaries, while $R^2 - backwards$ starts from the end of the chromosome providing the right centromeric and right telomeric boundaries. These R^2 values were calculated using the `rsq` package in R (Zhang, 2018). To compute R^2 cumulative vectors, `rsq` function is applied on the polynomial regression model. In fact, there is no such function for non-linear regression like Loess, because in such models, high R^2 does not always mean good fit. A sliding window is defined and applied on the R^2 vectors with the aim of precisely analysing

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

their variations (see details in the next step). In case of a telocentric chromosome, the position of the centromere is then deduced as the left or the right side of the arm, while in case of an atelocentric chromosome, the existence of a centromeric gap is investigated.

2.5 Step 4 - Identify centromeric boundaries

Since the centromeric region is known to present reduced recombination rates, the starting point for detecting its boundaries is the physical position corresponding to the smallest polynomial-based recombination rate value. Then, a sliding window is applied in order to expand the starting point into a region based on R^2 variations in two opposite directions. The size of the sliding window is automatically computed for each chromosome as the largest value of ranges between each two consecutive positions on the physical map (indicated as i and $i+1$ in Equation 1). After making sure the sliding window includes at least two data points, the mean of local growth rates inside the current window is computed and tested compared to zero. If it is positive (resp. negative) on the forward (resp. backwards) R^2 curve, the value corresponding to the window's ending edge is returned as the left (resp. right) boundary. Else, the window moves by a step value equal to its size.

$$sliding_window_size(chromosome) = max\{|physPos_{i+1} - physPos_i| : 1 \leq i \leq n-1\} \quad (1)$$

There are some cases where chromosome data present a centromeric gap. Such lack of data produces biased centromeric boundaries. To overcome this issue, chromosomes with a centromeric gap are handled with a slightly different approach: after comparing the mean of local rates of growth regarding to zero, accumulated slopes of all data points within the sliding window are computed adding one more point at a time. If the mean of accumulated slopes keeps the same variation direction as the mean of growth rates, the centromeric boundary is set as the ending edge of the window. Else, the window slides by the same step value as before (equal to its size). The difference between the two chromosome types is that for the telocentric case, only one sliding window is used, it's starting point is the centromeric side, and it moves away from it. As for the atelocentric case, two sliding windows are used (one on each R^2 curve), their starting point is the same, and they move in opposite directions to expand the centromere into a region.

2.6 Step 5 - Identify telomeric boundaries

Since telomeres are considered heterochromatin regions as well, they also tend to exhibit a low fitness between the regression model and the data points. More specifically, the accumulated R^2 curve tends to present a significant depletion around telomeres. Therefore, a telomeric boundary is defined here as the physical position of the most significant depletion corresponding to the smallest value of the R^2 curve. As such, in the telocentric case, only one R^2 curve is used and it gives one boundary of the telomeric region (the other boundary is defined by the beginning of the left telomere or the end of the right telomere). Whilst in the atelocentric case, where the are two telomeres, the depletion on $R^2 - forward$ detects the end of the left telomeric region and the depletion on $R^2 - backwards$ detects the beginning of the right telomeric region. The other two boundaries (the beginning of the left telomere and the end of the right telomere) are defined to be, respectively, the same values of the two markers with the smallest and the largest physical position available within the input data of the chromosome of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.7 Step 6 - Extrapolate the local recombination rate estimates and generate interactive plot

The extrapolation of recombination rate estimates within the identified centromeric and telomeric regions automatically performs an adjustment by resetting the initial biased values to zero along these heterochromatin ranges. Then, each of the above BREC outputs are combined to generate one interactive plot displayed for visualisation and download (see details in Section 4.5).

3 Materials and Methods

3.1 Validation data

The only input dataset to provide for BREC is genetic and physical maps one or several chromosomes. A simple CSV/TXT file with at least two columns for both maps is valid. If the dataset is for more than one chromosome or for the whole genome, a third column, with the chromosome identifier, is required.

Our results have been validated using the Release 5 of the fruit fly *D. melanogaster* (Hoskins *et al.*, 2007, 2015) genome as well as the domesticated tomato *Solanum lycopersicum* genome (version SL3.0).

We also tested BREC using other datasets of different species: house mouse (*Mus musculus castaneus*, MGI) chromosome 4 (Cox *et al.*, 2009), roundworm (*Caenorhabditis elegans*, ws170) chromosome 3 (Hillier *et al.*, 2008), zebrafish (*Danio rerio*, Zv6) chromosome 1 (Freeman *et al.*, 2007), respectively (see Figure S3), as samples from the multi-genome dataset included within BREC (see Table S3).

3.1.1 Fruit fly genome *D.melanogaster*

Physical and genetic maps are available for download from the FlyBase website (`http://flybase.org/`; Release 5) (Thurmond *et al.*, 2019). This genome is represented here with five chromosomal arms : 2L, 2R, 3L, 3R and X (see Table S1), for a total of 618 markers, 114.59Mb of physical map and 249.5cM of genetic map. This dataset is manually curated and is already clean from outliers. Therefore, the cleaning step offered within BREC was skipped.

3.1.2 Tomato genome *S. lycopersicum*

Domesticated tomato with 12 chromosomes has a genome size of approximately 900Mb. Based on the latest physical and genetic maps reported by the Tomato Genome Consortium (Sato *et al.*, 2012), we present both maps content (markers number, markers density, physical map length and genetic map length) for each chromosome in Table S2. For a total of 1957 markers, 752.47Mb of physical map and 1434.49cM of genetic map along the whole genome.

3.2 Simulated data for quality control testing

We call *data scenarios*, the layout in which the data markers are arranged along the physical map. Various *data scenarios*, for experimentally testing the limits of BREC, have been specifically designed based on *D. melanogaster* chromosomal arms (see Figure S4).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Markers density analysis In an attempt to investigate how markers density vary within and between the five chromosomal arms of *D. melanogaster* Release 5 genome, markers density is analyzed in two ways: locally (with 1-Mb bins) and globally (on the whole chromosome). Figure S5 shows the results of this investigation where each little box indicates how many markers are present within each bin of 1 Mb size on the physical map, while global markers density per chromosomes is represented by the mean value. Global markers density per chromosomes is also shown in Table S1 where the values are slightly different. This is due to computing markers density in two different ways with respect to the analysis. Table S1, presenting the genomic features of the validation dataset, shows markers density in Column 3, which is simply the result of the division of markers number (in column 2) by the physical map length (in Column 4). For example, in the case of chromosomal arm X, this gives $165/21.22 = 7.78 markers/Mb$. On the other hand, Figure S5, aimed for analysing the variation of local markers density, displays the mean of of all 1-Mb bins densities which is calculated as the sum of local densities divided by the number of bins, and this gives $165/22 = 7.5 markers/Mb$.

The exact same analysis has been conducted on the tomato genome *S. lycopersicum* where the only difference lies is using 5-Mb instead of 1-Mb bins, due to the larger size of its chromosomes (see Figure S6).

3.3 Validation metrics

The measure we used to evaluate the resolution of BREC's heterochromatin boundaries is called $shift$ hereafter. It is defined as the difference between the observed heterochromatin boundary ($observed_HCB$) and the expected one ($expected_HCB$) in terms of physical distance (in Mb)(see Equation 2).

$$shift = |observed_HCB - expected_HCB| \tag{2}$$

The $shift$ value is computed for each heterochromatin boundary independently. Therefore, we observe only two boundaries on a telocentric chromosome (one centromeric and one telomeric) while we observe four boundaries in case of an atelocentric chromosome (two centromeric giving the centromeric region and two telomeric giving each of the two telomeric regions).

The shift measure was introduced not only to validate BREC's results with the reference equivalents, but also to empirically calibrate the DQC module, where we are mostly interested in the variation of its value with respect to variations of the quality of input data.

3.4 Implementation and Analysis

The entire BREC project was developed using the R programming language (version 3.6.3 / 2020-02-29) (R Core Team, 2018) and the RStudio environment (version 1.2.5033) (RStudio Team, 2019). The graphical user interface is build using the shiny (RStudio, Inc, 2014) and shinydashboard (Chang and Borges Ribeiro, 2018) packages. The web-based interactive plots are generated by the plotly package. Data simulations, result analysis, reproducible reports and data visualizations are implemented using a large set of packages such as tidyverse, dplyr, R markdown, Sweave and knitr among others. The complete list of software resources used is available on the online version of BREC package accessible at `https://github.com/ymansour21/BREC`.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4 Results

In this section, we present the results obtained through the following validation process. First, we automatically re-identified heterochromatin boundaries with approximate resolution to the reference equivalents. Second, we tested the robustness of BREC method according to input data quality, using the well-studied *D. melanogaster* genome data, for which recombination rate and heterochromatin boundaries have already been accurately provided (Fiston-Lavier *et al.*, 2010; Comeron *et al.*, 2012; Chan *et al.*, 2012; Langley *et al.*, 2012)(Figure S7). In addition, we extended the robustness test to a completely different genome, the domesticated tomato *S. lycopersicum* (Sato *et al.*, 2012) to better interpret the study results. Even if the Loess span value does not impact the heterochromatin boundaries identification, but only the resulting recombination rate estimates, the span values used in this study are: 15% for *D. melanogaster* (for comparison purpose) and 25% for the rest of experiments. Our analysis shows that BREC is applicable on data from a various range of organisms, as long as the data quality is good enough. BREC is data-driven, thus, the outputs are strongly dependant of the markers density, distribution and chromosome type specified (automatically, or with the user's a priori knowledge).

4.1 Approximate, yet congruent heterochromatin boundaries

4.1.1 Fruit fly genome *D.melanogaster*

Our method for identifying heterochromatin boundaries has been primarily validated with cytological data experimentally generated on the *D. melanogaster* Release 5 genome (Riddle *et al.*, 2011; Chan *et al.*, 2012; Langley *et al.*, 2012; Thurmond *et al.*, 2019). For all five chromosomal arms (X, 2L, 2R, 3L, 3R). his genome presents a mean density of 5.39 markers/Mb and a mean physical map length of 22.92Mb. We obtained congruent heterochromatin boundaries with a good overlap and shift, distance between the physical position of the reference and BREC, from 20Kb to 4.58Mb (see Section 3). We did not observe a difference in terms of mean shift for the telomeric and centromeric BREC identification ($\chi^2 = 0.10$, df $= 1$, $p - value = 0.75$)(See Tables 1 and S1). We observe a lower resolution for the chromosomal arms 3L and 3R (see Figure 2). This suggests that the data for those two chromosomal arms might not present a quality as good as the rest of the genome. Interestingly, the local markers density for these two chromosomal arms shows a high variation, not like for the other chromosomal arms. For instance, the 2L for which BREC returns accurate results, shows a lower variation (see Figure S5). Without these two arms, the max shift for both centromeric and telomeric BREC boundaries is smaller than 1.54Mb with a mean shift decreasing from 1.43Mb to 0.71Mb.

This first analysis suggests that BREC method returns accurate results on this genome. However, the boundaries identification process appears very sensitive to the local density and distribution of the markers along a chromosome (see Figure 2). Therefore, we conducted further experiments on a different dataset, the tomato genome (see Figure S6).

4.1.2 Tomato genome *S. lycopersicum*

Results of experimenting BREC behaviour on all 12 chromosomes of *S. lycopersicum* genome (Sato *et al.*, 2012) are shown as values in Table 2 and as plots in Figure S8. This genome presents a mean density of 2.64 markers/Mb and a mean physical map length of 62.71Mb. We observe a variation in the shift value representing the difference on the physical map between reference heterochromatin boundaries and their equivalents returned by BREC. Unlike *D. melanogaster* genome which is of a smaller size, with five telocentric chromosomes (chromosomal arms) and a strongly different markers distribution, the tomato genome exhibits a

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

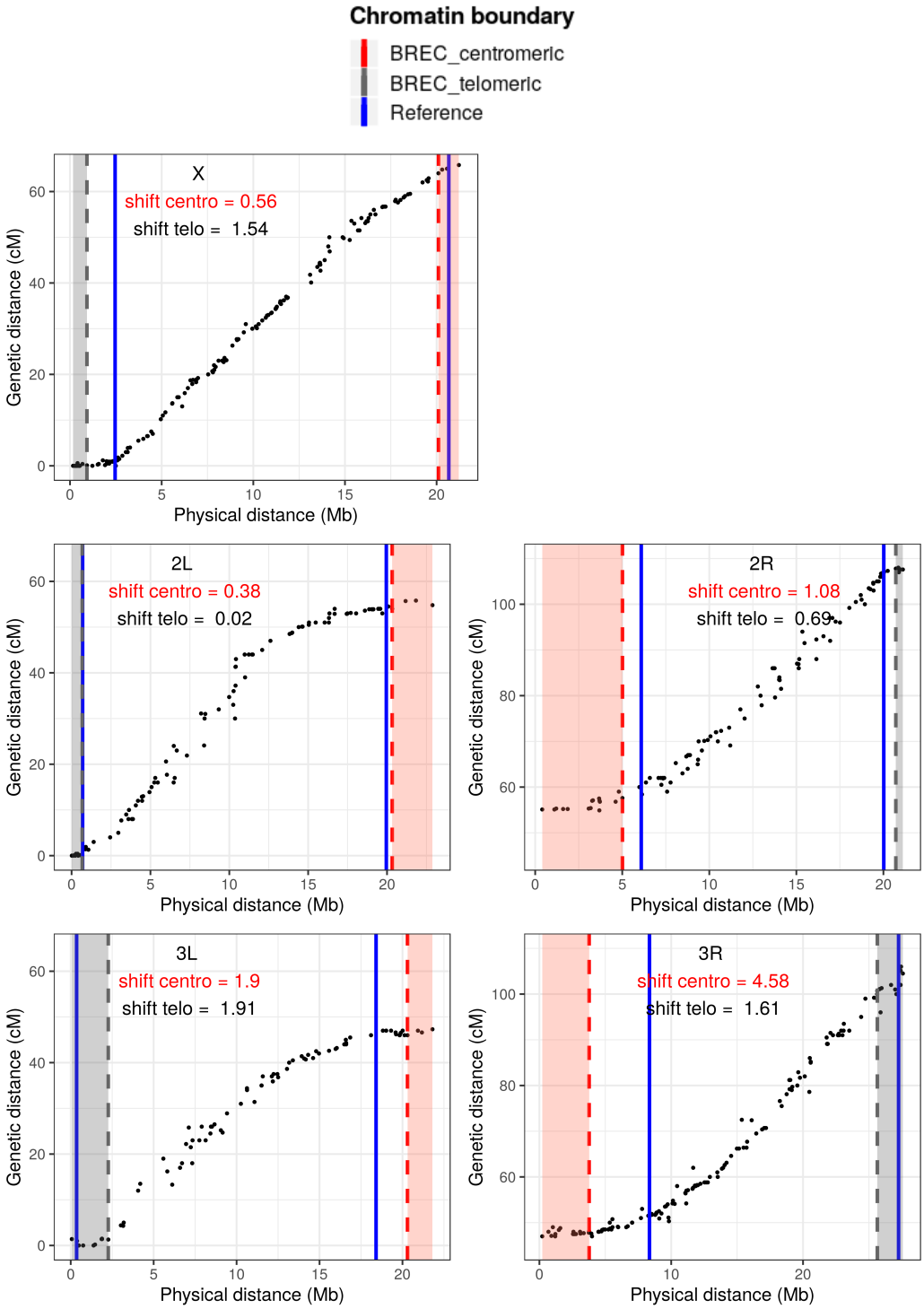


Figure 2: **Plots of the BREC heterochromatin boundaries and reference boundaries from the *D. melanogaster* genome**. The results are summarized in Table 1. From top to bottom are the five chromosomal arms X, 2L, 2R, 3L, 3R, respectively. Black dots represent genetic markers in ascendant order according to their physical position (in Mb). Vertical lines represent heterochromatin boundaries for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). For each chromosomal arm, two shift values of centromeric and telomeric boundaries are shown under the chromosome identifier.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Chromosomal arm	Centromeric (Mb)			Telomeric (Mb)		
	Boundaries		Shift	Boundaries		Shift
	Reference	BREC		Reference	BREC	
X	20.67	20.10	0.56	2.46	0.92	1.54
2L	19.95	20.33	0.38	0.70	0.68	0.02
2R	6.09	5.01	1.08	20.02	20.71	0.69
3L	18.41	20.30	1.90	0.36	2.26	1.91*
3R	8.35	3.77	4.58*	27.25	25.64	1.61
Min. shift	0.38			0.02		
Max. shift	4.58			1.91		
Mean shift	1.70			1.15		
Median shift	1.08			1.54		

Table 1: **BREC heterochromatin boundaries compared to reference boundaries from the reference genome of *D. melanogaster*.** The shift is the absolute value of the distance between the BREC and the reference physical heterochromatin boundary. The first five rows represent all chromosomal arms. Grouped columns present reference, BREC and shift values for the centromeric boundaries (Columns 2-4), and for the telomeric boundaries (Columns 4-6). Here the boundary values correspond to the internal heterochromatin boundaries. The external boundaries are represented by the physical positions of the first and the last markers of the chromosomes. All values are expressed in Megabase (Mb). The red asterisk indicates the largest shift value reported on centromeric and telomeric boundaries separately (see corresponding Figure 2). The last four rows represent general statistics on the shift value. From top to bottom, they are minimum, maximum, mean, and median respectively. See details on the shift metrics in Section 3.3.

completely different study case. This is a plant genome, with an approximately 8-fold bigger size genome. It is organized as twelve atelocentric chromosomes of a mean size of 60Mb except for chromosomes 2 and 6 which are more likely to be rather considered telocentric based on their markers distribution. Also, we observe a long plateau of markers along the centromeric region with a lower density than the rest of the chromosomes, something which highly differs from *D. melanogaster* data. We believe all these differences between both genomes gives a good validation but also evaluation for BREC behaviour towards various data quality scenarios. Furthermore, since BREC is a data-driven tool, these experiments help analysing data-related limitations that BREC could be facing while resolving differently. From another view point, BREC results on the tomato genome highlights the fact that markers distribution along heterochromatin regions, in particular, strongly impacts the identification of eu-heterochromatin boundaries, even when the density is of 2 markers/Mb or more.

4.2 Consistency despite the low data quality

We aim in this part to study to what extent BREC results are depending on the data quality.

4.2.1 BREC handles low markers density

We start by assessing the marker density on the BREC estimates. We generated simulated datasets with decreasing fractions of markers for each chromosomal arms (from 100% to 30%). For that, we randomly select a fraction of markers 30 times and compute the mean

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Chromosome	Centromeric left (Mb)			Centromeric right (Mb)		
	Boundaries		Shift	Boundaries		Shift
	Reference	BREC		Reference	BREC	
1	5.78	22.88	17.09	67.80	76.48	8.68
2	3.15	1.51	1.64	27.43	21.31	6.12
3	5.75	6.98	1.23	55.34	49.28	6.06
4	5.48	1.21	4.27	54.92	47.21	7.72
5	6.02	15.03	9.01	60.23	51.04	9.19
6	1.50	1.68	0.19	29.62	20.42	9.20
7	5.62	23.05	17.43	52.51	33.52	18.98*
8	5.10	22.87	17.77	51.73	43.96	7.77
9	4.38	32.51	28.12*	61.16	49.16	12.00
10	4.40	24.37	19.97	58.83	49.92	8.91
11	5.56	10.86	5.29	47.57	32.77	14.80
12	7.27	14.34	7.07	60.27	54.33	5.94
Min. shift	0.19			5.94		
Max. shift	28.12			18.98		
Mean shift	10.76			9.61		
Median shift	8.04			8.80		

Table 2: **BREC heterochromatin boundaries compared to reference boundaries from the reference genome of *S. lycopersicum*.** The shift is the absolute value of the distance between the BREC and the reference physical heterochromatin boundary. The first twelve rows represent all chromosomes. Grouped columns present reference, BREC and shift values for the left centromeric boundaries (Columns 2-4), and for the right centromeric boundaries (Columns 4-6). All values are expressed in Megabase (Mb). The red asterisk indicates the largest shift value reported on centromeric and telomeric boundaries separately (see corresponding Figure S8). The last four rows represent some general statistics on the shift value. From top to bottom, they are minimum, maximum, mean, and median respectively. See details on the shift metrics in Section 3.3.

shift between the BREC and the reference telomeric and centromeric boundaries. We note that BREC's resolution decreases drastically with the fraction and thus with the marker density (see Figure S9). However, BREC results appears stable until 70% of the data for all the chromosomal arms and more specifically for the telomeric boundary detection. Only for the centromeric boundary of the chromosomal arm 3R, we observe the opposite pattern: BREC returns more accurate telomeric boundary estimates when the number of markers decreases. This supports the low quality of the data around the 3R centromere.

This simulation process allowed to set a min density threshold representing the minimum value for data density in order to guarantee an accurate results of BREC estimates at 5 markers/Mb (fraction of around 70% of the data) on average in *D. melanogaster*. This analysis also supports that as the marker density alone can not explain the BREC resolution, BREC may be also sensitive to the marker distribution.

Figure S5 clearly shows that markers density varies within and between the five chromosomal arms with a mean of 4 to 8 markers/Mb. The variance is induced by the extreme values of local density, such as 0 or 24 markers/Mb on the chromosomal arm X. Still, the overall density is around 5 markers/Mb for the whole genome.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.2.2 BREC handles heterogeneous distribution

Along chromosomes, genetic markers are not homogeneously distributed. Therefore, to assess the impact of the markers distribution on BREC results, we designed different *data scenarios* with respect to reference data distribution (see Materials and Methods: 3.2). We choose as reference the chromosomal arms 2L and 2R of *D. melanogaster* as we obtained accurate results for these two chromosomal arms. After the concatenation of the two arms 2L and 2R, we ended up with a metacentric simulated chromosome as a starting simulation (total physical length of 44 Mb). While this length was kept unchanged, markers local density and distribution were modified (see Materials and Methods: 3.2; Figure S4).

One particular yet common case is the centromeric gap. Throughout our analysis, we consider that a chromosome presents a centromeric gap if its data exhibit a lack of genetic markers on a relatively large region on the physical map. As centromeric regions usually are less accessible to sequence due to its high compact state. Consequently, these regions are also hard to assemble and that is why a lot of genomes have chromosomes presenting a centromeric gap. It is important to know that a centromeric gap is not always exactly located on the middle of a chromosome. Instead, its physical location depends on the type of chromosome (see more details on Figure S2).

We also assess the veracity of BREC on datasets with variable distribution using simulated data with and without centromeric gap (see Figure S4).

For all six simulation datasets, BREC results overlap the reference boundaries. Thus BREC correctly handles the presence of a centromeric gap (see Figure S4: (a)(c)(e)). BREC stays robust to a non-uniform distribution of markers, under the condition that regions bordering the boundaries are greater than 2 markers/Mb (see Figure S10). In case of non-uniform distribution, BREC resolution is higher when the local density is stronger around heterochromatic regions (see Figure S4: (c)(d)(e)(f)). This suggests that low density on euchromatin regions far from the boundaries is not especially a problem either.

4.3 Accurate local recombination rate estimates

After the identification of heterochromatin boundaries, BREC provides optimized local estimates of recombination rate along the chromosome by taking into account the absence of recombination in heterochromatic regions. Recombination rates are set to zero across the centromeric and telomeric regions regardless of the regression model. To closely compare the third degree polynomial with Loess, using different span values, we experimented this aspect on *D. melanogaster* chromosomal arms and reported the results in Figure S11.

To assess the veracity of the recombination rates along the whole genome, we compared BREC results with previous recombination rate estimates (see Figure 3; (Chan *et al.*, 2012; Langley *et al.*, 2012)). BREC recombination rate estimates are significantly strongly correlated with reference data (Spearman's: $P << 0.001$) while the reference estimates fail in telomeric regions.

4.4 BREC is non-genome-specific

NGS, High Throughput Sequencing (HTS) technologies and numerous further computational advances are increasingly providing genetic and physical maps with more and more accessible markers along the centromeric regions. Such shift on the availability of data of poorly accessible genomic regions is a huge opportunity to shift our knowledge of the biology and dynamics of heterochromatin DNA sequences as Transposable Elements (TEs) for example. Therefore, BREC is not identifying centromeric gaps as centromeric regions as it might seem, instead, it is targeting centromeric as well as telomeric boundaries identification no matter of the presence or absence of markers neither of their density or distribution variations across

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

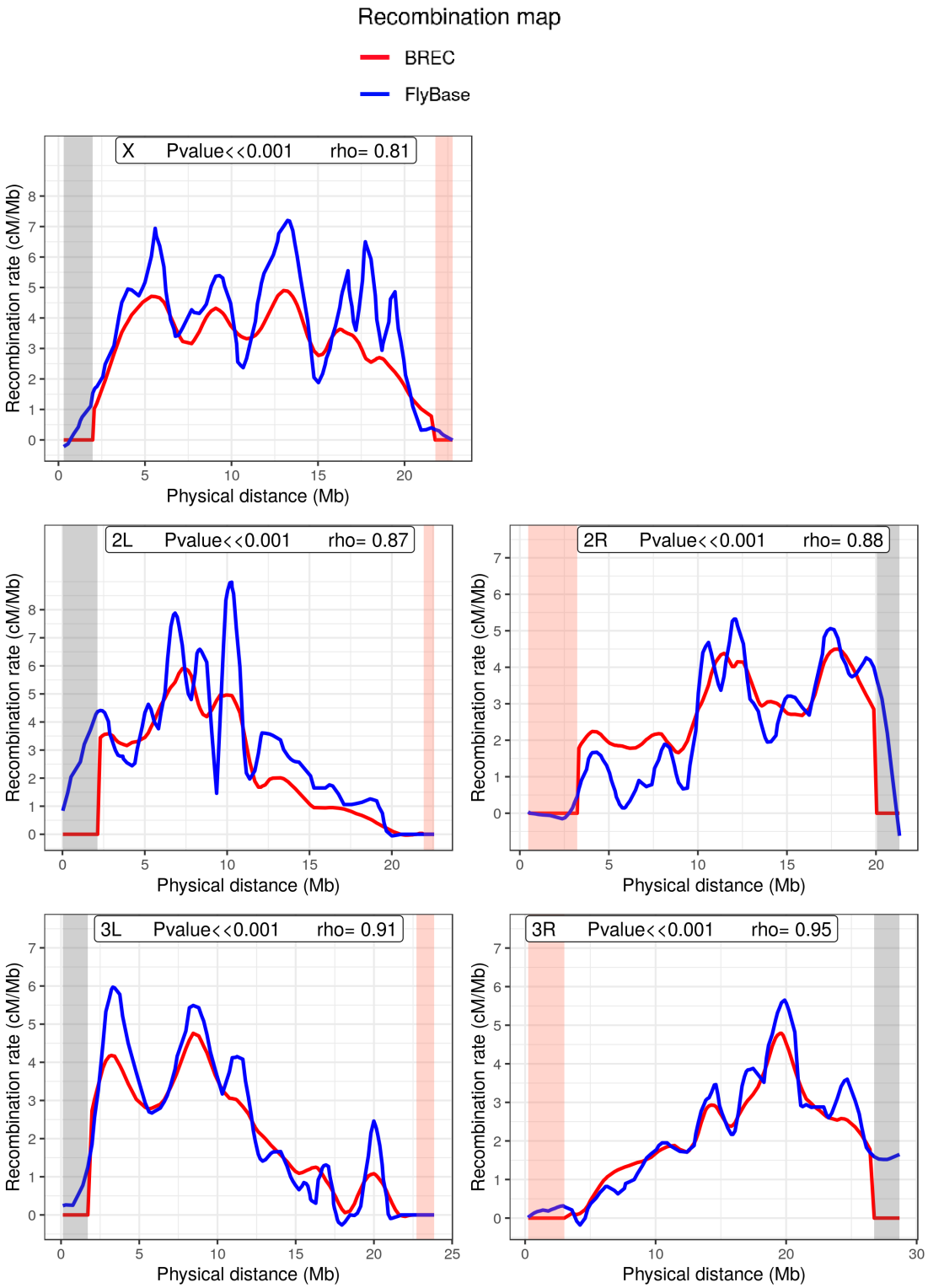


Figure 3: **Comparison of BREC *vs.* FlyBase recombination rate recombination rates along the five chromosomal arms (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5**. Both recombination maps are obtained using the same regression model: Loess with span 15%. The heterochromatin boundaries defined by BREC are represented in red and the reference data are in blue. Heterochromatic regions identified by BREC are highlighted in yellow.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

such complicated genomic regions (see Figure S3). Given BREC is non-genome-specific, applying heterochromatin boundaries identification on various genomes allows to widen the experimental design and to test more thoroughly how BREC responds to different *data scenarios*. Despite the several challenges due to data quality issues and following a data-driven approach, BREC is a non-genome-specific tool that aims to help tackling biological questions.

4.5 Easy, fast and accessible tool *via* an R-package and a Shiny user interface

BREC is an R-package entirely developed in R programming language (R Core Team, 2018). Current version of the package and documentation are available on the GitHub repository: `https://github.com/ymansour21/BREC`

In addition to the interactive visual results provided by BREC, the package comes with a web-based Graphical User Interface (GUI) build using the shiny and shinydashboard libraries (RStudio, Inc, 2014). The intuitive GUI makes it a lot easier to use BREC without struggling with the command line (see screenshots in Figures 4 and S12).

As for the speed aspect, BREC is quite fast when executing the main functions. We reported the running time for *D. melanogaster R5* and *S. lycopersicum* in Tables S1 and S2, respectively (plotting excluded). Nevertheless, when running BREC *via* the Shiny application, and due to the interactive plots displayed, it takes longer because of the plotly rendering. Still, it depends on the size of the genetic and physical maps used, as well as the markers density, as slightly appears in the same tables. The results presented from other species (see Figure S3) highlight better this dependence.

All BREC experiments have been carried out using a personal computer with the following specs:

- Processor: Intel® Core™ i7-7820HQ CPU @ 2.90GHz x 8
- Memory: 32Mo
- Hard disc: 512Go SSD
- Graphics: NV117 / Mesa Intel® HD Graphics 630 (KBL GT2)
- Operating system: 64-bit Ubuntu 20.04 LTS

From inside an R environment, the BREC package can be downloaded and installed using the command in the code chunk in Figure S13. In case of installation issues, further documentation is available online on the ReadMe page. If all runs correctly, the BREC shiny application will be launched on your default internet browser (see Shiny interface screenshots in Figure S12 and description of the build-in dataset as well as GUI elements in Supplementary materials).

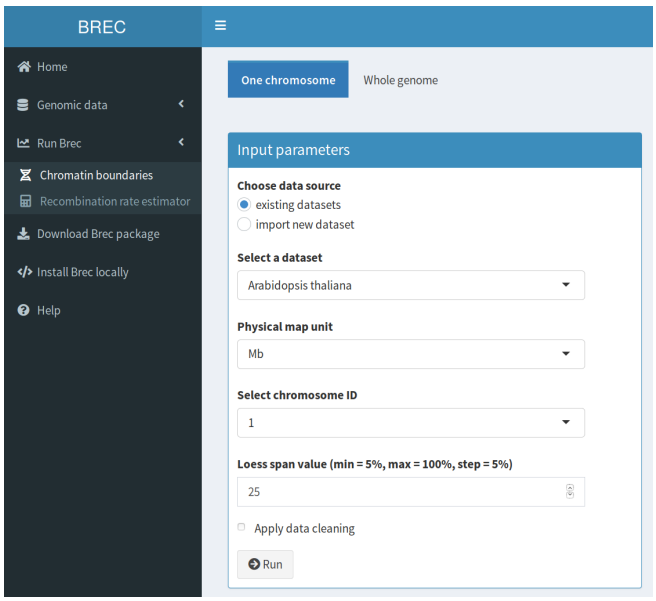
5 Discussion

The main two results of BREC are the eu-heterochromatin boundaries and the local recombination rate estimates (see Figures 2; 3).

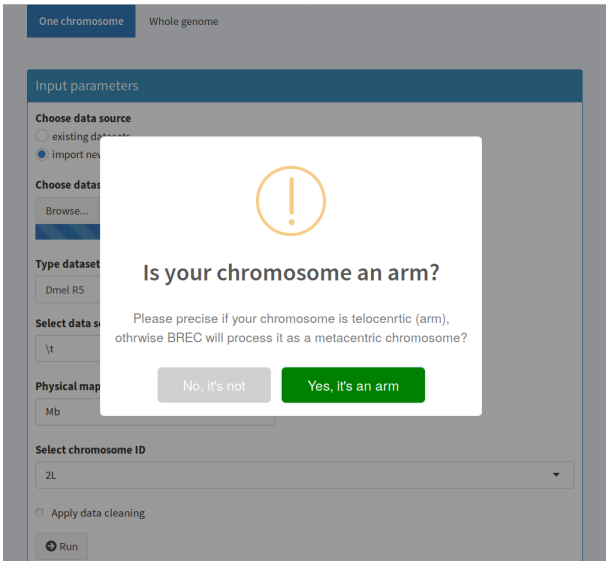
The heterochromatin boundaries algorithm, which identifies the location of centromeric and telomeric regions on the physical map, relies on the regression model obtained from correlation the physical distance and the genetic distance of each marker. Then, the goodness-of-fit measure, the R-squared, is used to obtain a curve upon which the transition between euchromatin and heterochromatin is detectable.

On the other hand, the recombination rate algorithm, which estimates local recombination

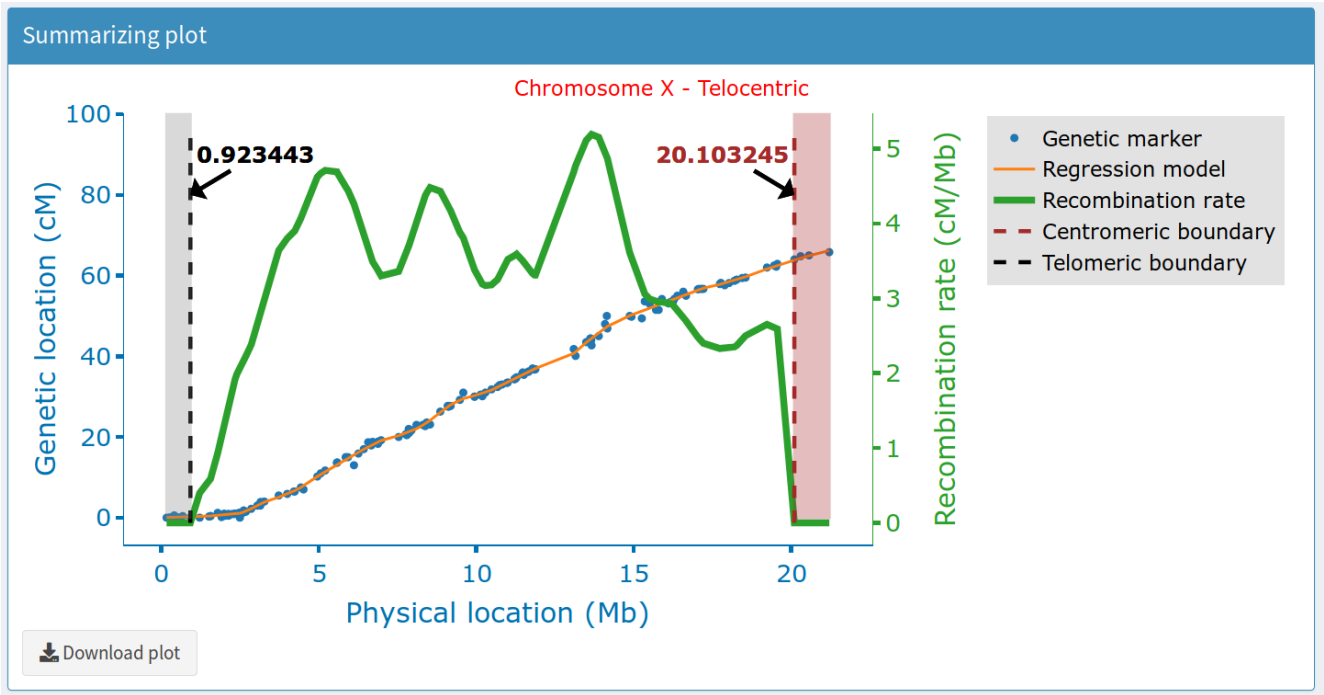
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) **Inputs - 1** Run BREC for heterochromatin boundaries page, indicated on the left dark panel.



(b) **Inputs - 2** After selecting input parameters and clicking the "Run" button, a popup alert is displayed to ask the user to confirm the chromosome type.



(c) **Outputs -** Here, the interactive summarizing plot of BREC main results is showing the telocentric chromosome X. Respectively with the plot legend order, it includes the input genetic markers (blue dots), the generated regression model (orange line), the local recombination rate estimates (green line), the centromeric boundary (dashed red vertical line on the right) delimiting the centromeric region (highlighted in light red), and the telomeric boundary (dashed black vertical line on the left) delimiting the telomeric region (highlighted in light grey)

Figure 4: **Screenshots of BREC web application - Run BREC web page** (4a) and (4b) show the inputs interface. (4c) shows the output of running BREC on the specified inputs, represented with an interactive web-based plot as a result.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

rates, returns the first derivative of the previous regression model as the recombination rates, then, resets the derivative values to zero along the heterochromatin regions identified (see Figure 1).

We validated the BREC method with a reference dataset known to be of high quality : *D. melanogaster*. While two distinct approaches were respectively implemented for the detection of telomeric and the centromeric regions, our results show a similar high resolution (see Table 1 and Figure 2). Then we analysed BREC's robustness using simulations of a progressive data degradation (see Figures S9; S10). Even if BREC is sensitive to the markers distribution and thus the local marker density, it can correctly handle a low global marker density. For *D. melanogaster* genome, a density of 5 markers/Mb seems to be sufficient to detect precisely the heterochromatin boundaries.

We also validated BREC using the domesticated tomato *S. lycopersicum* dataset (see Table 2 and Figure S8). At first glance, one might ask: why validating with this species when the results do not seem really congruent? In fact, we have decided to investigate this genome as it provides a more insightful understanding of the data-driven aspect of BREC and how data quality strongly impacts the heterochromatin identification algorithm. Variations in the local density of markers in this genome are particularly associated with the relatively large plateaued centromeric region representing more than 50% of the chromosome's length. Such *data scenario* is quite different compared to what we previously reported on the *D. melanogaster* chromosomal arms. This is partially the reason for which we chose this genome for testing BREC limits. While analysing the experiments more closely, we found that BREC processes some of the chromosomes as presenting a centromeric gap, while that is not actually the case. Thus, we forced the heterochromatin boundaries algorithm to automatically apply the *with-no-centromeric-gap-algorithm*, then, we were inspired to implement this option into the GUI in order to give the users the ability to take advantage of their *a priori* knowledge and by consequence to use BREC more efficiently. Meanwhile, we are considering how to make BREC completely automated regarding this point for an updated version later on. In addition, the reference heterochromatin results we used for the BREC validation are in fact rather an approximate than an exact indicator. The reference physical used correspond to the first and last markers tagged as "heterochromatic" on the spreadsheet file published by the Tomato Genome Consortium authors in (Sato *et al.*, 2012). However, we hesitated before validating BREC results with these approximate reference values due to the redundant existence of markers tagged as "euchromatic" directly before or after these reference positions. Unfortunately, we were not able to validate telomeric regions since the reference values were not available. As a result, we are convinced that BREC is approximating well enough in the face of all the disrupting factors mentioned above.

On the other hand, the ambition of this method is to escape species-dependence, which means it is conceived to be applicable to a various range of genomes. To test that, we thus also launched BREC on genomic data from different species (the house mouse's chromosome 4, roundworm's chromosome 3 and the chromosome 1 of zebrafish). Experiments on these whole genomes showed that BREC works as expected and identifies chromosome types in 95% of cases (see Figure S3).

One can assume, with the exponential increase of genomics resources associated with the revolution of the sequencing technologies, that more and more fine-scale genetic maps will be available. Therefore, BREC has quite the potential to widen the horizon of deployment of data science in the service of genome biology and evolution. It will be important to develop a dedicated database to store all these data.

BREC package and design offer numerous advantageous (see Table S4) compared to similar existing tools (Siberchicot *et al.*, 2017; Fiston-Lavier *et al.*, 2010). Thus, we believe our new computational solution will allow a large set of scientific questions, such as the ones raised by the authors of (Lenormand *et al.*, 2016; Stapley *et al.*, 2017), to be addressed more

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

confidently, considering model as well as non-model organisms, and with various perspectives.

6 Conclusion

We designed a user-friendly tool called BREC that analyses genomes on the chromosome scale, from the recombination point-of-view. BREC is a rapid and reliable method designed to determine euchromatin-heterochromatin frontiers on chromosomal arms or whole chromosomes (resp. telocentric or metacentric chromosomes). BREC also uses its heterochromatin boundary results to improve the recombination rate estimates along the chromosomes.

Whole genome version of BREC is a work in progress. Its will allow to run BREC on all the chromosomes of the genome of interest at ones. This version will also present the identified heterochromatin regions on chromosome ideograms. As short-term perspectives for this work, we may consider extending the robustness tests to other datasets with high quality and mandatory information (*e.g.* boundaries identified with cytological method, high quality maps). Retrieving such datasets seems to become less and less difficult. As well, we may improve the determination of boundaries with a finer analysis around them, for instance using an iterative multi-scale algorithm. Finally, we will be happy to take into account users feedback and improve the ergonomy and usability of the tool. As mid-term perspectives, we underline that BREC could integrate other algorithms aiming to provide further analysis options such as the comparison of heterochromatin regions between closely related species.

Acknowledgments

This work has been supported by the Algerian Ministry of Higher Education and Scientific Research as part of the Algerian Excellency Scholarship Program AVERROES. This work was also funded by the Junior Research Group (ERJ 2018) grant from Labex CeMEB (ANR). Our thanks go to the members of the "Phylogeny and Molecular Evolution" from the Institute of Sciences and Evolution of Montpellier (ISEM) and the "Methods and Algorithms in Bioinformatics" from the Laboratory of Computer Science, Robotics and Microelectronics of Montpellier (LIRMM).

References

Agresti, A. (2007). *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

Auton, A. and McVean, G. (2012). Estimating recombination rates from genetic variation in humans. *Methods in Molecular Biology*, **856**, 217–237.

Chakravarti, A. (1991). A graphical representation of genetic and physical maps: The Marey map. *Genomics*, **11**(1), 219–222.

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila melanogaster. *PLoS Genetics*, **8**(12), e1003090.

Chang, W. and Borges Ribeiro, B. (2018). *shinydashboard: Create Dashboards with 'Shiny'*. R package version 0.7.1.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, **83**(403), 596–610.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Cleveland, W. S. and Loader, C. (1996). Smoothing by local regression: Principles and methods. In W. Härdle and M. G. Schimek, editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 10–49, Heidelberg. Physica-Verlag HD.

Comeron, J. M., Ratnappan, R., and Bailin, S. (2012). The Many Landscapes of Recombination in Drosophila melanogaster. *PLoS Genet.*, **8**(10), e1002905.

Coop, G. and Przeworski, M. (2007). An evolutionary view of human recombination. *Nature Reviews Genetics*, **8**(1), 23–34.

Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLoS Biology*, **13**(4), e1002112.

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., Tsaih, S.-W., Churchill, G. A., and Broman, K. W. (2009). A New Standard Genetic Map for the Laboratory Mouse. *Genetics*, **182**(4), 1335–1344.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., Shamim, M. S., Machol, I., Lander, E. S., Aiden, A. P., and Aiden, E. L. (2017). De novo assembly of the aedes aegypti genome using hi-c yields chromosome-length scaffolds. *Science*, **356**(6333), 92–95.

Duret, L. and Galtier, N. (2009). Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annual Review of Genomics and Human Genetics*, **10**(1), 285–311.

Fiston-Lavier, A. S., Singh, N. D., Lipatov, M., and Petrov, D. A. (2010). Drosophila melanogaster recombination rate calculator. *Gene*, **463**(1-2), 18–20.

Freeman, J. L., Adeniyi, A., Banerjee, R., Dallaire, S., Maguire, S. F., Chi, J., Ng, B., Zepeda, C., Scott, C. E., Humphray, S., Rogers, J., Zhou, Y., Zon, L. I., Carter, N. P., Yang, F., and Lee, C. (2007). Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. *BMC Genomics*, **8**(1), 195.

Gui, S., Peng, J., Wang, X., Wu, Z., Cao, R., Salse, J., Zhang, H., Zhu, Z., Xia, Q., Quan, Z., Shu, L., Ke, W., and Ding, Y. (2018). Improving Nelumbo nucifera genome assemblies using high-resolution genetic maps and BioNano genome mapping reveals ancient chromosome rearrangements. *Plant Journal*, **94**(4), 721–734.

Hillier, L. D. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M., Tsung, E. F., Wylie, T., Schedl, T., Wilson, R. K., and Mardis, E. R. (2008). Whole-genome sequencing and variant discovery in C. elegans. *Nature Methods*, **5**(2), 183–188.

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K. H., Park, S., Mendez-Lago, M., Rossi, F., Villasante, A., Dimitri, P., Karpen, G. H., and Celniker, S. E. (2007). Sequence finishing and mapping of Drosophila melanogaster heterochromatin. *Science*, **316**(5831), 1625–1628.

Hoskins, R. A., Carlson, J. W., Wan, K. H., Park, S., Mendez, I., Galle, S. E., Booth, B. W., Pfeiffer, B. D., George, R. A., Svirskas, R., Krzywinski, M., Schein, J., Accardo, M. C., Damia, E., Messina, G., Méndez-Lago, M., De Pablos, B., Demakova, O. V., Andreyeva, E. N., Boldyreva, L. V., Marra, M., Carvalho, A. B., Dimitri, P., Villasante, A., Zhimulev, I. F., Rubin, G. M., Karpen, G. H., and Celniker, S. E. (2015). The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Research*, **25**(3), 445–458.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Jeffreys, A. J. (2000). High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Human Molecular Genetics*, **9**(5), 725–733.

Langley, C. H., Stevens, K., Cardeno, C., Lee, Y. C. G., Schrider, D. R., Pool, J. E., Langley, S. A., Suarez, C., Corbett-Detig, R. B., Kolaczkowski, B., Fang, S., Nista, P. M., Holloway, A. K., Kern, A. D., Dewey, C. N., Song, Y. S., Hahn, M. W., and Begun, D. J. (2012). Genomic variation in natural populations of Drosophila melanogaster. *Genetics*, **192**(2), 533–598.

Lenormand, T., Engelstädter, J., Johnston, S. E., Wijnker, E., and Haag, C. R. (2016). Evolutionary mysteries in meiosis.

Levan, A., Fredga, K., and Sandberg, A. A. (1964). Nomenclature for centromeric position on chromosomes. *Hereditas*, **52**(2), 201–220.

Lu, M. and He, X. (2019). Centromere repositioning causes inversion of meiosis and generates a reproductive barrier. *Proceedings of the National Academy of Sciences*, **116**(43), 21580–21591.

Morata, J., Tormo, M., Alexiou, K. G., Vives, C., Ramos-Onsins, S. E., Garcia-Mas, J., and Casacuberta, J. M. (2018). The evolutionary consequences of transposon-related pericentromer expansion in melon. *Genome Biology and Evolution*, **10**(6), 1584–1595.

Muller, H., Gil, J., and Drinnenberg, I. A. (2019). The Impact of Centromeres on Spatial Genome Architecture. *Trends in Genetics*, **35**(8), 565–578.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rezvoy, C., Charif, D., Guéguen, L., and Marais, G. A. (2007). MareyMap: An R-based tool with graphical interface for estimating recombination rates. *Bioinformatics*, **23**(16), 2188–2189.

Riddle, N. C., Minoda, A., Kharchenko, P. V., Alekseyenko, A. A., Schwartz, Y. B., Tolstorukov, M. Y., Gorchakov, A. A., Jaffe, J. D., Kennedy, C., Linder-Basso, D., Peach, S. E., Shanower, G., Zheng, H., Kuroda, M. I., Pirrotta, V., Park, P. J., Elgin, S. C. R., and Karpen, G. H. (2011). Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. *Genome Research*, **21**(2), 147–163.

Robert L. Nussbaum, M., McInnes, R., and Willard, H. (2015). *Thompson & Thompson Genetics in Medicine*. Saunders W.B. Elsevier Health Sciences.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R., and Weigel, D. (2019). An Ultra High-Density Arabidopsis thaliana Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics*, **213**(3), genetics.302406.2019.

RStudio, Inc (2014). *Shiny: Easy web applications in R*. URL: `http://shiny.rstudio.com`.

RStudio Team (2019). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Aoki, K., Egholm, M., Knight, J., Bogden, R., Li, C., Shuang, Y., Xu, X., Pan, S., Cheng, S., Liu, X., Ren, Y., Wang, J., Albiero, A., Dal Pero, F.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Todesco, S., Van Eck, J., Buels, R. M., Bombarely, A., Gosselin, J. R., Huang, M., Leto, J. A., Menda, N., Strickler, S., Mao, L., Gao, S., Tecle, I. Y., York, T., Zheng, Y., Vrebalov, J. T., Lee, J., Zhong, S., Mueller, L. A., Stiekema, W. J., Ribeca, P., Alioto, T., Yang, W., Huang, S., Du, Y., Zhang, Z., Gao, J., Guo, Y., Wang, X., Li, Y., He, J., Li, C., Cheng, Z., Zuo, J., Ren, J., Zhao, J., Yan, L., Jiang, H., Wang, B., Li, H., Li, Z., Fu, F., Chen, B., Han, B., Feng, Q., Fan, D., Wang, Y., Ling, H., Xue, Y., Ware, D., Richard McCombie, W., Lippman, Z. B., Chia, J. M., Jiang, K., Pasternak, S., Gelley, L., Kramer, M., Anderson, L. K., Chang, S. B., Royer, S. M., Shearer, L. A., Stack, S. M., Rose, J. K., Xu, Y., Eannetta, N., Matas, A. J., McQuinn, R., Tanksley, S. D., Camara, F., Guigó, R., Rombauts, S., Fawcett, J., Van De Peer, Y., Zamir, D., Liang, C., Spannagl, M., Gundlach, H., Bruggmann, R., Mayer, K., Jia, Z., Zhang, J., Ye, Z., Bishop, G. J., Butcher, S., Lopez-Cobollo, R., Buchan, D., Filippis, I., Abbott, J., Dixit, I. R., Singh, M., Singh, A., Pal, J. K., Pandit, A., Singh, P. K., Mahato, A. K., Dogra, V., Gaikwad, K., Sharma, T. R., Mohapatra, T., Singh, N. K., Causse, M., Rothan, C., Noirot, C., Bellec, A., Klopp, C., Delalande, C., Berges, H., Mariette, J., Frasse, P., Vautrin, S., Zouine, T. M., Latché, A., Rousseau, C., Regad, F., Pech, J. C., Philippot, M., Bouzayen, M., Pericard, P., Osorio, S., Del Carmen, A. F., Monforte, A., Granell, A., Fernandez-Muñoz, R., Conte, M., Lichtenstein, G., Carrari, F., De Bellis, G., Fuligni, F., Peano, C., Grandillo, S., Termolino, P., Pietrella, M., Fantini, E., Falcone, G., Fiore, A., Giuliano, G., Lopez, L., Facella, P., Perrotta, G., Daddiego, L., Bryan, G., Orozco, B. M., Pastor, X., Torrents, D., Van Schriek, M. G., Feron, R. M., Van Oeveren, J., De Heer, P., Da Ponte, L., Jacobs-Oomen, S., Cariaso, M., Prins, M., Van Eijk, M. J., Janssen, A., Van Haaren, J. J., HwanJo, S., Kim, J., Kwon, S. Y., Kim, S., Koo, D. H., Lee, S., Clouser, C., Rico, A., Hallab, A., Gebhardt, C., Klee, K., Jöcker, A., Warfsmann, J., Göbel, U., Kawamura, S., Yano, K., Sherman, J. D., Fukuoka, H., Negoro, S., Bhutty, S., Chowdhury, P., Chattopadhyay, D., Datema, E., Smit, S., Schijlen, E. G., Van De Belt, J., Van Haarst, J. C., Peters, S. A., Van Staveren, M. J., Henkens, M. H., Mooyman, P. J., Hesselink, T., Van Ham, R. C., Jiang, G., Droege, M., Choi, D., Kang, B. C., Kim, B. D., Park, M., Kim, S., Yeom, S. I., Lee, Y. H., Choi, Y. D., Li, G., Gao, J., Liu, Y., Huang, S., Fernandez-Pedrosa, V., Collado, C., Zuñ Iga, S., Wang, G., Cade, R., Dietrich, R. A., Rogers, J., Knapp, S., Fei, Z., White, R. A., Thannhauser, T. W., Giovannoni, J. J., Botella, M. A., Gilbert, L., Gonzalez, F. R., Goicoechea, J. L., Yu, Y., Kudrna, D., Collura, K., Wissotski, M., Wing, R., Meyers, B. C., Gurazada, A. B., Green, P. J., Mathur, S., Vyas, S., Solanke, A. U., Kumar, R., Gupta, V., Sharma, A. K., Khurana, P., Khurana, J. P., Tyagi, A. K., Dalmay, T., Mohorianu, I., Walts, B., Chamala, S., Barbazuk, W. B., Li, J., Guo, H., Lee, T. H., Wang, Y., Zhang, D., Paterson, A. H., Wang, X., Tang, H., Barone, A., Chiusano, M. L., Ercolano, M. R., D'Agostino, N., Di Filippo, M., Traini, A., Sanseverino, W., Frusciante, L., Seymour, G. B., Elharam, M., Fu, Y., Hua, A., Kenton, S., Lewis, J., Lin, S., Najar, F., Lai, H., Qin, B., Shi, R., Qu, C., White, D., White, J., Xing, Y., Yang, K., Yi, J., Yao, Z., Zhou, L., Roe, B. A., Vezzi, A., D'Angelo, M., Zimbello, R., Schiavon, R., Caniato, E., Rigobello, C., Campagna, D., Vitulo, N., Valle, G., Nelson, D. R., De Paoli, E., Szinay, D., De Jong, H. H., Bai, Y., Visser, R. G., Lankhorst, R. K., Beasley, H., McLaren, K., Nicholson, C., Riddle, C., and Gianese, G. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**(7400), 635–641.

Schueler, M. G., Higgins, A. W., Rudd, M. K., Gustashaw, K., and Willard, H. F. (2001). Genomic and genetic definition of a functional human centromere. *Science*, **294**(5540), 109–115.

Shen, C., Li, X., Zhang, R., and Lin, Z. (2017). Genome-wide recombination rate variation in a recombination map of cotton. *PLoS ONE*, **12**(11), e0188682.

Siberchicot, A., Bessy, A., Guéguen, L., and Marais, G. A. B. (2017). MareyMap Online: A

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

User-Friendly Web Application and Database Service for Estimating Recombination Rates Using Physical and Genetic Maps. *Genome Biol. Evol.*, **9**(10), 2506–2509.

Sievert, C. (2018). *plotly for R*.

Silva-Junior, O. B. and Grattapaglia, D. (2015). Genome-wide patterns of recombination, linkage disequilibrium and nucleotide diversity from pooled resequencing and single nucleotide polymorphism genotyping unlock the evolutionary history of Eucalyptus grandis. *New Phytologist*, **208**(3), 830–845.

Stapley, J., Feulner, P. G. D., Johnston, S. E., Santure, A. W., and Smadja, C. M. (2017). Variation in recombination frequency and distribution across eukaryotes: patterns and processes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **372**(1736), 20160455.

Stumpf, M. P. and McVean, G. A. (2003). Estimating recombination rates from population-genetic data. *Nat. Rev. Genet.*, **4**(12), 959–968.

Termolino, P., Cremona, G., Consiglio, M. F., and Conicella, C. (2016). Insights into epigenetic landscape of recombination-free regions. *Chromosoma*, **125**(2), 301–308.

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., Matthews, B. B., Millburn, G., Antonazzo, G., Trovisco, V., Kaufman, T. C., Calvi, B. R., Perrimon, N., Gelbart, S. R., Agapite, J., Broll, K., Crosby, L., Dos Santos, G., Emmert, D., Falls, K., Jenkins, V., Sutherland, C., Tabone, C., Zhou, P., Zytkovicz, M., Brown, N., Garapati, P., Holmes, A., Larkin, A., Pilgrim, C., Urbano, P., Czoch, B., Cripps, R., and Baker, P. (2019). FlyBase 2.0: The next generation. *Nucleic Acids Research*, **47**(D1), D759–D765.

Vanrobays, E., Thomas, M., and Tatout, C. (2017). Heterochromatin Positioning and Nuclear Architecture. In *Annual Plant Reviews online*, pages 157–190. John Wiley & Sons, Ltd, Chichester, UK.

Weinstock, G. M., Robinson, G. E., Gibbs, R. A., Worley, K. C., Evans, J. D., Maleszka, R., Robertson, H. M., Weaver, D. B., Beye, M., Bork, P., Elsik, C. G., Hartfelder, K., Hunt, G. J., Zdobnov, E. M., Amdam, G. V., Bitondi, M. M., Collins, A. M., Cristino, A. S., Lattorff, H. M. G., Lobo, C. H., Moritz, R. F., Nunes, F. M., Page, R. E., Simões, Z. L., Wheeler, D., Carninci, P., Fukuda, S., Hayashizaki, Y., Kai, C., Kawai, J., Sakazume, N., Sasaki, D., Tagami, M., Albert, S., Baggerman, G., Beggs, K. T., Bloch, G., Cazzamali, G., Cohen, M., Drapeau, M. D., Eisenhardt, D., Emore, C., Ewing, M. A., Fahrbach, S. E., Forêt, S., Grimmelikhuijzen, C. J., Hauser, F., Hummon, A. B., Huybrechts, J., Jones, A. K., Kadowaki, T., Kaplan, N., Kucharski, R., Leboulle, G., Linial, M., Littleton, J. T., Mercer, A. R., Richmond, T. A., Rodriguez-Zas, S. L., Rubin, E. B., Sattelle, D. B., Schlipalius, D., Schoofs, L., Shemesh, Y., Sweedler, J. V., Velarde, R., Verleyen, P., Vierstraete, E., Williamson, M. R., Ament, S. A., Brown, S. J., Corona, M., Dearden, P. K., Dunn, W. A., Elekonich, M. M., Fujiyuki, T., Gattermeier, I., Gempe, T., Hasselmann, M., Kadowaki, T., Kage, E., Kamikouchi, A., Kubo, T., Kucharski, R., Kunieda, T., Lorenzen, M., Milshina, N. V., Morioka, M., Ohashi, K., Overbeek, R., Ross, C. A., Schioett, M., Shippy, T., Takeuchi, H., Toth, A. L., Willis, J. H., Wilson, M. J., Gordon, K. H., Letunic, I., Hackett, K., Peterson, J., Felsenfeld, A., Guyer, M., Solignac, M., Agarwala, R., Cornuet, J. M., Monnerot, M., Mougel, F., Reese, J. T., Schlipalius, D., Vautrin, D., Gillespie, J. J., Cannone, J. J., Gutell, R. R., Johnston, J. S., Eisen, M. B., Iyer, V. N., Iyer, V., Kosarev, P., Mackey, A. J., Solovyev, V., Souvorov, A., Aronstein, K. A., Bilikova, K., Chen, Y. P., Clark, A. G., Decanini, L. I., Gelbart, W. M., Hetru, C., Hultmark, D., Imler, J. L.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Jiang, H., Kanost, M., Kimura, K., Lazzaro, B. P., Lopez, D. L., Simuth, J., Thompson, G. J., Zou, Z., De Jong, P., Sodergren, E., Csurös, M., Milosavljevic, A., Osoegawa, K., Richards, S., Shu, C. L., Duret, L., Elhaik, E., Graur, D., Anzola, J. M., Campbell, K. S., Childs, K. L., Collinge, D., Crosby, M. A., Dickens, C. M., Grametes, L. S., Grozinger, C. M., Jones, P. L., Jorda, M., Ling, X., Matthews, B. B., Miller, J., Mizzen, C., Peinado, M. A., Reid, J. G., Russo, S. M., Schroeder, A. J., St. Pierre, S. E., Wang, Y., Zhou, P., Jiang, H., Kitts, P., Ruef, B., Venkatraman, A., Zhang, L., Aquino-Perez, G., Whitfield, C. W., Behura, S. K., Berlocher, S. H., Sheppard, W. S., Smith, D. R., Suarez, A. V., Tsutsui, N. D., Wei, X., Wheeler, D., Havlak, P., Li, B., Liu, Y., Jovilet, A., Lee, S., Nazareth, L. V., Pu, L. L., Thorn, R., Stolc, V., Newman, T., Samanta, M., Tongprasit, W. A., Claudianos, C., Berenbaum, M. R., Biswas, S., De Graaf, D. C., Feyereisen, R., Johnson, R. M., Oakeshott, J. G., Ranson, H., Schuler, M. A., Muzny, D., Chacko, J., Davis, C., Dinh, H., Gill, R., Hernandez, J., Hines, S., Hume, J., Jackson, L. R., Kovar, C., Lewis, L., Miner, G., Morgan, M., Nguyen, N., Okwuonu, G., Paul, H., Santibanez, J., Savery, G., Svatek, A., Villasana, D., and Wright, R. (2006). Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **443**(7114), 931–949.

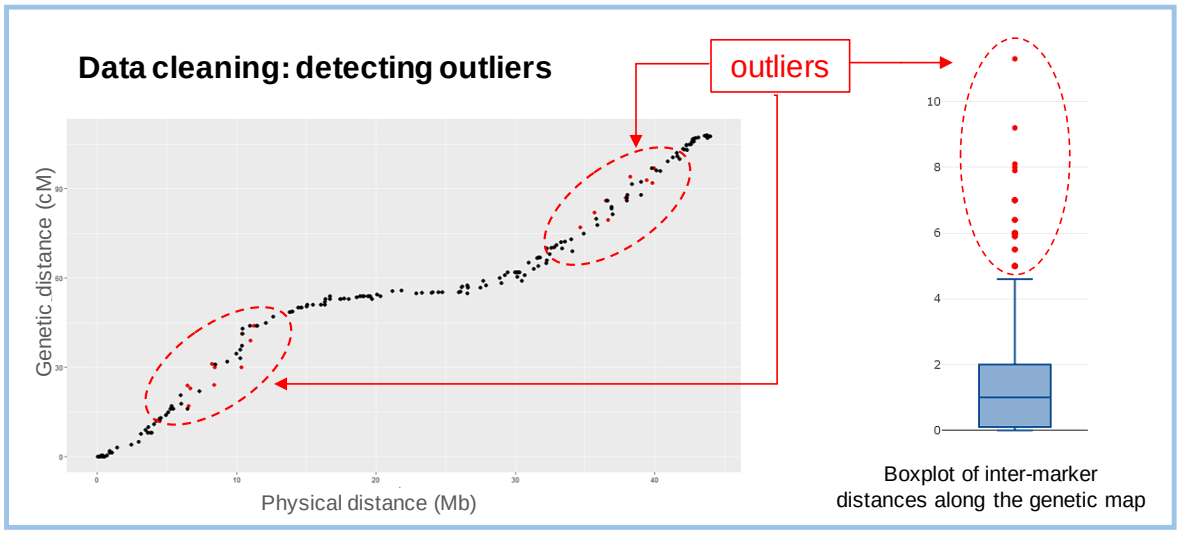
Zhang, D. (2017). A Coefficient of Determination for Generalized Linear Models. *American Statistician*, **71**(4), 310–316.

Zhang, D. (2018). *rsq: R-Squared and Related Measures*. R package version 1.1.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

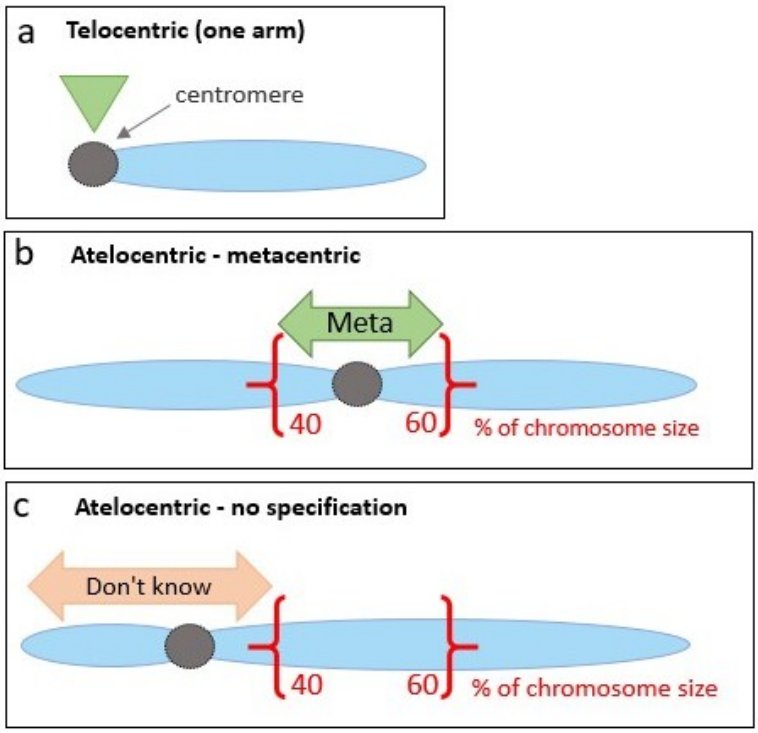
Supplementary materials

Figure S1: **The data cleaning process implemented within BREC**. Inter-marker distances (*i.e.* genetic distances between each two consecutive points along the genetic map) are represented using a boxplot in order to identify outliers and give the user the option to remove them. Here is an example showing raw data of a simulated chromosome (left) with the specific markers detected as outliers (red dots circled with red dashed ovals) and the corresponding genetic distances (also in red) on the boxplot (right).



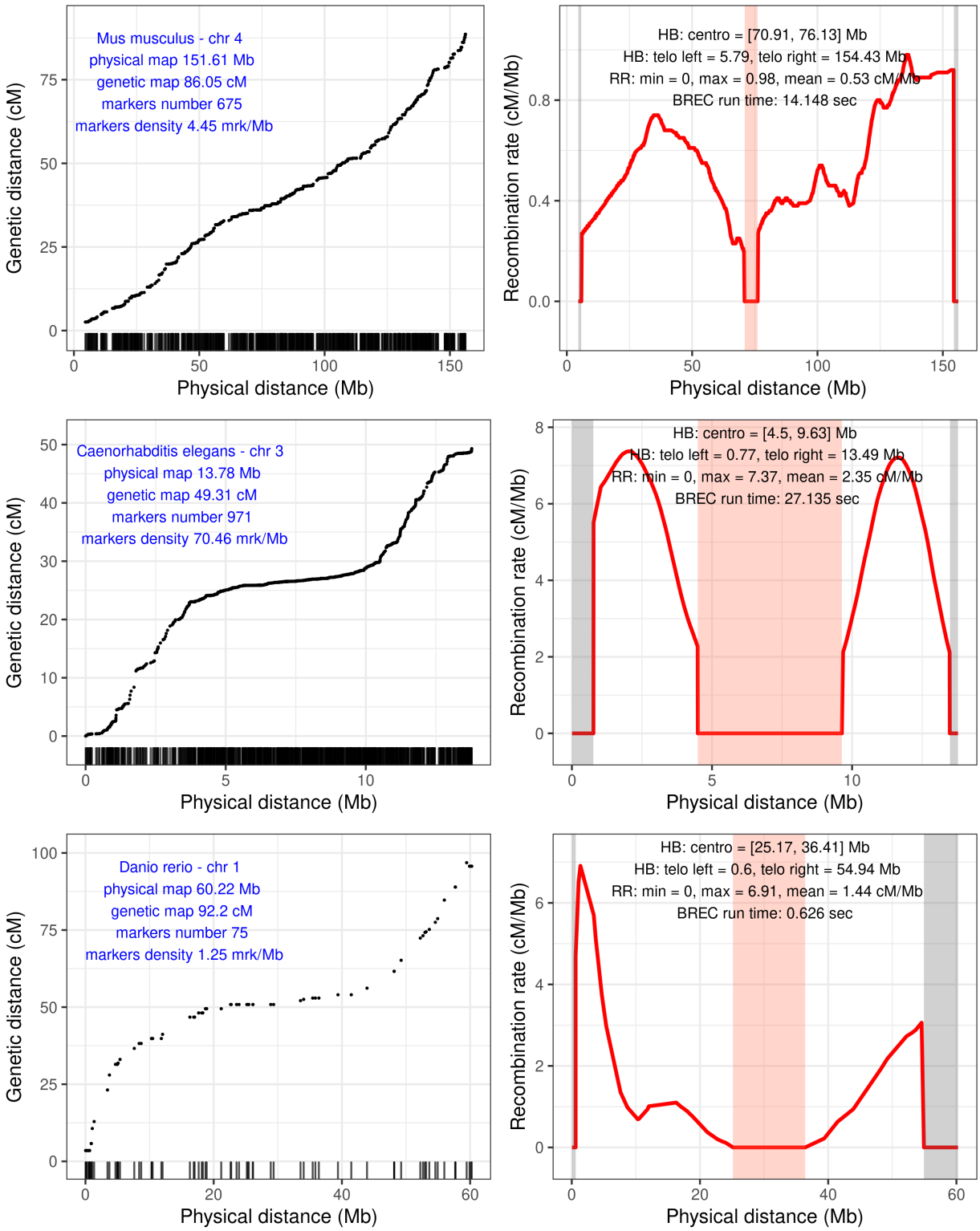
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S2: **A schematic description of the chromosome type identification process implemented within BREC.** (a) Telocentric chromosome type is when the centromere (the grey colored circle) is located on one of the chromosomal arm extremities (indicated with the green upside down triangle). (b) Atelocentric chromosome type -confirmed as metacentric- is when the centromere is located approximately on the middle of the chromosome, here showed within the physical positions 40% and 60% of the chromosome's size (delimited by the red brackets and indicated with the tag "Meta"). (c) Atelocentric chromosome type -with no specification- is when the centromere is located either inside the first arm (between the beginning of the chromosome and 40% of its size), or inside the second arm (between 60% and the end, indicated with the tag "Don't know").



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

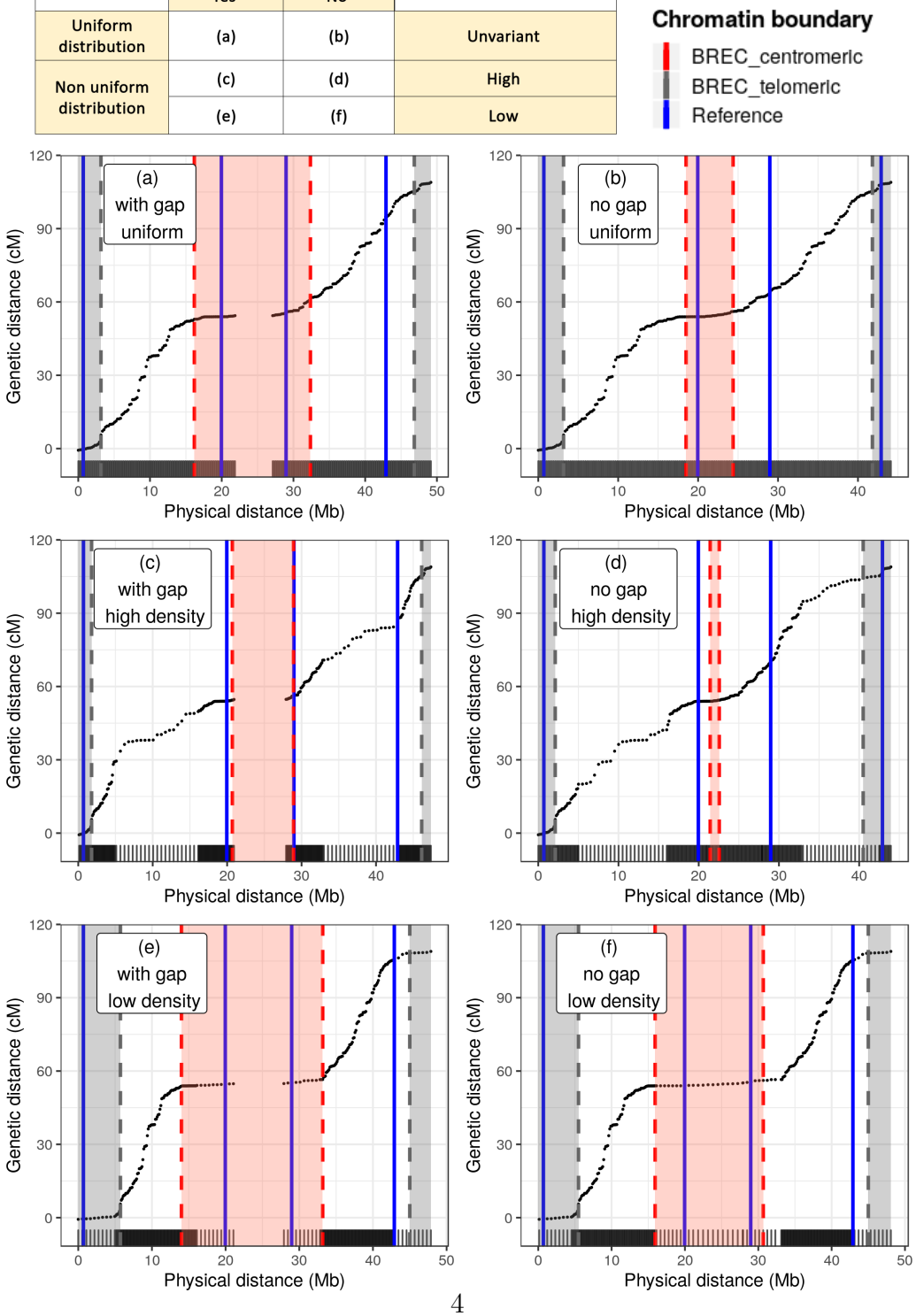
Figure S3: **BREC results on different species: from top to bottom are *M. musculus* (house mouse) chromosome 4, *C. elegans* (roundworm) chromosome 3, *D. rereo* (zebrafish) chromosome 1, respectively.** For each species, two plots are shown: on the left is the chromosome's genetic markers (black points), their distribution along the physical map (rug on the x-axis), and reported genomic features (label in blue). On the right is BREC results: (HB) heterochromatin boundaries for centromeric (red highlight) and telomeric (grey highlight) regions, (RR) local recombination rate estimates (red line), and the running time of BREC's algorithms to get these results (loading data and plotting are excluded).



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

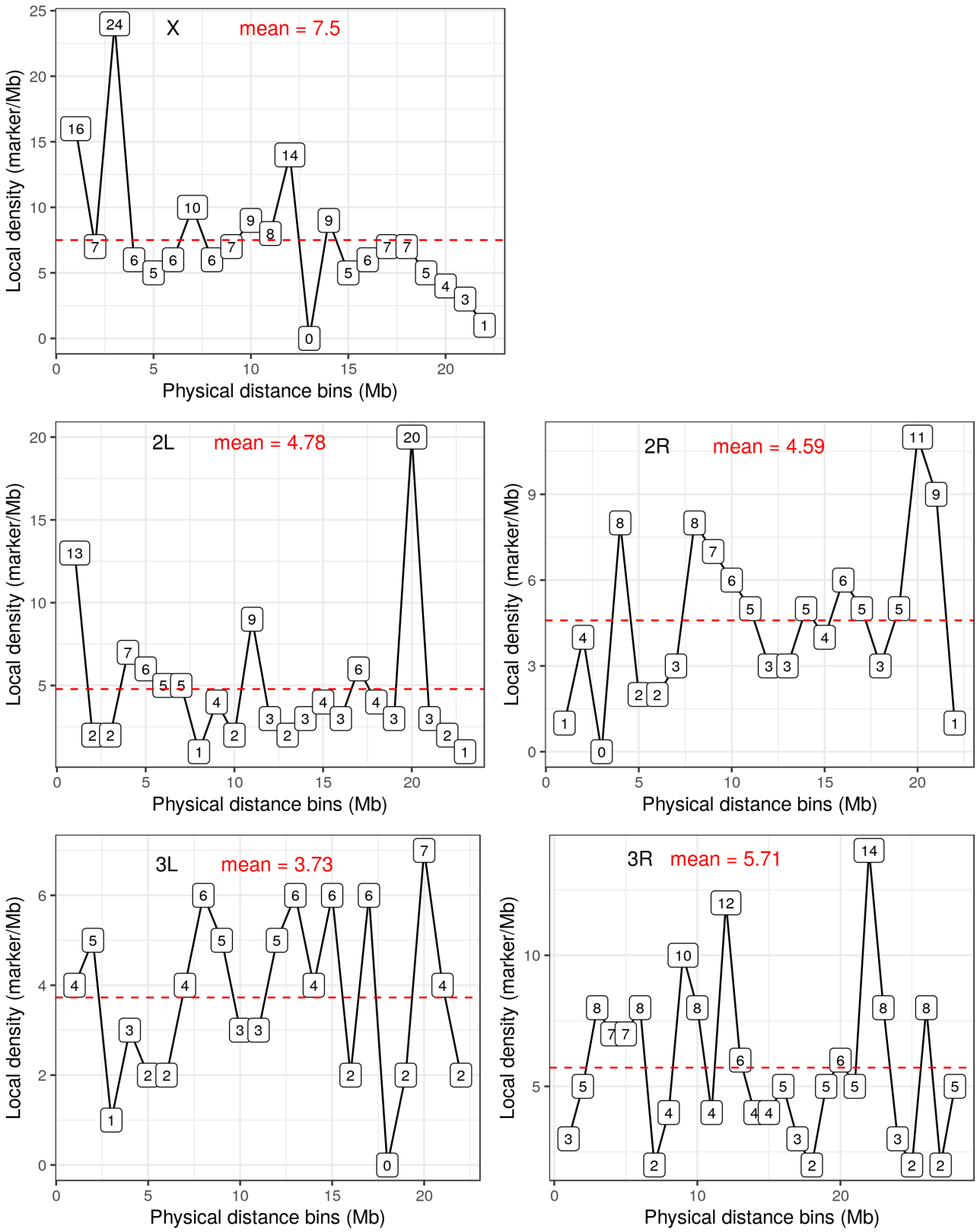
Figure S4: **Distribution simulations.** BREC results on the simulated chromosomes with different scenarios of markers distribution around heterochromatin regions, as presented in the table (top) . Plots (right after) are presenting the corresponding results for each simulation scenario. On the left, (a, c, e) show the cases with the existence of centromeric gap while the ones on the right (b, d, f) show the cases with no centromeric gap. From top to bottom, cases (a) and (b) show a uniform distributions while (c) to (f) are for non uniform distributions. Cases (c) and (d) show a higher density of markers around heterochromatin regions while cases (e) and (f) show a lower density on the same regions. Black dots represent genetic markers. Vertical lines represent heterochromatin boundaries for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). The rug plot, added on the x axis, shows more clearly the variation in markers density as well as the existence or not of the centromeric gap.

Simulation Data scenarios	Centromeric gap exists (lack of markers)		Markers density on heterochromatic regions
	Yes	No	
Uniform distribution	(a)	(b)	Unvariant
Non uniform distribution	(c)	(d)	High
	(e)	(f)	Low



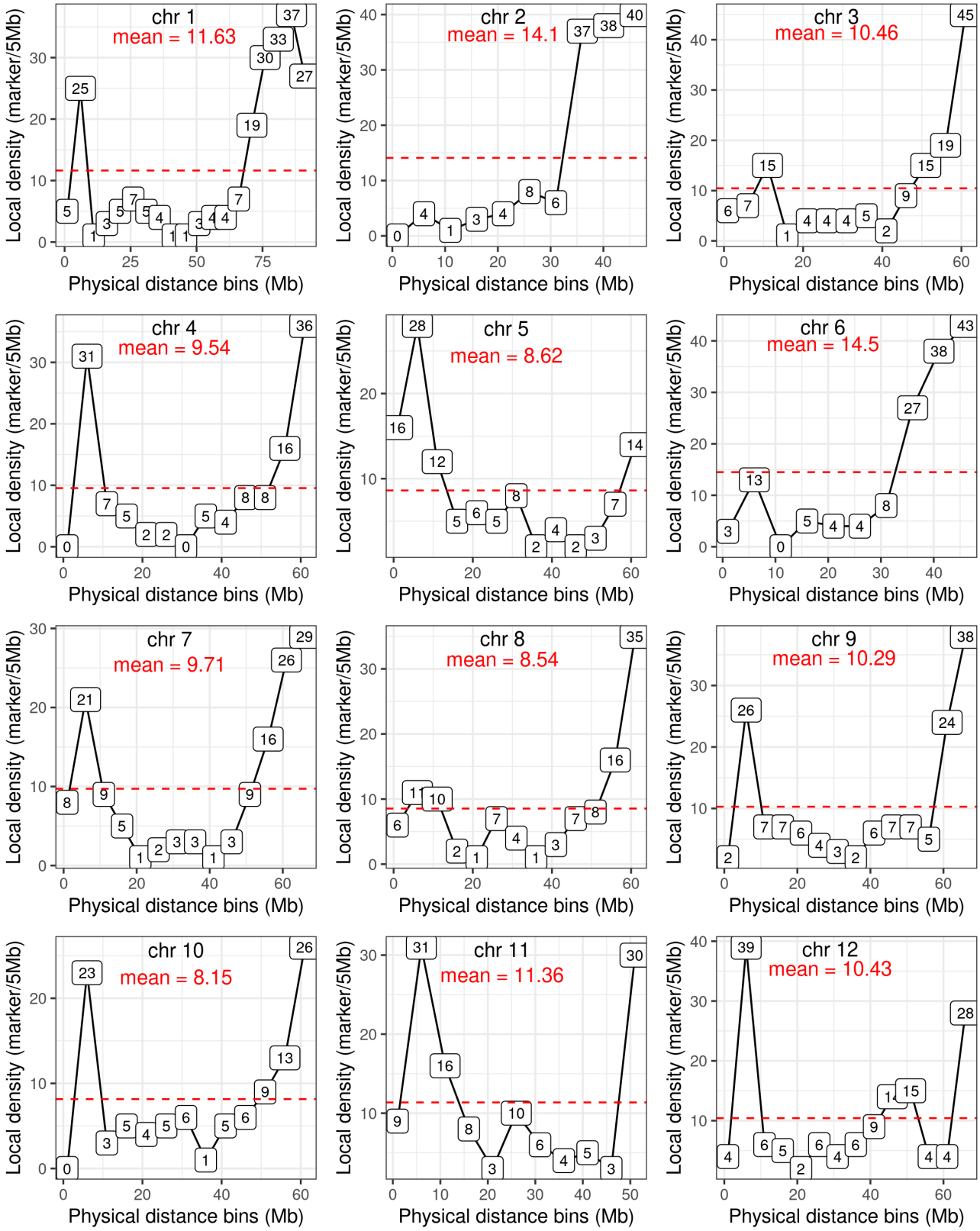
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S5: **Variations of markers local density per 1-Mb bins along *D. melanogaster* Release 5 chromosomal arms.** The red dashed line indicates the mean and represents the global density. Each bin indicates the number of markers it contains. Local density values are represented within the little boxes.



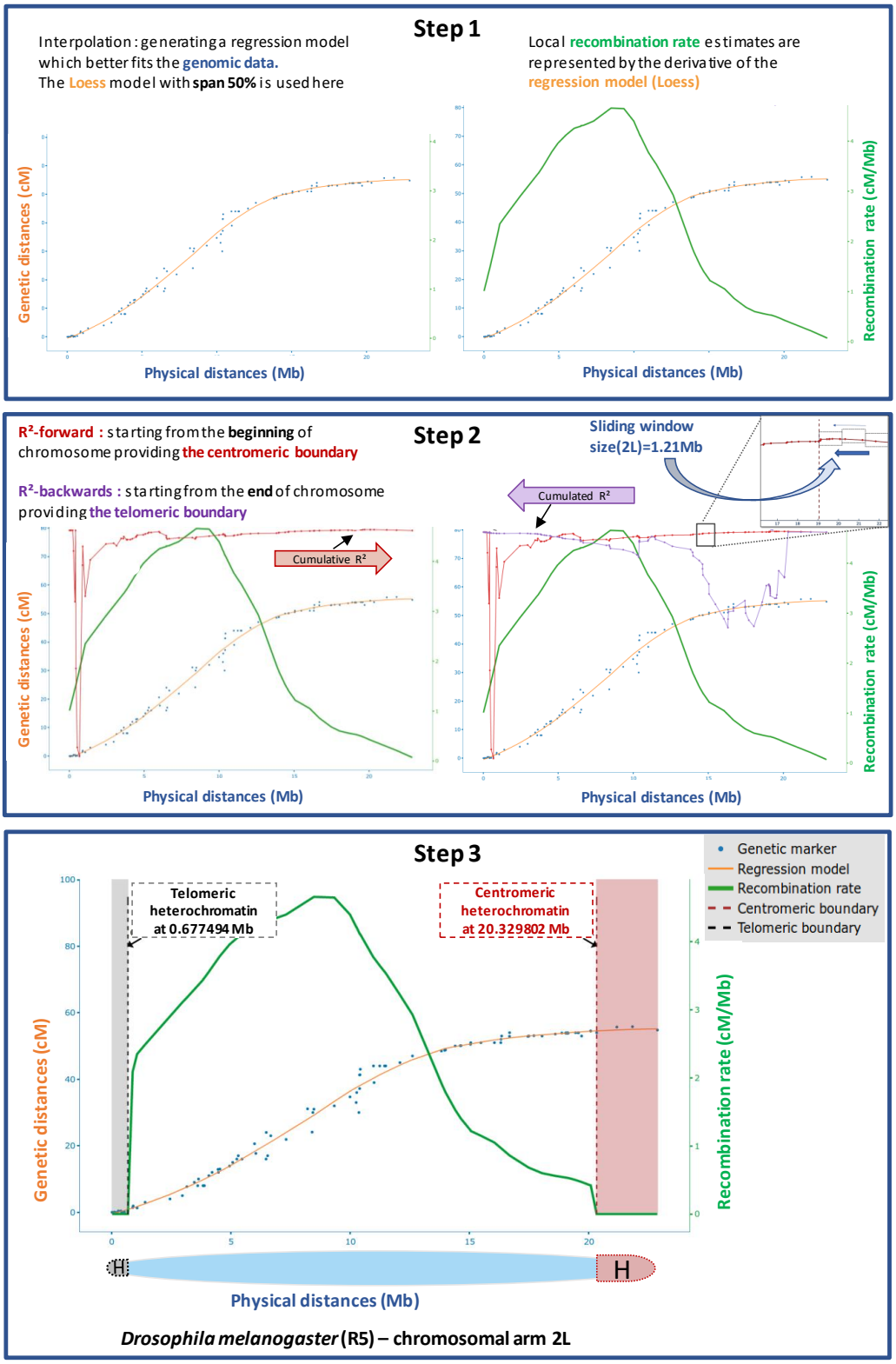
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S6: **Variations of markers local density per 5-Mb bins along the tomato genome *S. lycopersicum* 12 chromosomes.** The red dashed line indicates the mean and represents the global density. Each bin indicates the number of markers it contains. Local density values are represented within the little boxes.



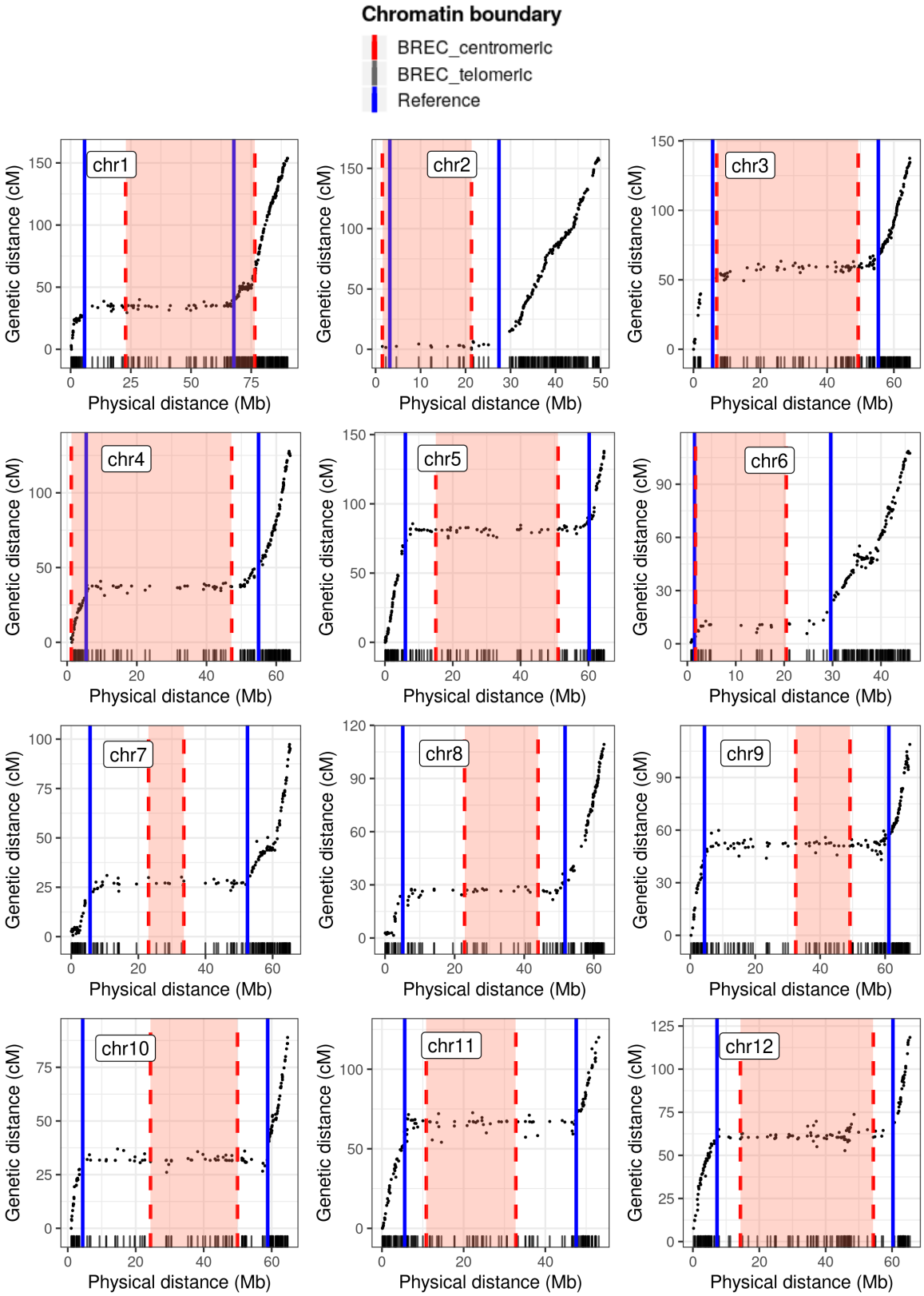
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S7: **BREC pipeline steps applied on chromosome 2L of *D. melanogaster* Release 5.** On each plot, the x-axis represents physical distances (Mb). The left y-axis represents genetic distances (cM) shared between markers (blue data points) and the regression model (orange line). The right y-axis represents recombination rates (cM/Mb) for local estimates (green line). R^2 values, varying between zero and one, are following $R^2 - forward$ (red line) and $R^2 - backwards$ (purple line). Left telomere and Right centromere (resp. black and purple dashed lines) indicate heterochromatin boundaries for the corresponding identified heterochromatin region.



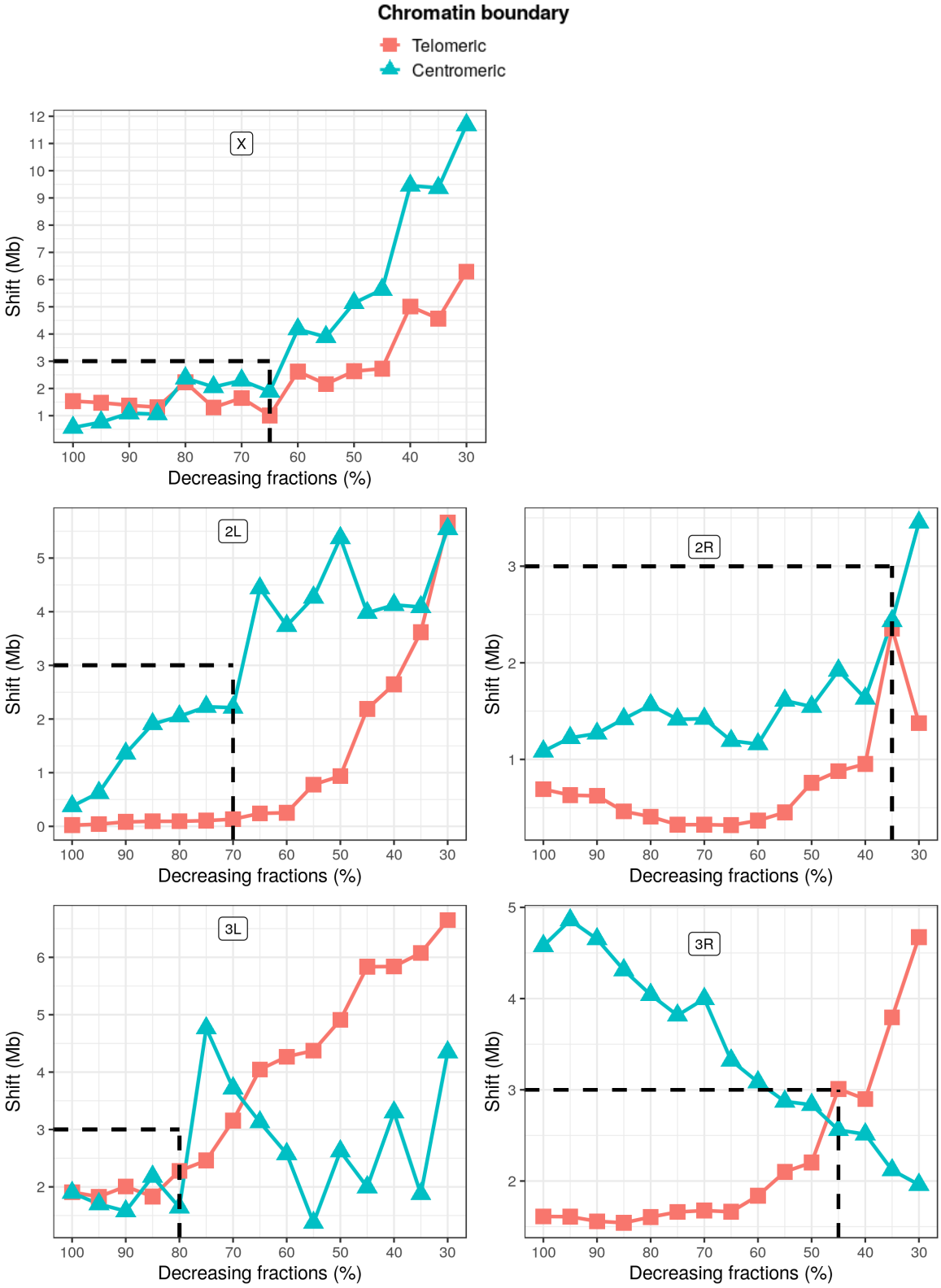
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S8: **Plots of the BREC heterochromatin boundaries and reference boundaries from the *S. lycopersicum* genome**. The results are summarized in Table 2. From top to bottom are the twelve chromosomes 1 to 12, respectively. Black dots represent genetic markers in ascendant order according to their physical position (in Mb). Vertical lines represent heterochromatin boundaries for BREC centromeres (in red dashed line), and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red). Rug plot on the x-axis represents the markers density according to the physical map.



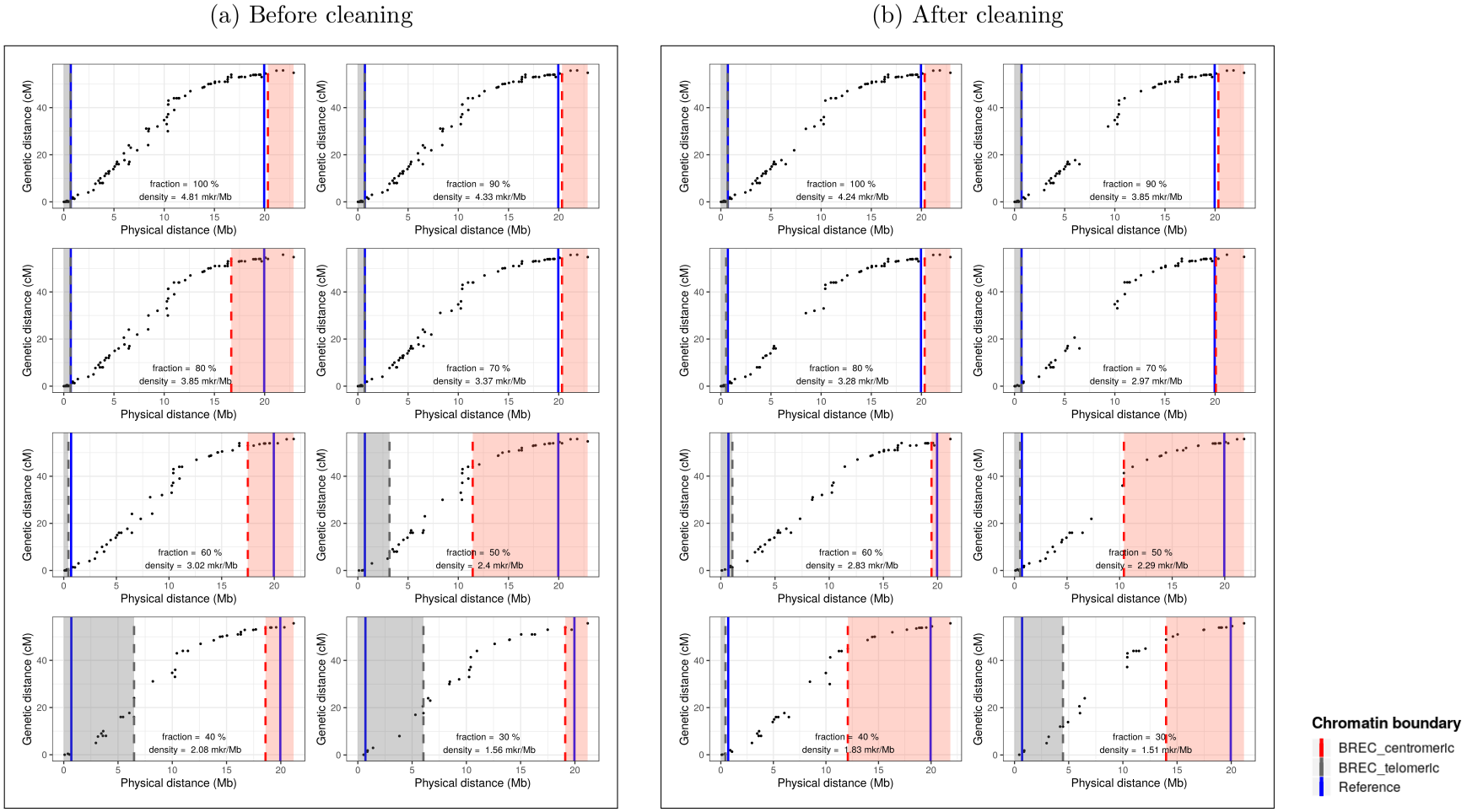
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S9: **The impact of decreasing markers density on the resolution of BREC's heterochromatin boundaries expressed by the shift value.** Here is an overview of the variation of shift values (see Equation 2) for BREC's heterochromatin boundaries compared to reference results for the five *D. melanogaster* chromosomal arms (X, 2L, 2R, 3L, 3R). For each arm, two heterochromatin boundaries are shown: squares (in red) for telomeric and triangles (in light blue) for centromeric boundaries. The horizontal dashed line (in black) delimits results smaller than a shift value of 3Mb for all arms while the vertical dashed line (in black) indicates up to which fraction the 3Mb shift is conserved on each chromosomal arm's simulations. Note that the x axis is reversed, so from left to right it goes from 100% to 30% with a step of -5%at each point. The simulation process is further clarified for one fraction on the chromosomal arm 2L and is illustrated in Figure S10.



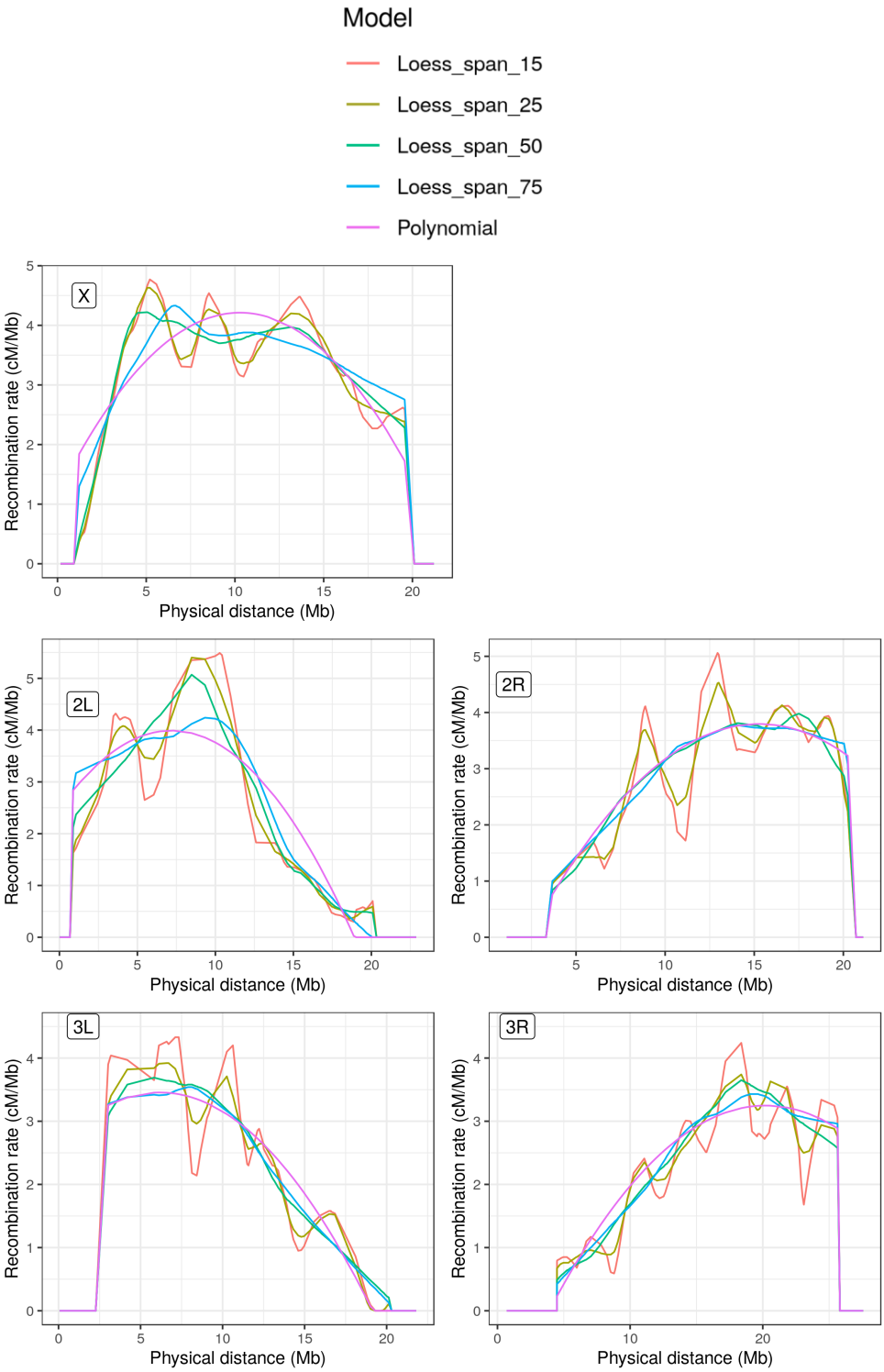
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S10: **Low density simulations** BREC results on the simulated telocentric chromosomes with different density scenarios. Simulating decreasing markers density going from 100% to 30% of the original chromosome 2L (of size 23Mb) of the *D. melanogaster* Release 5 genome. These simulations allow to study the impact of variable markers density on BREC results compared to reference heterochromatin boundaries. (a) on the left is before and (b) on the right is after the cleaning step. These simulations have been conducted on each of the five chromosomes (X, 2L, 2R, 3L, 3R) 30 times where the mean shift value is reported in Figure S9. Black dots represent genetic markers. Vertical lines represent heterochromatin boundaries for BREC centromeres (in red dashed line), for BREC telomeres (in grey dashed line) and for the reference (in solid blue line). The heterochromatin regions identified by BREC are highlighted for the centromere (in red) and the telomere (in grey). The corresponding fraction and markers density is shown on the top left of each simulation plot.



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

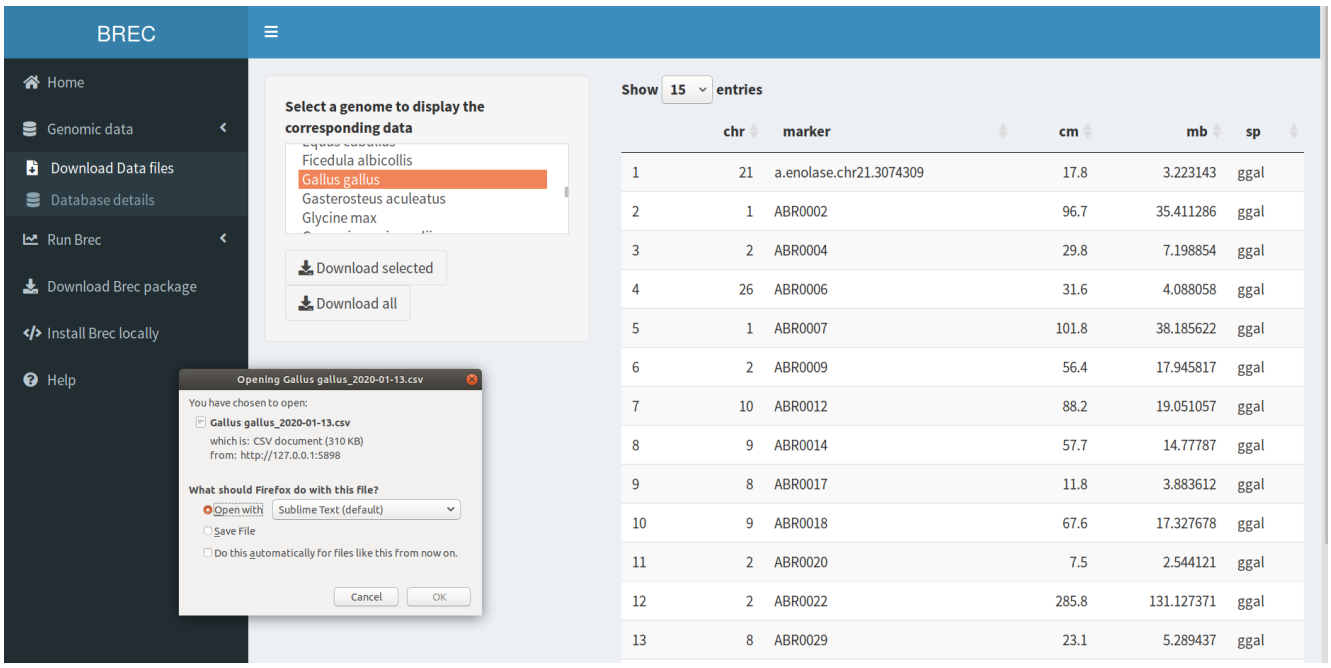
Figure S11: **Comparison of regression models for recombination rate estimates along the five chromosomes (X, 2L, 2R, 3L, 3R) of *D. melanogaster* Release 5**. Regression models used here are Loess with span values, 15%, 25%, 50%, 75% and third degree polynomial. The heterochromatin boundaries defined by BREC remain unchanged and only local recombination rates differ according to the model used to fit the genetic and physical maps. Recombination rate is represented by the derivative of the model. In case of two or more models yielding the same recombination rate estimates on the same physical position, the overlap results in only one curve line. Here, all curves show null recombination rate value on the centromeric and telomeric regions.



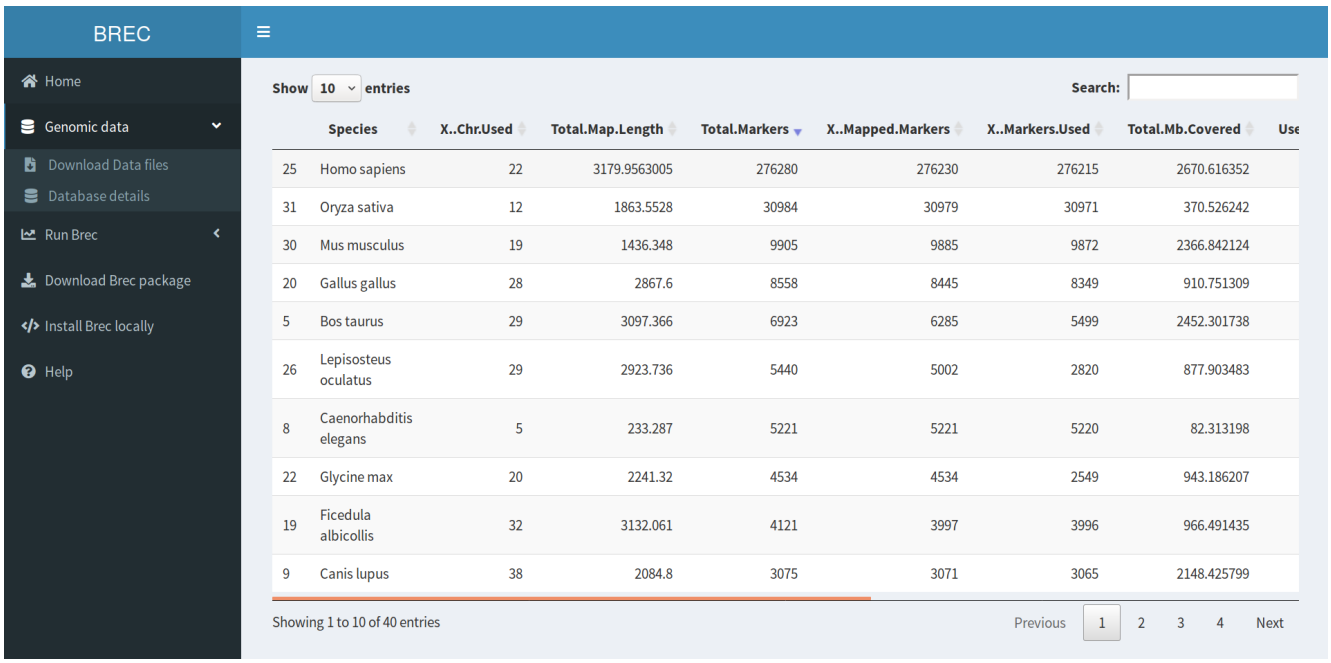
bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S12: **Screenshots of BREC web application - Genomic data web pages.**

(a) Download data files page from the Genomic data section, indicated on the left dark panel, is displayed Here, After selecting on the top list the *Gallus gallus* genome and clicking the "Download selected" button, a dialog box is open waiting for the user to specify the file path to save the selected data file.



(b) Dataset details page from the Genomic data section is showing a sample of ten available genomes provided within the BREC package. The table is intentionally sorted using the forth column values with descending number of "Total markers".



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Description of main components of the Shiny app

Build-in dataset Users can either run BREC on a dataset of 40 genomes, mainly imported from (Corbett-Detig *et al.*, 2015), enriched with two mosquito genomes from (Dudchenko *et al.*, 2017) and updated with *D. melanogaster* Release 6 from FlyBase (Thurmond *et al.*, 2019) (see Table S3), already available within the package, or, load new genomes data according to their own interest.

User-specific genomic data should be provided as inputs within at least a 3-column CSV/TXT file format including for each marker: chromosome identifier, genetic distance and physical distance respectively. On the other hand, outputs from BREC running results are mainly represented *via* interactive plots.

GUI inputs The BREC shiny interface provides the user with a set of options to select as parameters for a given dataset (see figure 4a). These options are mainly necessary in case the user works on his/her own dataset and this way the appropriate parameters would be available to choose from. First, a tab to specify the running mode (one chromosome). Then, a radio button group to choose the dataset source (existing within BREC or importing new dataset). For the existing datasets case, there is a drop-down scrolling list to select one of the available genomes (over 40 options), a second one for the corresponding physical map unit (Mb or pb) and a third one for the chromosome ID (available based on the dataset and not the genome biologically speaking). While for the import new dataset case, three more objects are added (see Figure 4b); a fileInput to select csv or txt data file, a textInput to enter the genome name (optional), and a drop-down scrolling list to select the data separator (comma , semicolon or tab character -set as the default-). As for the Loess regression model, the span parameter is required. It represents the percentage of how many markers to include in the local smoothing process. There is a numericInput object set by default at value 15% with an indication about the range of the span values allowed (min = 5%, max = 100%, step = 5%). The user should keep in mind that the span value actually goes from zero to one, yet, in a matter of simplification, BREC handles the conversion on it's own. Thus, for example, a value of zero basically means that no markers are used for the local smoothing process by Loess, and so, it will induce a running error. Lastly, there is a checkbox to apply data cleaning if checked. Otherwise, the cleaning step will be skipped. This options could save the user some running time if s/he already have a priori knowledge that a specific genome's dataset has already been manually curated). The user is then all set to hit the Run button. BREC will start processing the chromosome of interest by identifying its type (telocentric or atelocentric). Since this step is quite difficult to automatically get the correct result, the user might be invited to interfere *via* a popup alert asking for a chromosome type confirmation (see Figure 4b). As shown in Figure (S12a), all available genomes could be accessed from the left-hand panel (in dark grey) and specifically on the tab "Genomic data" where two pages are available: "Download data files" which provides a data table corresponding to the selected genome on a scrolling list along with download buttons, and "Dataset details" displaying a more global overview of the whole build-in aata repository (see Figure S12b). To give a glance at the GUI outputs, Figure 4c shows BREC results displayed within an interactive plot where the user will have the an interesting experience by hovering over the different plot lines and points, visualising markers labels, zooming in and out, saving a snapshot as a PNG image file, and many more available options thanks to the plotly package (Sievert, 2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure S13: **Download, install and launch BREC.** Code chunk showing the R commands allowing to download, install and run the BREC shiny application. The entire R package is available with open access on the indicated GitHub repository.

```
# Install devtools and shiny from CRAN
install.packages("devtools", "shiny")
# Load installed libraries
library(devtools, shiny)


# Download and install the BREC package from the GitHub repository
install_github("ymansour21/BREC")
# Load Brec and shiny
library(Brec)
library(shiny)
# Launch Brec graphical interface in your default internet browser
runApp("shinyApp/Brec_dashboard.R", launch.browser = TRUE)
```

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S1: **Genomic features and BREC running time for the *D. melanogaster* Release 5 genome**. The first five rows represent chromosomal arms. Columns represent the genome features as follows: (1) the names of chromosomal arms X, 2L, 2R, 3L, and 3R; (2) the markers number included in the study; (3) the markers density (in markers/Mb); and (4) the physical map length (in Mb). The last row summarizes the same features for the whole genome.

Chromosomal arms	Markers number	Markers density (marker/Mb)	Physical map length (Mb)	Genetic map length (cM)	BREC run time (sec)
X	165	7.78	21.22	65.8	1.278
2L	110	4.81	22.88	54.8	0.949
2R	101	4.78	21.12	52.5	0.821
3L	82	3.56	21.81	45.9	0.916
3R	160	5.80	27.57	57.5	1.379
Whole genome	618	5.39	114.59	276.5	5.343

Table S2: **Genomic features and BREC running time for the *S. lycopersicum*** . The first twelve rows represent chromosomes. Columns represent the genome features as follows: (1) the identifiers of chromosomes 1 to 12; (2) the markers number included in the study; (3) the markers density (in markers/Mb); (4) the physical map length (in Mb); and the elapsed time when running BREC. The last row summarizes the same features for the whole genome.

Chromosomes	Markers number	Markers density (marker/Mb)	Physical map length (Mb)	Genetic map length (cM)	BREC run time (sec)
1	232	2.58	89.85	150.72	2.164
2	176	3.66	48.10	154.58	1.391
3	184	2.84	64.77	134.52	1.434
4	160	2.55	62.79	122.64	1.295
5	150	2.32	64.52	137.91	1.098
6	151	3.34	45.20	106.63	1.197
7	145	2.22	65.18	92.48	1.102
8	144	2.29	62.87	106.63	1.047
9	171	2.54	67.37	108.90	1.357
10	148	2.32	63.66	88.92	1.095
11	142	2.68	52.98	119.99	1.081
12	154	2.36	65.18	110.72	1.221
Whole genome	1957	2.64	752.47	1434.49	15.479

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S3: **BREC's built-in dataset of genomic data.** The available genetic and physical maps for 40 species from (Corbett-Detig *et al.*, 2015), enriched with two recently assembled mosquito genomes: *Culex pipiens* and *Aedes aegypti* from (Dudchenko *et al.*, 2017), domesticated tomato *S. lycopersicum* from (Sato *et al.*, 2012), and *D. melanogaster* Release 6 (update) from FlyBase (Thurmond *et al.*, 2019). The species in red bold text are the one we use in BREC experiments. Since the data collection process is still ongoing, the current version of this dataset is continuously evolving.

Species	**Common Name**	**Kingdom and Subgroup**
Aedes aegypti	Yellow fever mosquito	Animal Invertebrate
Anopheles gambiae	African malaria mosquito	
Apis mellifera scutellata	Honeybee	
Bombyx mandarina	Silkworm	
Caenorhabditis briggsae	Roundworm	
Caenorhabditis elegans	Roundworm	
Culex pipiens	Common house mosquito	
Drosophila melanogaster R5	Fruit fly	
Drosophila melanogaster R6	Fruit fly	
Drosophila pseudoobscura	Fruit fly	
Heliconius melpomene melpomene	Postman butterfly	
Bos taurus	Cow	Vertebrate
Canis lupus	Wolf	
Cynoglossus semilaevis	Tongue sole	
Danio rerio	Zebrafish	
Equus ferus przewalskii	Prewalksii's horse	
Ficedula albicollis	Collared flycatcher	
Gallus gallus	Chicken	
Gasterosteus aculeatus	Stickleback	
Homo sapiens	Human	
Lepisosteus oculatus	Spotted gar	
Macaca mulatta	Rhesus macaque	
Meleagris gallopavo	Turkey	
Mus musculus castaneus	House mouse	
Oryzias latipes	Medaka	
Ovis canadensis	Bighorn sheep	
Papio anubis	Olive baboon	
Sus scrofa	Wild boar	
Citrus reticulata	Mandarin Orange	Plant Woody
Gossypium raimondii	New world cotton	
Populus trichocarpa	Black cottonwood	
Prunus davidiana	David's peach	
Arabidopsis thaliana	Thale cress	Plant Herbaceous
Brachypodium distachyon	Purple false brome	
Capsella rubella	Pink Shepherd's Purse	
Citrullus lanatus lanatus	Watermellon	
Cucumis sativus var. hardwickii	Cucumber	
Glycine soja	Wild soybean	
Medicago truncatula	Barrel medic	
Oryza rufipogon	Wild rice	
Setaria italica	Foxtail millet	
Sorghum bicolor subsp. verticilliflorum	Wild Sudan grass	
Solanum lycopersicum	Domesticated tomato	
Zea mays ssp parviglumis	Teosinte	

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.29.178095; this version posted June 30, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table S4: **Comparing BREC with similar widely used tools.** BREC's provided features and functionalities are compared along with the Recombination Rate Estimator (Fiston-Lavier *et al.*, 2010) and the MareyMapOnline (Siberchicot *et al.*, 2017), following a chronological order (the oldest first).

Features / Tool		**Recombination Rate Calculator (RRC)**	**MareyMap Online**	**BREC**
Publication year		2010	2017	2020
Genome-specific		D. melanogaster	all	all
interpolation method	Polynomial	yes	yes	no
	Loess	no	yes	yes
	Cubic spline	no	yes	no
Data cleaning		no	manually	automatically
Data Quality Control		no	no	yes
Chromatin boundaries identification		no	no	yes
Software	R package	no	yes	yes
	Web-based GUI	Perl CGI	Shiny	Shiny