Detection of large scale inversion location patterns with ordinal logistic regression

Ruth Gómez Graciani

First, we obtain the density distribution, and local minima and maxima for the CEU Spence recombination map (1).

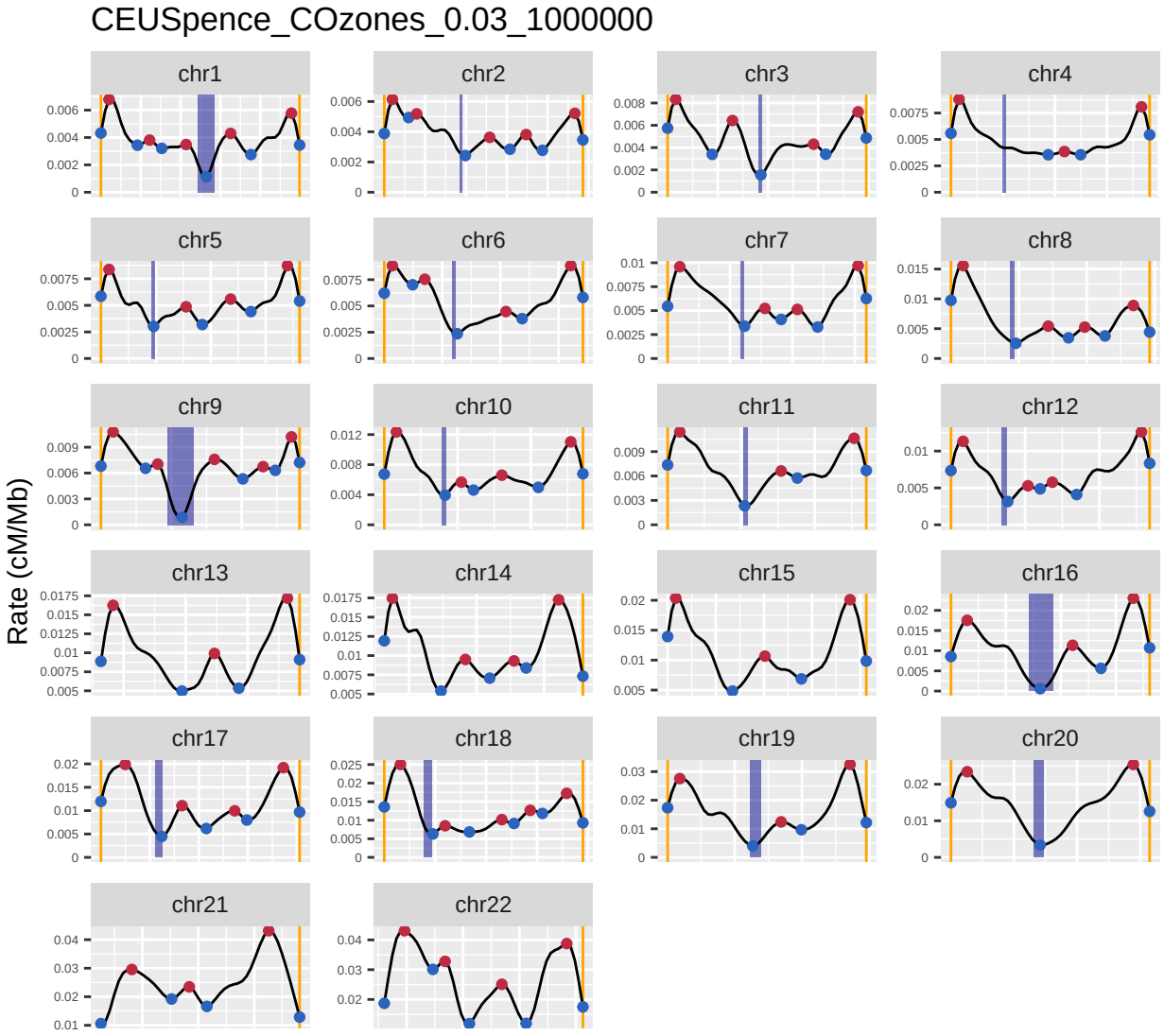

Figure 1: Black line is crossover density, blue and red points are local minima ans maxima respectively, centromeres in blue, chromosome limits in orange. Each chromosome has its own size and recombination scales. Telocentric chromosomes do not include p arm and centromere.

Next, we calculate telomeric and centromeric regions and divide the genome into windows accordingly (2). Telomeric regions are calculated from the extremes towards the centromere, starting at a chromosome limit and ending at the center point between the first local maximum and the next local minimum inwards. Centromeric regions are calculated from the centromere limits towards the extremes, starting at the centromere limit and ending at the center point between de limit and the next local maximum outwards. Centromeres are discarded because they tend to contain less reliable recombination estimates. The remaining spaces between telomeric and centromeric regions are divided into 2 windows and marked as arm regions. This last step is necessary, as arm regions are significantly larger than centromeric or telomeric regions if considered as a whole, which means that the levels of noise between the different chromosome regions would be different and could impact the result. In addition, this eliminates all the size outliers from the general distribution without actually deleting data (3).

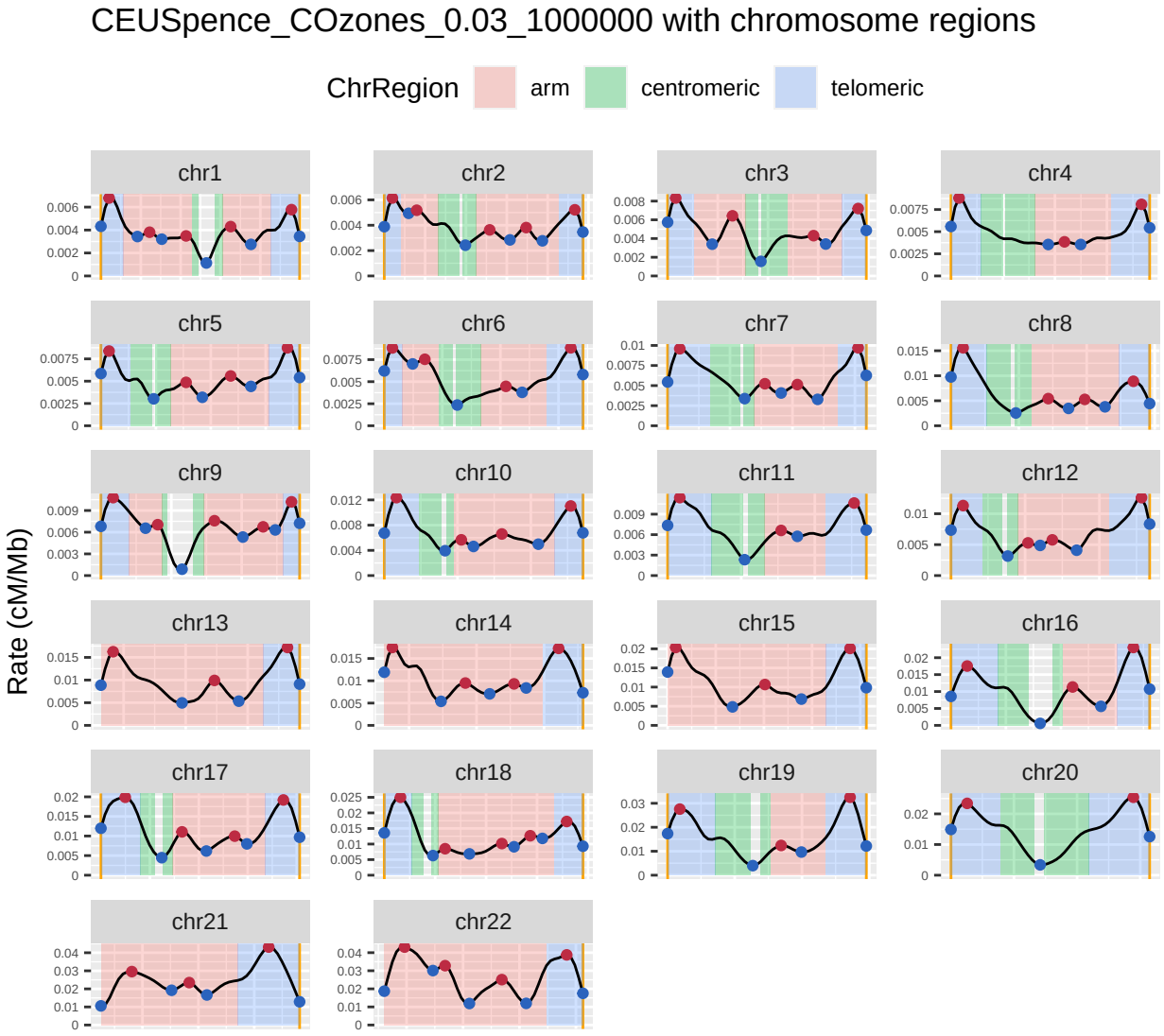

Figure 2: ChrRegion-coded windows for telomeric, centromeric and arm categories. Black line is crossover density, blue and red points are local minima ans maxima respectively, chromosome limits in orange. Each chromosome has its own size and recombination scales. Telocentric chromosomes do not include p arm and centromere.

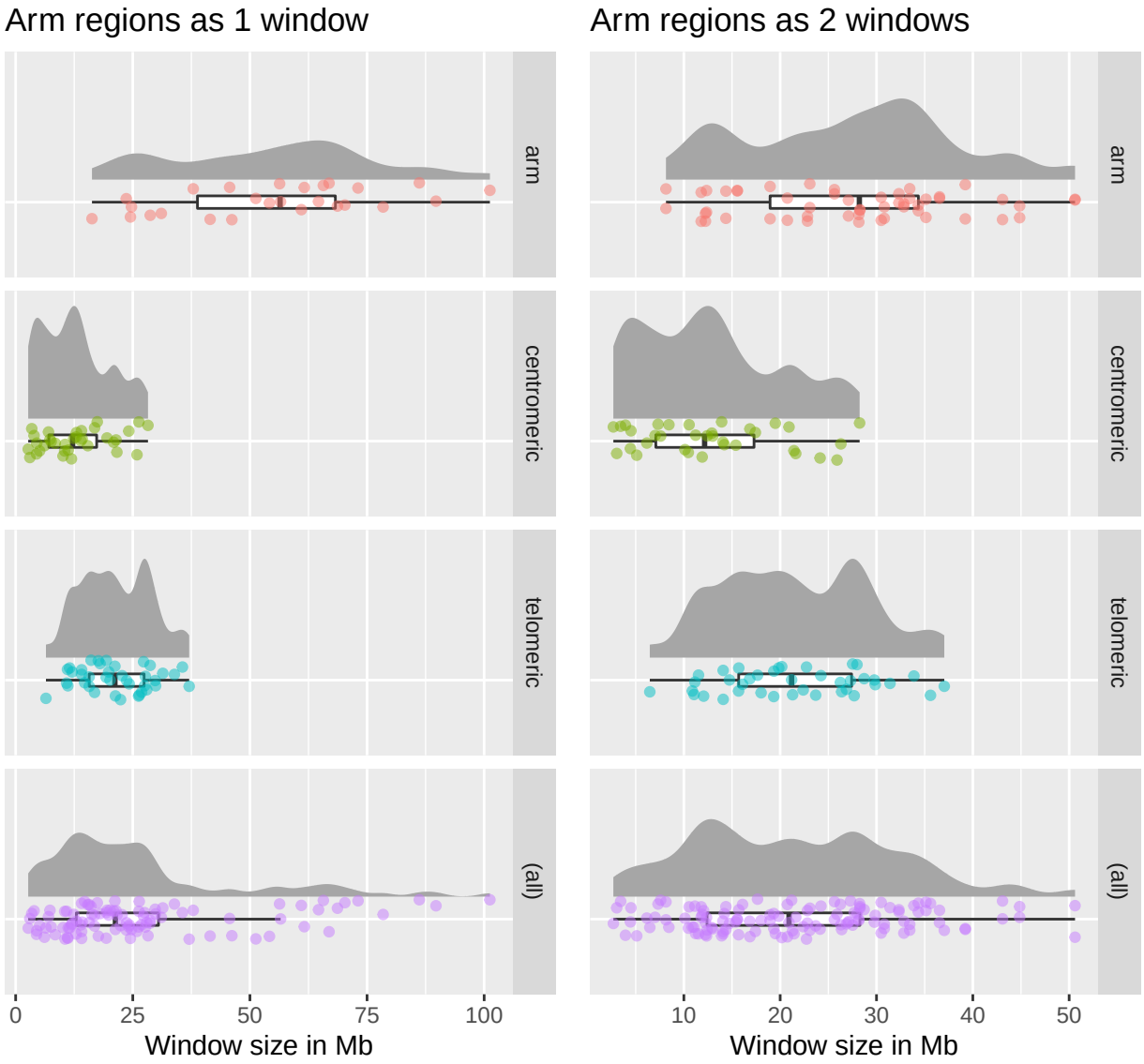

Figure 3: Raincloud plot for size distribution of windows depending on the treatment of arm regions. When arm regions are divided into 2 windows, the shape of the arm size distribution remains the same while the average is not significantly different from centromeric and telomeric regions. All outliers are removed from the general distribution without actually losing data.