ArmReject - improving compartment predictions

Team WB1: Barbara Jurzysta, Aleksandra Możwiłło, Natalia Rutecka

1 Method

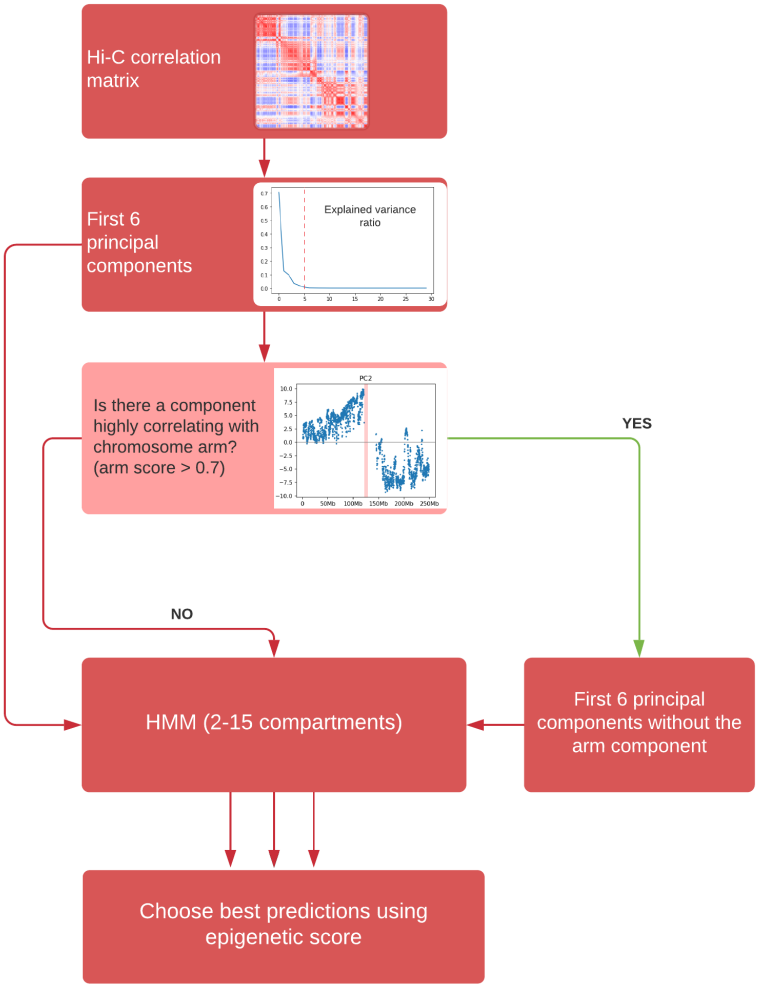

Figure 1: Workflow of ArmReject

We calculated correlation matrices on HiC data for each chromosome from 5 cell lines (GM12878, NHEK, IMR90, HUVEC and HMEC) and extracted the first 6 principal components. The choice of 6 components was based on percentage of variability explained by each PC (see Figure 1). We noticed that often there is a component, denoted the arm component, that is mostly positive on one chromosome arm and negative on the other. Since there seems to be no significant difference between the arms

in terms of epigenetic marks, we removed the arm component and trained Gaussian Hidden Markov Model on the remaining 5 components. We compared the results to results of training HMM on all 6 components. During training we used numbers of states ranging from 2 to 15, obtaining predictions for 2-15 compartments. In case of non-centromere-centric chromosomes and chromosomes without a clear arm component, we didn't remove any PC. Both the choice of the better method and the optimal number of compartments was done using self-developed score, called epi score. The aim of the epi score was to find the partition with highest chance of compartments having distinct biological properties. We did that by maximising mean difference in epigenetic marks distribution between the pairs of compartments. Finally, we accumulated the information from all chromosomes of a given cell line and tried to find one number of compartments that worked for all of them. Here we simply chose the number of compartments with the highest sum of epi scores among the chromosomes.

2 Results

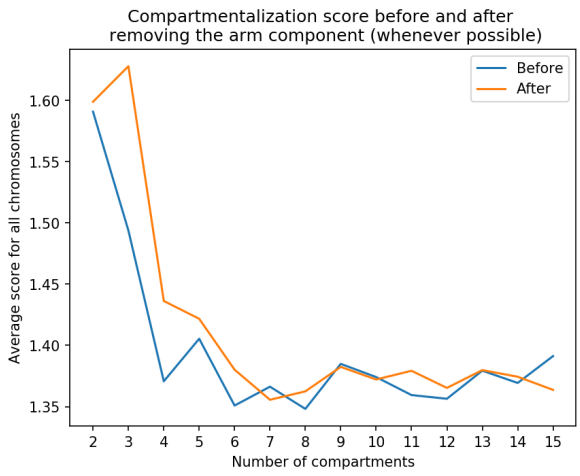

Figure 2: Average epi score across all chromosomes for prediction of 2-15 compartments on GM12878 cell line. Blue: HMM on 6 PC, orange: HMM on 5 PC (without the arm component)

We developed a new method, called ArmReject, based on filtering the noise introduced by position on chromosome arm. We found that in most cases it improved the compartment prediction, resulting in compartments with more distinct composition of epigenetic marks. The estimated optimal numbers of compartments was equal to 2 in 4 out of 5 analysed cell lines (NHEK, IMR90, HUVEC and HMEC) and 3 in GM12878 cell line. Notably, in GM12878 cell line the inferred optimal number of compartments changed after applying our component rejection method and in NHEK, IMR90 and HMEC it improved the epi score for the best number of compartments. However, in case of HUVEC cell line removing the arm component yielded prediction with lower score. Therefore, we suggest using both methods - with and without filtering and choosing the prediction with higher epi score. The combined runtime of both methods on a single cell line should not exceed 15 minutes.