Seminar Neuere Methoden 31.08.17:
Basic4CSim: Comparison of 4C-seq pipelines based on real and simulated data
Carolin Walter, Institut für Medizinische Informatik

Introduction

Circular chromosome conformation capture combined with high-throughput sequencing (4C-seq) is a powerful 3C-based method that provides information on chromatin organization. Since interpreting the raw 4C-seq data is difficult due to its inherent semi-quantitative fragmented structure and possible technical biases, special care has to be taken during the analysis [1]. Different algorithms based on different strategies have been published, but an unbiased benchmarking study that allows for comparisons between the programs is still missing.

Methods

We present Basic4CSim, an R-based simulator capable of generating single-sample and replicate 4C-seq data. Based on published experimental data sets from various species, e.g. Stadhouders' murine data [7] or Rodrigues' chicken experiments [8], Basic4CSim can create the typical fragment structure and characteristic background read distribution found in 4C-seq experiments. Different types of interactions can be simulated, with their structure adapted to those of validated biological interaction data, and the introduction of varying levels of background noise is possible. The degree of similarity for simulated samples can be customized by specifying the allowed variance between simulated interaction strengths and likelihoods, and the introduction of noise levels. Additionally, quality controls (e.g. read distributions, general density, viewpoint region coverage) and visualization routines for the viewpoint region, trend lines for multiple samples or difference plots offer further insight into the data. Scale-space visualisation, a multi-scale visualisation technique, shows the main features of a chosen 4C-seq sample, and allows the comparison of the structure of different samples, thus presenting a new metric to assess the quality of a simulated sample's near-cis region.
We combine simulated data sets and published data with validated interaction sites to create a substantial base for assessing the precision and recall of five 4C-seq pipelines, from Splinter's algorithm to Raviram's 4C-ker [2,3,4,5,6].

Results and discussion

Our benchmarking shows the effect of different parameters on detected candidate interactions, with 'window size' having the most notable effect, and unveils the effects of filter options, identifying a heuristic evaluation step featured in one algorithm as especially effective. The results underline the importance of filter steps and sensible parameter choices.

References

[1] van de Werken et al. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions, Nature Methods, 9, 969-972.
[2] Splinter, E. et al. (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation, Methods, 58, 221-230.
[3] Thongjuea et al. (2013) r3Cseq: an R/Bioconductor package for the discovery of long-range genomic interactions from chromosome conformation capture and next-generation sequencing data, Nucleid Acids Res., e132, 1-12.
[4] Williams, R. L. et al. (2014) fourSig: a method for determining chromosomal interactions in 4C-seq data, Nucleid Acids Res., e68, 1-16.
[5] Klein et al. (2015) FourCSeq: analysis of 4C sequencing data. *Bioinformatics* 31 (19): 3085-3091
[6] Raviram et al (2016) 4C-ker: a method to reproducibly identify genome-wide interactions captured by 4C-seq experiments, PLoS Comput Biol, 12 (3), e1004780.
[7] Stadhouders et al. (2012) Dynamic long-range chromatin interactions control Myb proto-oncogene transcription during erythroid development., EMBO, 31, 986-999.
[8] Rodrigues et al. (2017) Integration of Shh and Fgf signaling in controlling Hox gene expression in cultured limb cells, PNAS, 114 (12), 3139-3144.