Bioinformatics, 31(19), 2015, 3085–3091
doi: 10.1093/bioinformatics/btv335
Advance Access Publication Date: 1 June 2015
Original Paper

Genome analysis

FourCSeq: analysis of 4C sequencing data

Felix A. Klein*, Tibor Pakozdi, Simon Anders, Yad Ghavi-Helm, Eileen E. M. Furlong and Wolfgang Huber*

European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Meyerhofstr. 1, 69117 Heidelberg, Germany

*To whom correspondence should be addressed.

Associate Editor: Gunnar Ratsch

Received on August 13, 2014; revised on May 8, 2015; accepted on May 26, 2015

Abstract

Motivation: Circularized Chromosome Conformation Capture (4C) is a powerful technique for studying the spatial interactions of a specific genomic region called the 'viewpoint' with the rest of the genome, both in a single condition or comparing different experimental conditions or cell types. Observed ligation frequencies typically show a strong, regular dependence on genomic distance from the viewpoint, on top of which specific interaction peaks are superimposed. Here, we address the computational task to find these specific peaks and to detect changes between different biological conditions.

Results: We model the overall trend of decreasing interaction frequency with genomic distance by fitting a smooth monotonically decreasing function to suitably transformed count data. Based on the fit, *z*-scores are calculated from the residuals, and high *z*-scores are interpreted as peaks providing evidence for specific interactions. To compare different conditions, we normalize fragment counts between samples, and call for differential contact frequencies using the statistical method **DESeq2** adapted from RNA-Seq analysis.

Availability and implementation: A full end-to-end analysis pipeline is implemented in the R package **FourCSeq** available at www.bioconductor.org.

Contact: felix.klein@embl.de or whuber@embl.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Circularized Chromosome Conformation Capture (4C) couples the low-throughput Chromosome Conformation Capture (3C) technique (Dekker *et al.*, 2002) for studying chromatin–chromatin interactions with high-throughput sequencing (Simonis *et al.*, 2006; Stadhouders *et al.*, 2013). 4C detects the contacts of a chosen viewpoint with, in principle, the entire genome. The 4C protocol consists of six main steps (Stadhouders *et al.*, 2013). First, the chromatin is cross-linked with formaldehyde to fix DNA-protein complexes, thereby capturing DNA sequences that are in close spatial proximity. In the next step, the cross-linked chromatin is digested with a restriction enzyme. In the third step, the fragment ends from the digestion treatment are ligated under dilute conditions to favor intra-complex ligation, ligating DNA sequences that have been in close spatial proximity. After this, the cross-linking is reversed, followed by a second round of digestion with a different restriction enzyme to obtain smaller DNA molecules. These molecules are then circularized and amplified by polymerase chain reaction (PCR). The resulting library is sequenced. The possibility to multiplex several viewpoints in one sequencing library further increases the throughput.

As result, the distribution of reads from a 4C sequencing library throughout the genome provides an estimate of the contact frequencies of the viewpoint with the rest of the genome. Overall, the 4C signal decreases with genomic distance from the viewpoint and reaches a constant level of noise for large distances. Specific interactions of DNA elements sit on top of this overall trend. One task is to identify positions that stand out from the general trend. Moreover, if 4C has been performed on samples with different cell types, developmental stages or experimental treatments, a second task is the detection of changes in interaction frequencies between the sample groups.

Several analysis approaches for the first task, detection of interactions, have already been developed for 4C sequencing data. The approach by Thongjuea *et al.* (2013) uses a non-parametric smoothing spline on library-size normalized count data to estimate the signal decrease with distance to the viewpoint and detects interactions by calculating *z*-scores from the residuals of this fit. Another approach, used by van de Werken *et al.* (2012) and Splinter *et al.* (2012) employs two complementary arms: in the proximity of the viewpoint, multi-scale visualization of a semi-quantitative contact map, remote from the viewpoint, an empirically estimated contact background model of binary contact profiles combined with a window-based enrichment and permutation analysis.

Currently, methods are missing that use replicate information as the basis for data-driven error modeling to detect consistent peaks and to statistically infer changes in contact frequencies between different conditions.

We address these needs with the following approach. We use a distance-dependent monotone fit to estimate the signal decay with increasing distance from the viewpoint, since the unspecific component of the signal decreases monotonically. As input to the fit we use variance-stabilized read count data (Anders and Huber, 2010). To detect strong interactions, we calculate *z*-scores from the fit residuals and associated *P*-values. For the comparison of different conditions we use the methods implemented in the **DESeq2** package (Love *et al.*, 2014).

2 Materials and methods

2.1 Data preprocessing

The data processing pipeline (Fig. 1) starts from the reads of the 4C library. If several 4C libraries were multiplexed, the viewpoint primer sequences and, if present, additional barcodes, are used to demultiplex the sample and trim of the primer sequences. For the demultiplexing and trimming of primer sequences a Python script is included in the package. The remaining sequences are aligned to the full reference genome using a standard alignment tool.

The analysis pipeline of our R package starts with the binary alignment/map (BAM) files output from the alignment. The following steps are now described in more detail.

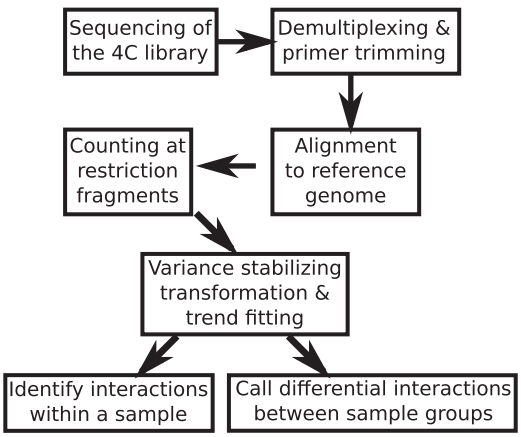

Fig. 1. Overall workflow of steps described in this paper

2.1.1 Cutting the reference genome

The input to the statistical analysis is a count table, with one row for each restriction fragment, and one column for each sample, with the table entries indicating how many reads have been assigned to each restriction fragment in each sample. By restriction fragment, we mean the sequences between the cutting sites of the first restriction enzyme, because this first digestion defines the resolution at which interactions can be seen in 4C. To define fragments, we cut the reference genome *in silico* using the recognition sequence of the first cutter. Fragments are delimited by adjacent cutting sites of the first restriction enzyme. The second restriction enzyme is used to reduce the size of the fragments for efficient circularization and PCR amplification. Correspondingly, fragment ends are defined as the genomic region between the start/end position of the fragment and the closest cutting site of the second enzyme (Fig. 2a).

Because mainly fragments that contain a site for the second cutter are efficiently amplified, a fragment is considered valid if it contains at least one cutting site of the second enzyme and has long enough fragments ends. By default, we use a threshold of 20 nt.

2.1.2 Mapping of primer sequences

The primer sequence of the viewpoint is mapped to the reference genome to find the fragment that contains the viewpoint. This fragment is used to calculate the genomic distance to fragments on the same chromosome. More precisely, we use the genomic distance between the middle of the viewpoint fragment and the middle of the other fragment.

2.1.3 Mapping reads to fragment ends

To filter out non-informative reads, we use the following criteria, motivated by the 4C protocol. Only reads that fulfill the criteria are mapped to a fragment end. A first condition is that reads should start directly at a restriction enzyme cutting site. Additionally, the

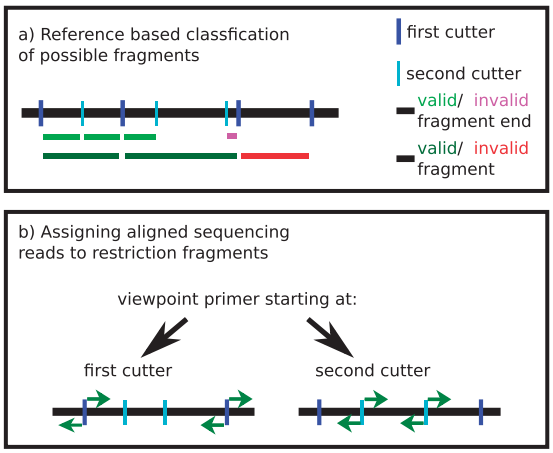

Fig. 2. (**a**) Schematic of the rules to define valid fragments, i.e. fragments that are used subsequently in the analysis. The pink fragment end is smaller than the defined threshold, but since the other fragment end is valid, the fragment is kept for analysis. The red fragment is invalid because it does not contain a cutting site of the second restriction enzyme, and it is removed from the analysis. (**b**) If the sequencing primer starts at the first restriction enzyme cutting site, reads (green arrows) that start at the fragment ends and are oriented toward the fragment middle are kept for analysis. If the sequencing primer starts at the second restriction enzyme cutting site, reads (green arrows) that start right next to the cutting site of the second restriction enzyme and are directed toward the ends of the fragment are kept for analysis

orientation of the read at the fragment end is important and defined by the protocol (Fig. 2b). If the sequencing library was prepared with a primer starting next to a cutting site of the first cutting enzyme, reads should be directed toward the middle of the fragment. If instead the primer starts next to a cutting site of the second cutter, reads should be directed toward the fragment ends. The reads mapped to both fragment ends are combined for subsequent analysis. To check for consistency between replicates, we visualize scatter plots of count values (Fig. 3).

2.2 Detecting interactions

2.2.1 Variance-stabilizing transformation

The count values usually span several orders of magnitude. If a logarithmic transformation were used for the count values, low abundance fragments would tend to show large standard deviations across samples. On the other hand, if untransformed data were used, the standard deviations across samples would be large for high abundance fragments. Such heteroscedasticity would skew the analysis toward either the fragments far from or close to the viewpoint. Therefore, we use the variance-stabilizing transformation v as introduced by Anders and Huber (2010) and implemented in the **DESeq2** package (Love *et al.*, 2014) to transform the count k_{ij} of fragment i in sample j to $v(k_{ij})$. After transformation the standard deviations show less dependence on the fragment abundance (Fig. 4).

2.2.2 Trend fitting

The 4C signal decays with genomic distance from the viewpoint and converges toward a constant level of background. This decay trend $f_j(d_i)$ is fitted using the transformed count values $v(k_{ij})$ as a function of the logarithm of the genomic distance d_i from each fragment i to the viewpoint.

The **FourCSeq** package offers two choices for the distance dependence fit. Using the smooth monotone fit function of the **fda** package (Ramsay *et al.*, 2014), we may choose to assume that the trend is symmetric around the viewpoint and fit a symmetric monotone curve on the combined data from both sides. Alternatively, we perform a monotone fit separately for each side of the viewpoint.

The second option can be useful if one is interested in finding asymmetries in the interaction profiles of a viewpoint, which might be of particular interest at boundaries of topological domains (Dixon *et al.*, 2012). For both methods, we provide standard parameters that work for a wide range of data and which can be adjusted by the user if necessary.

An example of a symmetric monotone fit is shown in Figure 5.

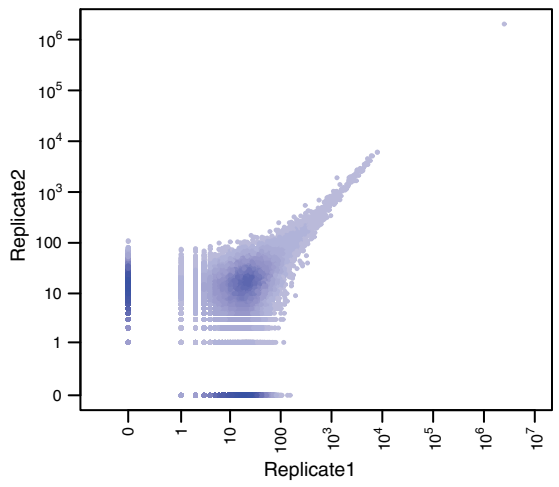

Fig. 3. Correlation between two biological replicates of the *apterous* CRM viewpoint for whole embryo tissue at 6–8 h after fertilization. In the plot, the pairwise distribution of count values per fragment is shown. The *x*- and *y*-axes (drawn in logarithm-like scale, with zero) correspond to the counts for the fragments in two biological replicate libraries for the same viewpoint and biological condition. The replicates show good concordance for higher count values. Fragments with 0 counts for both replicates are not shown

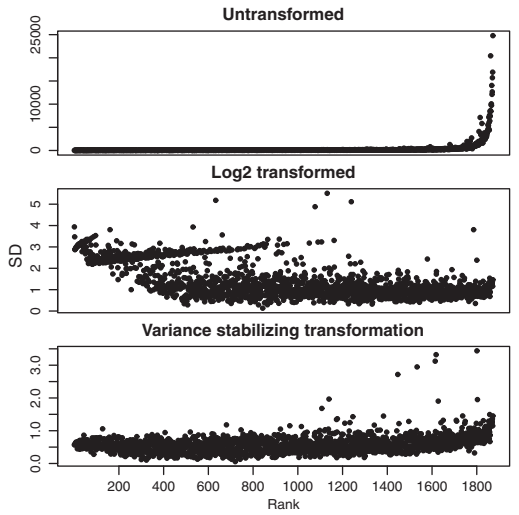

Fig. 4. Variance-stabilizing transformation. For each fragment, the mean and standard deviation of its count data were computed across all samples for the *apterous* CRM viewpoint. The plots visualize the distributions of these values for all fragments. Fragments close to the viewpoint are on the right side with higher count values. When the untransformed count data are considered (upper panel), the standard deviations are very large for high abundance fragments (close to the viewpoint). When the count data are considered on the logarithmic scale (middle panel), the standard deviations are large for low abundance fragments (far from the viewpoint). Both effects would make the analysis overly susceptible to noise either close to or far from the viewpoint. When a variance stabilizing transformation is applied, the standard deviations show less dependence on the fragment abundance, facilitating a more consistent statistical treatment across the whole dynamic range of the data

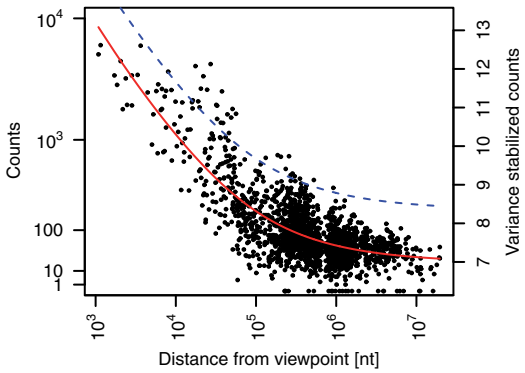

Fig. 5. Symmetric monotone fit of the variance-stabilized count data over the logarithm of the distance from viewpoint for the *apterous* CRM viewpoint. The solid line shows the fit and the dashed line is the fit plus 3σ

2.2.3 *z*-scores of residuals

To find specific interactions, i.e. fragments that show interaction frequencies higher than expected at a given distance from the viewpoint, we calculate *z*-scores from the residuals of the fit:

$$z_{ij} = \frac{v(k_{ij}) - f_j(d_i)}{\sigma_j}, \quad (1)$$

where $\sigma_j = \mathrm{mad}_i(v(k_{ij}) - f_j(d_i))$ is the median absolute deviation (a robust estimator of scale), i runs over all fragments and j over all samples. In principle, users can call specific interactions by looking for large, positive values of the *z*-score. To select the threshold, they can use known positive and likely negative control regions. A potential disadvantage of this approach is that no type I error control in the face of multiple testing is provided. Therefore, by default **FourCSeq** performs the following additional steps. The *z*-scores are converted into one-sided *P*-values using the standard Normal cumulative distribution function, and these are adjusted for multiple testing using the method of Benjamini and Hochberg (1995). In this way, control of the false discovery rate (FDR) is provided. Specific interactions are then found by looking for fragments with small adjusted *P*-values; a large enough value of the effect size, *z* may be an additional requirement (Section 3.2). For the *P*-values to be well-calibrated in this approach, the *z*-scores should follow a standard Normal distribution under the null hypothesis, corresponding to fragments that are not affected by an interaction with the viewpoint. In the data that we examined (Ghavi-Helm *et al.*, 2014), this assumption appeared reasonable; for their own data, users are advised to inspect quantile–quantile plots of z against $N(0, 1)$, and histograms of the unadjusted *P*-values to asses the calibration. Example visualizations are provided in the package vignette.

2.3 Differences between conditions

We have observed the distance dependence of the signal to be variable between samples, and this needs to be taken into account for comparisons. Therefore, we calculate a matrix of normalization factors n_{ij}, such that the scaled read counts $n_{ij}k_{ij}$ for fragment i become comparable across the samples j. Moreover, we need the normalization factors to represent the fitted distance dependence on the scale of the raw counts. Hence, we back-transform the fitted values f_{ij} to the scale of raw counts and scale them to have unit geometric mean across samples to obtain the normalization factors:

$$n_{ij} = \frac{v^{-1}(f_j(d_i))}{\sqrt[J]{\prod_{j=1}^{J} v^{-1}(f_j(d_i))}}, \quad (2)$$

where n_{ij} is the normalization factor and $v^{-1}(f_j(d_i))$ is the back transformed fitted value at the genomic distance d_i. The index i runs over all fragments and j over all samples.

To detect differences between conditions, we apply the methods implemented in the **DESeq2** package to the counts k_{ij} together with the normalization factors n_{ij} (Love *et al.*, 2014). **DESeq2** is a statistical method for differential analysis of count data. Originally established for RNA-Seq (Anders and Huber, 2010), it has in the meanwhile also been shown to be useful for other sequencing-based assays, including ChIP-Seq (Ross-Innes *et al.*, 2012), CLIP-Seq (König *et al.*, 2011) and Hi-C (Pekowska *et al.*, 2014). **DESeq2** allows analysis-of-variance (ANOVA) type analyses—including the simple pairwise comparison between two conditions—with an error model adapted to the technical and biological variability of the data. Technically, it does this via generalized linear models (GLMs) of the Negative Binomial (NB) family. To facilitate the application of NB-GLMs to experiments with small numbers of replicates, **DESeq2** uses an Empirical Bayes method to shrink (and stabilize) its estimates of dispersion and effect size parameters.

For each fragment, a significant interaction change is called when the observed change between conditions is significantly stronger than what is expected from the size of the changes seen between replicates.

3 Results

To illustrate our approach, we use a 4C dataset of developing *Drosophila melanogaster* embryos (Ghavi-Helm *et al.*, 2014). In this dataset, 103 viewpoints were selected throughout the *D. melanogaster* genome, with a focus on *cis*-regulatory modules (CRMs). Samples were taken from embryos at 2–4 h and 6–8 h after fertilization either using whole embryos or mesoderm-specific cells (Ghavi-Helm *et al.*, 2014).

3.1 Preprocessing

Starting from FASTQ, files we used a Python program included in the **FourCSeq** package to demultiplex the libraries and trim off bar codes and adapters. Next, we aligned the reads to the dm3 reference genome with Novoalign (http://www.novocraft.com).

For short restriction fragments, we observed the problem that reads contained the whole fragment and then continued through the cutting site of the second restriction enzyme into the ligated fragments (in most cases the viewpoint fragment). This often resulted in two possible alignments causing the reads to be reported as not uniquely mapping. To address this problem and rescue some of the shorter fragments we checked whether the restriction enzyme cutting site was found within unaligned reads. In such a case the end of the read was trimmed at the restriction enzyme cutting site and alignment was attempted again.

We then generated a fragment reference and mapped the aligned reads to these fragments as described in Sections 2.1.1 and 2.1.3.

For quality control, the percentage of reads mapping to valid fragments from all aligned reads was calculated. For our data this value was around 70–95% in most cases. A value in that range should be obtained for a 4C library. If the percentage is much smaller, the first region that should be investigated is the region around the viewpoint, where a single fragment can pile up a high percentage of the reads. Other possible reasons for low mapping percentages might be reads that map either to invalid fragments, which have been removed from analysis, or to new fragments, created by mutations relative to the reference genome.

To check whether technical and biological replicates gave a similar signal, scatter plots of the replicates were generated. For our dataset, these plots showed good agreement for higher count values in most cases. However, at lower count values, the replicates showed higher relative variation, as is expected from Poisson noise. An example is shown in Figure 3.

3.2 Detecting interactions

First, to reduce noise in the data, we removed fragments that had less than a median of 40 counts across all samples for one viewpoint. Second, we set aside fragments that were too close to the viewpoint, because the high ligation frequencies seen for these fragments tend to obscure any specific signal. The package used a heuristic to identify the boundaries of this masked out region around the viewpoint as those fragments where the observed signal for the first time

increased between successive fragments. The parameters of the variance-stabilizing transformation were fitted on the count values of the remaining fragments (Fig. 4). Next, the decay trend was fitted on the transformed scale using a monotone symmetric fit. The fit is shown in Figure 5. *z*-scores and associated *P*-values were calculated from the fit residuals. Interactions were found by looking for fragments with *z*-scores larger than 3 in both replicates and an adjusted *P*-values smaller than 0.01 in at least one replicate. Figure 7 shows the results for one of the viewpoints in our dataset, which is located in a CRM close to the *apterous* (*ap*) gene. The fragments that showed an interaction are highlighted by red or orange dots.

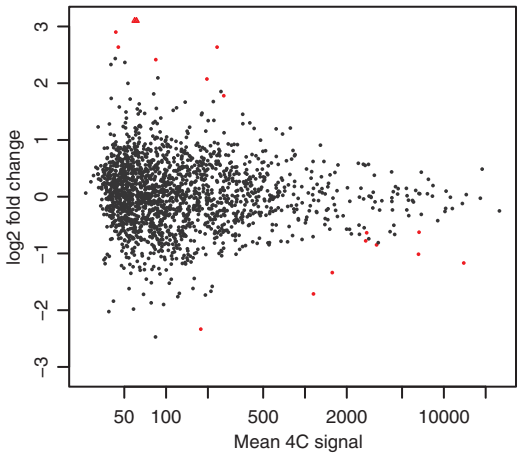

Fig. 6. MA plot of the *apterous* CRM viewpoint 4C profile comparison between *Drosophila* embryo mesoderm tissue and whole embryo 6–8 h after fertilization. The *y*-axis shows the difference between log interaction counts for a given fragment which is plotted against the average log interaction counts per fragment on the *x*-axis. Red dots represent fragments that show differential interactions (*p*-adjusted < 0.01)

In mesoderm specific and whole embryo tissue at 6–8 h after fertilization the interaction of the viewpoint with the *ap* gene promoter on the right side of the viewpoint was captured. Further interactions were found as well, but could not be directly attributed to a specific genomic element. In general, we were able to detect interactions between 10 known enhancer-promoter pairs and many more interactions throughout the set of 103 viewpoint (Ghavi-Helm *et al.*, 2014).

3.3 Differences between conditions

To detect differences between conditions, we used the method described in Section 2.3. Figure 6 shows the MA plot comparing mesoderm tissue and whole embryo for *Drosophila* embryos 6–8 h after fertilization, and an along-genome visualization of the results for the same viewpoint is shown in Figure 7. Fragments that had an adjusted *P*-value of less than 0.01 in the Wald test are highlighted by blue points, or by orange points, if they additionally were called as an interaction in the depicted sample.

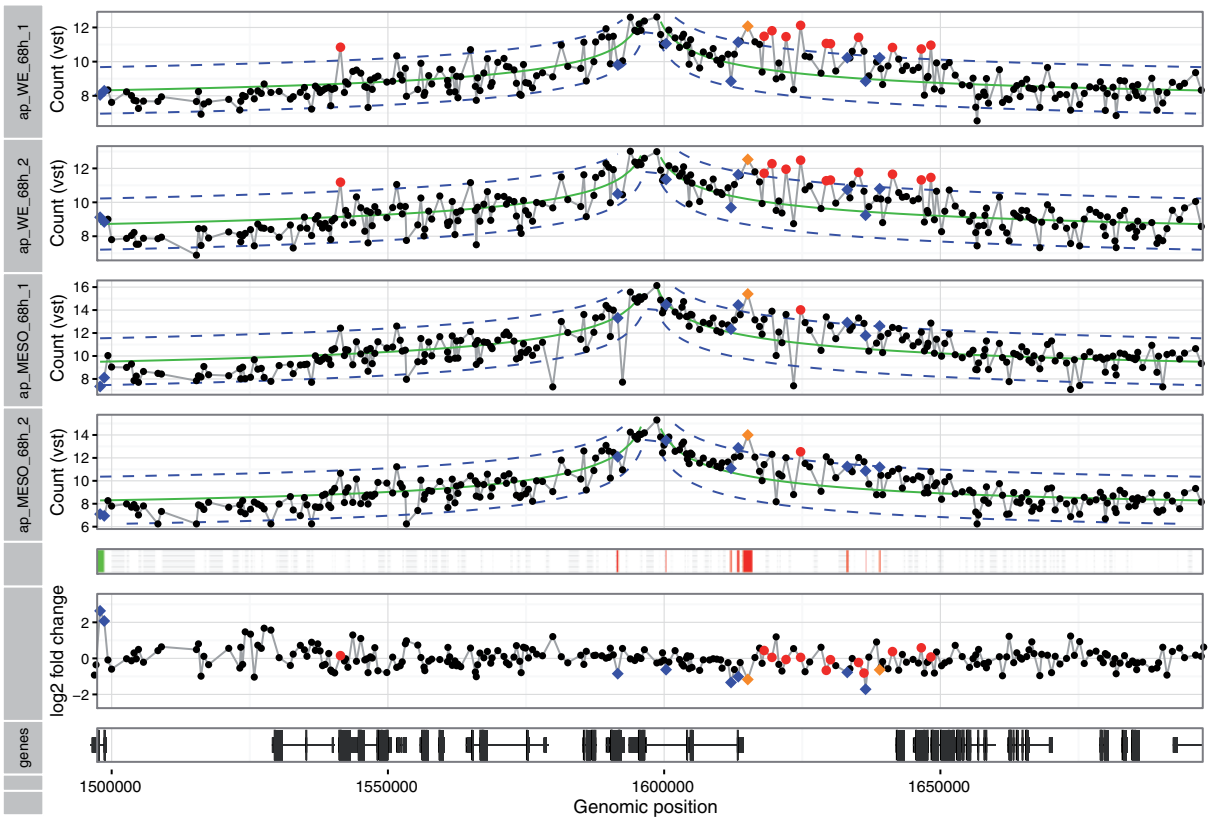

Fig. 7. Detection of interactions and differences: The figure shows the plot generated by the **FourCSeq** to visualize the results. The upper four wide tracks show the variance-stabilized counts for two biological replicates of *Drosophila* embryo mesoderm tissue and whole embryo 6–8 h after fertilization for the *apterous* CRM viewpoint. The fit of the distance dependence is shown as solid green line and the dashed blue lines represent the fit $\pm 3\sigma$. Interactions detected by *z*-score > 3 in both replicates and *p*-adjusted < 0.01 for at least one replicate are shown as red or orange points. Fragments represented by orange points additionally show a differential interactions (*p*-adjusted < 0.01, differential Wald test). Differential changes in the contact profile that are not called as interactions are shown as blue points (*p*-adjusted < 0.01, differential Wald test). The color bar below the 4C profiles shows whether the upper condition (green) or the lower condition (red) has the higher signal for the detected differences (*p*-adjusted < 0.01). The calculated log2 fold change of the differential testing per fragment are shown above the track at the bottom, which shows the gene model of the region

In general one can observe that the effect sizes for differential changes are very small, and the overall pattern of the interaction profiles remains largely unchanged, as we recently reported (Ghavi-Helm *et al.*, 2014). Only 6% of identified interactions showed evidence of interaction changes across time and tissue context (Ghavi-Helm *et al.*, 2014). However, for the strong interaction at the *ap* promoter we estimated a fold change of 2.25 between the conditions. Stronger contacts in the mesoderm tissue could be due to the fact that the *ap* gene is only expressed in the mesoderm.

4 Discussion

Our approach to detect peaks is broadly similar to that of the **r3Cseq** package (Thongjuea *et al.*, 2013). However, while **r3Cseq** performs the fit on raw count scale, we use a variance-stabilizing transformation on the data to reduce biases deriving from the large dynamic range of the count data. To detect specific interactions, we fit the decay of the variance-stabilized 4C signal with distance from the viewpoint and calculate *z*-scores from the fit residuals. With this approach, we were able to detect long-range chromatin interactions that spanned genomic distances > 100 kb in the compact *Drosophila* genome (Ghavi-Helm *et al.*, 2014). A direct comparison of **r3Cseq** and **FourCSeq** on two mouse datasets from different labs, the Myb data of Thongjuea *et al.* (2013) and the Ap2c data of Tsujimura *et al.* (2015), is provided in Supplementary File S1. It shows that **r3Cseq** is prone to over-calling interactions when the library has good coverage, while **FourCSeq** identifies interactions between the *Tfap2c* promoter and an annotated DNAseI hypersensitive region containing a brain enhancer with high specificity. On the other hand, for data with low agreement of fragment counts between replicates—possibly due to high rates of PCR duplicates—the statistical model **FourCSeq** does not call significant interactions, while **r3Cseq** reports a large set of peaks.

Instead of only looking at fold changes from single or merged samples between conditions as in **r3Cseq**, we make use of the framework for differential expression analysis implemented in the **DESeq2** package to detect differences between groups of samples in different experimental conditions. With this approach, we take the variability between replicates of the data for each genomic position into account for the quantitative comparison of the fragment counts between conditions. The fold change between conditions is compared with the variability of the data between biological replicates, and differential interaction are called statistically significant only if the observed fold change between conditions is significantly higher than what it is expected based on the noise level in the data.

Our implementation allows the use of any FASTA file as reference genome. For example, the dm3 genome was used for the data shown in Section 3. In contrast, the **r3Cseq** package is currently limited to the mm9, hg18 and hg19 genomes.

The method of van de Werken *et al.* (2012) uses a customized approach for aligning reads to a reference of fragment ends. The resulting coverage profiles can be further normalized and visualized with the tool that they provide. The results are plots of contact profiles and contact domainograms generated by analyzing the data with different window sizes. However, with this approach, comparisons between interaction profiles are only made qualitatively, and no statistical framework is provided.

To integrate called interactions and differences with other genomic data the results from our package can be used within the Bioconductor framework of *GenomicRanges* (Lawrence *et al.*, 2013). Furthermore, we provide the possibility to export the interaction profiles as bigWig files for visual inspection in a genome browser along with other tracks of interest.

Our approach looks for localized, specific interactions and treats large-scale patterns that decrease with distance from the viewpoint as background (Section 2.2.3). In particular, changes in the background between conditions will be absorbed by the normalization (2). Although these choices are reasonable for analyses such those reported in Ghavi-Helm *et al.* (2014), studies that investigate large-scale reorganization of chromosomal structure will need different analytical approaches.

Potential for improvement might lie in methods that adjust the 4C signal for the influences of fragment size, GC content and mappability. Such models exist for Hi-C and ChIA-PET data (Imakaev *et al.*, 2012; Yaffe and Tanay, 2011). However, due to the much smaller amount and viewpoint-centric nature of 4C data, the correct estimation of these biases and deconfounding them from the dominant distance-related effects is difficult. Moreover, in the differential analysis of interactions, much of the per-fragment biases will cancel out, as we essentially consider ratios.

Our method has been thoroughly validated in the course of the analysis of more than 100 viewpoints in the developing *Drosophila* embryo. Positive controls of enhancer-promoter interactions were confirmed, and several newly detected long-range interactions were validated by DNA-FISH (Ghavi-Helm *et al.*, 2014). We calculated the overlap between the identified interactions of all viewpoints from this 4C dataset and regulatory regions identified by DNaseI hypersensitivity sites (Thomas *et al.*, 2011). This analysis showed significant enrichment overlap compared with a background set (Ghavi-Helm *et al.*, 2014).

In summary, our package provides the tools to analyze 4C sequencing data and integrate the results with other genomic features. Its use will help to further investigate and understand the role of chromatin 3D structure in biological processes such as gene regulation and embryogenesis.

Acknowledgements

The authors thank the members of the Huber group for discussion and comments, in particular A. Pękowska and B. Klaus for their valuable suggestions.

Funding

F.A.K., S.A. and W.H were supported by the EC FP7 Collaborative research project RADIANT (no. 305626). T.P. was supported by an ITN grant (EvoNet) and DFG grant (FU750). E.E.M.F. was supported by a DFG (FU 750) grant and Y.G.-H. by an EMBO post-doctoral fellowship.

Conflict of Interest: none declared.

References

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Dekker,J. *et al.* (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Ghavi-Helm,Y. *et al.* (2014) Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, **512**, 96–100.

Imakaev,M. *et al.* (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.

König,J. *et al*. (2011) iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *J. Vis. Exp. JoVE*, **50**, 2638.

Lawrence,M. *et al*. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

Love,M.I. *et al*. (2014) Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.*, **15**, 550.

Pekowska,A. *et al*. (2014) Neural lineage induction reveals multi-scale dynamics of 3d chromatin organization. *bioRxiv*.

Ramsay,J.O. *et al*. (2014) *fda: Functional Data Analysis*. R Package Version 2.4.3.

Ross-Innes,C.S. *et al*. (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.

Simonis,M. *et al*. (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). *Nat. Genet.*, **38**, 1348–1354.

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.

Stadhouders,R. *et al*. (2013) Multiplexed chromosome conformation capture sequencing for rapid genome-scale high-resolution detection of long-range chromatin interactions. *Nat. Protoc.*, **8**, 509–524.

Thomas,S. *et al*. (2011) Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. *Genome Biol.*, **12**, R43.

Thongjuea,S. *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. *Nucleic Acids Res.*, **41**, e132.

Tsujimura,T. *et al*. (2015) A discrete transition zone organizes the topological and regulatory autonomy of the adjacent *Tfap2c* and *Bmp7* genes. *PLoS Genet.*, **11**, e1004897.

van de Werken,H.J.G. *et al*. (2012) Robust 4C-seq data analysis to screen for regulatory DNA interactions. *Nat. Methods*, **9**, 969–972.

Yaffe,E. and Tanay,A. (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat. Genet.*, **43**, 1059–1065.