Bioinformatics, 35(14), 2019, 2354–2361
doi: 10.1093/bioinformatics/bty1014
Advance Access Publication Date: 7 December 2018
Original Paper

Genome analysis

A novel measure of non-coding genome conservation identifies genomic regulatory blocks within primates

Alexander J. Nash [1,2] **and Boris Lenhard** [1,2,3,*]

[1]Computational Regulatory Genomics Group, MRC London Institute of Medical Sciences, [2]Faculty of Medicine, Institute of Clinical Sciences, Imperial College London, Hammersmith Campus, London W12 0NN, UK and [3]Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on September 21, 2018; revised on November 26, 2018; editorial decision on December 4, 2018; accepted on December 6, 2018

Abstract

Motivation: Clusters of extremely conserved non-coding elements (CNEs) mark genomic regions devoted to cis-regulation of key developmental genes in Metazoa. We have recently shown that their span coincides with that of topologically associating domains (TADs), making them useful for estimating conserved TAD boundaries in the absence of Hi-C data. The standard approach—detecting CNEs in genome alignments and then establishing the boundaries of their clusters—requires tuning of several parameters and breaks down when comparing closely related genomes.

Results: We present a novel, kurtosis-based measure of pairwise non-coding conservation that requires no pre-set thresholds for conservation level and length of CNEs. We show that it performs robustly across a large span of evolutionary distances, including across the closely related genomes of primates for which standard approaches fail. The method is straightforward to implement and enables detection and comparison of clusters of CNEs and estimation of underlying TADs across a vastly increased range of Metazoan genomes.

Availability and implementation: The data generated for this study, and the scripts used to generate the data, can be found at https://github.com/alexander-nash/kurtosis_conservation.

Contact: b.lenhard@imperial.ac.uk

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

1 Introduction

Regulation of developmental genes requires intricate control of the timing, location and magnitude of their gene expression. This fine level of control is primarily provided by multiple enhancers that act together to establish highly specific spatiotemporal expression patterns. In Metazoan genomes, many of these genes are maintained within syntenic arrays of evolutionarily conserved enhancers; these enhancers are known as conserved non-coding elements (CNEs) (Bejerano *et al.*, 2004; Sandelin *et al.*, 2004; Woolfe *et al.*, 2005). CNEs are extremely highly conserved, containing stretches of tens to hundreds of base pairs of nearly perfectly conserved sequence between humans and teleost fish, surviving ~450 million years of evolutionary separation (Woolfe *et al.*, 2005). The vast majority of tested CNEs have been shown to act as transcriptional enhancers which are individually capable of driving highly specific gene expression, and together recapitulate the complex temporal and spatial expression patterns of developmental genes (Bhatia *et al.*, 2014; Kimura-Yoshida *et al.*, 2004; Navratilova *et al.*, 2009; Pennacchio *et al.*, 2006; Spieler *et al.*, 2014).

The requirement for an enhancer to remain in physical proximity with the gene it regulates has constrained the evolution of vertebrate genomes, resulting in long syntenic arrays of CNEs clustered around their target genes. Each of these arrays is a functional, long-range regulatory unit known as a genomic regulatory block (GRB) (Engström *et al.*, 2007; Kikuta *et al.*, 2007; Ritter *et al.*, 2010). Most GRBs regulate a single target gene, but often span additional

bystander genes. Bystander genes are unresponsive to this form of long-range regulation, primarily due to differences in promoter structure and associated epigenetic modifications (Akalin *et al.*, 2009; Engström *et al.*, 2007; Zabidi *et al.*, 2015). For a GRB to regulate its target gene, there must be frequent physical interaction between CNEs and the promoter of the gene they regulate. Indeed, topologically associating domains (TADs), regions of the genome that preferentially interact with themselves (Dixon *et al.*, 2012; Rao *et al.*, 2014) have been shown to coincide with GRBs, suggesting they are two manifestations of the same underlying phenomenon (Harmston *et al.*, 2017). The physical location of TAD and GRB boundaries are strongly correlated in both vertebrates and invertebrates, suggesting that the insulation provided by TAD boundaries might serve to prevent the ectopic interaction of CNEs within a TAD with developmental genes in neighbouring TADs (Harmston *et al.*, 2017). Further, it was shown that TADs that are associated with GRBs have stronger self-interaction than those that are not. This is consistent with the observation that developmental genes require complex gene regulation mediated by frequent enhancer—promoter contacts, while other classes of genes, such as housekeeping genes, do not (Harmston *et al.*, 2017; Zabidi *et al.*, 2015).

Studying the dynamics of GRB evolution and the functional relationship between GRBs and TADs relies on robust methods to identify GRBs across a wide range of evolutionary timescales. Currently, CNE identification, and therefore GRB identification, hinges tightly on the selection of a threshold beyond which a conserved region is defined as a CNE. While GRB boundaries in distantly related species are robust to the CNE identification threshold used (Harmston *et al.*, 2017), in closely related species this approach breaks down, as the neighbouring, neutrally evolving sequence has not diverged enough to be able to non-arbitrarily define CNE identification thresholds. Due to the resulting increasing average length of conserved sequences, it is often necessary to choose very long thresholds for minimal CNE length (>400 bp), thereby casting doubt on the biological relevance of comparing the distribution of such elements with those identified in more distant comparisons.

Since GRBs reveal the span of regions populated by enhancers targeting specific developmental genes, and can serve as sequence-based proxies for TADs, their comparative studies are of paramount importance for understanding long-range gene regulation and the 3D chromatin structure that supports it. At the genome level, the critical GRB features to determine are their boundaries. In this paper we address this problem by defining and exploring a threshold-free measure of pairwise sequence conservation based on the kurtosis of the distribution of the lengths of all sequences perfectly conserved between two genomes. Kurtosis measures the movement of probability mass from the shoulders of a distribution into its centre and tails, and thus a distribution with high kurtosis can be considered 'fat tailed' (Balanda and Macgillivray, 1988; DeCarlo, 1997). We use kurtosis to measure the effect of the number of extreme observations on the distribution of the lengths of runs of perfect sequence identity between two genomes. We show that this measure is highly correlated with CNE density and can be effectively used to predict high quality GRBs for the species comparisons used in Harmston *et al.* (2017). Further, we use this kurtosis-based measure to predict GRBs from genome alignments between human and non-human primates, and show that it is superior to CNE density at these short evolutionary distances. The ability of our method to detect GRBs across close evolutionary distances, without the requirement for arbitrary conservation thresholds, will enable the study of GRB evolution and the detection of recent lineage-specific changes in gross GRB structure.

2 Materials and methods

2.1 CNE identification

CNEs were identified using the R Bioconductor package, CNEr (Tan, 2017, https://github.com/ge11232002/CNEr). The standard pipeline, described in the CNEr vignette, was followed. CNE density across the genome was calculated by running a 300 kb sliding window across the genome, with 30 kb steps, and calculating the number of CNEs per kb in each window.

The minimum length and identity thresholds for CNE identification must be adjusted for each species comparison due to the continuous divergence of CNEs since the last common ancestor of the two species being compared. The identification thresholds used for each species comparison are listed in Supplementary Table S1.

2.2 CNE-based GRB identification

CNE-dense regions of the genome were identified using an unsupervised two-state hidden Markov model that partitions the genome into high and low CNE density regions (as described in Harmston *et al.* (2017). In brief, the genome was segmented into high- and low-density regions, and those CNEs within the high-density regions, which were separated by less than a pre-defined genomic distance, were merged to form blocks. Human—rhesus monkey and human—gorilla GRBs were generated for this paper, while previously published human—opossum GRBs were retrieved from Harmston *et al.* (2017).

2.3 Genome-wide kurtosis calculation

For each species comparison, the kurtosis of the distribution of the lengths of all identical sequences was calculated in bins across the genome. Initially, all runs of 100% sequence identity were extracted from the pairwise whole-genome alignment and filtered for annotated repeats and exonic sequences. The genome was then divided into 30 kb bins and the lengths of all runs of identity within each bin were calculated. Windows of 30kb were used as this is the window size we traditionally use for CNE density calculation, thereby maximizing the comparability of the two approaches. The kurtosis of the distribution of lengths in each bin was then calculated as follows:

$$R(F) = \frac{q_{0.99}(F) - q_{0.01}(F)}{G_{50}}$$

where F is the distribution of the lengths of runs of perfect sequence identity in a bin, and G_{50} is the range of the middle 50% of the distribution of lengths of all runs of identity, from all bins (background distribution); calculated as follows:

$$G_{50} = q_{0.75}(J) - q_{0.25}(J)$$

where J is the distribution of the lengths of runs of perfect sequence identity across the whole genome. For each bin, $R(F)$ is a ratio of the range of 99% of all lengths of runs of identical sequence, in a bin, to the range of 50% of all lengths of runs of identity for the whole genome. In practice it measures the number, and extremity, of long runs of perfect identity, in each bin, compared to the background conservation for the whole genome. This is an adaptation of the robust kurtosis measure proposed in Ruppert (1987).

2.4 Correlation of kurtosis and CNE density

Maximum kurtosis and CNE density was calculated in 90 kb windows across the genome, with 1000 windows randomly sampled from previously defined human—opossum GRBs and 1000 from non-GRB regions. This was performed for human to dog, chicken

and spotted gar comparisons at each CNE identification threshold listed in Supplementary Table S1. The Spearman's correlation coefficient between maximum scores in each window was then calculated. For the purpose of visualization, a linear model was fitted to the data for each comparison at each CNE identification threshold.

2.5 Kurtosis-based GRB identification

Kurtosis-based GRBs were generated by using the change point modelling approach to identify change points in the binned kurtosis data, indicating a shift to higher mean kurtosis values (Ross, 2015). Under this framework, kurtosis values in bins across the genome are treated as a series of n independent observations $x_i, \ldots, x_n$. The assumption that all observations (genomic windows) are identically distributed according to an undefined distribution F_0, can then be tested by choosing between the following hypotheses:

$$H_0: X_i \sim F_0(x_i;\theta_0), i = 1,2, \ldots, n,$$

$$H_1: X_i \sim \begin{cases} F_0(x;\theta_0), i = 1,2,\ldots,k, \\ F_1(x;\theta_1), i = k+1, k+2,\ldots,n, \end{cases}$$

where θ_i represents the unknown parameters of each distribution. In this scenario the two distributions F_0 and F_1 represent the distribution of values coming from non-GRB and GRB regions of the genome, respectively. The presence of a change point can be tested using a two-sampled Mann–Whitney test and the null hypothesis rejected if the test statistic exceeds a pre-defined cut-off. For a series of observations $x_i, \ldots, x_t$ the test statistic is calculated at every x_k, for $1 < k < t$, and the maximum test statistic obtained for all values of k is used. As successive observations are made (successive windows along the genome), the test statistic is calculated again at every x_k but now for $1 < k < t+1$. If no significant change point is detected, the next observation, x_{t+2} is received and the testing is performed again on $x_i, \ldots, x_{t+2}$. However if a change is detected at x_k the process begins again with x_{k+1} as the first observation in the new series of observations to be tested. For further details refer to Ross (2015). This analysis was performed using the cpm package in R, and the ARL0 parameter was set to 370. This is the least stringent ARL0 value implemented in the package and ensures that all potential change points are detected, at the risk of including more false positives. Greater sensitivity combined with a merging step (described below) was preferred to stringent change point detection that potentially misses GRB boundaries.

Once significant change points in the binned kurtosis values have been identified, these are treated as potential GRB boundaries. The mean kurtosis within each range is then calculated, and adjacent ranges are merged if the mean kurtosis in both is above a specified quantile of all binned kurtosis values. The quantile used was determined empirically based on the predicted GRBs ability to recapitulate known GRB boundaries. For all species comparisons, the quantile used was 0.7.

2.6 Hi-C data processing

hESC and IMR90 Hi-C data were obtained from the Gene Expression Omnibus (GSE35156) and processed as described in Harmston *et al.* (2017). To visualize how well kurtosis-based GRBs recapitulate TAD boundaries, we produced heatmaps of Hi-C directionality index (DI) within genomic windows centred on GRBs and ordered from largest to smallest GRBs.

3 Results

3.1 CNE identification

To test the improvement in performance of the kurtosis-based approach, we first produced CNE sets for standard GRB span detection and its comparison with our new method. The first step in this analysis was to identify CNEs between human and a range of species chosen to represent distinct vertebrate lineages. For each species comparison we used multiple CNE identification thresholds to facilitate a comprehensive comparison between the proposed kurtosis-based conservation measure and CNE density calculated for a range of conservation thresholds. The results of CNE identification are presented in Supplementary Table S2. As expected, the number of CNEs identified for each species comparison decreases as the stringency of the threshold increases. The mean width of the elements identified also decreases as the minimum required identity is increased. In general, the stringency of the threshold used for CNE identification is reduced as the evolutionary distance between the species compared increases. This is to account for the continual sequence divergence in conserved regions during the time that the two genomes have been evolving independently. The effect of this sequence divergence is clearly discernible from the number of CNEs identified in dog, opossum, chicken and spotted gar at 80% identity over 50 bp (30 bp in spotted gar). The divergence time between human and each of these species ranges from 96 to 435 million years, and with the increasing time so the number of CNEs identified drops from 3 763 684 to just 33 172.

3.2 Kurtosis-based conservation is strongly correlated with CNE density

Next, for each of the same species comparisons, we calculated the kurtosis of the distribution of the lengths of perfectly conserved sequences (i.e. of gapless alignment blocks with no substitutions) in bins across the genome. Figure 1A shows the distribution of kurtosis values across the genome for each comparison. The distributions are very similar across species comparisons, illustrating that the results of the method are comparable across a wide range of evolutionary distances. In the closer species comparisons (gorilla to opossum) the distributions of kurtosis values are centred on 4.5. As the evolutionary distance of the comparison increases, so the number of bins containing a value of zero increases, and the median kurtosis value drops. The increasing number of zero bins is due to an increasing number of bins that do not contain any alignable sequence. This trend shows that, as expected, with increasing evolutionary distance there will be larger portions of the genome that are unalignable due to continual sequence divergence. Since the kurtosis calculation does not depend on the total amount of aligned sequence per bin, the kurtosis-based measure is also robust to variation in proportion of other unalignable sequences along the genome, such as repetitive elements. It is also striking that the range of the kurtosis values increases with evolutionary distance. This trend reflects the potential for more extreme outliers relative to the genomic background in more distant comparisons. These extreme outliers are the CNEs that we normally identify using traditional CNE identification approaches. Another notable feature of the distributions is the increased dispersion with increasing evolutionary distance. This is most likely due to the increased variability in the number and length of runs of sequence identity from bin to bin in the more distant comparisons.

To compare kurtosis and CNE density across the genome, we sampled 1000 random 90 kb windows from previously defined CNE-based human—opossum GRBs, and non-GRB regions of the genome. We then calculated the maximum kurtosis and CNE density in each window. We repeated this using CNE density calculated

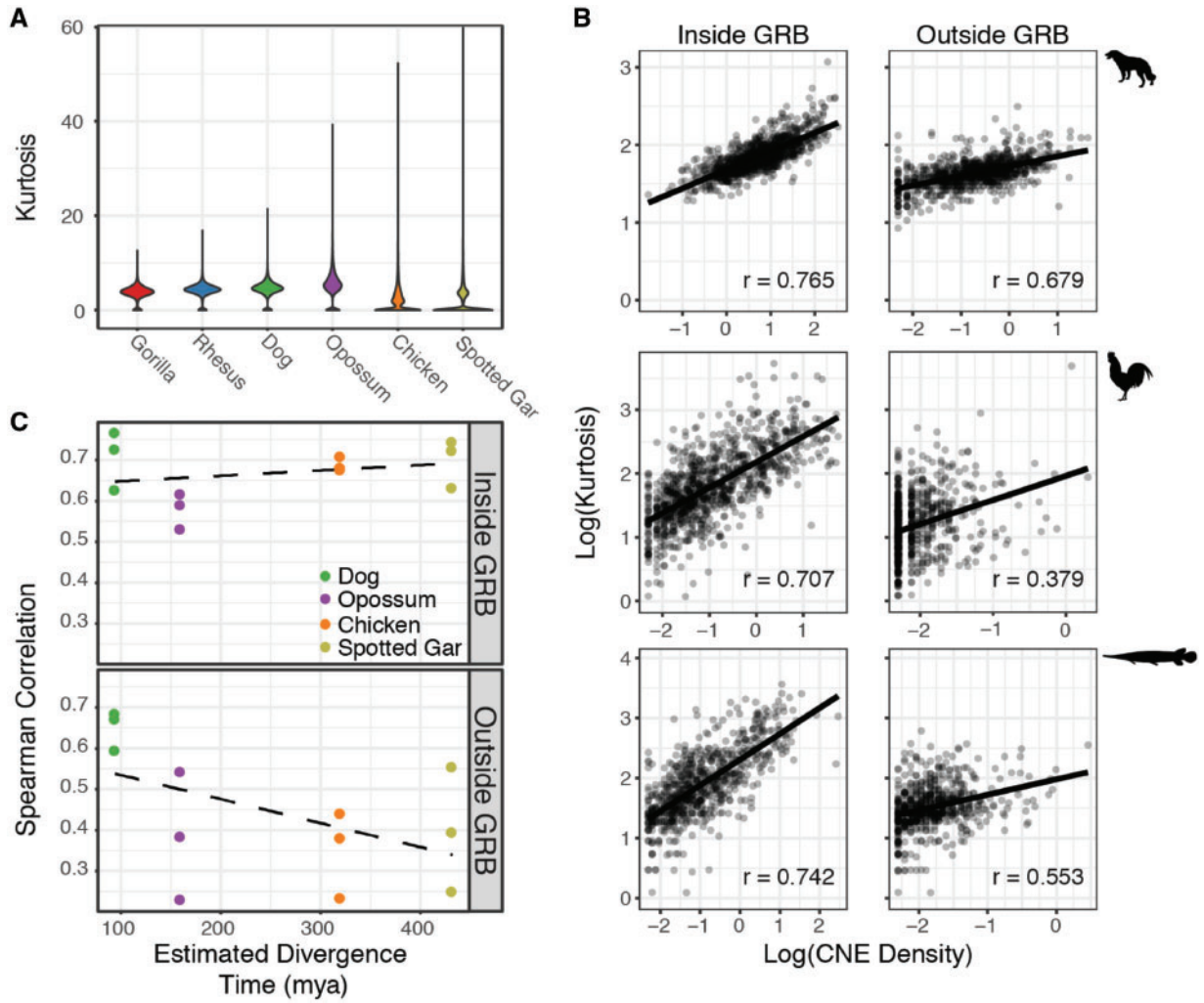

Fig. 1. Kurtosis and CNE density are highly correlated. (**A**) The distribution of the kurtosis values calculated for human to each other species. The kurtosis values were calculated for the distribution of the lengths of perfectly conserved sequences in bins across the genome for each species comparison separately. (**B**) The correlation between CNE density and kurtosis inside and outside of CNE-based GRBs. (**C**) The correlation of CNE density (calculated for multiple conservation thresholds) and kurtosis inside and outside of GRBs as a function of evolutionary distance. Here, each point for a species represents the correlation of kurtosis-based conservation with CNE density at a different CNE identification threshold. CNE density and kurtosis are highly correlated within GRBs, regardless of the evolutionary distance of the comparison. Outside of GRBs there is a decreasing linear relationship between the correlation and the evolutionary distance of the comparison

at multiple thresholds for each species comparison. Next, we calculated the Spearman's correlation coefficient between kurtosis and CNE density for each species comparison inside and outside of GRBs. There is a strong correlation between kurtosis and CNE density, and this correlation is greater within GRBs than outside GRBs (Fig. 1B). This trend is confirmed in Figure 1C, which shows the Spearman's correlation coeffcient between kurtosis and CNE density, calculated for all CNE identification thresholds used for each species comparison. This data are also presented in detail in Supplementary Table S3. It is striking that regardless of the evolutionary distance of the comparison, kurtosis and CNE density values are similarly correlated within GRBs, whereas outside of GRBs it appears that the correlation drops with increasing evolutionary distance. The reduced correlation outside of GRBs may be caused by multiple properties of kurtosis and CNE density:

1. Outside of GRBs, CNE density is consistently either zero or close to zero, while kurtosis fluctuates around 4.5 from bin to bin, thereby reducing the correlation. Within GRBs, both the CNE density and kurtosis will be high in the majority of bins.
2. Outside of GRBs, there may be stretches of identical non-coding sequence that are shorter than the minimum length of the threshold used for calling CNEs, and are therefore not identified. These stretches will still result in distributions with relatively high kurtosis. Within GRBs there are many identifiable CNEs and thus both CNE density and kurtosis will be high.

Overall, the consistency of the distribution of kurtosis values for species comparisons spanning vastly different evolutionary timescales, and its high correlation with CNE density in conserved regions of the genome, suggest that the kurtosis of the lengths of runs of sequence identity can be used as an effective threshold-free proxy for sequence conservation in genomic windows.

3.3 Kurtosis-based GRB identification in moderately to distantly related species

In the past, GRB identification has succeeded for moderate to distant evolutionary comparisons because the CNE density across the genome forms discrete peaks that are easily distinguished from the

Table 1. Kurtosis-based GRB number and size

Query species	Number	Mean width (kb)
Dog	559	1233.1
Opossum	487	1195
Chicken	426	978.7
Spotted Gar	400	804.8

genomic background (Akalin *et al.*, 2009; Engström *et al.*, 2007; Harmston *et al.*, 2017; Kikuta *et al.*, 2007). To test how well the kurtosis-based measure of conservation can discriminate highly conserved regions of the genome from non-conserved regions, we used binned kurtosis values to identify GRBs from human to moderately and distantly related species which have previously been used for CNE-based GRB prediction (Harmston *et al.*, 2017). The number and size of GRBs identified for each comparison are presented in Table 1. The number of GRBs identified in each comparison is similar, but there is a slight decrease in total GRBs as the evolutionary distance of the comparison increases. The average width of the identified GRBs also decreases with increasing evolutionary distance. The decreasing average width reflects the erosion of sequence conservation over time, making accurate prediction of GRB boundaries diffcult over large evolutionary distances. This effect is also observed in GRBs identified using CNE density and has been previously described (Harmston *et al.*, 2017). The decreasing number of GRBs may be due to the identification of relatively rapidly evolving GRBs in the closer comparisons that are not identifiable in the more distant comparisons.

As an initial assessment of the quality of the identified GRBs, we visualized CNE density within genomic windows centred on the kurtosis-based GRBs for each species comparison (Fig. 2A). GRBs were ordered by width, and thus any feature that is enriched within GRBs forms a characteristic funnel pattern. From these heatmaps, it is immediately apparent that there is a very strong enrichment of CNE density within kurtosis-based GRBs. This enrichment is robust to the CNE identification threshold used (Supplementary Fig. S1). Interestingly, as the stringency of the CNE identification threshold is increased, there are an increasing number of GRBs that contain no enrichment for CNE density (Supplementary Fig. S1). This likely reflects the ability of the kurtosis-based measure to identify runs of non-coding identity that fail to pass the more stringent CNE identification thresholds.

As with CNE-based GRBs, the boundaries of kurtosis-based GRBs are very similar between species comparisons (Supplementary Fig. S2), and there is significant overlap between the kurtosis-based GRBs and those predicted using CNE density (Supplementary Fig. S3). In general, the kurtosis-based method predicts fewer GRBs than the CNE-based method, with the former appearing to be a high-confidence subset of the latter.

To further evaluate the accuracy of the kurtosis-based GRB boundaries, we took advantage of the fact that GRB boundaries frequently coincide with TAD boundaries (Harmston *et al.*, 2017), and hypothesized that a better prediction of GRB boundaries should result in an increased agreement with TAD boundaries. To test this, we plotted the Hi-C DI from hESC cells within the same GRB-containing genomic windows (Fig. 2B). In these plots the intensity of red and blue in a region shows the frequency with which this region interacts with downstream and upstream loci, respectively. Visualized this way, TADs appear as a span of red followed by a span of blue. For GRBs defined from human to dog, opossum and chicken, there is a very clear funnel present in the DI heatmaps. The funnels have a well-defined red boundary followed by a well-defined blue boundary, indicating that the GRBs coincide well with TADs. There is no visible funnel in the human to spotted gar GRBs, with only a hint of a funnel visible in the very largest GRBs, many of which are also the most deeply conserved (Fig. 2A). While these GRBs clearly do not coincide with TADs, it is possible that at greater evolutionary distances, the kurtosis-based conservation measure is only identifying the core, highly conserved regions of each GRB, and thus underestimating their true extent—possibly because of some turnover of the boundary positions themselves. Based on the concordance between the kurtosis-based GRB predictions and the CNE density for all species comparisons, it is likely that CNE-based GRB prediction will suffer from the same problem.

Taken together, the concordance between kurtosis-based GRB predictions and CNE density, the high degree of overlap between kurtosis-based and CNE-based GRBs and the strong correlation between GRB and TAD boundaries, suggests that kurtosis-based conservation can be used to accurately predict high quality GRBs.

3.4 Kurtosis-based GRB identification in alignments between human and non-human primates

CNE identification thresholds necessitate the implementation of an arbitrary cut-off for what is defined as a CNE and what is not. At the edge of the threshold, a single mismatch in two aligned sequences is suffcient for an otherwise highly conserved region to be declared non-conserved. In the context of GRB identification, this is seldom a problem for evolutionarily distant species comparisons, but at shorter evolutionary timescales it becomes increasingly diffcult to determine how long a stretch of perfect sequence identity should be for the region to be declared a CNE. This is the context in which kurtosis-based conservation may see the most utility. By its nature, kurtosis-based conservation takes into account the background level of conservation for a particular species comparison, and only defines those regions with unexpectedly long runs of identity as highly conserved.

We predicted GRBs for human to two non-human primates, the rhesus monkey and the gorilla, to test the limits of kurtosis-based conservation for GRB detection. Humans and rhesus monkeys (referred to as rhesus from here on) diverged ~30 million years ago, while humans and gorillas diverged only 8.6 million years ago. Using kurtosis-based conservation, we predicted 523 human—rhesus GRBs (mean width =1279.9 kb) and 483 human—gorilla GRBs (mean width =1242.9 kb). This is a reassuringly similar number of GRBs to the sets identified by comparison to more distant vertebrates, suggesting that even at such short evolutionary timescales the method can predict comparable GRB sets.

To assess the quality of these GRBs, we plotted CNE density and Hi-C DI across genomic windows centred on the kurtosis-based GRB predictions, as previously described (Fig. 3A–C). For the human-rhesus GRBs there is a strong enrichment of CNE density within the predicted GRBs, indicating that for this species comparison kurtosis is a good proxy for conserved non-coding conservation. For the human to gorilla comparison there is also a visible CNE density enrichment within the predicted GRBs, but the strength of the enrichment is much reduced. The average mismatch rate between human and gorilla is only 1.75%, and therefore it is very surprising that there is any CNE density enrichment within the kurtosis-based GRBs (Scally *et al.*, 2012). This result is strong evidence that kurtosis-based conservation can identify highly conserved regions of the genome. Examining the DI heatmaps, it is clear that the rhesus GRBs have a visible funnel, although it is not as strong as in the more distant comparisons. The largest rhesus GRBs have the weakest correspondence with the DI, and appear to span multiple TADs. These are probably physically close GRBs that have been merged by the GRB prediction. This may also account for the increased mean GRB width in this set. Separating

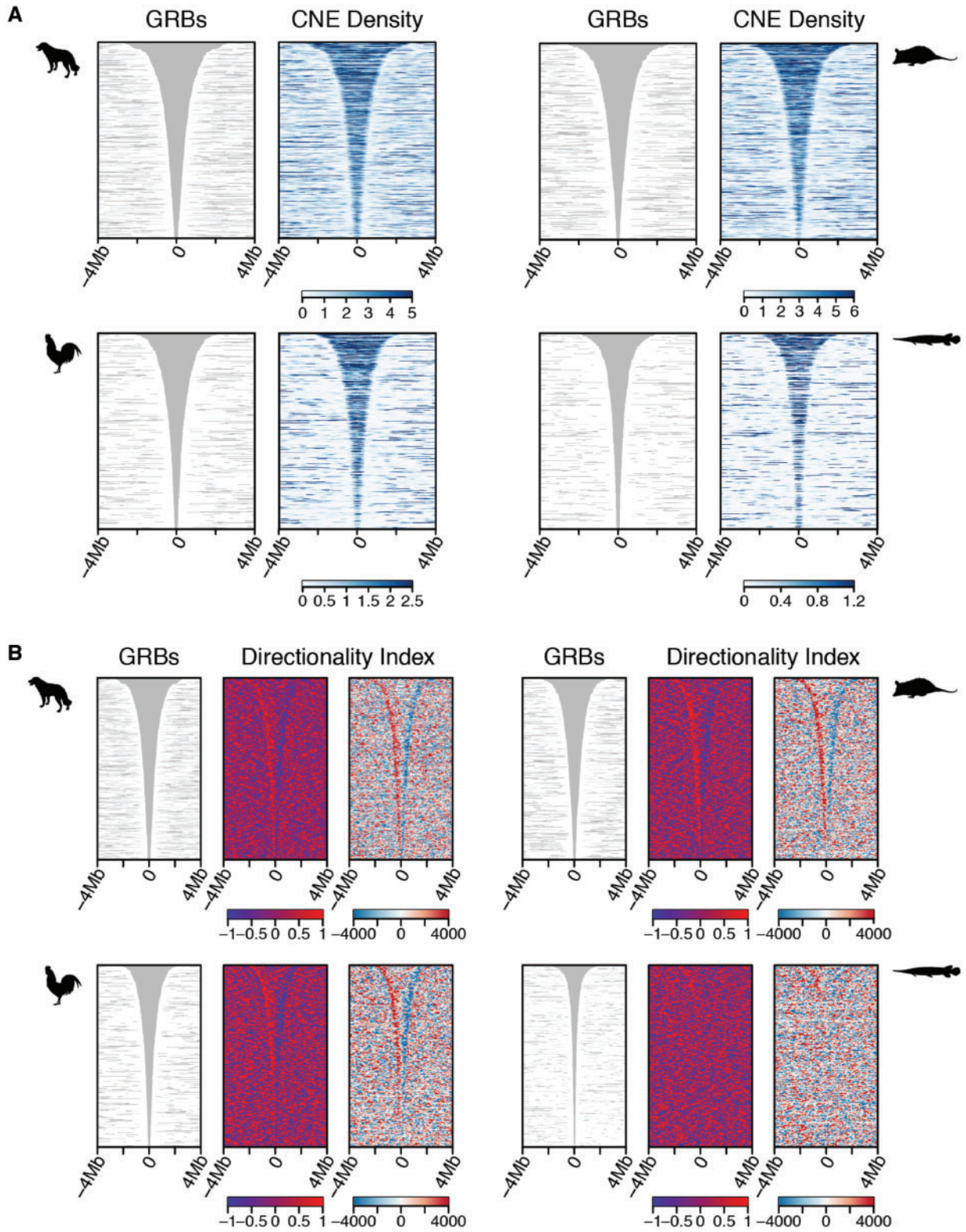

Fig. 2. CNE density and Hi-C directionality index within kurtosis-based GRBs. GRBs were predicted from human to dog, opossum, chicken and spotted gar using the kurtosis-based measure of conservation. Grey heatmaps represent the extent of the predicted GRBs. (**A**) CNE density within genomic windows centred on predicted GRBs. (**B**) hESC derived Hi-C DI within genomic windows centred on predicted GRBs. In each case, the left-hand most heatmap shows the sign of the DI, representing the direction of interaction bias, while the right-hand most heatmap shows the DI value, representing the strength of the interaction bias

adjacent synteny blocks using sequence conservation alone is a known diffculty in GRB prediction (Harmston *et al.*, 2017). Since kurtosis-based GRB prediction is also sequence-based, it is not immune to this problem. For the human-gorilla GRBs a similar issue is visible. The largest third of GRBs display no visible funnel in the DI heatmaps, however there is a noisy funnel visible in the rest of the GRBs. Overall

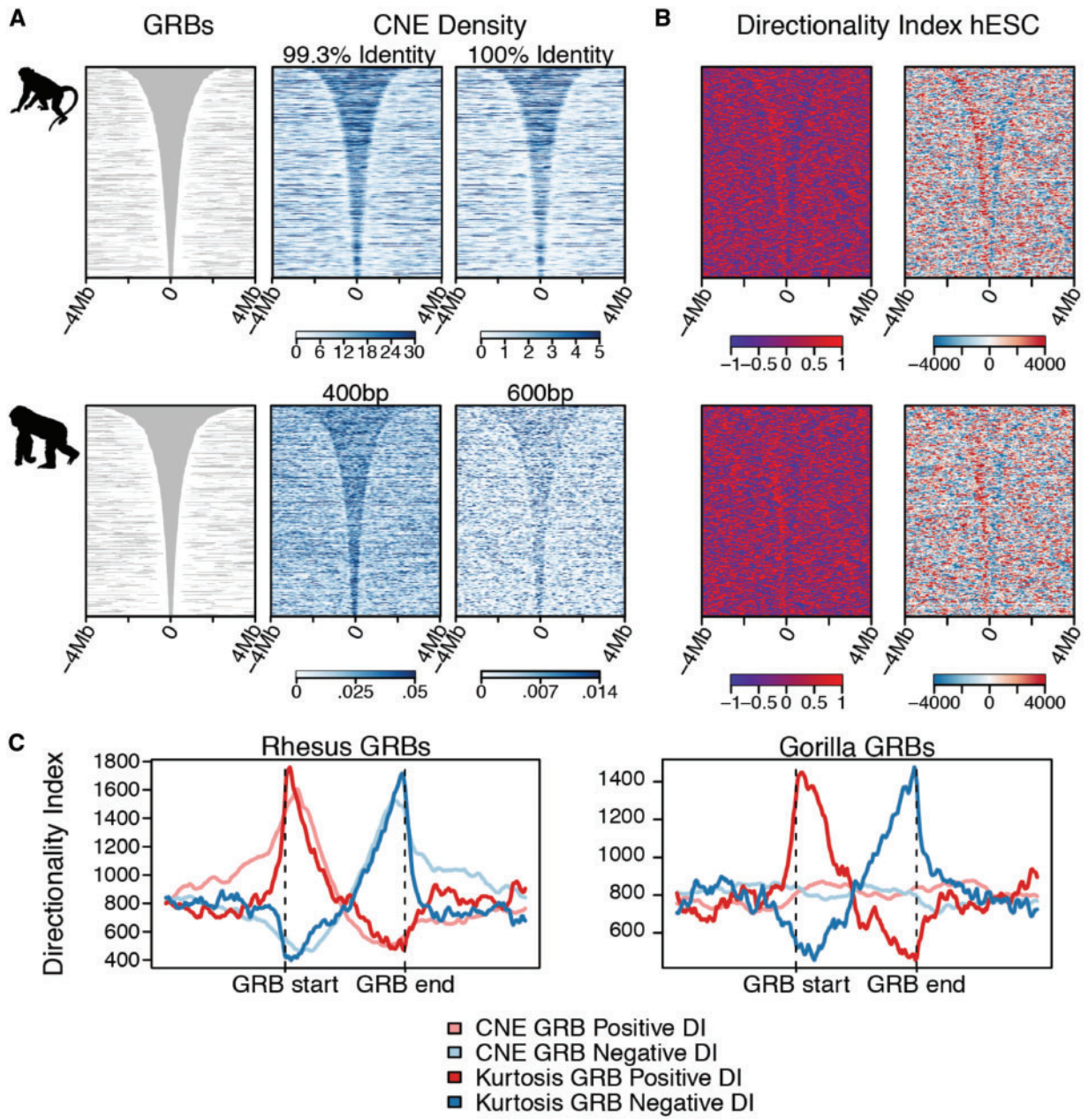

Fig. 3. Kurtosis-based GRBs in primates. Kurtosis-based GRBs were identified from human to rhesus monkey and gorilla. Grey heatmaps represent the extent of the predicted GRBs. (**A**) CNE density within genomic windows centred on predicted GRBs. The CNE density is shown for two conservation thresholds. In the case of human-gorilla CNEs all CNEs are identified at 100% identity and therefore to increase stringency, the minimum length is increased. (**B**) Hi-C DI within genomic windows centred on predicted GRBs. In each case the left-hand most heatmap shows sign of the DI representing the direction of interaction bias, while the right-hand most heatmap shows the DI value, representing the strength of the interaction bias. (**C**) Average Hi-C DI strength within kurtosis- and CNE-based GRBs for human-rhesus monkey and human-gorilla

these results suggest that kurtosis-based conservation can identify signatures of non-coding conservation in very closely related species, but that GRB boundary prediction becomes less precise in the most closely related comparisons.

Next, we compared our kurtosis-based GRBs to GRBs identified using the CNE-based approach described in Harmston *et al.* (2017). The CNE-based GRB prediction yielded 744 human to rhesus GRBs with a mean width of 482.9 kb and 2220 human to gorilla GRBs with a mean width of 504.4 kb. The number of GRBs identified in human-rhesus is greater than for the other species comparisons used so far, but not exceedingly so. For the human-gorilla comparison, however, there were a very large number of predicted GRBs. In Figure 3C, the average Hi-C DI is plotted across the predicted GRBs from both sets. We can clearly see that, for the human-rhesus comparison, the kurtosis-based GRBs have a stronger peak of the positive and negative DI (at their starts and ends, respectively) than the CNE-based GRBs. There is also a much sharper boundary effect in the kurtosis-based GRBs, with the DI signal spreading well beyond the boundaries of the CNE-based GRBs. In the human-gorilla comparison the kurtosis-based GRBs boundaries also coincide with peaks in the positive and negative DI, while the CNE-based GRBs show no enrichment of DI score at either boundary.

These results conclusively demonstrate that the kurtosis-based conservation measure can identify highly conserved regions of the genome, even in very closely related species, and that kurtosis-based GRB predictions recapitulate TAD boundaries better than the CNE-based GRB predictions at these evolutionary timescales.

4 Discussion

In this paper we have defined a novel measure of pairwise sequence conservation based on the kurtosis of the distribution of the lengths of sequences perfectly conserved between two genomes. We have shown that the kurtosis-based measure is highly correlated with CNE density and can be used to generate high quality GRB predictions for moderate to distant species comparisons. Importantly, our method enables accurate prediction of GRB-scale regulatory domains, but does not identify the individual conserved elements themselves. This presents the potential for complementary use of kurtosis-based GRB identification and traditional CNE identification in future analyses.

We have also shown that kurtosis-based GRB prediction far outperforms CNE-based GRB prediction in closely related species. The identification of GRBs between human and gorilla is a surprising result as previously it has been impossible to define conserved regulatory domains between such closely related species. Humans and gorillas share over 98% of their genome sequence, and so to be able to use sequence conservation to define regulatory regions that coincide with TADs is strong testament to our method's ability to account for the general background conservation between two genomes.

Most importantly, unlike CNE-based conservation analysis, our method works without requiring any pre-defined minimum length or sequence identity thresholds for a sequence to be considered conserved. Having a threshold-free approach for measuring conservation allows us to directly compare the results of species comparisons spanning a range of evolutionary distances. This feature, combined with the success we have had in identifying GRB-like structures in extremely closely related species, opens up the possibility of systematically investigating the evolutionary dynamics of GRBs in multiple closely related Metazoan lineages, potentially yielding a greater understanding of the origin and evolution of long-range gene regulation in Metazoan genomes.

Further, our method may have utility in the analysis of GRB developmental gene regulation in species that have undergone extreme genome compaction such as the puffer fish, *Tetraodon nigroviridis*, and the sea squirt, *Oikopleura dioica* (Denoeud *et al.*, 2010). The tiny size of these genomes makes it very difficult to define the minimum length a stretch of conserved sequence should be to be considered a conserved element, and as described above, comparing the results of this analysis with those performed in larger genomes is problematic. Our method may provide the ability to accurately define GRB boundaries in compact genomes and therefore deliver insights into the effects of genome compaction of long-range gene regulation.

Acknowledgements

We thank Dr Ge Tan for generating a number of the CNE datasets used in this analysis, and Dr Nathan Harmston for processing the Hi-C data. We are also grateful to Dr Leonie Roos, Dr Anja Baresic, Dr Sasha Murrell and Dr Ben Murrell for comments on the manuscript.

Funding

This work was supported by the Medical Research Council [MC UP 1102/1 to B.L., 1584095 to A.J.N.

Conflict of Interest: none declared.

References

Akalin,A. *et al.* (2009) Transcriptional features of genomic regulatory blocks. *Genome Biol.*, **10**, R38.

Balanda,K.P. and Macgillivray,H.L. (1988) Kurtosis: a critical review. *Am. Stat.*, **42**, 111–119.

Bejerano,G. *et al.* (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.

Bhatia,S. *et al.* (2014) A survey of ancient conserved non-coding elements in the PAX6 locus reveals a landscape of interdigitated cis-regulatory archipelagos. *Dev. Biol.*, **387**, 214–228.

DeCarlo,L.T. (1997) On the meaning and use of kurtosis. *Psychol. Methods*, **2**, 292–307.

Denoeud,F. *et al.* (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, **330**, 1381–1385.

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

Engström,P.G. *et al.* (2007) Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.*, **17**, 1898–1908.

Harmston,N. *et al.* (2017) Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.*, **8**, 441.

Kikuta,H. *et al.* (2007) Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.*, **17**, 545–555.

Kimura-Yoshida,C. *et al.* (2004) Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, **131**, 57–71.

Navratilova,P. *et al.* (2009) Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.*, **327**, 526–540.

Pennacchio,L. *et al.* (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature*, **444**, 499–502.

Rao,S.S.P. *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.

Ritter,D.I. *et al.* (2010) The Importance of Being Cis: evolution of Orthologous Fish and Mammalian Enhancer Activity Research article. *Mol. Biol. Evol.*, **27**, 2322–2332.

Ross,G.J. (2015) Parametric and nonparametric sequential change detection in R: the cpm package. *J. Stat. Softw.*, **66**, 1–20.

Ruppert,D. (1987) What is kurtosis? An influence function approach. *Am. Stat.*, **41**, 1–5.

Sandelin,A. *et al.* (2004) Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, **5**, 99.

Scally,A. *et al.* (2012) Insights into hominid evolution from the gorilla genome sequence. *Nature*, **483**, 169–175.

Spieler,D. *et al.* (2014) Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. *Genome Res.*, **24**, 592–603.

Tan,G. (2017) CNEr: cNE detection and visualization. *R Package Version 1.16.0*.

Woolfe,A. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.

Zabidi,M. *et al.* (2015) Enhancer—core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, **518**, 556–559.