BIOINFORMATICS

Vol. 00 no. 00
Pages 1–3

Constitution Analysis

A [1,3,*], B [2] and C

[1] Laboratory of Molecular Immunology, The Rockefeller University, 1230 York Avenue, New York, NY 10065
[2] Departamento de Computação e Matemática, Universidade de São Paulo. Av. Bandeirantes, 3900, Ribeirão Preto, CEP 14049-901, SP, Brazil

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT
Motivation:
Results: https://github.com/someone/rich.
Contact: someone@somewhere.world

1 INTRODUCTION

The genome is targeted by a sophisticated and highly coordinated series of molecular events. Among these events, aberrant DNA methylation patterns in human malignancy De *et al.* (2013), somatic retrotransposition in human cancers Lee *et al.* (2012), AID-dependent chromosomal translocations Klein *et al.* (2011) and HIV integration Craigie and Bushman (2012), which arrives throughout DNA, are not randomly distributed but instead associated with chromosomal regions and contributes to disrupt the integrity of the genome and human disease.

As result, these regions represents a genomic context in which are associate with multiple underlying mechanisms. The motif-based sequence analysis is the starting point to aim potential binding site of cis-regulatory elements associated. Nevertheless, the inherent low signal/noise ratio in sequence-based motif discovery is a limitation to detect a nucleic acid sequence pattern that has some biological significance Hu *et al.* (2005). Moreover, these events may recognise a structural feature, rather than a specific sequence motif.

However, how exactly the pattern nucleotide composition could influence the selection of target site selection are not well understood. To further characterize at a genome-wide scale these regions, we introduce a new method (*k-enrich*) to provide a quantitative measure of the differential spectra of k-mers (DNA 'words' of length k) throughout target DNA.

2 METHOD

Let $\mathcal{A} = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ and $\mathcal{S} \in \mathcal{A}^{\ell}$, be a given specific string of length $\ell \geq 1$. In what follows, we describe a method to study the profile of $\mathcal{S}$ along a region of interest such as those defined by viral insertion or translocation breakpoint hotspots. This provides the means to asses the significance of a differently distributed profiles along two functionally defined regions. We specialise genomic regions with translocations hotspots as described by Klein *et al.* (2011), but the scope is clearly not restricted to this particular application.

Let $h = \{h_1, \dots, h_n\}$ bet a set of translocation breakpoint hotspots, namely a set of DNA segments characterized by having a substantially high density of translocations events. Let w be the length of the longest of such segments. The segments $h_1, \dots, h_n$ are aligned with respect to their central base and then extended at both ends to have length w. Next, consider the partition of resulting set of segments into k evenly spaced bins of length $\ell = w/k$. Denote by h_{ij}, $1 \leq j \leq k$, the jth bin of the ith segment. Consider now the set $r = \{r_1, \dots, r_n\}$ of segments of width w that are either at the left or at the right of any one segment in h. Likewise, let r_{ij}, $1 \leq i \leq n$, $1 \leq j \leq k$, be the matrix formed by bins of length ℓ that result by partitioning the elements of r. For any $j = 1, \dots, k-1$ and $i = 1, \dots, n$, let ξ_{ij} and η_{ij} be the following Bernoulli random variables

$$\xi_{ij} = \begin{cases} 1, & \text{if } \mathcal{S} \in h_{ij} \text{ and } \mathcal{S} \notin h_{i,j+1} \\ 1, & \text{if } \mathcal{S} \in h_{ij} \text{ and } \mathcal{S} \in h_{i,j+1} \\ 0, & \text{otherwhise} \end{cases},$$

$$\eta_{ij} = \begin{cases} 1, & \text{if } \mathcal{S} \in r_{ij} \text{ and } \mathcal{S} \notin r_{i,j+1} \\ 1, & \text{if } \mathcal{S} \in r_{ij} \text{ and } \mathcal{S} \in r_{i,j+1} \\ 0, & \text{otherwhise} \end{cases}.$$

Set $\xi_{ik} = 1$ if $\mathcal{S} \notin h_{i,k-1}$ and $\mathcal{S} \in h_{i,k-1}$, and $\xi_{ik} = 0$ otherwhise. Similarly define η_{ik} by using the information in r_{ik}. Finaly, let

$$\xi_j = \sum_{i=1}^{n} \xi_{ij}, \qquad \eta_j = \sum_{i=1}^{n} \eta_{ij}.$$

The variables ξ_j and η_j, $1 \leq j \leq k$, count the number of times that the string $\mathcal{S}$ occurs along of a hotspot region and of a reference region respectively. A schematic representation of this procedure is shown in Figure 1.

The basic question we like to address is wether the distribution profile of the string $\mathcal{S}$ is significatively different along a typical hotspot region and a reference region. This may be assessed by considering the following $2 \times k$ contingency table

$$\begin{bmatrix} \xi_1 & \xi_2 & \dots & \xi_k \\ \eta_1 & \eta_2 & \dots & \eta_k \end{bmatrix},$$

obtained by merging the two vectors of counts ξ_j and η_j. Provided the number of counts in each of the cells of this table is sufficiently

[*] to whom correspondence should be addressed

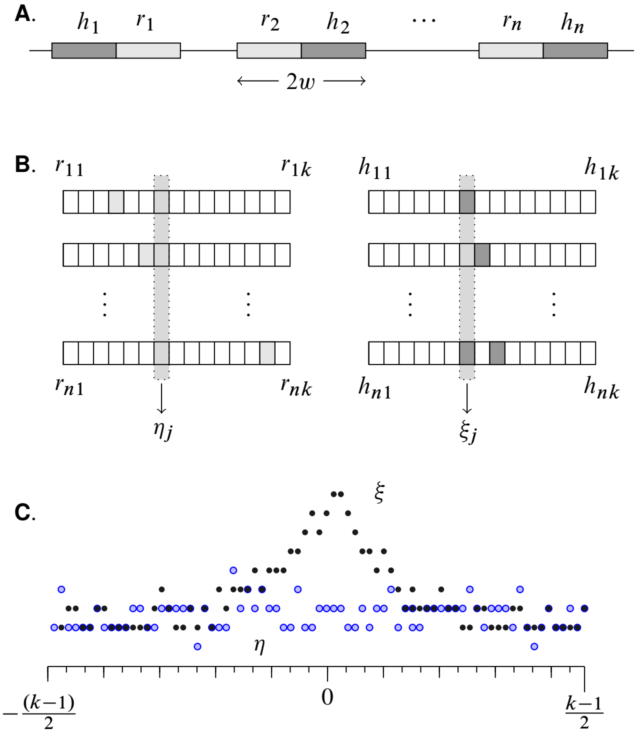

Fig. 1. A: Input data segments $h_1, \dots, h_n$ containing the occurence of a string $\mathcal{S}$ and reference segments $r_1, \dots, r_n$. B: matrizes of counts for a particular realization of the random variables η_{ij}, ξ_{ij}. C: Profile distribution for the occurence of $\mathcal{S}$ along a hotspot and a reference region.

large, the significance of a differential profile can be determined by using Pearson's X^2 statistic, which is distributed according to a χ^2 density with $k - 1$ degrees of freedom. Other alternatives for the large sample case exist, see for instance Read and Cressie (1988), but we do not pursue this further here. It is well known that this procedure can give a poor approximation when several cells present low counts (smaller than 10). This may be the case in the current setting when analysing the profile distribution of longer strings with $\ell \geq 10$ or even smaller but rarely occuring strings. In these situations the significance for a differential profile is more appropriately determined by using Fisher's exact test, see for instance Agresti (2012). The computations necessary to derive a p-value are not feasible because of the large number of contingency tables that have to be considered as a reference set when k is large. The significance may however be approximately computed by considering a permutation test using the method in Patefield (1981). We found that R's implementation via `fisher.test` takes only few secconds for relatively large tables, for instance with $k = 1000$.

We provide examples for the two scenarios just mentioned by considering strings formed by a single base and strings defined by longer motivs with $\ell = 15$. The former provides an example where the X^2 statistic is a appropriate and the latter one that is amenable to the analysis with Fisher's exact test.

2.1 TC-Seq libraries

The TC-Seq datasets analyzed here are those described by Klein *et al.* (2011). These are deposited at Sequence Read Archive (SRA, `http://www.ncbi.nlm.nih.gov/sra`) under accession numbers SRA061477. These datasets are from three different translocation libraries: (i) a library from activated B cells infected with AID-expressing retrovirus (denoted hereafter as AID^{rv}), (ii) a library from AID-deficient B cells (denoted as AID$^{-/-}$) and (iii) a library from activated B cells expressing wild-type levels (AID^{wt}). Three set of curately hotspots were defined from these samples: (i) 59 physiological hotspots in Ig-genes (AID^{rv} and AID^{wt} samples), (ii) 157 off-target hotspots (in non Ig-genes from AID^{rv} and AID^{wt} samples) and iii) 34 hotspots from AID-deficient B cells (AID−/−).[1]

3 RESULTS & DISCUSSION

Common uses of k-mers include counting all distinct k-mers in genome and transcriptome assembly, variants detection and read error correction Rizk *et al.* (2013). Here, we address the problem to detect functional regions across the entire genome by searching k-mer enrichment on DNA regions.

We have applied our method to further understand the genomic complexity at recurrent translocations hotspots locus induced by activation-induced cytidine deaminase (AID). Recurrent chromosomal translocations are associated with hematopoietic malignancies such as leukemia and lymphoma and with some sarcomas and carcinomas Nussenzweig and Nussenzweig (2010).

Although AID specifically targets the immunoglobulin genes loci (*IgH*, *Igl* and *Ig*λ), it also targets a array of non-immunoglobulin genes and how nucleotide composition could impact the formation of translocations require better understanding. In this work we examine the landscape of k-mers across the physiological and off-targets translocation hotspots from AID^{wt}, AID^{rv} and AID^{ko} samples.[2]

[1] The actuall choice of the data sets and the objective for this is not clear. I am really confused. We have 3 conditions: (i)=AID^{rv}, (ii)=AID$^{-/-}$=AID^{ko}, (iii)=AID^{wt}. Fine. Now, we make use the HS of these data sets: 59 of these are in Ig genes, by joining HS from (i) and (iii), that is by gathering the HS from both conditions that include Ig genes – is this right?; 157 are in non Ig genes (again data is from the (i) and (iii) conditions); and then 34 HS from the (ii) condition. What is the purpose of this: namely to see if there is any k-mer difference between Ig HS & non-Ig HS? There could be another objective: are there any differences between a k-mer composition across (i), (ii) and (iii)? This is all not made clear and has to be explicited. What I mean is that we have to say what we are actually wanting to compare –what is the objective of all this?

[2] OK, this is the sort of phrase that holds the key to what we want to show with these examples. Still, it is not clear enough. For instance, if we want to compare k-mer profiles in HS non-HS regions in each (i), (ii) and (iii) as you state in this paragraph, why do we join the HS of *wt* and *rv*? Your phrase is OK but we need to say perhaps something more. Some further questions:

1. what happens if we do not join *wt* with *rv* and analyse the HS & non-HS profiles for the CG-mer? Is there any difference between these two groups?
2. Are the data in Figure 1(A) (i.e. for the *ko* condition) for Ig or non-Ig or both?

The results obtained by analyzing the three hotspot sets (Section 2.1) are presented in Figure 1 using 1-mer (A-mer, C-mer, G-mer and T-mer) and 2-mer (CG-mer). The nucleotide composition for each sample are remarkably different. The Figures 1D and 1F exhibit 1-mer inrichment throughout of hotspots (C and G nucleotides goes up and A and T nucleotides goes down, see Table S1 for p-values), but only in the physiological targets (Figure 1D) the enrichment is sharply in the middle of hotspots while in the off-targets is broad around the center of hotspots (Figure 1F). Interestingly, when we search for 2-mer (CG-mer), off-targets hostpots is marked by a high degree of CG-mer (p-value = 4.67179250921589e-136). This effect in off-targets AID is consistent with the theoretical mechanism for CpG-type Double Strand Breakage proposed in Tsai *et al.* (2008), when a slippage event between the top and bottom strand would place the CpG within a loop, thereby making it vulnerable to AID.[3]

According to Figure 1B, we can not observe specific preference across hotspots from AID^{ko} sample, although occurs nucleotide composition enrichment for $A|C|G|T$-mer (see Table S1 for p-values).

Following these observations, these findings strongly suggest that either $C|G$-mer or CG-mer are markers for distinct sequence-dependent mechanisms that attract the AID under physiological and overexpressed levels of AID respectively.

4 CONCLUSION

We propose a standalone method (*k-enrich*) to calculate enrichment distribution of k-mer throughout of target DNA. We use a few examples to demonstrate that *k-enrich* provide a way to investigate the genomic complexity, although our method can be applied to any nominal variable data.

REFERENCES

Agresti, A. (2012). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 3rd edition.

Craigie, R. and Bushman, F. D. (2012). HIV DNA integration. *Cold Spring Harb Perspect Med*, **2**(7), a006890.

De, S., Shaknovich, R., Riester, M., Elemento, O., Geng, H., Kormaksson, M., Jiang, Y., Woolcock, B., Johnson, N., Polo, J. M., Cerchietti, L., Gascoyne, R. D., Melnick, A., and Michor, F. (2013). Aberration in DNA methylation in B-cell lymphomas has a complex origin and increases with disease severity. *PLoS Genet.*, **9**(1), e1003137.

Hu, J., Li, B., and Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.*, **33**(15), 4899–4913.

Klein, I. A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D. F., Casellas, R., and Nussenzweig, M. C. (2011). Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell*, **147**(1), 95–106.

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L. J., Lohr, J. G., Harris, C. C., Ding, L., Wilson, R. K., Wheeler, D. A., Gibbs, R. A., Kucherlapati, R., Lee, C., Kharchenko, P. V., and Park, P. J. (2012). Landscape of somatic retrotransposition in human cancers. *Science*, **337**(6097), 967–971.

Nussenzweig, A. and Nussenzweig, M. C. (2010). Origin of chromosomal translocations in lymphoid cancer. *Cell*, **141**(1), 27–38.

Patefield, W. M. (1981). Algorithm AS 159: An efficient method of generating random $R \times C$ tables with given row and column totals. *J. Appl. Stat*, **30**(1), 91–97.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*. Springer Series in Statistics. Springer Verlag, New York.

Rizk, G., Lavenier, D., and Chikhi, R. (2013). DSK: *k*-mer counting with very low memory usage. *Bioinformatics*, **29**(5), 652–653.

Tsai, A. G., Lu, H., Raghavan, S. C., Muschen, M., Hsieh, C. L., and Lieber, M. R. (2008). Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. *Cell*, **135**(6), 1130–1142.

[3] OK, now comes the interpretation of all of this. My very imediate conclusion/questions from the data in Figure 1 are as follows.

1. CG-mer analysis: Fig (C) and Fig (E) show that non-IG AID breaks are indeed enriched in CG. This does not happens in the physiological target. Nice. But how to compare these two with Fig (A)? Again, is (A) for Ig HS, or not?
2. The description of the A, C, G, T-mers results (i.e. Fig (D) & Fig (F)) is fine! And it also connects with the profile analysis of CG for IG and non-IG.
3. To supoort further the CpG theory, would it be interesting to look after longer CG repets, i.e. for k-mers like CGCG or CGCGCG or even CGCGCGCG?
4. How does the *ko* data fits into the analysis. It is hard to tell. Maybe, we could leave this out of the paper?