Journal of Biomedical Informatics 60 (2016) 77–83

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A knowledgebase of the human Alu repetitive elements

Izaskun Mallona *, Mireia Jordà, Miguel A. Peinado

Institute of Predictive and Personalized Medicine of Cancer (IMPPC) and Health Research Institute Germans Trias i Pujol (IGTP), Can Ruti Campus. Ctra. de Can Ruti, camí de les escoles, s/n, 08916 Badalona, Spain

ARTICLE INFO

Article history:
Received 10 August 2015
Revised 20 January 2016
Accepted 22 January 2016
Available online 28 January 2016

Keywords:
Alu
Repetitive element
Knowledgebase
Ontology

ABSTRACT

Alu elements are the most abundant retrotransposons in the human genome with more than one million copies. Alu repeats have been reported to participate in multiple processes related with genome regulation and compartmentalization. Moreover, they have been involved in the facilitation of pathological mutations in many diseases, including cancer. The contribution of Alus and other repeats in genomic regulation is often overlooked because their study poses technical and analytical challenges hardly attainable with conventional strategies. Here we propose the integration of ontology-based semantic methods to query a knowledgebase for the human Alus.

The knowledgebase for the human Alus leverages Sequence (SO) and Gene Ontologies (GO) and is devoted to address functional and genetic information in the genomic context of the Alus. For each Alu element, the closest gene and transcript are stored, as well their functional annotation according to GO, the state of the chromatin and the transcription factors binding sites inside the Alu. The model uses Web Ontology Language (OWL) and Semantic Web Rule Language (SWRL). As a case of use and to illustrate the utility of the tool, we have evaluated the epigenetic states of Alu repeats associated with gene promoters according to their transcriptional activity.

The ontology is easily extendable, offering a scaffold for the inclusion of new experimental data. The RDF/XML formalization is freely available at http://aluontology.sourceforge.net/.

1. Introduction

A striking feature of most eukaryote genomes is the abundance of repetitive elements, reaching up to 80% of the total DNA content in some plants. In humans, about half of the genome is derived from repetitive elements, whereas protein coding sequences represent less than the 2%. The study of repetitive elements provides important clues on evolution and the underlying genetic mechanisms, but their functional impact on genome structure and regulation is still a matter of controversy. Moreover, genome-scale studies often overlook these elements, as they are intrinsically difficult to sequence and map. As an example, the ENCODE consortium flag paper claiming that 80% of the human genome was functional [9] disregarded the contribution of the repeats. Excellent reviews on the classification, structure and function of repeat elements have been published [35,31,15,5,3,26].

Alu is the most frequent repeat element in the human genome with more than one million copies per haploid genome. Alu elements are members of short interspersed repetitive elements (SINE). Being non autonomous retrotransposons, they produce RNA species during their life cycle and rely on other repeats to be retroprocessed. They are small ($\simeq$ 300 bp) and carry a PolIII promoter in their 5'. They harbor polyA elements and CpG domains. Their origin is the 7SL tRNA. They are flanked by short direct repeats [40].

Interestingly, Alus are not randomly distributed within the human genome, as they tend to accumulate in GC-rich regions [26] and participate in the architecture of the genome by delimiting the active/inactive domains and the epigenetic landscape [6] and gene regulation at different levels [3,4].

The advent of next generation sequencing technologies and their application to profile the genome and the epigenome of literally thousands of experimental settings offers a new opportunity to explore the structural and functional properties of Alus. Here we propose the use of ontologies to address this issue. Ontologies model the knowledge of realm while defining it in a formal manner [17] by providing a controlled vocabulary to refer explicitly to its subjects [1]. Indeed, they describe the inner properties of the system, such as the relationships between subjects. The annotated data can be stored in different exchangeable formats allowing semantically rich queries [21]. Finally, ontologies can be used for hypothesis evaluation [37].

* Corresponding author.
E-mail address: imallona@imppc.org (I. Mallona).

http://dx.doi.org/10.1016/j.jbi.2016.01.010
1532-0464/

The usage of ontologies is an emerging field in biology and biomedicine, although some controlled vocabularies are widely used. For instance, Sequence Ontology (SO) offers a hierarchy of concepts and relationships to be used to annotate genomic data; and Gene Ontology (GO) provide a set of terms to describe molecular functions, biological processes and cellular locations of genes and gene products. In this paper we report a biological ontology of human Alu repetitive elements covering their physical characteristics, their epigenetic status and the functional annotation of their nearby elements.

2. Materials and methods

The UCSC repository was queried through its MySQL public interface for Ensembl Genes, Repeat Masker, Gene Ontology, and Chromatin State Segmentation using Hidden Markov Model (HMM) from ENCODE/Broad [10]. Methylation data was retrieved from the Lister's whole genome bisulphite sequencing (WGBS) data [29]. A summary of the data origins is available as Fig. 1 and supplementary files S1 and S2.

GO Slim generic as provided by Open Biomedical Ontologies (OBO) [45] was downloaded from Gene Ontology Consortium [14]. Sequence Ontology [8] version 2.5.1 was retrieved from Sequence Ontology Consortium [42].

The OWL/RDF ontology was modelled with Protégé v4.0.2 [27] and populated with a set of custom-made bash and python scripts. Source code is available at Mallona, I. [30] under the GPL v2 terms.

DL Queries were run in a Fedora Core 14 Linux Workstation with Intel Xeon at 2.40 GHz processor and 16 GB of memory.

Statistical analysis were performed under the R environment v3.1.1. Genome-wide validations were performed using BedTools v2.19.1.

3. Approach

3.1. Ontology scope

During primates evolution, Alu elements were inserted at different evolutionary time frames. The majority of human Alus were incorporated before the divergence of human and non-human primates and are said to belong to old subfamilies [41], but others are restricted to the human lineage and some are still being amplified. As retrotransposition events are a source of genomic variation with enormous impact, we hypothesize that insertion permissiveness might be related to the genomic landscape as shaped by genetic elements and the epigenetic code. This trait is difficult to model as many features might shape the Alu insertion and selection dynamics. In our opinion, an integrative Alu knowledgebase may help to elucidate the functional implications of the Alu distribution along the genome. With this purpose in mind, we gathered structural and epigenetic properties of the human Alus.

Chromatin functional status has been described as the result of the crosstalk of epigenetic modifications (mainly histone modifications) [10], so we took the chromatin states of each Alu as a proxy to its putative functional properties. Given that Alu elements harbor active, protein-recruiting domain, we introduced sequence-based predictions of transcription factor binding sites (TFBS). Methylation status was also included as it is associated with functional repression. We took the Ensembl gene set, which includes protein-coding genes, non-coding RNAs and automatically-annotated pseudogenes, and assigned the closest one to each Alu regardless of the distance between them, as Alu tend to accumulate in GC-rich regions [26]. Finally, we also recovered from Ensembl the Gene Ontology annotation of these genes and gene products.

3.2. Modeling and formalization

Ontologies are commonly represented by using the Web Ontology Language (OWL). OWL uses formal semantics and represents them using RDF/XML-based schemata. The World Wide Web Consortium (W3C) endorses OWL and is a standard for ontology dissemination [33]; along with the Open Biomedical Ontologies (OBO) format, OWL is widely used in the biomedical field [19]. As a result, the ontology formalization produces a text file with a machine-readable syntax; moreover, the semantics is stated in such a manner that is also readable by computers.

As the repetitive elements are really abundant, the data annotated by the ontology, even regarding only the Alu elements, involve over a million instances. This fact implies that the usage of ontology viewers and reasoners might require a noticeable amount of computational resources for genome-wide queries. On the other hand, Alu elements are not randomly distributed along the genome, showing a noticeable heterogeneity between chromosomes [22]. Therefore, we splitted the Alu Ontology as a set of 24 OWL ontologies serialized in XML/RFDL, one for each canonical chromosome. As indicated at supplementary file S3, the resulting subontologies range from a couple of hundreds (chrY) to nearly half a million individuals (chr1), covering the Alu elements, genes, chromatin colors, etc. The ontologies are fully compatible with ontology viewers and formal reasoners like Pellet [20]. Splitting the information into chromosome-centered ontologies does not undermine whole-genome analysis, as the viewers and reasoners can serialize multiple ontologies at once; and allows the user to focus on a subset of chromosomes matching further requirements, such as gene content or autosomal nature.

Finally, we took advantage of the OBO initiative [20] mature ontologies to describe sequences and functional annotations,

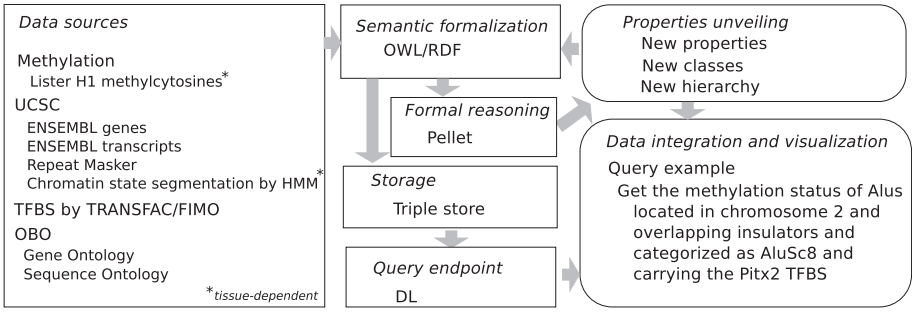

Fig. 1. The data sources include: UCSC's information on Ensembl Genes, Gene Ontology, RepeatMasker, Broad ChromHMM; Lister's data DNA methylomes at base resolution; and TRANSFAC-based FIMO transcription factor binding sites predictions. H1 hESC data was retrieved from Lister's and ChromHMM sources since methylation and chromatin states are cell-type-dependent. To perform the semantic formalization, OWL/RDF and SWRL (Semantic Web Rule Language) have been used.

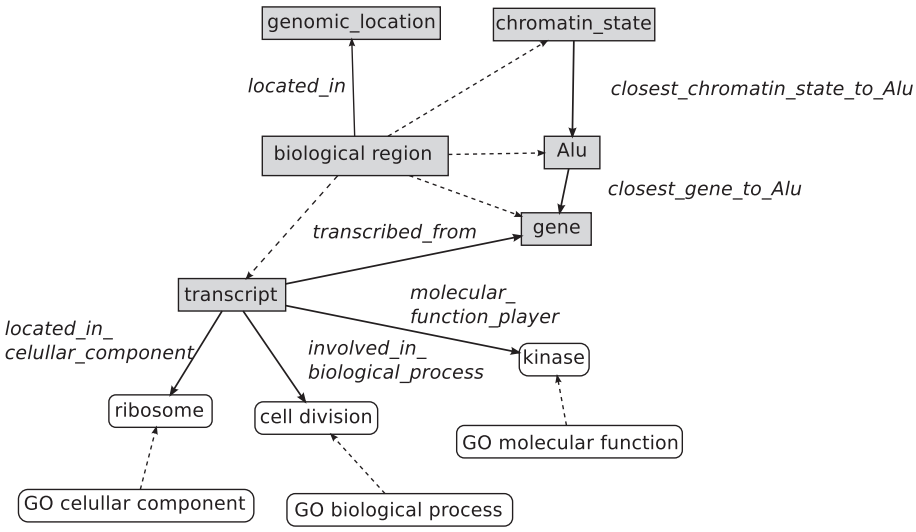

Fig. 2. Relationships between some classes of the ontology. Dashed lines indicate subclass relationships. Solid lines reflect property relationships. Italics text indicates properties. Those elements inside grey boxes are assimilable to SO terms and thus are circumscribed to genomic coordinates; round white boxes correspond to GO terms and therefore describe gene products.

incorporating subsets of the published Sequence and Gene ontologies (see Fig. 2).

3.3. Data integration

The Alu knowledgebase gathers three major types of information: data constrained to genomic coordinates that are shared across humans, the gene ontology terms, and functional data retrieved experimentally from specific conditions. The released version of the ontology covers human stem cells hESC data for the latter.

Regarding attributes primarily linked to the genomic position, we fetched the SOFA (Sequence Ontology Feature Annotation) reduced flavour of the Sequence Ontology (SO). The vocabulary addresses features of biological sequences. SOFA v 1.275 has 2432 terms that were included into the Alu ontology. SO describes the non LTR transposons (*non_LTR_retrotransposon*, SO:0000189) as a group of retrotransposons with three children, one of which, short interspersed elements (*SINE_element*, SO:0000206) is the group the Alus belong to. However, SINE elements are a terminal leaf in Sequence Ontology, meaning that they do not have descendants. Therefore, we defined *de novo* the Alu element class as a child of SINE elements.

Gene Ontology is an effort to provide a vocabulary for gene products annotation [1]. The Alu Ontology incorporates the basic Goslim that contains 527 terms; this pruned version covers biological processes, molecular functions and cellular locations.

Once defined the vocabulary for sequences and gene product functions, we integrated the genes and transcripts as defined by the Ensembl knowledgebase [13]. Ensembl offers the transcripts functional annotation using GO terms. As a gene can be transcribed to different transcripts that are annotated differently, we aggregated all the transcripts annotations and assigned them to the parent gene. The gene genomic coordinates were defined as to span all the transcribed region, thus melding the overall functional annotation.

Regarding transcription factor binding sites (TFBSs), we selected sequence-based predictions rather than tissue-specific experimental measurements. Although there are ChIP-Seq datasets that quantify transcription factor binding to the genome, the pipeline used to call the binding peaks removes repetitive elements with low mappability, including some Alu elements [48]. TFBSs prediction is based upon scanning a set of known motifs over a database of sequences. The Alu Ontology employs a whole-genome approach described by WS Noble and coworkers at [34] that used the TRANSFAC 10.12 motifs database [32] and treated each motif independently during the scanning with FIMO [16]. Although the Alu elements tend to possess histone modifications associated with open chromatin and enhancers [46], we note that the presence of the TFBS motifs solely does not ensure transcription factor binding.

We integrated two cell-line specific data layers for human stem cells addressing chromatin and methylation statuses. Chromatin functionality is said to be determined by the crosstalk of different chromatin modifications, such as the different types of histone acetylation. Ernst and Kellis [10] trained a Hidden Markov Model (HMM) segmenting the chromatin into 15 states of distinct biological functionality using epigenetic information. The HMM does not mask repetitive DNA, explicitly including it into the model. We fetched the human stem cells' HMM data and assigned to each Alu as many chromatin states as they overlap to the Alu. Methylation data was retrieved from Lister et al. [29]. Briefly, for each cytosine we calculated a methylation β value dividing the number of reads reporting a methylated status (mc value) against the number of both methylated and unmethylated reads (h value).

A summary of the data resources is available as supplementary files S1 and S2.

3.4. Production rules

Taking advantage of the Semantic Web Rule Language (SWRL), a powerful framework that associates the rule language RuleML to the OWL ontology language, we extended the axioms definition by using production rules. That is, conditional statements that build new properties chaining previously defined ones.

For instance, according to Ensembl each transcript has a list of Gene Ontology terms describing its molecular functions, biological processes and cellular locations. Given that a gene may be transcribed to one transcript or more, each of which may have different annotations, the gene receive through inheritance the annotations of all its transcripts. Finally, the Alu gathers these attributes from its closest gene. The first part of the property chain, which links genes and transcripts, is represented in Eq. (1).

$$gene(?x) \wedge transcribedTo(?x, ?y) \wedge involvedInMolecularFunction(?y, ?z) \Rightarrow involvedInMolecularFunction(?x, ?z) \quad (1)$$

The logics behind the properties and production rules affect the reasoning capabilities over the ontology. According to the presence of transitive, inverse, hierarchical and complex properties, the Alu Ontology Description Logic (DL) expressivity is SRIF(D).

4. Results

4.1. Query simplicity using DL Query

The DL syntax for data query allows easy retrieval of information from Alu instances [2]. Using a graphical user interface for ontology visualization such as Protégé, the user can ask the knowledgebase for rather elaborated queries. For instance, retrieving the Alus sitting on poised promoters containing the predicted transcription factor binding site *Pitx2* and located at chromosome 21 requires loading the chromosome ontology and querying the following:

Alu and (2)
closest_chromatin_state_to_Alu some 3_*Poised_Promoter*
and has_tfbs value "*Pitx2*"

That retrieves 7 individuals for chromosome 21; for instance, the Alu located at chr21:34481306–34481616.

This query may be expanded to retrieve those Alus that have a nearby gene with at least a transcript with intracellular signal transducer activity. The query in this case is:

Alu and (3)
closest_chromatin_state_to_Alu some 3_*Poised_Promoter*
and has_tfbs value "*Pitx2*"
and closest_gene_to_Alu some
(*gene and involved_in_molecular_function value GO* : 0005216)

That, when querying chromosome 21, produces only a hit, the Alu located at chr21:34601288–34601597.

4.2. Alus on active promoters show a distinctive methylation pattern

According to the chromatin states segmentation, promoters are categorized in three groups: active promoters, whose genes are readily transcribed; weak promoters, which control gene expression at a lower level of transcription; and poised promoters, which are key regulatory regions that readily change their transcriptional state under certain conditions. We queried the ontology for the Alus sitting on each one of the promoter types and placed at ±1 kbp of the closest gene transcription start site (taking into account the strand). The formalization of the query for poised promoters is represented in Eq. (4).

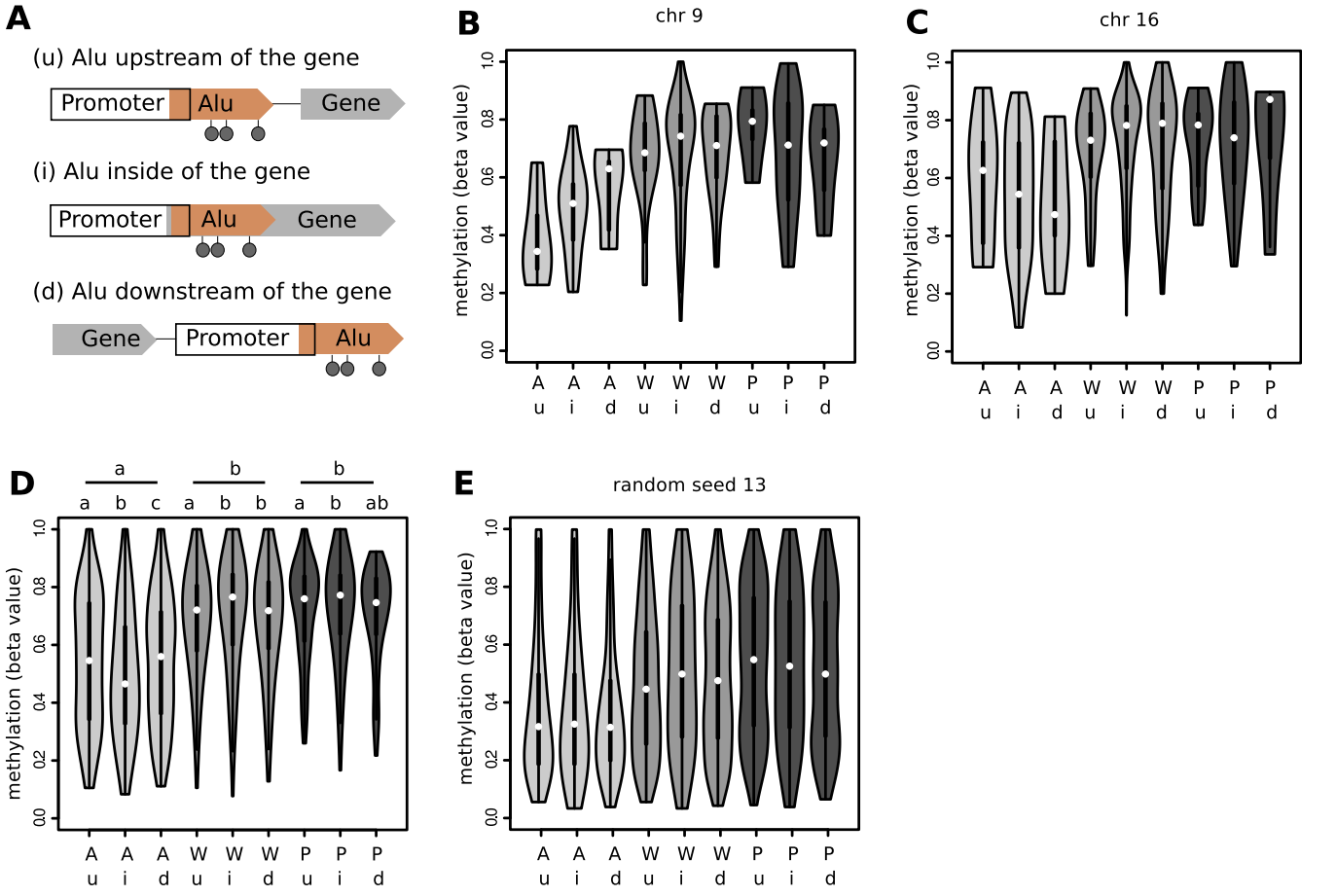

Fig. 3. Methylation of Alus in promoter chromatin states show differences according to the promoter type and location relative to the closest gene. A, Alus can be classified according to their location upstream, downstream or inside genes; and by their overlap with each of the promoter types. B and C, the Alus in each of the arrangements show intermediate methylation levels (violinplots for two randomly picked chromosomes, chr9 and chr16). D, Alus on active promoters are less methylated than those in poised and weak promoters; post hoc ranked Wilcoxon test, distinct letters describe categories with detectable differences. E, background for Alu-free regions; that is, genome-wide results after randomly allocating Alu elements along the genome (further, independent randomizations are available as supplementary file S5). Abbreviations: *A*, active promoter; *W*, weak promoter; *P*, poised promoter, *u* upstream, *i* inside, *d* downstream.

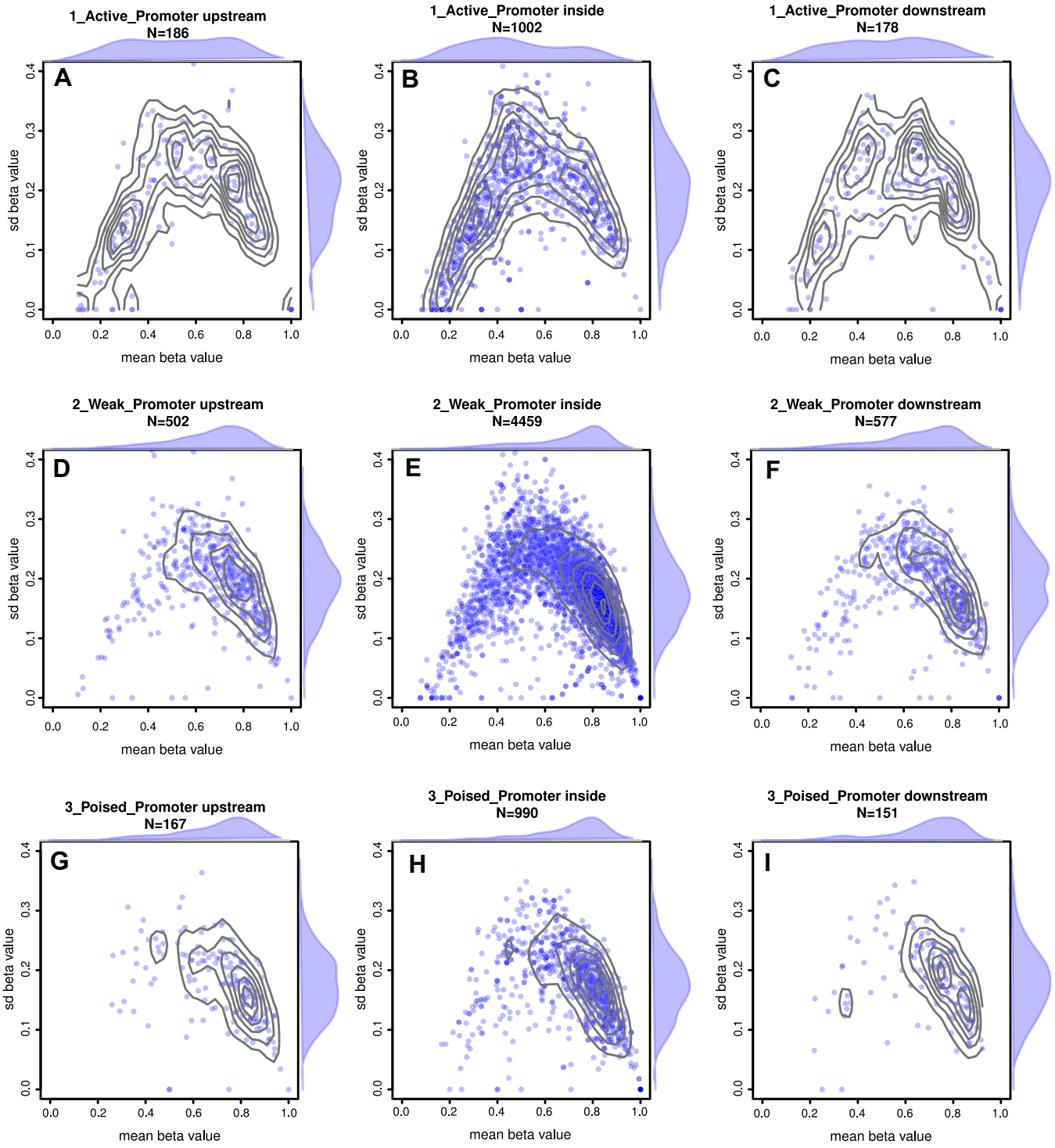

Fig. 4. Variation on DNA methylation of Alus in promoter chromatin states depends on the promoter type. The comparison of the standard deviation and the mean of the methylation of each Alu results in a bell-shaped curve, with higher variability of moderately methylated Alus and lower variability on extreme methylation values. Alus on active promoters (A, B, C) are in a continuum of lowly to high methylation statuses. However, Alus on weak (D, E, F) and poised promoters (G, H, I) show a shift towards high methylation levels. Regarding the arrangement of gene and Alu pair, upstream (A, D, G) and downstream features (C, F, I) are less represented than those in which the Alu overlaps the gene body (B, E, H). The Alu location (inside, upstream or downstream of the closest gene) has less influence in its methylation heterogeneity than the promoter type. The Alu-free background, in which the Alu elements were randomly shuffled along the genome, is available as figure S6.

Alu and
(*closest chromatin state to Alu some* 3_*Poised*_*Promoter*)
and (*has*_*distance*_*to*_*nearest*_*gene some integer* [$>$ -1000])
and (*has*_*distance*_*to*_*nearest*_*gene some integer* [$<$ 1000]) (4)

Picking two chromosomes randomly, we found Alus matching the criteria for active, weak and poised promoters (Fig. 3). The methylation of these populations of Alus did not show the characteristic bimodality of full (β value of 0.8–1) and low (β value of 0–0.2) methylation, being noticeable the abundance of CpGs with

intermediate methylation values. Alus on active promoters presented lower methylation levels, whereas those sitting on poised and weak promoters were slightly more methylated. Although noisy, the pattern was consistent across chromosomes (supplementary file S4).

In order to further explore the trend, we extended the analysis to the whole genome and detected differences between each of the Alu/gene arrangements in our query (being the promoter-flavoured Alu elements those up to 1 kbp up- or downstream, or inside a gene) (Fig. 3D). Focusing on the active promoters subgroup, the lowest methylation was found for the promoter-flavored Alus located inside a gene ($\beta = 0.56 \pm 0.29$), an intermediate for those upstream ($\beta = 0.61 \pm 0.29$) and the highest for the downstream conformation ($\beta = 0.62 \pm 0.28$) (Figs. 3 and 4). Alus on active promoters showed a significantly lower methylation than those on weak or poised promoters (post hoc pairwise Wilcoxon rank sum test, $p < 0.05$).

To obtain a random background, we randomly placed the Alu elements along the genome. More concretely, we shuffled genomic intervals matching those of the Alu elements and allocated them randomly, although impeding them to overlap and avoiding the original Alu positions. As shown in Fig. 3 and supplementary files S5 and S6, shuffled Alu elements show a marked demethylation status in active promoters, while poised and weak promoters show intermediate values. We repeated the random allocation 100 times with similar results (supplementary files S5 and S6). For the sake of reducing the computational cost and easing statistical analysis, we recodified the query with BedTools for the genome-wide analysis [36].

5. Discussion

Research projects are currently collecting data from high-throughput sequencing technologies and sharing them publicly. Data availability raises the challenge of a precise and effective integration across different data types. This demands an unambiguous mapping of the terms used, but also consistency checking. A proved useful approach to deal with this is by means of knowledgebases, that are data storages intended to offer the information in a structured manner. For instance, the European Bioinformatics Institute (EBI) is switching to semantic platforms to access its data and services [25].

The ontology presented in this work offers a machine-readable summary of the human Aluome. The structural information, i.e. the genomic coordinates, is enriched by functional data, such as DNA methylation, chromatin state, gene products function and putative TFBS. Due to the modelling procedure, the ontology offers [17,18]:

- Verification of the consistency of both data model (coherence) and data (satisfiability). For instance, the axiom 'transcripts are transcribed from transcripts' is incoherent because it does not adhere to the formal specification of the central dogma of molecular biology.
- Easy retrieval of data based in the crossmapping of the features. Gene, chromatin state or transcript-based queries are allowed, even though the knowledgebase is Alu-centered.
- A scaffold for upgrades. The highly structured nature of ontologies facilitates including further data, such as that coming from different tissues or cell lines.

We note that, as the elements included into the model (i.e. genes, Alus...) are just sequence features characterized by a genomic location (a chromosome, a start, an end), the software that populates the ontology might be run with virtually any discrete genomic feature. For instance, Alus might be replaced by LINE elements or ChIP-seq called peaks, as long as data properties remain biologically meaningful.

The Alu ontology is characterized by SRIF(D) DL expressivity and therefore is decidable [23]. However, the complexity of reasoning problems over it has a theoretical worst-case of NExpTime-hard. That implies that querying the ontology will produce theoretically an answer but with quite computational effort. For this reason, as the lower the number of instances, the lower the cost, we provided a set of ontologies rather than a whole-genome one. Nevertheless, modern DL reasoners are able to produce inferences on these NP-hard problems using optimization methods and heuristics [44,43,47]. Regarding updatability, the annotation of Alu data by the ontology is performed by a set of python scripts upon queries to the UCSC MySQL repository. Thus, the ontology can be updated with almost no effort as soon as the database at UCSC upgrades its content. As for knowledge reusability, the Alu ontology takes advantage of the well-stablished Sequence and Gene ontologies and thus is designed to readily acquire new capabilities from these sources.

Alus might contain regulatory features that alter the expression and regulation of the contiguous *loci*, even acting as promoters or enhancers [40]. As a proof of concept, we took advantage of the compartmentalization by Ernst and Kellis [10] of promoters into three major types: active, weak and poised. We queried the methylation status of those that harbor Alus and thats are close to the transcription start site. The overall trend of Alus belonging to promoter chromatin states to be partly methylated might reflect a fine regulation of repetitive elements placed in non heterochromatic DNA. Further differences in methylation of active, weak and poised promoters might indicate the complexity of Alu internalization dynamics. The low number of results for active and poised promoters matches to the expected unadaptative role of Alu insertions on active genomic compartments [39]. However, the abundance of lowly methylated Alus within active promoters compared to those in weak and poised promoters might reflect that Alu insertions can be functionalized and result in operative epigenetic signatures. It is of note the extensive DNA demethylation of repeat elements as a general feature of most cancers [38,12,24,11,28,7], and a better understanding of the contribution of Alus and other retrotransposons in genome regulation is likely to uncover new strategies for the prevention, detection and management of these diseases.

6. Conclusion

We have generated a knowledgebase integrating epigenetic and functional annotation information for human Alu repeat elements, and provided a framework to investigate the contribution of repeat elements in genome biology. The ontology can be freely accessed at http://aluontology.sourceforge.net/; the source code is available under the GPL v2 terms.

Competing interests

MAP is cofounder and equity holder of Aniling, a biotech company with no interests in this paper. The other authors declare that no competing interests exist.

Acknowledgements

The authors thank Judith Flo for her excellent technical assistance, Joan Anton Perez Braña and Jordi Conesa Caralt for their comments during ontology development and the two anonymous reviewers for their evaluation of the ontology and manuscript.

This work was supported by FEDER; the Spanish Ministry of Economy and Competitiveness [SAF2011/23638 to M.A.P.]; the Instituto de Salud Carlos III [PI14/00308 to M.J.]; and Fundació Olga Torres [to M.J.].

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2016.01.010.

References

[1] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, Gene ontology: tool for the unification of biology, Nat. Genet. 25 (1) (2000) 25–29.
[2] F. Baader, The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, 2003.
[3] R. Cordaux, M.A. Batzer, The impact of retrotransposons on human genome evolution, Nat. Rev. Genet. 10 (10) (2009) 691–703.
[4] C. Daniel, G. Silberberg, M. Behm, M. Ohman, Alu elements shape the primate transcriptome by cis-regulation of rna editing, Genome Biol. 15 (2014) R28.
[5] P.L. Deininger, J.V. Moran, M.A. Batzer, H.H. Kazazian, Mobile elements and mammalian genome evolution, Curr. Opinion Genet. Develop. 13 (6) (2003) 651–658.
[6] J.R. Edwards, A.H. O'Donnell, R.A. Rollins, H.E. Peckham, C. Lee, M.H. Milekic, B. Chanrion, Y. Fu, T. Su, H. Hibshoosh, Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns, Genome Res. 20 (7) (2010) 972–980.
[7] M. Ehrlich, Dna hypomethylation in cancer cells, Epigenomics 1 (2) (2009) 239–259.
[8] K. Eilbeck, S.E. Lewis, C.J. Mungall, M. Yandell, L. Stein, R. Durbin, M. Ashburner, The sequence ontology: a tool for the unification of genome annotations, Genome Biol. 6 (5) (2005) R44.
[9] ENCODE Project Consortium, An integrated encyclopedia of dna elements in the human genome, Nature 489 (7414) (2012) 57–74.
[10] J. Ernst, M. Kellis, Discovery and characterization of chromatin states for systematic annotation of the human genome, Nat. Biotechnol. 28 (8) (2010) 817–825.
[11] M. Esteller, Cancer epigenomics: dna methylomes and histone-modification maps, Nat. Rev. Genet. 8 (4) (2007) 286–298.
[12] A.P. Feinberg, B. Tycko, The history of cancer epigenetics, Nat. Rev. Cancer 4 (2) (2004) 143–153.
[13] P. Flicek, B. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, Ensembls 10th year, Nucl. Acids Res. 38 (Suppl. 1) (2010) D557–D562.
[14] Gene Ontology Consortium, Generic goslim by obo, 2015. <http://www.geneontology.org/GO_slims/goslim_generic.obo> (accessed: 2015-03-27).
[15] J.L. Goodier, H.H. Kazazian, Retrotransposons revisited: the restraint and rehabilitation of parasites, Cell 135 (1) (2008) 23–35.
[16] C.E. Grant, T.L. Bailey, W.S. Noble, Fimo: scanning for occurrences of a given motif, Bioinformatics 27 (7) (2011) 1017–1018.
[17] T.R. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (2) (1993) 199–220.
[18] R. Hoehndorf, M. Dumontier, G.V. Gkoutos, Evaluation of research in biomedical ontologies, Briefings Bioinform. 14 (6) (2013) 696–712.
[19] R. Hoehndorf, A. Oellrich, M. Dumontier, J. Kelso, D. Rebholz-Schuhmann, H. Herre, Relations as patterns: bridging the gap between obo and owl, BMC Bioinform. 11 (1) (2010) 441.
[20] R. Hoehndorf, K. Prufer, M. Backhaus, H. Herre, J. Kelso, F. Loebe, J. Visagie, A proposal for a gene functions wiki, in: Proceedings of OTM 2006 Workshops, Montpellier, France, October 29–November 3, Part I, Workshop Knowledge Systems in Bioinformatics, KSinBIT of Lecture Notes in Computer Science 4277, 2006, pp. 669–678.
[21] M.E. Holford, E. Khurana, K.-H. Cheung, M. Gerstein, Using semantic web rules to reason on an ontology of pseudogenes, Bioinformatics 26 (12) (2010) i71–i78.
[22] D. Holste, I. Grosse, S. Beirer, P. Schieg, H. Herzel, Repeats and correlations in human dna sequences, Phys. Rev. E 67 (6) (2003) 061913.
[23] I. Horrocks, O. Kutz, U. Sattler, The even more irresistible sroiq, KR 6 (2006) 57–67.
[24] P.A. Jones, S.B. Baylin, The epigenomics of cancer, Cell 128 (4) (2007) 683–692.
[25] S. Jupp, J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, The ebi rdf platform: linked open data for the life sciences, Bioinformatics 30 (9) (2014) 1338–1339.
[26] J. Jurka, V.V. Kapitonov, O. Kohany, M.V. Jurka, Repetitive sequences in complex genomes: structure and evolution, Annu. Rev. Genom. Hum. Genet. 8 (2007) 241–259.
[27] H. Knublauch, R.W. Fergerson, N.F. Noy, M.A. Musen, The protégé owl plugin: an open development environment for semantic web applications, in: The Semantic Web–ISWC 2004, Springer, 2004, pp. 229–243.
[28] P.W. Laird, Cancer epigenetics, Human Molec. Genet. 14 (Suppl. 1) (2005) R65–R76.
[29] R. Lister, M. Pelizzola, R.H. Dowen, R.D. Hawkins, G. Hon, J. Tonti-Filippini, J.R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, Human dna methylomes at base resolution show widespread epigenomic differences, Nature 462 (7271) (2009) 315–322.
[30] I. Mallona, Alu ontology repository, 2015. <http://sourceforge.net/projects/aluontology/> (accessed: 2015-03-27).
[31] P.K. Mandal, H.H. Kazazian, Snapshot: vertebrate transposons, Cell 135 (1) (2008). 192-192.
[32] V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, Transfac®: transcriptional regulation, from patterns to profiles, Nucl. Acids Res. 31 (1) (2003) 374–378.
[33] D.L. McGuinness, F. Van Harmelen, et al., Owl web ontology language overview, W3C Recommend. 10 (10) (2004) 2004.
[34] Noble lab, Fimo transfac prediction ws noble, 2015. <http://noble.gs.washington.edu/custom-tracks/fimo.transfac.description.html> (accessed: 2015-03-27).
[35] A.F. Palazzo, T.R. Gregory, The case for junk dna, PLoS Genet. 10 (5) (2014) e1004351.
[36] A.R. Quinlan, I.M. Hall, Bedtools: a flexible suite of utilities for comparing genomic features, Bioinformatics 26 (6) (2010) 841–842.
[37] S.A. Racunas, N. Shah, I. Albert, N.V. Fedoroff, Hybrow: a prototype system for computer-aided hypothesis evaluation, Bioinformatics 20 (Suppl. 1) (2004) i257–i264.
[38] J. Rodriguez, L. Vives, M. Jordà, C. Morales, M. Muñoz, E. Vendrell, M.A. Peinado, Genome-wide tracking of unmethylated dna alu repeats in normal and cancer cells, Nucl. Acids Res. 36 (3) (2008) 770–784.
[39] R.A. Rollins, F. Haghighi, J.R. Edwards, R. Das, M.Q. Zhang, J. Ju, T.H. Bestor, Large-scale structure of genomic methylation patterns, Genome Res. 16 (2) (2006) 157–163.
[40] D.J. Rowold, R.J. Herrera, Alu elements and the human genome, Genetica 108 (1) (2000) 57–72.
[41] A.M. Roy-Engel, M.L. Carroll, M. El-Sawy, A.-H. Salem, R.K. Garber, S.V. Nguyen, P.L. Deininger, M.A. Batzer, Non-traditional alu evolution and primate genomic diversity, J. Molec. Biol. 316 (5) (2002) 1033–1040.
[42] Sequence Ontology Consortium, Sequence ontology repository, 2015. <https://sourceforge.net/projects/alu_ontology/> (accessed: 2015-03-27).
[43] R. Shearer, B. Motik, I. Horrocks, HermiT: a highly-efficient OWL reasoner, in: A. Ruttenberg, U. Sattler, C. Dolbear (Eds.), Proc. of the 5th Int. Workshop on OWL: Experiences and Directions (OWLED 2008 EU), Karlsruhe, Germany, October 26–27, 2008.
[44] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, Y. Katz, Pellet: a practical owl-dl reasoner, Web Semant.: Sci. Services Agents World Wide Web 5 (2) (2007) 51–53.
[45] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, The obo foundry: coordinated evolution of ontologies to support biomedical data integration, Nat. Biotechnol. 25 (11) (2007) 1251–1255.
[46] M. Su, D. Han, J. Boyd-Kirkup, X. Yu, J.-D. Han, Evolution of alu elements toward enhancers, Cell Rep. 7 (2) (2014) 376–385.
[47] D. Tsarkov, I. Horrocks, Fact++ description logic reasoner: system description, in: Automated Reasoning, Springer, 2006, pp. 292–297.
[48] K. Yip, C. Cheng, N. Bhardwaj, J. Brown, J. Leng, A. Kundaje, J. Rozowsky, E. Birney, P. Bickel, M. Snyder, M. Gerstein, Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors, Genome Biol. 13 (9) (2012) R48.