Identifying Significant Correlations Between Histone Modifications and Ultraconserved Genomic Regions

Armando A. Banuelos[*,1], **Alexandra M. Porter**[†], **Anjini Karthik**[‡], **Effie Nehoran**[§] **and Laura Miron**[**]

[*]B.S. Computer Science with specialization in AI at Stanford University, [†]Ph.D. Computer Science at Stanford University, [‡]B.S. Computer Science with specialization in Biocomputation at Stanford University, [§]B.S. Computer Science with specialization in AI at Stanford University, [**]M.S. Computer Science with specialization in AI at Stanford

ABSTRACT Investigating the significant correlation between histone modification regions and ultraconserved genomic regions via four statistical evaluative mediums (Simple Counting, Uniform Random Permutation, K-Clusertering, and Synthetic Data Generation) proposed the idea that conservation of histone H327AC is highly correlated to DNA promoter binding sites, in particular with the H3K36me3 promoter binding site. As indicated by previous studies, promoter binding site conservation is due to conservation of transcription factor binding sites (TFBS) across species. This paper attempts to cross-validate and expand on previous literature by examining conservation across the five varied cell types in the hg19 human genome, Mm mice genome, and Can canine genome along with the intersectionality between H327AC and H3K36me3 using UCSC Genome Browser Tracks.

KEYWORDS histone modifications; ultra-conserved regions; k-clustering, simple counting, uniform random permutation, synthetic data generation

Introduction

Scholarly research investigating the conservation of histone modifications stems from the need to decode the evolution of disease and viruses within and outside clades of species. Examples include using comparative genomics to understand how coding (exonic) and non-coding (non-exonic) areas on the human genome formulate strong correlations to psychiatric disorders such as Rett syndrome[1]. Decoding the non-coding genome and elucidating genetic risk or disease from correlation is not the only way to discover conservation of histone modifications. In-vivo modifcations of embryonic eukaryotic organisms are also pathways of research in this area. One such study explored the conservation of histone modifications in zebrafish embryos by measuring how known disease-associated variants impacted transference of histone modifications in the development of said organism[2]. Despite tampering with disease associated variants, 60% of conserved enhancers expressed similar chromatin expression indicating strong conservation of histone modifications. Conservation of histone modification is also explored through locating how particular histones interact with specific areas on the known human genome that is known to express conservation. For instance, being able to see how proximity to transcription factor binding sites of certain histone modifications leads to their subsequent conservation[3]. This paper attempts to use all three pathways of research to investigate whether promoter binding sites correlate significantly with the conservation of histone modifications.

Abstract

By investigating the intersection of five differentiated cell types, namely muscle cells (muschlehsmm), umbilical cells (umbilicalhuvec), brain stem cells (stemh1), lung cells (lungnhlf) and arbitrary epidermal cells (skinnhek) on chromosome 17 for mice, canines, and humans we will derive statistical correlation between said intersection via four evaluative measures – Uniform Random Permutation, Synthetic Data Creation, K-Clustering, and Simple Counting. As a biological confirmation of statistical correlation from the previous step, this research paper intends to explore how these identified correlations relates to correlation yielded from the intersection of the H327AC histone modification and the H3K36me3 promoter binding site.

[1]Correspondence: abanuelo@stanford.edu

[1] Barr, C. L., and V. L. Misener. "Decoding the Non-Coding Genome: Elucidating Genetic Risk Outside the Coding Genome." Genes, brain, and behavior 15.1 (2016): 187–204. PMC. Web. 10 Mar. 2018.

[2] Miguel Escalada, Irene (2014) Functional analysis of human enhancers using the zebrafish embryo. Ph.D. thesis, University of Birmingham.

[3] Yong H. Woo, Wen-Hsiung Li; Evolutionary Conservation of Histone Modifications in Mammals, Molecular Biology and Evolution, Volume 29, Issue 7, 1 July 2012, Pages 1757–1767, https://doi.org/10.1093/molbev/mss022

Methods

Since four different methods of statistical evaluation were used to measure correlation between histone modifications (H327AC), differentiated cell types in humans, mice, and canines, and promoter binding sites (H3K36me3), this paper will walk though each method as following: (1) Creation of Synthetic Data, (2) K-Clustering, (3) Uniform Random Permutation, and (4) Simple Counting.

Creation of Synthetic Data

Creation of synthetic human genome chromosome data allows for the randomization of upstream and downstream cross-sections of the human genome. For evaluative performance, hg19.bed human chromosome 17 (chrm17) was extracted using the UCSC Genome Browser Table Browser tool. Using starting and ending coordinates of the hg19 chrm17, the following pseudo-algorithm was implemented to randomize nucleotide base windows across the chromosome to normalize for outliers across the chromosome when extracting correlation between the synthetic data and the histone modification region H327AC.

```
$\epsilon$ = 1
    $k$ = randint(0, 200)
for each chrm start coord ($i$) do
if $i \geq maxval$ then
    $i \leftarrow 0$
else
    if $i + k \leq maxval$ then
        $i \leftarrow i + k$ or $i \leftarrow i - k$
$\epsilon$ = rand(0,1)
$i$ = $\epsilon$ X $i$
return $i$
```

The synthetic creation algorithm above works by either stretching or shrinking the nucleotide base pair window for the start coordinate and randomizing its location across chromosome 17 via value ϵ.

K Clustering

It is important to indicate what statistical analysis has been performed; not just the name of the software and options selected, but the method and model applied. In the case of many genes being examined simultaneously, or many phenotypes, a multiple comparison correction should be used to control the type I error rate, or a rationale for not applying a correction must be provided. The type of correction applied should be clearly stated. It should also be clear whether the p-values reported are raw, or after correction. Corrected p-values are often appropriate, but raw p-values should be available in the supporting materials so that others may perform their own corrections. In large scale data exploration studies (e.g. genome wide expression studies) a clear and complete description of the replication structure must be provided.

Data Availability

At the end of the Materials and Methods section, include a statement on reagent and data availability. Please read the Data and Reagent Policy before writing the statement. Make sure to list the accession numbers or DOIs of any data you have placed in public repositories. List the file names and descriptions of any data you will upload as supplemental information. The statement should also include any applicable IRB numbers. You may include specifications for how to properly acknowledge or cite the data.

For example: Strains are available upon request. File S1 contains detailed descriptions of all supplemental files. File S2 contains SNP ID numbers and locations. File S3 contains genotypes for each individual. Sequence data are available at GenBank and the accession numbers are listed in File S3. Gene expression data are available at GEO with the accession number: GDS1234. Code used to generate the simulated data is provided in file S4.

Results and Discussion

The results and discussion should not be repetitive. The results section should give a factual presentation of the data and all tables and figures should be referenced; the discussion should not summarize the results but provide an interpretation of the results, and should clearly delineate between the findings of the particular study and the possible impact of those findings in a larger context. Authors are encouraged to cite recent work relevant to their interpretations. Present and discuss results only once, not in both the Results and Discussion sections. It is sometimes acceptable to combine results and discussion. The text should be as succinct as possible. Heed Strunk and White's dictum: "Omit needless words!"

Additional guidelines

Numbers

In the text, write out numbers nine or less except as part of a date, a fraction or decimal, a percentage, or a unit of measurement. Use Arabic numbers for those larger than nine, except as the first word of a sentence; however, try to avoid starting a sentence with such a number.

Units

Use abbreviations of the customary units of measurement only when they are preceded by a number: "3 min" but "several minutes". Write "percent" as one word, except when used with a number: "several percent" but "75%." To indicate temperature in centigrade, use ° (for example, 37°); include a letter after the degree symbol only when some other scale is intended (for example, 45°K).

Nomenclature and Italicization

Italicize names of organisms even when when the species is not indicated. Italicize the first three letters of the names of restriction enzyme cleavage sites, as in HindIII. Write the names of strains in roman except when incorporating specific genotypic designations. Italicize genotype names and symbols, including all components of alleles, but not when the name of a gene is the same as the name of an enzyme. Do not use "+" to indicate wild type. Carefully distinguish between genotype (italicized) and phenotype (not italicized) in both the writing and the symbolism.

Cross References

Use the `\nameref` command with the `\label` command to insert cross-references to section headings. For example, a `\label` has been defined in the section **??**.

In-text Citations

Add citations using the `\citep{}` command, for example (Neher and Hallatschek 2013) or for multiple citations, (Neher and Hallatschek 2013; Rödelsperger *et al.* 2014)

Examples of Article Components

The sections below show examples of different header levels, which you can use in the primary sections of the manuscript (Results, Discussion, etc.) to organize your content.

First level section header

Use this level to group two or more closely related headings in a long article.

Second level section header

Second level section text.

Third level section header: Third level section text. These headings may be numbered, but only when the numbers must be cited in the text.

Figures and Tables

Figures and Tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

Sample Figure

Figure 1 shows an example figure.

Sample Video

Figure 2 shows how to include a video in your manuscript.

Sample Table

Table 1 shows an example table. Avoid shading, color type, line drawings, graphics, or other illustrations within tables. Use tables for data only; present drawings, graphics, and illustrations as separate figures. Histograms should not be used to present data that can be captured easily in text or small tables, as they take up much more space.

Tables numbers are given in Arabic numerals. Tables should not be numbered 1A, 1B, etc., but if necessary, interior parts of the table can be labeled A, B, etc. for easy reference in the text.

Sample Equation

Let $X_1, X_2, \dots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i \tag{1}$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

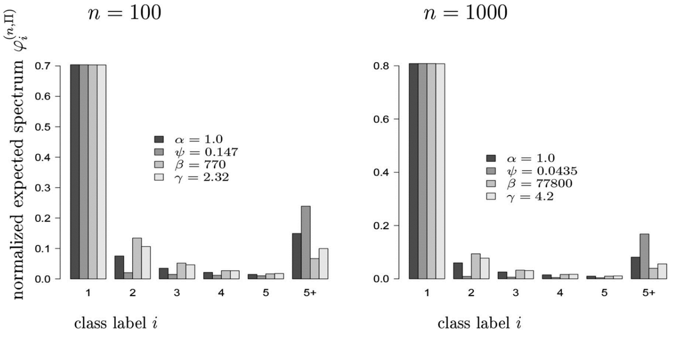

Figure 1 Example figure from 10.1534/genetics.114.173807. Please include your figures in the manuscript for the review process. You can upload figures to Overleaf via the Project menu. Upon acceptance, we'll ask for your figure files to be uploaded in any of the following formats: TIFF (.tiff), JPEG (.jpg), Microsoft PowerPoint (.ppt), EPS (.eps), or Adobe Illustrator (.ai). Images should be a minimum of 300 dpi in resolution and 500 dpi minimum if line art images. RGB, CMYK, and Grayscale are all acceptable. Halftones should be high contrast with sharp detail, because some loss of detail and contrast is inevitable in the production process. Figures should be 10-20 cm in width and 1-25 cm in height. Graph axes must be exactly perpendicular and all lines of equal density. Label multiple figure parts with A, B, etc. in bolded type, and use Arrows and numbers to draw attention to areas you want to highlight. Legends should start with a brief title and should be a self-contained description of the content of the figure that provides enough detail to fully understand the data presented. All conventional symbols used to indicate figure data points are available for typesetting; unconventional symbols should not be used. Italicize all mathematical variables (both in the figure legend and figure) , genotypes, and additional symbols that are normally italicized.

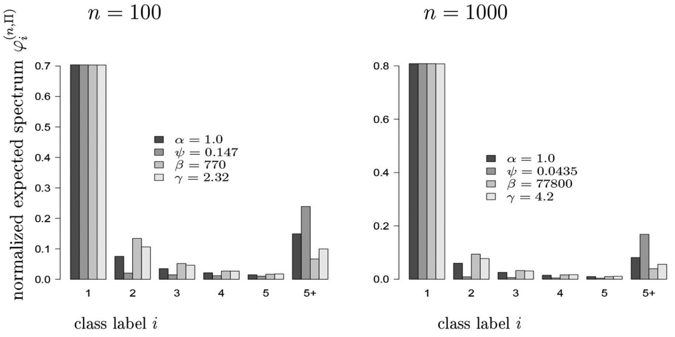

Figure 2 Example movie (the figure file above is used as a placeholder for this example). *GENETICS* supports video and movie files that can be linked from any portion of the article - including the abstract. Acceptable formats include .asf, avi, .wav, and all types of Windows Media files.

Table 1 **Students and their grades**

Student	Grade[a]	Rank	Notes
Alice	82%	1	Performed very well.
Bob	65%	3	Not up to his usual standard.
Charlie	73%	2	A good attempt.

[a] This is an example of a footnote in a table. Lowercase, superscript italic letters (a, b, c, etc.) are used by default. You can also use *, **, and *** to indicate conventional levels of statistical significance, explained below the table.

Literature Cited

Neher, R. A. and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proceedings of the National Academy of Sciences **110**: 437–442.

Rödelsperger, C., R. A. Neher, A. M. Weller, G. Eberhardt, H. Witte, *et al.*, 2014 Characterization of genetic diversity in the nematode pristionchus pacificus from population-scale resequencing data. Genetics **196**: 1153–1165.