RESEARCH ARTICLE

Chromas from chromatin: sonification of the epigenome [version 1; referees: 1 approved, 2 approved with reservations]

Davide Cittaro[1], Dejan Lazarevic[1], Paolo Provero[1,2]

[1]Center for Translational Genomics and Bioinformatics, San Raffaele Hospital- via olgettina 58, 20138 Milano, Italy
[2]Department of Molecular Biotechnology and Life Sciences, University of Turin - via Nizza 52, 10126 Torino, Italy

v1 **First published:** 03 Mar 2016, **5**:274 (doi: 10.12688/f1000research.8001.1)
Latest published: 03 Mar 2016, **5**:274 (doi: 10.12688/f1000research.8001.1)

Abstract

The epigenetic modifications are organized in patterns determining the functional properties of the underlying genome. Such patterns, typically measured by ChIP-seq assays of histone modifications, can be combined and translated into musical scores, summarizing multiple signals into a single waveform. As music is recognized as a universal way to convey meaningful information, we wanted to investigate properties of music obtained by sonification of ChIP-seq data. We show that the music produced by such quantitative signals is perceived by human listeners as more pleasant than that produced from randomized signals. Moreover, the waveform can be analyzed to predict phenotypic properties, such as differential gene expression.

Open Peer Review

Referee Status: ? ? ✓

	Invited Referees 1	2	3
version 1 published 03 Mar 2016	? report	? report	✓ report

1 **Tao Liu**, University at Buffalo USA, University at Buffalo USA

2 **Ho-Ryun Chung**, Max Planck Institute for Molecular Genetics Germany

3 **Federico M Giorgi**, Columbia University USA

Discuss this article

Comments (0)

Corresponding author: Davide Cittaro (cittaro.davide@hsr.it)

How to cite this article: Cittaro D, Lazarevic D and Provero P. **Chromas from chromatin: sonification of the epigenome [version 1; referees: 1 approved, 2 approved with reservations]** *F1000Research* 2016, **5**:274 (doi: 10.12688/f1000research.8001.1)

Grant information: The author(s) declared that no grants were involved in supporting this work.

Competing interests: No competing interests were disclosed.

First published: 03 Mar 2016, **5**:274 (doi: 10.12688/f1000research.8001.1)

Introduction

Sonification is the process of converting data into sound. Sonification itself has a long, yet punctuated, story of applications in molecular biology, several algorithms to translate DNA[1] or protein sequences[2,3] to musical scores have been proposed. The same principles have also been extended to the analysis of complex data[4] showing that, all in all, sonification can be used to describe and classify data. This approach is sustained by the idea that music is acknowledged as a way to deliver information[5]. Indeed, the very same procedures may also be applied for recreational purposes.

One of the limitations of sonification of actual DNA and protein sequences is their intrinsic conservative nature. Assuming the differences in two individual genomes are, on average, one nucleotide every kilobase[6], the corresponding musical scores would have little differences.

On the contrary, dynamic ranges typical of transcriptomic and epigenomic data may provide a richer source for sonification.

In this work we describe an approach to convert ChIP-seq signals, and in principle any quantitative genomic feature, into a musical score. We started working on our approach for amusement mainly, and we realized that the sonificated chromatin signals were surprisingly harmonious. We then tried to assess some properties of the music tracks we were able to generate. We show that the emerging sounds are not random and instead appear more melodious and tuneful than music generated from randomized notes. We also show that different ChIP-seq signals can be combined into a single musical track and that tracks representing different conditions can be compared allowing for the prediction of differentially expressed genes.

Examples of sonification for various genomic loci are available at https://soundcloud.com/davide-cittaro/sets/k562.

Definitions

MIDI: MIDI (Musical Instrument Digital Interface) is a standard that describes protocols for data exchange among a variety of digital musical instruments, computers and related devices. MIDI format encodes information about note notation, pitch, velocity and other parameters controlling note execution (*e.g.* volume and signals for synchronization).

MIDI file format: a binary format representing MIDI data in a hierarchical set of objects. At the top of hierarchy there is a Pattern, which contains a list of Tracks. A Track is a list of MIDI events, encoding for note properties. MIDI events happen at specific time, which is always relative to the start of the track.

MIDI Resolution: resolution sets the number of times the status byte is sent for a quarter note. The higher the resolution, the more natural the sound is perceived. Resolution is the number of Ticks per quarter note. At a specific resolution R, Tick duration in microseconds T is related to tempo (expressed in Beats per Minute, BPM) by the following equation

$$T = \frac{60R}{BPM}$$

Results

Approach

In order to translate a single ChIP-seq signal track to music we bin the signal over a specified genomic interval (*i.e.* chrom:start-end) into fixed-size windows (*e.g.* 300 bp) and note duration will be proportional to the size of such windows. As we are dealing with MIDI standard, we let the user specify track resolution and the number of ticks per window (see Definitions); the combination of these parameters defines the duration of a single note. The default parameters associate a bin of 300 bp with one quaver (1/8 note).

In order to define the note pitch, we take the logarithm of the average intensity of the ChIP-seq signal in a genomic bin. The sounding range of the whole signal is discretized in a predefined number of semitones. At default parameters, the range is binned into 52 semitones, covering four octaves. In order to introduce pauses, the lowest bin of the signal range represents a rest. If two consecutive notes or rests fall in the same bin, we merge them in one note doubling its duration.

Using this approach, any ChIP-seq signal can be mapped to a chromatic scale. We implemented the possibility to map a signal on a different scale (major, minor, pentatonic…); to this end, intensity bin boundaries are merged according to the definition of a specific scale (Figure 1). MIDI tracks produced in this way can be then imported into a sequencer software where they can be further processed, setting tempo and time signature.

Music produced from chromatin marks is not perceived as a random pattern

In order to test whether sonification of chromatin marks are perceived as random patterns, we selected ten genomic regions and generated corresponding tracks based on the following histone modifications: H3K27me3, H3K27ac, H3K9ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3 (Supporting Audio files S1.1 to S10.1). For the same regions, we randomized genomic signal at base and bin level (Supporting Audio files S1.2 to S10.2). When data are randomized at base level, the average intensity is uniforms across the bins, resulting in a repeated note; this is largely expected as ChIP-seq signals are distributed on the genome according to a Poisson law[7] or, more precisely, to a Negative Binomial law[8].

Randomization at bin level, instead, equals to shuffling notes during the execution. We administrated a questionnaire to a set of volunteers (n=8) not previously tested for education in music. Volunteers were asked to listen to each pair of original/random track and choose which track they felt was more appealing. Track order was randomized when testing different volunteers. Notably, in the majority of the cases (62/80) the music generated from genomic signal without randomization was judged more appealing. Results are significant to a Fisher-exact test (p=1.95e-3), suggesting that genomic signals contain information that can be recognized by human ear. The number of correct answers for each volunteer ranged from 5 to 10, with a median value of 8.

Differences in musical tracks reflect differences in gene expression

Once we assessed the existence of musical patterns in genomics signals, we were keen to explore if this kind of information could

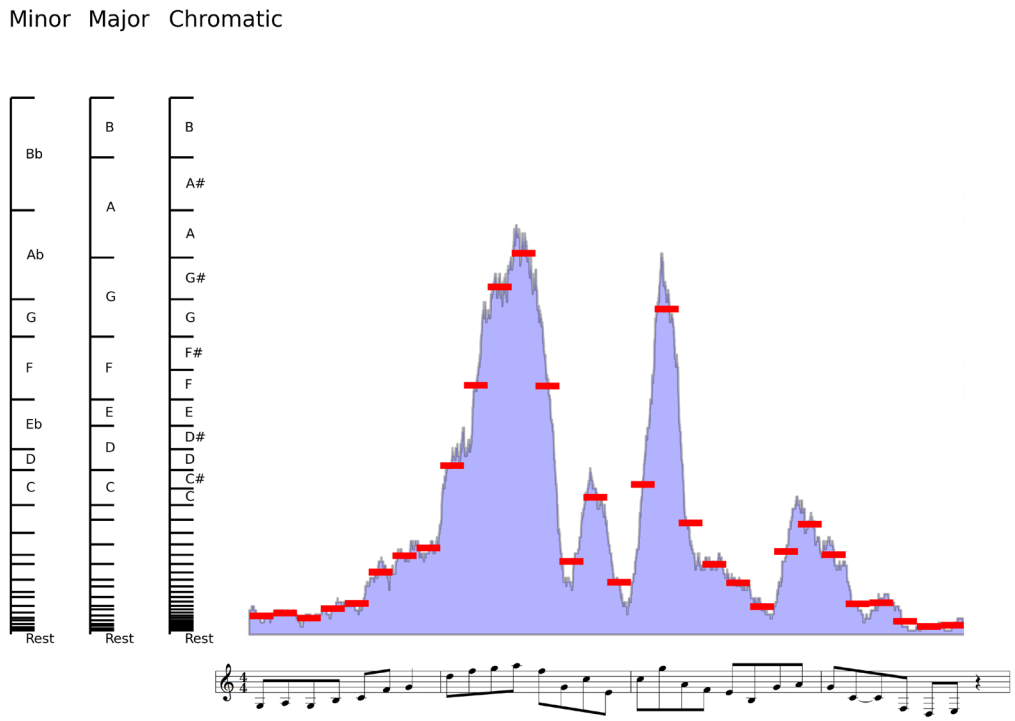

Figure 1. Graphical representation of the approach used to transform quantitative signals to music. ChIP-seq values (H3K4me3 in the example) are binned in fixed-size intervals over the genome. Each interval corresponds to a 1/8 note. Average values of log-transform of read counts in each genomic bin (red lines) are matched into predefined number of semitones (chromatic scale). Notes may be mapped to a specified scale (major and minor scales are exemplified in the figure). Consecutive equal notes are merged in single note with double duration. Values falling in the first bin are considered rests.

be exploited to identify biological features of samples. Since the epigenetic DNA modifications reflected by histone marks influence gene expression[9], we tested if differences in musical tracks generated from various ChIP-seq signals reflects differences in gene expression of the corresponding loci. To this end, we downloaded ChIP-seq marks (H3K27me3, H3K27ac, H3K9ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, Pol2b) and RNA-seq data for K562 and NHEK cell lines from the ENCODE project[10]. For each RefSeq locus we converted ChIP-seq signals to music with fixed parameters (see Methods). RNA-seq data were used to identify genes that are differentially expressed between the two cell lines, under a p-value <0.01 and $|\log FC|>1$, according to recent SEQC recommendations[11].

A common way to classify music is based on summarization of track features after spectral analysis[12,13]. Such approach involves the summarization of track as Mel-Frequency Cepstral Coefficients (MFCC) that are subsequently clustered using Gaussian Mixture Models (GMM). A distance between tracks can then be defined as described in 14, who used it as a classifier for musical genres.

We tested if a similar approach could be used to develop a predictor of differential expression based on the distance between musical tracks generated from two cell lines.

We defined a distance between songs as described in methods and we optimized the parameters using as a training set the 250 genes with the most significant differential expression p-value and as many genes with the least significant p-value according to RNA-seq (Figure 2). We found that optimal performance is at MFCC=30 and GMM=10, with an AUC=0.609.

We summarized tracks representing all RefSeq genes using such parameters, we then compared distances with differential expression performing a ROC analysis. Our results indicate that differences in information contained in musical representation of chromatin signals may be linked to differential expression, although power of prediction is limited (AUC=0.5184, p=1.4597e-03).

Similarity between musical tracks overlaps similar biological properties

An additional issue we wanted to assess was if similarities between musical representation of chromatin status may be linked to the biology of the underlying genes. To this end, we calculated pairwise distances for all regions using parameters identified above on K562 cell line. Hierarchical clustering of the distance matrix identifies eight major clusters (Figure 3, left). We performed Gene Ontology Enrichment analysis on each cluster, here represented as word cloud of significant terms (Figure 3, center, Supplementary table 1); we

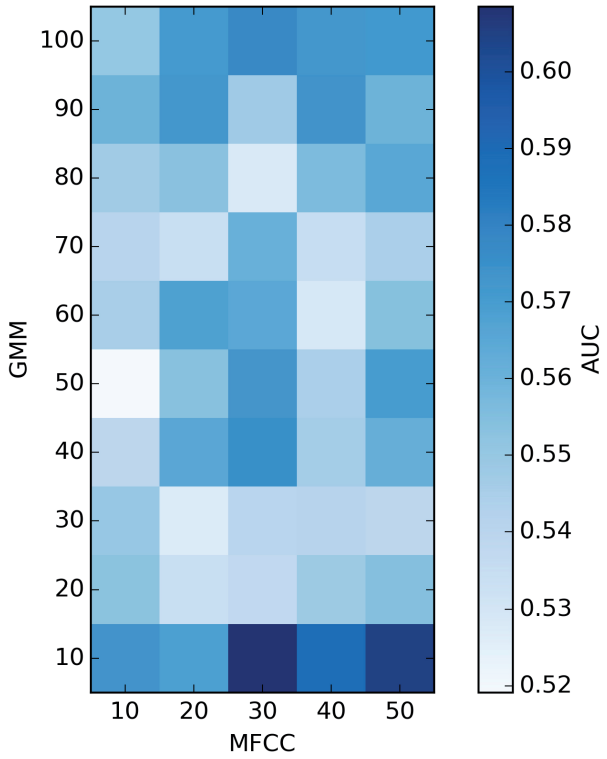

Figure 2. Evaluation of different combination of parameters in predicting differential gene expression on a gold-standard subset of 500 genes. Each square corresponds to a number of Mel-Frequency Cepstral Coefficients (MFCC) used to summarize signal and a number of centers for Gaussian Mixture Model (GMM). Colors are given by the corresponding Area Under the Curve (AUC).

found that different clusters are linked to genes showing different biological properties. For example, some clusters (6, 7 and 8) were linked to regulation of cell cycle, others were linked to metabolic processes (2 and 5) or vesicle transport (3 and 4). We also evaluated the distribution of expression (expressed as log(RPKM)) of the underlying genes (Figure 3, right); we found that regions clustered by the distance between musical tracks broadly reflects groups of genes with different level of expression, spotting clusters of higher expression (cluster 5) or lower expression (clusters 2 and 3); assessment of statistical significance of differences in distribution of gene expression values among clusters is presented in Table 1.

Discussion

Chromatin shape and genome function are governed, among several factors, by the coordinated organization of epigenetic marks[15]. Modifications of such marks are dynamic and are fine-tuned during the life of a cell or an organism. Analysis of histone modifications, as well as transcription factors and other proteins binding DNA, by ChIP-seq already described patterns of enrichment that are specific to their relative function[16,17]. Analysis of combinatorial patterns of histone modifications already unveiled its potential in understanding functional properties of the genome[18,19] and the cross-talk among multiple chromatin marks[20].

We show, in this work, that the information carried by multiple histone modifications can be caught in a human-friendly way by translating ChIP-seq signals into musical scores. Although the investigation of the psychological factors that underlie tuneful perception of sonificated genomic signals is out of the scope of this manuscript, our results suggest that human hearing is able to perceive patterns conveying information encoded in ChIP-seq data analyzed and to distinguish from random noise.

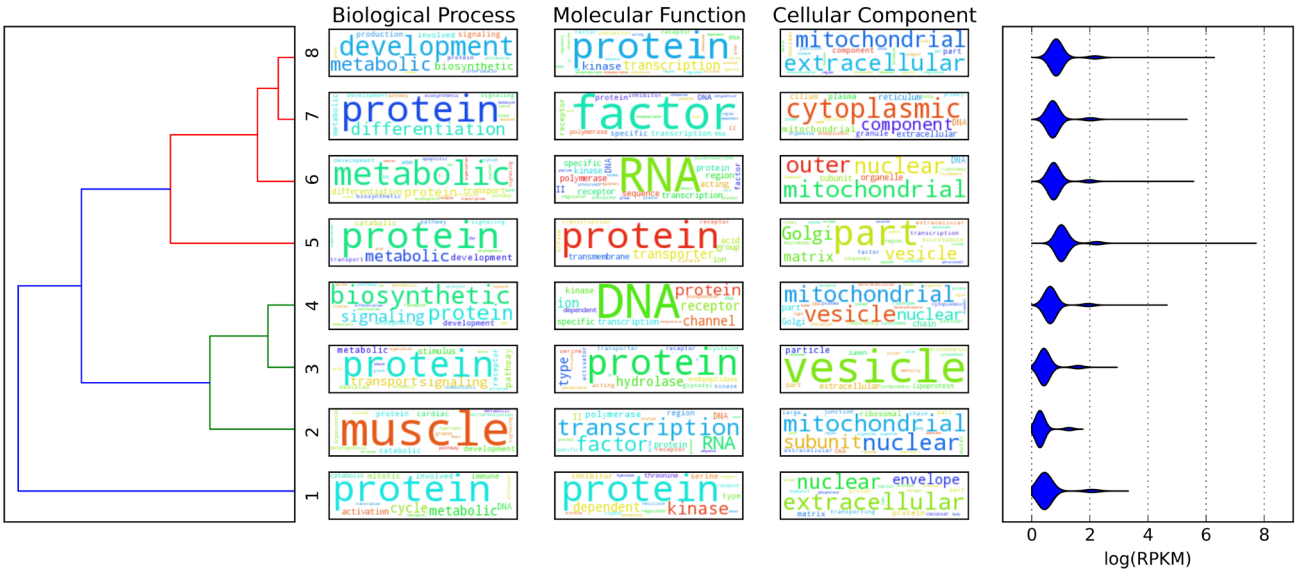

Figure 3. Hierarchical clustering of genomic regions identifies 8 main clusters (left). Each cluster broadly corresponds to specific biological properties according to Gene Ontology enriched terms (middle). Level of expression of genes included in each cluster show specific distributions (right).

Table 1. *p*-values of Mann-Whitney U-test for differences in distribution of expression between clusters. Tests showing significant difference ($p \leq 0.05$) are presented in bold face.

	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
Cluster 1	**2.635e-23**	**3.548e-18**	**8.127e-37**	**8.627e-125**	**2.859e-83**	**3.231e-71**	**3.169e-63**
Cluster 2		1.068e-01	2.008e-01	4.005e-01	2.801e-01	7.907e-02	9.467e-02
Cluster 3			2.880e-01	**2.225e-02**	1.399e-01	3.910e-01	3.745e-01
Cluster 4				**4.159e-02**	3.025e-01	2.614e-01	2.821e-01
Cluster 5					**5.098e-04**	**3.898e-09**	**7.586e-07**
Cluster 6						**1.259e-02**	**2.866e-02**
Cluster 7							4.545e-01

We automated the analysis of differences between musical tracks using an established method based on summarization of spectral data. By this approach, we investigated the possible link between differences in ways chromatin sounds and phenotypic features. Our results suggest that differences in transcript levels can be predicted by the differences of sonificated genomic regions, although performances of such approach are limited. We reasoned that many factors may explain such poor results: first of all there is a vast space of parameters that can be tuned to create a single musical track and we still lack methods to explore it efficiently. In addition, the Mel scale used to summarize audio signal has been developed to match human capabilities to perceive sound[21], hence it may not be optimal for the comparison of the tracks generated in this work.

It has already been shown that it is possible to predict levels of gene expression starting from chromatin states, although the method used to perform chromatin segmentation has a large impact on such predictions[22]. In this work we found that differences in chromatin-derived music reflects, to some extent, differences in the level of expression of underlying genes and their related biology.

To conclude, although we cannot advocate the usage of musical analysis as universal tool to analyze biological data yet, we confirm that quantitative features on the genome are patterned and contain information, hence can be converted into sounds that are perceived as musical. We limited our analysis on specific chromatin modifications, but in principle any quantitative genomic feature can be converted and integrated into a musical track. The choice of parameters and instruments has been standardized for the analysis presented, for illustrative purpose we show that different signals from the same region can be combined using different instruments (https://soundcloud.com/davide-cittaro/random-locus-blues) and signals from different genomic regions can be merged (https://soundcloud.com/davide-cittaro/non-homologous-end-joining).

Materials and methods

Sonification of ENCODE ChIP-seq data

Raw data for various modifications were downloaded from GEO database (GSE26320, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26320). Read tags were aligned to human genome (hg19) using BWA aligner v0.7.2[23]. Alignments were converted to bigwig tracks[24] after filtering for duplicates and quality score higher than 15:

```
for file in *.bam
do
  samtools view -q 15 -F 0x400 -u $file | \
  bedtools bamtobed -i stdin | \
  bedtools slop -i stdin -l 0 -r 250 -g hg19.chromSizes | \
  bedtools genomecov -g hg19.chromSizes -i stdin -bg | \
  wigToBigWig  stdin  hg19.chromSizes ${file%%.bam}.bigwig
done
```

In order to define regions to be converted to music scores, we selected intervals around RefSeq gene definition, from 1kb upstream of TSS to 2kb downstream of TES. ChIP-seq signals were firstly converted to MIDI using custom scripts (https://bitbucket.org/dawe/enconcert) according to parameters defined in Table 2. MIDI tracks belonging to the same region from the same sample were merged into a single MIDI file, converted to WAV format using timidity software v2.14.0 (http://timidity.sourceforge.net), with the exception of tracks presented as Supplementary audio files which have

Table 2. Parameters used to convert different ChIP-seq signals into corresponding musical tracks.

Antibody	Scale	Octave	Key	Tick Size	Bin Size
H3K27me3	minor	4	B	600	400
H3K27ac	minor	3	B	900	600
H3K9ac	minor	3	B	1200	800
H3K36me3	minor	4	B	300	200
H3K4me1	minor	3	B	1200	800
H3K4me2	minor	3	B	300	200
H3K4me3	minor	3	B	300	200
H3K9me3	minor	4	B	300	200
Pol2	minor	4	B	600	400

been processed with GarageBand software v10.1.0 (Apple Inc., Cupertino, USA).

Comparison of WAV tracks

In order to compare four samples for each converted genomic region, we extracted MFCC using python_speech_features library (https://github.com/jameslyons/python_speech_features). Selected components were then clustered using Gaussian Mixture Models, implemented in scikit-learn python library 0.15.2 (http://scikit-learn.org). Distances between two tracks were evaluated using Hausdorff distance (H) between GMM clusters. Briefly, we first calculate all pairwise distances between GMM clusters using Bhattacharyya distance (B) for multivariate normal distributions as

$$B = \frac{1}{8}(\mu_0 - \mu_1)^T P^{-1}(\mu_0 - \mu_1) + \frac{1}{2} log \frac{|P|}{\sqrt{|\Sigma_0||\Sigma_1|}}$$

where

$$P = \frac{\Sigma_0 + \Sigma_1}{2}$$

then, as GMM are not ordered, we take the Hausdorff distance (H) as the maximum between the row-wise and column-wise minimum of the pairwise distances between two GMM sets. ROC analysis on music distances was performed over the value of D, defined as

$$D = log(1 + \bar{b}) - log(1 + \bar{w})$$

where

$$\bar{w} = \frac{H(K562_a, K562_b) + H(NHEK_a, NHEK_b)}{2}$$

is the average of distances between replicates and

$$\bar{b} = \frac{H(K562_a, NHEK_a) + H(K562_a, NHEK_b) + H(K562_b, NHEK_a) + H(K562_b, NHEK_b)}{4}$$

is the average of pairwise distances among different cell lines.

Assessment of differentially expressed genes

RNA-seq tags were downloaded from GEO archive (GSE30567, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30567) and aligned to human reference genome (hg19) using STAR aligner v2.3.0[25]. Read counts over RefSeq intervals were extracted using bedtools v.2.24.0[26]. Discrete counts were normalized with TMM[27], differential gene expression was evaluated using the voom function implemented in limma v.3.26.7[28] with a simple contrast between two cell lines. Genes were considered differentially expressed under a p-value lower than 0.01 and absolute logarithm Fold Change higher than 1.

Cluster analysis

Cluster analysis was performed on replicate 1 of K562 dataset. We calculated all pairwise Hausdorff distances among genomic loci as defined above. Data were clustered using the Ward method. Enrichment analysis was performed using online Enrichr suite[29]. Word clouds were created with world_cloud python package (https://github.com/amueller/word_cloud) using text description of ontologies having positive Enrichr combined score. Differential expression among clusters was evaluated using Mann-Whitney U-test.

Consent

Written informed consent for publication was obtained from the study participants.

Data availability

Figshare: Supplementary audio files for 'Chromas from chromatin: sonification of the epigenome', doi: 10.6084/m9.figshare.3079540[30]

Figshare: Supplementary table for 'Chromas from chromatin: sonification of the epigenome', doi: 10.6084/m9.figshare.3079543[31]

Software availability

Latest source code

https://bitbucket.org/dawe/enconcert

Archived source code at time of publication

https://zenodo.org/record/45943[32]

License

MIT License https://opensource.org/licenses/MIT

Author contributions

DC, DL and PP designed the experiment and wrote the manuscript; DC performed the analysis.

Competing interests

No competing interests were disclosed.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgements

Authors would like to acknowledge Piotr Traczyk (Soltan Institute for Nuclear Studies, Warsaw, Poland) who inspired our research with his sonification of the Higgs Boson. Authors would like to thank all collaborators and relatives who kindly sacrificed their time to listen to music generated while this work was developed.

References

1. Ohno S, Ohno M: **The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition.** *Immunogenetics.* 1986; **24**(2): 71–78.
PubMed Abstract | Publisher Full Text
2. King RD, Angus CG: **PM--protein music.** *Comput Appl Biosci.* 1996; **12**(3): 251–252.
PubMed Abstract | Publisher Full Text
3. Takahashi R, Miller JH: **Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns.** *Genome Biol.* 2007; **8**(5): 405.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Larsen P, Gilbert J: **Microbial bebop: creating music from complex dynamics in microbial ecology.** *PLoS One.* 2013; **8**(3): e58119.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Juslin PN: **What does music express? Basic emotions and beyond.** *Front Psychol.* 2013; **4**: 596.
PubMed Abstract | Publisher Full Text | Free Full Text
6. 1000 Genomes Project Consortium, Abecasis GR, Auton A, *et al.*: **An integrated map of genetic variation from 1,092 human genomes.** *Nature.* 2012; **491**(7422): 56–65.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Zhang Y, Liu T, Meyer CA, *et al.*: **Model-based analysis of ChIP-Seq (MACS).** *Genome Biol.* 2008; **9**(9): R137.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics.* 2010; **26**(1): 139–40.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Kouzarides T: **Chromatin modifications and their function.** *Cell.* 2007; **128**(4): 693–705.
PubMed Abstract | Publisher Full Text
10. ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome.** *Nature.* 2012; **489**(7414): 57–74.
PubMed Abstract | Publisher Full Text | Free Full Text
11. SEQC/MAQC-III Consortium: **A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium.** *Nat Biotechnol.* 2014; **32**(9): 903–14.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Logan B: **Mel Frequency Cepstral Coefficients for Music Modeling.** *ISMIR.* 2000.
Reference Source
13. Dopler M, Schedl M, Pohle T, *et al.*: **Accessing Music Collections Via Representative Cluster Prototypes in a Hierarchical Organization Scheme.** *ISMIR.* 2008.
Reference Source
14. Aucouturier JJ, Pachet F, Sandler M: **"The way it Sounds": timbre models for analysis and retrieval of music signals.** *IEEE Trans Multimedia.* 2005; **7**(6): 1028–1035.
Publisher Full Text
15. Zhou VW, Goren A, Bernstein BE: **Charting histone modifications and the functional organization of mammalian genomes.** *Nat Rev Genet.* 2011; **12**(1): 7–18.
PubMed Abstract | Publisher Full Text
16. Barski A, Cuddapah S, Cui K, *et al.*: **High-resolution profiling of histone methylations in the human genome.** *Cell.* 2007; **129**(4): 823–837.
PubMed Abstract | Publisher Full Text
17. Valouev A, Johnson DS, Sundquist A, *et al.*: **Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data.** *Nat Methods.* 2008; **5**(9): 829–834.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Ernst J, Kellis M: **ChromHMM: automating chromatin-state discovery and characterization.** *Nat Methods.* 2012; **9**(3): 215–216.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Song J, Chen KC: **Spectacle: fast chromatin state annotation using spectral learning.** *Genome Biol.* 2015; **16**(1): 33.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Perner J, Lasserre J, Kinkley S, *et al.*: **Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling.** *Nucleic Acids Res.* 2014; **42**(22): 13689–13695.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Stevens SS, Volkmann J, Newman EB: **A Scale for the Measurement of the Psychological Magnitude Pitch.** *J Acoust Soc Am.* 1937; **8**(3): 185–190.
Publisher Full Text
22. Hoffman MM, Ernst J, Wilder SP, *et al.*: **Integrative annotation of chromatin elements from ENCODE data.** *Nucleic Acids Res.* 2013; **41**(2): 827–841.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Li H, Durbin R: **Fast and accurate long-read alignment with Burrows-Wheeler transform.** *Bioinformatics.* 2010; **26**(5): 589–595.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Kent WJ, Zweig AS, Barber G, *et al.*: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics.* 2010; **26**(17): 2204–7.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner**. *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Quinlan AR: **BEDTools: The Swiss-Army Tool for Genome Feature Analysis.** *Curr Protoc Bioinformatics.* 2014; **47**: 11.12.1–11.12.34.
PubMed Abstract | Publisher Full Text | Free Full Text
27. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol.* 2010; **11**(3): R25.
PubMed Abstract | Publisher Full Text | Free Full Text
28. Ritchie ME, Phipson B, Wu D, *et al.*: ***limma* powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Res.* 2015; **43**(7): e47.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Chen EY, Tan CM, Kou Y, *et al.*: **Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.** *BMC Bioinformatics.* 2013; **14**: 128.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Cittaro D, Lazarevic D, Provero P: **Supplementary audio files for 'Chromas from chromatin: sonification of the epigenome'.** *Figshare.* 2016.
Data Source
31. Cittaro D, Lazarevic D, Provero P: **Supplementary table for 'Chromas from chromatin: sonification of the epigenome'.** *Figshare.* 2016.
Data Source
32. Cittaro D: **Enconcert.** *Zenodo.* 2016.
Data Source

Open Peer Review

Current Referee Status:   

Version 1

Referee Report 22 March 2016

doi:10.5256/f1000research.8610.r12734

Federico M Giorgi
Department of Systems Biology, Columbia University, New York, NY, USA

The article "Chromas from chromatin: sonification of the epigenome" proposes a new and original method to sonify (i.e. convert into music) data information coming from the analysis of the epigenome.

The paper is extremely well written and it should be indexed on the basis of being the first to try this kind of review. The results of Cittaro *et al.*'s work, although without an immediate applicability in the field of biomedicine, is by itself a small scientifical breakthrough.

My only concerns regard the applicability of the sonification method as a truly alternative way to detect biological properties.

The authors imply that similar biological features (e.g. belonging to a particular pathway, differential expression) could be discerned by listening to gene-centered audio tracks. I listened to such tracks and I concur, but to be scientifically complete, the authors may think (even as a follow up paper) to test this on a wider subject set. I would propose a psychological study with a blind panel of human subjects, who are then asked, after listening to some biological properties on a "training set", to find genes with similar properties on a "test set".

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Competing Interests: No competing interests were disclosed.

Referee Report 18 March 2016

doi:10.5256/f1000research.8610.r12735

Ho-Ryun Chung
Otto-Warburg-Laboratory, Max Planck Institute for Molecular Genetics, Berlin, Germany

“Chromas from chromatin: sonification of the epigenome” by Cittaro *et al*. deals with the transformation of ChiP-seq profiles into music. The authors transform BigWig files into notes by discretizing the logarithm of the signal intensity. Test individuals perceive the resulting music as more appealing than randomized controls. The authors explore the possibility to use the music to discern differentially expressed genes in

two cell lines from unchanged ones. The classification accuracy is a little bit better than random guessing. They also use the similarity of the music to cluster genes. They claim that the clustering reveals genes with similar biological properties. They also show that the clustering is correlated to the gene expression level.

I find the idea really interesting. It may help the vision impaired to gather information about ChIP-seq tracks or other quantitative vector like information. It may be used to train individuals to recognize certain epigenomic features, like promoters and enhancers.

I find it not so surprising that the chromatin music is more appealing to test persons than randomized controls, because of the smoothness of the ChIP-seq profile data. Smoothness guaranties that subsequent notes are close by forming a more melodious line. Once randomized it results in random notes, which may be far apart and are perceived by most as not so appealing. Thus, humans can distinguish between smooth signals and signals that change abruptly and perceive the former more appealing than the latter. But clearly there is information in the music that can be exploited by the human ear.

The analyses about differential gene expression and gene clustering use the music as a feature in computer-aided classification tasks. These analyses show that there is some information in the music that a computer can recognize and use for classification. However, a base line using just the ChIP-seq profiles without turning them to music is missing. Without such a base line it remains hard to judge whether the reported results are meaningful or not. Moreover, I think that the classification results shown in Figure 3 are likely to reflect gene expression differences rather than gene function clustering.

Are the p-values reported in Table 1 for a one- or two-sided test? I do not understand why cluster 2 is not different from all the others except cluster 1.

I think the real potential in chromatin music is not so much in its use in machine learning approaches – it is much better suited for humans learning chromatin states and the like. Humans may be able to recognize patterns that a machine cannot. It would be really interesting to compare a human made chromatin segmentations using chromatin music with a segmentation generated from the ChIP-seq signals by a computer.

Finally, the proposed tool can be used to reach out to the public and demonstrate ideas about epigenomic states etc. using an easily accessible medium such as music.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.

Referee Report 15 March 2016

doi:10.5256/f1000research.8610.r12733

Tao Liu[1,2]

[1] Department of Biochemistry, University at Buffalo, Buffalo, NY, USA

[2] Department of Biostatistics, University at Buffalo, Buffalo, NY, USA

To translate functional genomics data in ChIP-seq to music pieces is entertaining. Cittaro *et al*'s work showed that, the chromatin data which has wider dynamic range is suitable (in my opinion perhaps more suitable than DNA/protein sequences in spite of lacking supporting evidence) for making appealing music. I feel excited while reading the manuscript and listening to the music pieces made by the authors at the same time.This work definitely has the novelty and importance especially for education and popular science. I believe more students and general public will be attracted by the chromatin songs then start to learn the science behind them. Because I don't have the expertise of music theory, I won't comment on the quality of the chromatin music or the methods to tune the music. Since authors showed the music patterns may also reflect the biology on the chromatin and may be associated with gene expression. I will focus on these.

1. Authors translated chromatin marks ChIP-Seq data from human K562 and NHEK cell lines at refSeq gene bodies from upstream 1k of TSS to 2kb downstream of TES, then tried to study if the music patterns can match the differential gene expression inferred from RNA-Seq. The prediction power was quite low, as pointed out by the authors, with an AUC of 0.52 (just a little better than random) with the optimal combination of parameters got from a subset of genes. A natural question is that how well the prediction is by directly using chromatin signals, such as the tags pileup. By comparing with this, we will see whether the underlying biology has been kept or lost during the approach.

2. Authors showed that the similarities, in terms of hierarchical clustering, of musical representation at gene bodies, from the K562 cell line, can be linked to gene functions and similar gene expression. However I found it's hard to see the consistency of the similarities of music with the similarities of gene function annotations. For example, the cluster 7 and 8 are the closet pair of clusters, although authors claimed that clusters 6,7,8 were linked to regulation of cell cycle, I can't see such words from the word clouds of the middle panel of figure 3. Instead, I can see that the cluster 1, 3, 5 and 7 all have the same term of biological function 'protein' although they are apart according to the hierarchical clustering. The gene expression analysis is also confusing while looking at the right panel of figure 3 and table 1. It seems that for all the clusters there are two major distributions of gene expression levels from the violin plots -- a big one at log(RPKM) less than 1 and a small one around log(RPKM) of 2. Does that mean the big population of each cluster are just random noises of weakly expressed genes (logRPKM<1)? The numbers of genes in each cluster were also missing in the manuscript. I wonder perhaps the cluster 2 contains very few genes since although visually it seems the distribution of gene expression of cluster 2 is quite different with all the other clusters, the p-values of the row 2 of table 1 (cluster 2 against 3, 4, 5, 6,7 and 8) are all very small, indicating no significant difference. Additionally, I should bring my comment of point 1 here as well. How the clustering works while checking only the raw ChIP-Seq pileup?

3. Comments on method section:

- For the equations of 'comparison of WAV track', the notations were not explained clearly.

- Description on how the RPKM was calculated is missing.

- The bases of logarithm functions are missing for the 'log(RPKM)' of gene expression level and the 'absolute logarithm fold change' used to define differential expressed genes.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Competing Interests: No competing interests were disclosed.