1   **Chromes from Chromatin: Sonification of the Epigenome**

2

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

4

5   *Author affiliations*

6   1: Center for Translational Genomics and Bioinformatics, San Raffaele Hospital- via Olgettina 58, 20138

7   Milano, Italy

8   2: Department of Molecular Biotechnology and Life Sciences, University of Turin – via Nizza 52, 10126

9   Torino, Italy

10

11   *Corresponding author*

12   Davide Cittaro; Center for Translational Genomics and Bioinformatics, San Raffaele Hospital- via

13   olgettina 58, 20138 Milano, Italy; telephone: +39 02 26439137; email: cittaro.davide@hsr.it

14

15   *Keywords*

16   Sonification; Chromatin; Music; Information content

17

1 **Abstract**

2 The epigenetic modifications are organized in patterns determining the functional properties of the

3 underlying genome. Such patterns, typically measured by ChIP-seq assays of histone modifications, can

4 be combined and translated into musical scores, summarizing multiple signals into a single waveform.

5 As music is recognized as a universal way to convey meaningful information (1), we wanted to

6 investigate properties of music obtained by sonification of ChIP-seq data. We show that the music

7 produced by such quantitative signals is perceived by human listener as more pleasant than that

8 produced from randomized signals. Moreover, the waveform can be analyzed to predict phenotypic

9 properties, such as differential gene expression.

10

11 **Significance Statement**

12 Music is recognized as universal way to communicate emotions and, more in general, meaningful

13 information. Various sources of information can be translated into music or sounds, mostly for

14 recreational purposes. It has been shown that human ear can classify information encoded into sounds.

15 Quantitative genomic features, and in particular epigenetic marks, do represent functional information

16 that is exploited by cells to drive biological processes. We test a method to translate such information

17 into music and we study some properties of the sonicated chromatin marks. We show that not only

18 musical representation of epigenetic marks has intrinsic musicality, but also that differences in musical

19 representation of genomic loci reflect differences of the RNA levels of the underlying genes.

20

21 **Introduction**

22 Sonification is the process of converting data into sound. Sonification itself has a long, yet punctuated,

23 story of applications in molecular biology, several algorithms to translate DNA (2) or protein sequences

24 (3, 4) to musical scores have been proposed. The same principles have also been extended to the

25 analysis of complex data (5) showing that, all in all, sonification can be used to describe and classify

26 data. Indeed, the very same procedures may also be applied for recreational purposes.

27

28 One of the limitations of sonification of actual DNA and protein sequences is their intrinsic conservative

29 nature. Assuming the differences in two individual genomes are, on average, one nucleotide every

30 kilobase (6), the corresponding musical scores would have little differences.

1   On the contrary, dynamic ranges typical of transcriptomic and epigenomic data may provide a richer

2   source for sonification.

3

4   In this work we describe an approach to convert ChIP-seq signals, and in principle any quantitative

5   genomic feature, into a musical score. We started working on our approach for amusement mainly, and

6   we realized that the sonificated chromatin signal were surprisingly harmonious. We then tried to assess

7   some properties of the music tracks we were able to generate. We show that the emerging sounds are

8   not random and instead appear more melodious and tuneful than music generated from randomized

9   notes. We also show that different ChIP-seq signals can be combined into a single musical track and

10   that tracks representing different conditions can be compared allowing for the prediction of

11   differentially expressed genes.

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

13   cittaro/sets/k562

14

15   **Definitions**

16   MIDI: MIDI (Musical Instrument Digital Interface) is a standard that describes protocols for data

17   exchange among a variety of digital musical instruments, computers and related devices. MIDI format

18   encodes information about note notation, pitch, velocity and other parameters controlling note

19   execution (*e.g.* volume and signals for synchronization).

20   MIDI file format: a binary format representing MIDI data in a hierarchical set of objects. At the top of

21   hierarchy there is a Pattern, which contains a list of Tracks. A Track is a list of MIDI events, encoding for

22   note properties. MIDI events happen at specific time, which is always relative to the start of the track.

23   MIDI Resolution: resolution sets the number of times the status byte is sent for a quarter note. The

24   higher the resolution, the more natural the sound is perceived. Resolution is the number of Ticks per

25   quarter note. At a specific resolution R, Tick duration in microseconds T is related to tempo (expressed

26   in Beats per Minute, BPM) by the following equation

27

28
$$T = \frac{60R}{BPM}$$

29

30

1 **Results**

2 *Approach*

3 In order to translate a single ChIP-seq signal track to music we bin the signal over a specified genomic

4 interval (*i.e.* chrom:start-end) into fixed-size windows (*e.g.* 300 bp) and note duration will be

5 proportional to the size of such windows. As we are dealing with MIDI standard, we let the user specify

6 track resolution and the number of ticks per window (see definitions); the combination of these

7 parameters defines the duration of a single note. The default parameters associate a bin of 300 bp with

8 one quaver (1/8 note).

9

10 In order to define the note pitch, we take the logarithm of the average intensity of the ChIP-seq signal

11 in a genomic bin. The sounding range of the whole signal is discretized in a predefined number of

12 semitones. At default parameters, the range is binned into 52 semitones, covering four octaves. In order

13 to introduce pauses, the lowest bin of the signal range represents a rest. If two consecutive notes or

14 rests fall in the same bin, we merge them in one note doubling its duration.

15

16 Using this approach, any ChIP-seq signal can be mapped to a chromatic scale. We implemented the

17 possibility to map a signal on a different scale (major, minor, pentatonic…); to this end, intensity bin

18 boundaries are merged according to the definition of a specific scale (Figure 1). MIDI tracks produced

19 in this way can be then imported into a sequencer software where they can be further processed,

20 setting tempo and time signature.

21

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

23 In order to test whether sonification of chromatin marks are perceived as random patterns, we selected

24 ten genomic regions and generated corresponding tracks based on the following histone modifications:

25 H3K27me3, H3K27ac, H3K9ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3

26 (Supporting Audio files S1.1 to S10.1). For the same regions, we randomized genomic signal at base

27 and bin level (Supporting Audio files S1.2 to S10.2). When data are randomized at base level, the

28 average intensity is uniforms across the bins, resulting in a repeated note (data not shown); this is

29 largely expected as ChIP-seq signals are distributed on the genome according to a Poisson law (7) or,

30 more precisely, to a Negative Binomial law (8).

1

2 Randomization at bin level, instead, equals to shuffling notes during the execution. We administrated

3 a questionnaire to a set of volunteers (n=8) not previously tested for education in music. Volunteers

4 were asked to listen to each pair of original/random track and choose which track they felt was more

5 appealing. Track order was randomized when testing different volunteers. Notably, in the majority of

6 the cases (62/80) the music generated from genomic signal without randomization was judged more

7 appealing. Results are significant to a Fisher-exact test (p=1.95e-3), suggesting that genomic signals

8 contain information that can be recognized by human ear. The number of correct answers for each

9 volunteer ranged from 5 to 10, with a median value of 8.

10

11 *Differences in musical tracks reflect differences in gene expression*

12 Once we assessed the existence of musical pattern in genomics signals, we were keen to explore if this

13 kind of information could be exploited to identify biological features of samples. Since the epigenetic

14 DNA modifications reflected by histone marks influence gene expression (9), we tested if differences in

15 musical tracks generated from various ChIP-seq signals reflects differences in gene expression of the

16 corresponding loci. To this end, we downloaded ChIP-seq marks (H3K27me3, H3K27ac, H3K9ac,

17 H3K36me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, Pol2b) and RNA-seq data for K562 and NHEK

18 cell lines from the ENCODE project (10). For each RefSeq locus we converted ChIP-seq signals to music

19 with fixed parameters (see methods). RNA-seq data were used to identify genes that are differentially

20 expressed between the two cell lines, under a p-value < 0.01 and |logFC|>1, according to recent SEQC

21 recommendations (11).

22

23 A common way to classify music is based on summarization of track features after spectral analysis (12,

24 13). Such approach involves the summarization of track as Mel-Frequency Cepstral Coefficients (MFCC)

25 that are subsequently clustered using Gaussian Mixture Models (GMM). A distance between tracks can

26 then be defined as described in (14), who used it as a classifier for musical genres.

27

28 We tested if a similar approach could be used to develop a predictor of differential expression based

29 on the distance between musical tracks generated from two cell lines.

1   We defined a distance between songs as described in methods and we optimized the parameters using

2   as a training set the 250 genes with the most significant differential expression *p*-value and as many

3   genes with the least significant *p*-value according to RNA-seq (Figure 2). We found that optimal

4   performance is at MFCC=30 and GMM=10, with an AUC=0.609.

5

6   We summarized tracks representing all RefSeq genes using such parameters, we then compared

7   distances with differential expression performing a ROC analysis. Our results indicate that differences

8   in information contained in musical representation of chromatin signals may be linked to differential

9   expression, although power of prediction is limited (AUC=0.5184, *p*=1.4597e-03).

10

11 *Similarity between musical tracks overlaps similar biological properties*

12 As additional issue we wanted to assess if similarities between musical representation of chromatin

13 status may be linked to the biology of the underlying genes. To this end, we calculated pairwise

14 distances for all regions using parameters identified above on K562 cell line. Hierarchical clustering of

15 the distance matrix identifies eight major clusters (Figure 3, left). We performed Gene Ontology

16 Enrichment analysis on each cluster, here represented as word cloud of significant terms (Figure 3,

17 center, supplementary Table S2); we found that different clusters are linked to genes showing different

18 biological properties. For example, some clusters (6, 7 and 8) were linked to regulation of cell cycle,

19 others were linked to metabolic processes (2 and 5) or vesicle transport (3 and 4). We also evaluated

20 the distribution of expression (expressed as log(RPKM)) of the underlying genes (Figure 3, right); we

21 found that regions clustered by the distance between musical tracks broadly reflects groups of genes

22 with different level of expression, spotting clusters of higher expression (cluster 5) or lower expression

23 (clusers 2 and 3); assessment of statistical significance of differences in distribution of gene expression

24 values among clusters is presented in Table 1.

25

26 **Discussion**

27 Chromatin shape and genome function are governed, among several factors, by the coordinated

28 organization of epigenetic marks (15). Modifications of such marks are dynamic and are fine-tuned

29 during the life of a cell or an organism. Analysis of histone modifications, as well as transcription factors

30 and other proteins binding DNA, by ChIP-seq already described patterns of enrichment that are specific

1   to their relative function (16, 17). Analysis of combinatorial patterns of histone modifications already

2   unveiled its potential in understanding functional properties of the genome (18, 19) and the cross-talk

3   among multiple chromatin marks (20).

4

5   We show, in this work, that the information carried by multiple histone modifications can be caught in

6   a human-friendly way by translating ChIP-seq signals into musical scores. Although the investigation of

7   the psychological factors that underlie tuneful perception of sonificated genomic signals is out of the

8   scope of this manuscript, our results suggest that human hearing is able to perceive patterns conveying

9   information encoded in ChIP-seq data analyzed and to distinguish from random noise.

10

11   We automated the analysis of differences between musical tracks using an established method based

12   on summarization of spectral data. By this approach, we investigated the possible link between

13   differences in ways chromatin sounds and phenotypic features. Our results suggest that differences in

14   transcript levels can be predicted by the differences of sonificated genomic regions, although

15   performances of such approach are limited. We reasoned that many factors may explain such poor

16   results: first of all there is a vast space of parameters that can be tuned to create a single musical track

17   and we still lack methods to explore it efficiently. In addition, the Mel scale used to summarize audio

18   signal has been developed to match human capabilities to perceive sound (21), hence it may not be

19   optimal for the comparison of the tracks generated in this work.

20

21   It has already been shown that it is possible to predict levels of gene expression starting from chromatin

22   states, although the method used to perform chromatin segmentation has a large impact on such

23   predictions (22). In this work we found that differences in chromatin-derived music reflects, to some

24   extent, differences in level of expression of underlying genes and their related biology.

25

26   To conclude, although we cannot advocate the usage of musical analysis as universal tool to analyze

27   biological data yet, we confirm that quantitative features on the genome are patterned and contain

28   information, hence can be converted into sounds that are perceived as musical. We limited our analysis

29   on specific chromatin modifications, but in principle any quantitative genomic feature can be converted

30   and integrated into a musical track. The choice of parameters and instruments has been standardized

1   for the analysis presented, for illustrative purpose we show that different signals from the same region

2   can be combined using different instruments (https://soundcloud.com/davide-cittaro/random-locus-

3   blues) and signals from different genomic regions can be merged (https://soundcloud.com/davide-

4   cittaro/non-homologous-end-joining).

5

6

7   **Materials and Methods**

8

9   *Sonification of ENCODE ChIP-seq data*

10   Raw data for various modifications were downloaded from (GSE26320). Read tags were aligned to

11   human genome (hg19) using bwa aligner (23). Alignments were converted to bigwig tracks (24) after

12   filtering for duplicates and quality score higher than 15.

13   In order to define regions to be converted to music scores, we selected intervals around RefSeq gene

14   definition, from 1kb upstream of TSS to 2kb downstream of TES. ChIP-seq signal were firstly converted

15   to MIDI using custom scripts (https://bitbucket.org/dawe/enconcert) according to parameters defined

16   in Table 2. MIDI tracks belonging to the same region from the same sample were merged into a single

17   MIDI file, converted to WAV format using timidity software (http://timidity.sourceforge.net), with the

18   exception of tracks presented as supplementary information which have been processed with Apple

19   GarageBand software.

20

21   *Comparison of WAV tracks*

22   In order to compare four samples for each converted genomic region, we extracted MFCC using

23   python_speech_features library (https://github.com/jameslyons/python_speech_features). Selected

24   components were then clustered using Gaussian Mixture Models, implemented in scikit-learn python

25   library 0.15.2 (http://scikit-learn.org). Distance between two tracks were evaluated using Hausdorff

26   distance (*H*) between GMM clusters. Briefly, we first calculate all pairwise distances between GMM

27   clusters using Bhattacharyya distance (*B*) for multivariate normal distributions as

28

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

1

2 where

3

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

5

6 then, as GMM are not ordered, we take the Hausdorff distance (H) as the maximum between the row-

7 wise and column-wise minimum of the pairwise distances between two GMM sets. ROC analysis on

8 music distances was performed over the value of D, defined as

9

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

11

12 where

13

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

15

16 is the average of distances between replicates and

17

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

19

20 is the average of pairwise distances among different cell lines.

21

22 *Assessment of differentially expressed genes*

23 RNA-seq tags were aligned to reference genome using STAR aligner (25). Read counts over RefSeq

24 intervals were extracted using bedtools. Discrete counts were normalized with TMM (26), differential

25 gene expression was evaluated using the voom function implemented in limma (27) with a simple

26 contrast between two cell lines. Genes were considered differentially expressed under a *p*-value lower

27 than 0.01 and absolute logarithm Fold Change higher than 1.

28

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 Enrichr (28). 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.

Acknowledgements

References

1. Juslin PN (2013) What does music express? Basic emotions and beyond. *Front Psychol* 4:596.

2. Ohno S, Ohno M (1986) The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition. *Immunogenetics* 24(2):71–78.

3. King RD, Angus CG (1996) PM--protein music. *Comput Appl Biosci* 12(3):251–252.

4. Takahashi R, Miller JH (2007) Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns. *Genome Biol* 8(5):405.

5. Larsen P, Gilbert J (2013) Microbial Bebop: Creating Music from Complex Dynamics in Microbial Ecology. *PLoS ONE* 8(3):e58119.

6. Consortium T1GP, et al. (2013) An integrated map of genetic variation from 1,092 human genomes. *Nature* 490(7422):56–65.

7. Zhang Y, et al. (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137.

8. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. doi:10.1093/bioinformatics/btp616.

9. Kouzarides T (2007) Chromatin Modifications and Their Function. *Cell* 128(4):693–705.

1
2
10. ENCODE Project Consortium, Consortium TEP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489(7414):57–74.

3
4
5
11. SEQC/MAQC-III Consortium, SEQC/MAQC-III Consortium (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. doi:10.1038/nbt.2957.

6
12. Logan B (2000) Mel Frequency Cepstral Coefficients for Music Modeling. *ISMIR*.

7
8
13. Dopler M, Schedl M, Pohle T, Knees P (2008) Accessing Music Collections Via Representative Cluster Prototypes in a Hierarchical Organization Scheme. *ISMIR*.

9
10
14. Aucouturier JJ, Pachet F, Sandler M (2005) "The way it Sounds": timbre models for analysis and retrieval of music signals. *IEEE Trans Multimedia* 7(6):1028–1035.

11
12
15. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet* 12(1):7–18.

13
14
16. Barski A, et al. (2007) High-resolution profiling of histone methylations in the human genome. *Cell* 129(4):823–837.

15
16
17. Valouev A, et al. (2008) Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth* 5(9):829–834.

17
18
18. Ernst J, Kellis M (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat Meth* 9(3):215–216.

19
20
19. Song J, Chen KC (2015) Spectacle: fast chromatin state annotation using spectral learning. *Genome Biol* 16(1):33.

21
22
23
20. Perner J, Lasserre J, Kinkley S, Vingron M, Chung HR (2014) Inference of interactions between chromatin modifiers and histone modifications: from ChIP-Seq data to chromatin-signaling. *Nucleic Acids Res* 42(22):13689–13695.

24
25
21. Stevens SS, Volkmann J, Newman EB (1937) A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America* 8(3):185–190.

26
27
22. Hoffman MM, et al. (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* 41(2):827–841.

28
29
23. Li H, Durbin R (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. 26(5):589–595.

30
31
24. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D (2010) BigWig and BigBed: enabling browsing of large distributed data sets. doi:10.1093/bioinformatics/btq351.

32
25. Dobin A, et al. (2012) STAR: ultrafast universal RNA-seq aligner. 29(1):15–21.

1 26. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression
2   analysis of RNA-seq data. *Genome Biol* 11(3):R25.

3 27. Ritchie ME, et al. (2015) limma powers differential expression analyses for RNA-sequencing and
4   microarray studies. *Nucleic Acids Res*. doi:10.1093/nar/gkv007.

5 28. Chen EY, et al. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment
6   analysis tool. *BMC Bioinformatics* 14:128.

7
8

1   **Figure Legends**

2   Figure 1: Graphical representation o the approach used to transform quantitative signals to music. ChIP-

3   seq values (H3K4me3 in the example) are binned in fixed-size intervals over the genome. Each interval

4   corresponds to a 1/8 note. Average values of log-transform of read counts in each genomic bin (red

5   lines) are matched into predefined number of semitones (chromatic scale). Notes may be mapped to a

6   specified scale (major and minor scales are exemplified in the figure). Consecutive equal notes are

7   merged in single note with double duration. Values falling in the first bin are considered rests.

8

9   Figure 2: Evaluation of different combination of parameters in predicting differential gene expression

10 on a gold-standard subset of 500 genes. Each square corresponds to a number of Mel-Frequency

11 Cepstral Coefficients (MFCC) used to summarize signal and a number of centers for Gaussian Mixture

12 Model (GMM). Colors are given by the corresponding Area Under the Curve (AUC)

13 Figure 3: Hierarchical clustering of genomic regions identifies 8 main clusters (left). Each cluster broadly

14 corresponds to specific biological properties according to Gene Ontology enriched terms (middle). Level

15 of expression of genes included in each cluster show specific distributions (right).

16

	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

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

3

1

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

2 *Table 2 – Parameters used to convert different ChIP-seq signals into corresponding musical trakcs*

3
4