# SONIFICATION OF GENOME ANNOTATIONS VIA THE SEQUENCE ONTOLOGY

*Kitty Gu*

University of California
Dept of Bioengineering
Stanley Hall
Berkeley, California
kittyguz@berkeley.edu

*Ian Holmes*

University of California
Dept of Bioengineering
Stanley Hall
Berkeley, California
ihh@berkeley.edu

## ABSTRACT

Genome annotations, specifying the location of genes and other features on DNA, are an essential tool to interpret and analyze genomic sequence data. Such annotations are typically visualized by biologists using a genome browser. We hypothesize that auditory feedback might enhance the user experience of such browsers, increasing awareness and navigational continuity. Previous audio representations of DNA have focused on sonifying the sequence itself; in contrast, we sonify a GFF genome annotation file by converting it to a MIDI file, representing gene features as musical notes. We provide a preliminary demonstrative sonification of the Major Histocompatibility Complex region on chromosome 6 of the human genome.

## 1. INTRODUCTION

Genome annotations are a layer of curated data that help biologists understand genomic DNA. They include both curated annotations, and the primary evidence that supports those annotations. Curated features encompass genes (such as protein-coding genes and non-coding RNA genes), the internal structure of genes (e.g. introns, exons, and untranslated regions), binding sites for various proteins (e.g. transcription factors), and variants. The primary evidence supporting these curated annotations typically takes the form of "reads" from a DNA sequencer. Together, these features — particularly curated gene features — are the primary landmarks and signposts used to navigate the alien space of the genome.

A wide variety of genomic workflows take advantage of these annotations. They are used to predict the function (or dysfunction) of genome sequences, to compare different genomes for medical or evolutionary purposes, to develop clinical genomic diagnostics, and to perform genetic engineering. Visualization is an important part of these workflows. The most common software application for this kind of visualization is the *genome browser*, a map-like application in which the genome sequence is represented horizontally and annotation "tracks" containing various kinds of annotation data are stacked beneath. An example screenshot of one such genome browser, JBrowse, is shown in Figure 1.

Because genome annotations are somewhat abstract data (albeit rooted in objectively testable reality), successfully navigating them requires numerous subtle hints and cues for the user. This has given rise to a set of visual conventions for representing annotation data, including various cartoons and glyphs for different types of gene, as well as user interface conventions for genome browsers. The vertical stacking of tracks is one such convention; another is the use of animated transitions to preserve navigational continuity.

We hypothesized that auditory display of these annotations would provide additional immersion and navigational orientation, improving user efficiency. Our idea was to use the genome annotations to create a unique musical "signature" for every region of the genome. We note that this idea is distinct from previous explorations of auditory display of biological sequences, which have focused on sonification of the sequence data itself, such as the amino acid sequence of various proteins [1, 2] or the DNA sequence of the coronavirus genome [3]. To our knowledge there has been no previous attempt to sonify the annotations themselves, despite the fact that these annotations are relatively stable digital objects, directly supported by material experimental results.

In this paper we report the prototyping of a simple system for converting a genome annotation database into a sequence of MIDI notes which yields an audible signal. We provide a simple demonstration using the Major Histocompatibility Complex (MHC) region of chromosome 6 of the human genome.

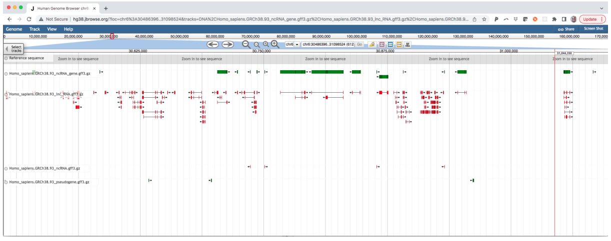

Figure 1: A portion of the human Major Histocompatibility Complex (MHC) Region, shown in the JBrowse web-based genome browser with four annotation tracks [4].

## 2. METHODS

Our goal was to produce a unique audio signature for any region on an annotated genome. Our approach is to convert a set of gene locations, specified by a genome annotation file in the standard General Feature Format (GFF), into a sequence of notes encoded by a file in Musical Instrument Digital Interface (MIDI) format.

https://doi.org/10.21785/icad2022.013

| Feature type | Count | Note |
|---|---|---|
| protein_coding_gene | 19982 | C4 |
| lncRNA | 17755 | E4 |
| processed_pseudogene | 10157 | G4 |
| non_processed_pseudogene | 2605 | C5 |
| misc_RNA | 2212 | E5 |
| snRNA | 1901 | G5 |
| miRNA | 1879 | C3 |
| TEC | 1056 | E3 |
| transcribed_unprocessed_pseudogene | 950 | G3 |
| snoRNA | 943 | C6 |
| transcribed_processed_pseudogene | 503 | E6 |
| rRNA Pseudogene | 497 | G6 |
| IG_V_pseudogene | 187 | C2 |
| transcribed_unitary_pseudogene | 146 | E2 |
| IG_V_gene | 145 | G2 |
| TR_V_Gene | 106 | C7 |
| unitary_pseudogene | 97 | E7 |
| TR_J_Gene | 79 | G7 |
| polymorphic pseudogene | 50 | C1 |
| scaRNA | 49 | E1 |
| rRNA | 47 | G1 |
| IG_D_gene | 37 | D4 |
| TR_V_pseudogene | 33 | F4♯ |
| Mt_tRNA | 22 | A4 |

Table 1: Note assignments for some of the most frequently occurring feature types in our dataset. The **Count** column shows how many times each type occurs in the data.[6]

Our prototype implementation converts GFF to an intermediate Comma-Separated Value (CSV) format, using the csvmidi utility to perform the subsequent conversion from CSV to MIDI.

To convert GFF features to MIDI notes, we began by downloading NCBI GenBank gene annotations for the human genome from the UCSC Genome Browser website [5] and extracted every line whose feature type (i.e. the text in the "type" field of the GFF record) is "gene". We further restricted the search to features in the range from nucleotides 28,477,797 to 33,448,354 of chromosome 6 of the hg38 assembly of the human genome, lying within the MHC region containing genes involved in adaptive immunity.

Inspired by the Protein Music sonification [1], we began our exploration by assigning each "gene type" (as further annotated in the "gene_type" subfield of the "group" field in the GFF file) to a different pitch of note. These gene types map approximately to terms from the Sequence Ontology (SO), an hierarchical description logic for genome annotation features [6]. We made manual assignments of the most common terms to notes in the key of C, as shown in Table 1. Our choices were guided by the frequency of occurrence of the corresponding SO terms; we also made some informed choices about the level of granularity of the SO hierarchy to keep (e.g. we lumped all processed pseudogenes together in one category). The start time of each note is proportional to the start position of the corresponding annotation feature, and the note length similarly proportional to the feature length.

The rationale behind the note assignment in Table 1 is that frequently occurred SO terms will be playing C major triads that are relatively closer to the mid-range of the keyboard, while less common terms will play an octave higher or lower, and rare terms will play discordant notes corresponding to higher harmonic overtones of the fundamental C note.

Our accompanying demonstration includes a musical rendition of the annotations shown in the MHC region of Figure 1, playing the notes with a preset of the Serum software synthesizer in Logic Pro X. The demonstration is around 123 seconds long, which represents a genome traversal rate of roughly 40kb per second. We used 20,000 as the MIDI timer division setting for the demonstration music file, with the playback tempo at 120bpm.

The demonstration shows that different genome regions clearly produce distinct audio signatures. Open questions remain as to how to make these signatures recognizable at different "zoom" settings, i.e. different playback speeds, that can vary over orders of magnitude. A few ways that the project could be extended include assigning abstracted sound samples to the different feature types (instead of pitched notes), reflecting the hierarchical nature of the Sequence Ontology in the assignment of similar sounds to subtypes of the same basal type (so e.g. the various types of pseudogene might have similar but distinct audio profiles), and exploiting generative procedural music algorithms to create more complex signatures based on the individual profiles of genes (e.g. signatures with harmonic and melodic structure).

The scale of the approximately 5-megabase region of our sonification can be contrasted with previous sonifications of individual human proteins such as the serotonin receptor 5-HT2 (roughly 500 amino acids) [1], or of viral genomes such as coronavirus (roughly 30 kilobases) [3]. We have by no means solved the problem of generating recognizable sonic signatures for biological sequences over a range of length scales; however, the nature of annotation metadata (which are routinely visualized at a range of zoom levels) is such that they readily lend themselves towards abstract sonifications, perhaps more easily than the primary DNA or protein sequences to which the annotations refer.

The source code and audio file for the software used to make our demonstration, and the audio file produced, are freely available at `github.com/kittyguz/Sonification_Project`

## 3. ACKNOWLEDGMENT

This work was supported by NIH/NHGRI grant R01-HG004483.

## 4. REFERENCES

[1] R.D. King and C.G. Angus, "PM–Protein Music," in *Comput Appl Biosci*, 1996, 12:3, pp. 251–252.

[2] N.W. Tay et al, "Protein music of enhanced musicality by music style guided exploration of diverse amino acid properties," in *Heliyon*, 2021, 7:9.

[3] M.D. Temple, "Real-time audio and visual display of the Coronavirus genome," in *BMC Bioinform*, 2020, 21:1, 431.

[4] R. Buels et al, "JBrowse: a dynamic web platform for genome visualization and analysis," in *Genome Biol*, 2016, 17.

[5] B.T. Lee et al, "The UCSC Genome Browser database: 2022 update," in *Nucleic Acids Res*, 2022, 50:D1, 1115-1122.

[6] K. Eilbeck et al, "The Sequence Ontology: a tool for the unification of genome annotations," in *Genome Biol*, 2005, 6:5, R44.