Genome analysis

**Gene Tracer: a smart, interactive, voice-controlled Alexa skill** For gene information retrieval and browsing, mutation annotation and network visualization

Shaoke Lou\(^1,2\), Tianxiao Li \(\odot \, 1,2\), Jason Liu\(^1,2\) and Mark Gerstein \(\odot \, 1,2,*\)

\(^1\)Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA and \(^2\)Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Associate Editor: Pier Luigi Martelli

Received on March 6, 2020; revised on December 29, 2020; editorial decision on December 30, 2020; accepted on March 31, 2021

**Abstract**

**Motivation:** Traditionally, an individual can only query and retrieve information from a genome browser by using accessories such as a mouse and keyboard. However, technology has changed the way that people interact with their screens. We hypothesized that we could leverage technological advances to use voice recognition as an interactive input to query and visualize genomic information.

**Results:** We developed an Amazon Alexa skill called Gene Tracer that allows users to use their voice to find disease-associated gene information, deleterious mutations and gene networks, while simultaneously enjoy a genome browser-like visualization experience on their screen. As the voice can be well recognized and understood, Gene Tracer provides users with more flexibility to acquire knowledge and is broadly applicable to other scenarios.

**Availability and implementation:** Alexa skill store (https://www.amazon.com/LT-Gene-tracer/dp/B08HCL1V68/) and a demonstration video (https://youtu.be/XbDbx7JDKmI).

**Contact:** mark@gersteinlab.org

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

Recent advances in machine learning, especially deep-learning and natural language processing (NLP) algorithms, have not only broadened the application areas of computing algorithms, but also accelerated the practical use of new technology. With the large-scale and deep-learning capabilities of NLP, voice recognition has successfully been brought to the everyday life of humans. Siri (Apple), Alexa (Amazon) and Google Voice have dominated the smart artificial intelligence voice assistant market. These smart assistants enable users to efficiently interact with their devices and can be used as a hub for home entertainment and research purposes.

The technology revolution has changed the traditional interaction between humans and devices. In the 1980s, the mouse and keyboard were introduced as an easier and more efficient way to connect people to personal computers. These accessories allow users to control a computer by clicking keys, dragging/clicking a mouse, and, eventually, using a touch pad. In the biological domain, researchers need to interact with devices in more complex ways. For instance, genome browsers or viewers can allow users to browse or search through genome information, mutation information and network interactions. However, this approach has limitations in certain application scenarios. For instance, during a presentation/demonstration, a speaker has limited ability to show online information in front of a computer; similarly, a surgical doctor may need to find medical information when they do not have a free hand. A solution for these situations could be to use voice recognition tools. The voice can be precisely recognized and analyzed, free human hands, and provide a more convenient way to get useful information.

To provide a new convenient interacting way and an educational tool for researchers and general users to retrieve useful biomedical and genetic information, we developed a voice-enhanced Amazon Alexa skill called Gene Tracer. With NLP and voice recognition technology as the backbone, Gene Tracer defines a comprehensive language model to analyze users’ intent and retrieve gene knowledge from NCBI, mutation effect prioritization by FunSeq2 (Fu, et al., 2014), and visualize protein-protein interaction network information. Gene Tracer is on-demand and provides a genome browser-like visualization experience on the screen and can further display an interesting gene region with information on histone marks, transcription factor (TF) binding and deleterious gene mutations predicted by FunSeq2 pipeline. Scientists in laboratories will find Gene Tracer very helpful in finding information quickly using voice-based...
operations, especially when they are wearing gloves during their experiments. Gene Tracer is available from the Alexa skill store by searching 'gene tracer' or directly going to https://www.amazon.com/LT-Gene-tracer/dp/B08HCL1V68/.

2 Materials and methods

Hardware and server. We deployed a back-end server using the Amazon EC2 cloud server. We used Nodejs v0.12 to build the server-side services.

Dataset. On the server side, we integrated GENCODE-annotated files (Frankish, et al., 2019), transcription factor chromatin immunoprecipitation sequencing (ChIP-Seq) data files, Funseq2 annotations and protein-protein interaction information. Gene Tracer also includes scripts and code used to generate the visualization. We used NCBI e-utility to retrieve gene information (Sayers, 2018) and pyGenomeTrack (Ramirez, et al., 2018) to generate the tracks of the region.

We used the Alexa skill kit v1.0+ for voice processing and recognition. We built the model for Gene Tracer using Amazon develop portal and the Amazon AWS Lambda server.

The framework of Gene Tracer is shown on Figure 1.

3 Results

Gene information retrieval. The user can provide an exact gene name or a fuzzy query term, such as an associated disease or some other related properties. The server will output the relevant genes based on the gene’s summary from NCBI and show it on the screen. For example, if a user wants to find the top genes related to breast cancer, Gene Tracer will show a screen like in Figure 1B.

Genomic visualization of a region related to a gene. Since it is not practical for a user to give the exact genomic location, we implemented and used a gene-centered visualization. The user can visualize the genome information like a genome browser, including the histone marks signal, the TF binding signal and other useful information. The user can also move forward and backward along the genome to search the surrounding regions of a gene. By providing a gene name, or selecting a gene from the gene query output, Gene Tracer can display the genomic landscape view around that gene (Fig. 1C).

Mutation view to indicate the most deleterious mutations. The user can also use Gene Tracer to annotate and display mutations to find those that have highly deleterious effects, through a mutation view for browsing. Slightly different from the gene-centered browsing, this view allows the user to find the coding and non-coding mutations from ClinVAR (Landrum, et al., 2020; Landrum, et al., 2014) that strongly affect a gene’s function or regulation according to FunSeq2’s annotation (Dhingra, et al., 2017; Fu, et al., 2014).

Network visualization. The user can also visualize the genes and their interacting partners using a network view (Fig. 1D). The server side includes a protein-protein interaction dataset (Stark, 2006), and the R scripts depend on the igraph package (Nepusz, 2006).

Exporting and revisiting the query history. The user can access their queried figures once the session ends (a device with a screen is required). The link will be presented on the screen and is accessible later on any desktop.

4 Discussion

In summary, we developed the first voice-controlled smart genome browser called Gene Tracer that integrates large scale ChIP-Seq data from ENCODE, a variant prioritization scheme from the FunSeq2 pipeline, web development, NLP and voice recognition techniques. Gene Tracer works on an Alexa-enabled device, especially one with a screen. For a device without a screen, the user can also get a voice response for each query.

Acknowledgements

The authors thank all the comments from the reviews.

Funding

This work was supported by a National Institute on Drug Abuse/NIH/DHHS [R01DA051906].
Conflict of Interest

none declared.

References

Dhingra P. et al. (2017) Using FunSeq2 for coding and non-coding variant annotation and prioritization. Curr. Protoc. Bioinf., 57, 15 11 11–15 11 17.
Frankish A. et al. (2019) GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res., 47, D766–D773.
Fu Y. et al. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. Genome Biol., 15, 480.
Landrum M.J. et al. (2020) ClinVar: improvements to accessing data. Nucleic Acids Res., 48, D835–D844.
Landrum M.J. et al. (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res., 42, D980–985.
Nepusz GCaT. (2006) The igraph software package for complex network research. Int. J., 1695, 1–9.
Ramirez F. et al. (2018) High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat. Commun., 9, 189.
Sayers E. (2018) Entrez Programming Utilities Help. National Center for Biotechnology Information, Bethesda. This citation should be changed to: Bethesda (MD): National Center for Biotechnology Information (US); Entrez Programming Utilities Help [Internet], 2010–. Available from: https://www.ncbi.nlm.nih.gov/books/NBK25501/
Stark C. (2006) BioGRID: a general repository for interaction datasets. Nucleic Acids Res., 34, D535–539.