Genotify: Fast, lightweight gene lookup and summarization

Jared M. Andrews[1], **Mohamed El-Alawi**[2], **and Jacqueline E. Payton**[1]

1 Washington University School of Medicine, Department of Pathology and Immunology **2** None

Motivation

With the advent of low-cost, massively parallel sequencing, researchers are often faced with the task of manually curating lists of significant genes to find points of biological interest for further study. Determining the protein product function and biological significance in the context of the study for these genes can consume significant time and effort. Despite dozens of data sources that provide gene annotations, including genomic mapping, aliases, expression, function, disease associations, ontology terms, and more, accessing this information requires combing through these databases as well as the knowledge of their existence. While many databases provide APIs for high-throughput annotation (e.g. UniProtKB (Bateman et al. 2017), NCBI Gene (Brown et al. 2015), and Ensembl (Zerbino et al. 2018)), there exist few non-programmatic options for querying and collating information from multiple databases for everyday use. Genotify addresses this unmet need, providing an intuitive GUI with flexible search options that intelligently queries both general and species-specific databases to expedite manual curation and enable convenient routine gene lookup.

Summary

Genotify is a lightweight desktop application that provides rapid gene lookup and summarized annotations from dozens of major and specialized data sources (Figure 1). Initial gene queries are submitted to the MyGene.info API, which collects gene annotation data from over 30 data sources (Xin et al. 2016). Additional API calls are made depending on species and the accessions returned by the MyGene.info API. Gene symbols, names, chromosomal coordinates or IDs (Entrez, Ensembl, etc) are all viable query terms, and the results returned in JSON format are quickly parsed and displayed to the user. Results are sortable, searchable, and navigable with a single click. Genotify supports queries for all species, though the information available for each differs significantly. The UniProtKB API is used to collect additional functional information from the curated Swiss-Prot database when available (Bateman et al. 2017). The EBI Expression Atlas widget displays interactive expression data for several species (Petryszak et al. 2016), and the ProtVista protein viewer provides a wealth of interactive protein data including domains, post-translation modifications, and impact of known genetic variants (Watkins et al. 2017). Disease associations for human genes are collected with the Comparative Toxicogenomics Database (CTDbase) API (A. P. Davis et al. 2017). Organism-specific databases like WormBase and CTDbase are utilized when appropriate (Lee et al. 2018). Importantly, directly querying major databases ensures that the information Genotify returns is always up to date and removes the need for manual updates of locally stored flat database files.
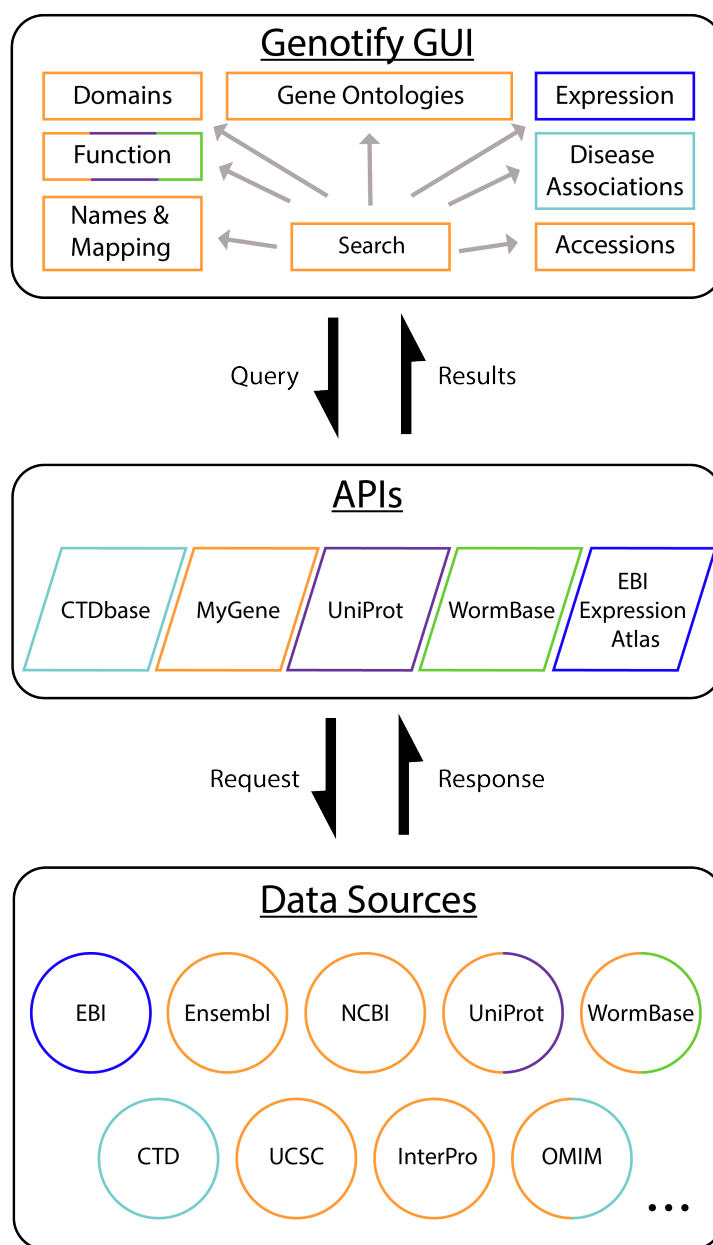
Figure 1: Diagram of Genotify workflow and data sources

Figure 1. Schematic showing design and workflow of Genotify. Colors indicate connections between results fields, APIs, and data sources. API queries and data sources are determined dynamically based on availability and species.

Genotify is a GUI desktop application built on the Electron Javascript framework, which allows for inherent cross-platform deployment to 32 or 64-bit linux, OSX, and Windows systems. Genotify's use of existing APIs means no data sources are downloaded, saving disk space and making installation simple. Users can limit their search to one or many species, and a hotkey command can directly query terms from the clipboard for ease of use.

Use cases

We designed Genotify for experimentalists and bioinformaticists who need an up-to-date, comprehensive summary of a gene's annotation, function, expression, ontology, and disease associations in a single location. Our group uses it daily to facilitate: * rapid, efficient lookup of genes while reviewing literature or curating lists of significant genes, * close investigation of families of related genes, * quick ascertainment of the biological significance of differentially expressed genes or associating proteins, * determination of known disease associations, * exploration of protein structure, modifications, and variants, * comparison of mRNA expression of a queried gene across diverse tissues, cell types, and species.

Availability

Genotify is released under the GPL-3.0 license with source code and binaries freely available at https://github.com/j-andrews7/Genotify, implemented as a desktop application built on the Electron framework and supported on linux, OS X, and MS Windows.

Funding

Acknowledgements

We would like to thank Sarah Pyfrom for code testing and feedback regarding UI design.

References

Bateman, Alex, Maria Jesus Martin, Claire O'Donovan, Michele Magrane, Emanuele Alpi, Ricardo Antunes, Benoit Bely, et al. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45 (D1):D158–D169. https://doi.org/10.1093/nar/gkw1099.

Brown, Garth R., Vichet Hem, Kenneth S. Katz, Michael Ovetsky, Craig Wallin, Olga Ermolaeva, Igor Tolstoy, et al. 2015. "Gene: A Gene-Centered Information Resource at NCBI." *Nucleic Acids Research* 43 (Database issue):D36–42. https://doi.org/10.1093/nar/gku1055.

Davis, Allan Peter, Cynthia J. Grondin, Robin J. Johnson, Daniela Sciaky, Benjamin L. King, Roy McMorran, Jolene Wiegers, Thomas C. Wiegers, and Carolyn J. Mattingly. 2017. "The Comparative Toxicogenomics Database: Update 2017." *Nucleic Acids Research* 45 (D1):D972–D978. https://doi.org/10.1093/nar/gkw838.

Lee, Raymond Y. N., Kevin L. Howe, Todd W. Harris, Valerio Arnaboldi, Scott Cain, Juancarlos Chan, Wen J. Chen, et al. 2018. "WormBase 2017: Molting into a New Stage." *Nucleic Acids Research* 46 (D1):D869–D874. https://doi.org/10.1093/nar/gkx998.

Petryszak, Robert, Maria Keays, Y. Amy Tang, Nuno A. Fonseca, Elisabet Barrera, Tony Burdett, Anja Füllgrabe, et al. 2016. "Expression Atlas Update—an Integrated Database of Gene and Protein Expression in Humans, Animals and Plants." *Nucleic Acids Research* 44 (D1):D746–D752. https://doi.org/10.1093/nar/gkv1045.

Watkins, Xavier, Leyla J. Garcia, Sangya Pundir, Maria J. Martin, and Alfonso Valencia. 2017. "ProtVista: Visualization of Protein Sequence Annotations." *Bioinformatics* 33 (13):2040–1. https://doi.org/10.1093/bioinformatics/btx120.

Xin, Jiwen, Adam Mark, Cyrus Afrasiabi, Ginger Tsueng, Moritz Juchler, Nikhil Gopal, Gregory S. Stupp, et al. 2016. "High-Performance Web Services for Querying Gene and Variant Annotation." *Genome Biology* 17 (May):91. https://doi.org/10.1186/s13059-016-0953-9.

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. "Ensembl 2018." *Nucleic Acids Research* 46 (D1):D754–D761. https://doi.org/10.1093/nar/gkx1098.