GENESTATION: THE GENOMIC SEARCH ENGINE TOOLKIT

By

Mara Kim

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

In partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

In

Biological Sciences

October 31, 2018

Approved:

Douglas K. Abbot, Ph.D.

Antonis Rokas, Ph.D.

Ge Zhang, M.D., Ph.D.

John A. Capra, Ph.D.

Thomas A. Lasko, M.D., Ph.D.

DEDICATION

To my mother, who showed me the wonders of art.

To my father, who taught me the beauty of science.

To my sisters that made me who I am.

To Andrea, who inspires me to be my best.

I love you.

ACKNOWLEDGEMENTS

This work was supported by Vanderbilt University, the Gisela Mosig Travel Fund, and the March of Dimes Ohio Collaborative. I would like to thank my collaborators Kris McGary, Antonis Rokas, Patrick Abbot, Ken Petren, Tony Capra, Lou Muglia, Scott Williams, Sashank Nutakki, Jibril Hirbo, Haley Eidem, Julie Phillips, Rohit Venkat, and Brian Cooper. I would also like to recognize Abigail Lind, David Rinker, Jen Wisecaver, Corinne Simonti, and Ling Chen for their help in developing ideas and techniques for this work.

I would like to thank Katherine Friedman and Steve Baskauf for their mentorship during my time as both an undergraduate and graduate student at Vanderbilt, as well as Alicia Goostree, Leslie Maxwell, Angela Titus, Christopher Patterson, and LaDonna Smith for their patience and assistance with the logistics of graduate school.

I am grateful to my committee, Patrick Abbot, Tony Capra, Tom Lasko, Ge Zhang, and Antonis Rokas, for the chance to do this work and pushing me to accomplish as much as possible. Their ideas and criticism were instrumental in ensuring the vision behind the project became a reality.

Finally, I would like to thank my advisor, Antonis Rokas, a fantastic scientist and teacher, for taking me in as an undergraduate researcher. He always believed in me and encouraged me in all my efforts. Without his mentorship and support, this work would not have been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

Introduction

A BRIEF HISTORY OF GENOMICS

Since the discovery of DNA as the carrier of genetic information in 1944 by Oswald Avery, Colin MacLeod, and Maclyn McCarty in bacterial transformation[1] and the subsequent characterization of its double helical structure in 1953 by Rosalind Franklin, James Watson, and Francis Crick[2], biologists have striven to understand the information encoded within this molecule. The elucidation of the "first sequence" by Fred Sanger[3] in 1955 sparked a revolution in the nascent field of genomics. Previous to this discovery, many biologists believed proteins to be an "ill-defined amorphous mixture"[4], with no definite structure. However, Sanger demonstrated that in fact proteins were composed of discrete amino acids in sequence, and three years later, Francis Crick proposed the Central Dogma of Molecular Biology[5], formalizing the hypothesis that DNA gave rise to RNA, which in turn gave rise to protein. Following this theory, researchers raced to develop a method of sequencing the other macromolecules. In 1965, Robert Holley published the first nucleic acid sequence[6] along with a description of the structure of ribonucleic acid[7], and Nirenberg and Leder described the RNA triplet code[8]. These discoveries culminated in the determination of the first nucleotide sequence for a gene in 1972 by Walter Fiers[9].

The invention of the first generation of DNA sequencing techniques in 1975-77, Sanger sequencing via chain termination[10], and Maxam-Gilbert sequencing via chemical modification[11], ushered in a new era of genomics. These technologies allowed for the accurate determination of sequences hundreds of bases long, and led to the publication of the first complete genome of the bacteriophage φX174 in 1977[12]. In 1981, researchers published the sequence of the human mitochondrial genome[13]. The proposal of construction of a human genetic map using DNA polymorphisms in 1980[13,14] and its application in 1983 to determine the locus responsible for

Huntington's Disease[15] would set sights on the assembly of the nuclear genome and its potential impacts on human health and disease. While at the time, the thought of sequencing a genome of 3 billion base pairs seemed an impossible dream, the possibility remained at the forefront of genomic research.

The next major advances in genomics came about with the application of computational techniques to sequencing. The first step came in 1986 with an improvement to the Sanger method using fluorescent dyes[16] eventually leading to the development of the first automated DNA sequencer by Applied Biosystems and its use demonstrated 1987[17]. This new technology made the idea of Whole Genome Sequencing (WGS) a reality, and in 1990 the US Department of Energy and the National Institute of Health announced the launch of the Human Genome Project. The development of shotgun sequencing[18] and paired end sequencing[19] provided new techniques that facilitated the assembly of the genomes of the first free-living organism *Haemophilus influenzae* in 1995[20] and the first eukaryote *Saccharomyces cerevisiae* in 1996[21]. These achievements were followed by the first metazoan genomes, the nematode *Caenorhabditis elegans* in 1998[22] and the fruit fly *Drosophila melanogaster* in 2000[23]. The tools created during the assembly and annotation of these "model organisms" paved the way for the completion of Human Genome Project in 2001[24].

The development of high-throughput sequencing technologies, such as 454 Sequencing[25] and Illumina Dye Sequencing[26], heralded the beginning of a new era of genomics. The commercial availability of these techniques in the mid 2000s has greatly reduced the cost of sequencing, enabling a new class of genomic analyses, such as RNA-Seq[27]. In the modern age of sequencing, it has become possible to complete massive sequencing projects on an unprecedented scale. The 1000 Genomes Project has assembled a deep catalog of human variation, recording the sequences of 2504 individuals across 26 populations in its Phase 3 dataset[28], and the Encyclopedia of DNA Elements (ENCODE) has annotated functional

elements using 1640 datasets in 147 cell types in its Phase 2 release[29]. As of today, there are 23,460 completed genomes available on NCBI[30] and over 77 million human polymorphisms recorded in dbSNP[31].

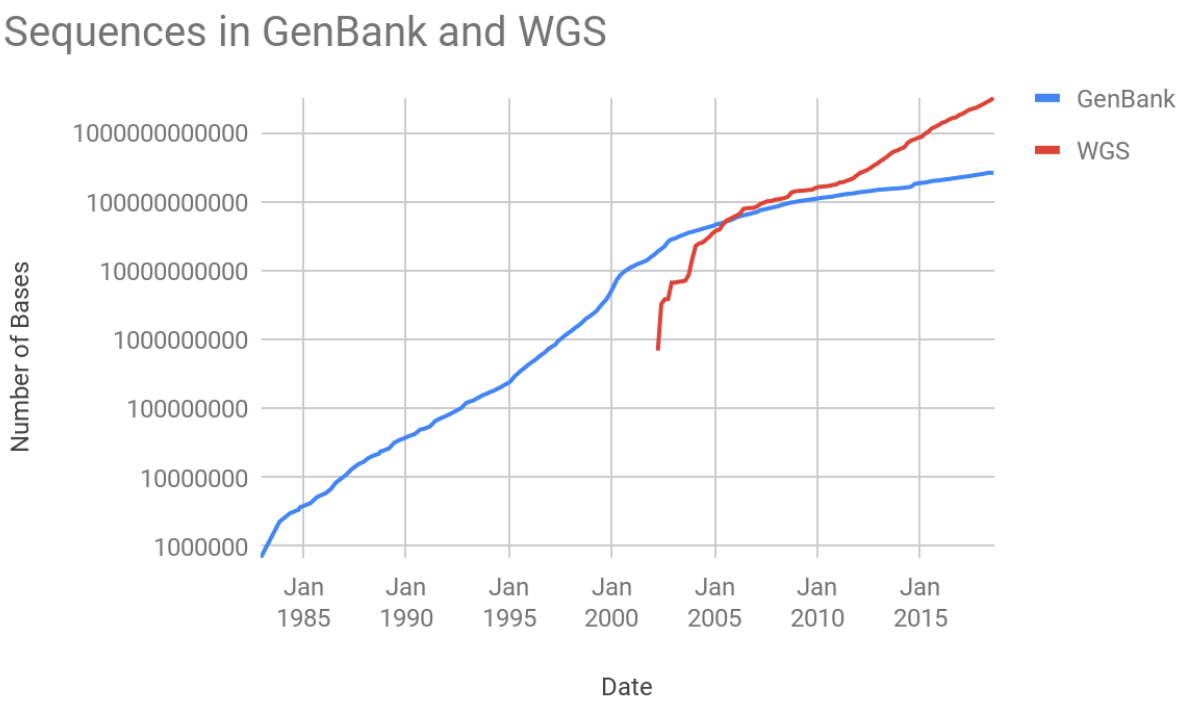

Figure 1. Sequences in GenBank since its inception to present day.

Genomic Databases

The realization that the structure and organization of the components and processes of the cell were defined by discrete and measurable units of information spurred the investigation of these topics with a newfound level of precision[4]. As the number of known sequences began to grow in size, efforts[32,33] to compile a definitive sequence database culminated in the foundation of the GenBank nucleic acid sequence database[34]. Since its inception, this sequence database has grown exponentially with each release (Figure 1). The availability of sequence data spurred a profound leap in the study of molecular evolution with the proposal of the molecular clock[35], the

formulation of the neutral theory of evolution[36], and the development of new methods of estimating population structure[37], adaptive evolution[38,39], sequence similarity search[40–42], and Genome Wide Association Studies[43].

The explosion of the amount of genomic sequence data, annotations, and analytical results has necessitated the development of systems to allow researchers to manage and query this massive amount of information. One of the earliest efforts to create a repository of genomic data was the Atlas of Protein Sequence and Structure first published in 1965 by Margaret Dayhoff[33], managed mostly by hand and expert curation. In 1979, Walter Goad established the Los Alamos Sequence Library[32] which eventually evolved to become GenBank in 1982. This database grew rapidly, doubling in size approximately every 18 months[44]. Outside the US, the European Nucleotide Archive[45] and the DNA Data Bank of Japan[46] serve as centralized repositories for sequence data.

The generation of non-sequence data, including functional annotations, expression data, and evolutionary analyses necessitated the development of repositories for these datasets. These specialized databases were adequate for those wishing to investigate these topics individually, but a growing desire to perform integrative analyses spurred the development of meta-databases. In 1991, the National Center for Biotechnology Information (NCBI) Entrez[47] was released with capability of searching across the many databases of NCBI simultaneously. Model organism databases, such as the *Saccharomyces* Genome Database[48], Flybase[49], and the Mouse Genome Database[50], were created to serve their respective research communities with a focus on collecting the annotations and analyses most relevant to their investigative audience. These databases are able to provide sophisticated tools and datasets tailored for their research focus, and more recently the field has seen the development of even more specialized resources for specific diseases, such as The Cancer Genome Atlas[51]. In 2018, *Nucleic Acids Research* Molecular Biology Database Collection counts 1737 entries in its annual

database issue[52].

GEneSTATION

In Chapters 2 and 3, I describe my work on GEneSTATION, a human pregnancy research database (NAR Molecular Biology Database Collection entry number 1881)[53]. This project was developed with the aim to investigate the genetic basis of preterm birth (PTB). Defined as birth before 37 weeks gestation, PTB is a complex, multifactorial syndrome whose onset is triggered via the sub-clinical dysregulation of the Common Pathway of Partruition[54]: myometrial contractility, cervical dilatation, and rupture of the chorioamniotic membranes. While many mammals experience early birth[55], it is theorized that in humans birth timing has had greater selective constraints due to pleiotropic effects from changes of the shape of the pelvic inlet during the evolution of bipedalism and increasing head-size along the human lineage. These effects are thought to impose a cephalopelvic constraint on fetal development known as the Obstetric Dilemma[56,57]. In 2016, babies born to non-Hispanic black mothers were 39.8% more likely to be born preterm--an effect that may be socio-economic or due to differences in genetic background[58].

In order to capture the complex possible etiologies of PTB, I compiled a diverse set of pregnancy specific functional and evolutionary data, and I created an interactive web encyclopedia with the ability to perform advanced searches and queries. Version 1.0 of the database was gene focused, and I was able to utilize the Generic Model Organism Database (GMOD) project's Chado ontological schema[59]. In Version 2.0 of the database, I expanded the focus to include sequence variation data, enhancers, and other non-coding elements of the genome. This expanded the size of the datasets by several orders of magnitude, and eventually I needed to transition to an entirely novel ElasticSearch database solution.

THE GENOMIC SEARCH ENGINE

In Chapter 4, I discuss my work in designing the ElasticSearch database and the associated tools I developed for the GEneSTATION encyclopedia. The Genome Feature Object specification is an ElasticSearch schema I created for the storage of genomic data in a generic and extensible manner. The use of this new database technology was necessitated by failure of a traditional relational database structure to deliver acceptable performance when operating over millions of genomic elements during the development of GEneSTATION 2.0. Because the original 1.0 version of the database implemented the GMOD Project's Chado schema, I needed the database to continue to support the same data and queries as this advanced schema. By using the Chado schema as a guide, I was able to ensure that the Genome Feature Object could represent complex biological data and support similar ontological queries.

The Genome Feature Object provides a consistent interface for computational tools to query and analyze biological data in a non-relational database. In order to facilitate the usage of this specification, I created the Genestation Command Line Interface (CLI) which is able to load GFF, FASTA, VCF, and generic tab delimited data into ElasticSearch following this specification. In addition, the Genestation CLI is able to perform queries and operations in ElasticSearch and integrate into data pipelines for further analysis.

SynTHy and GeneViewer

In Chapter 6 and 7, I detail two web based tools designed to query and interact with Genome Feature Objects. The Synthesis and Test of Hypotheses (SynTHy) is an advanced search builder which interactively visualizes the distribution of data values and the overlap of datasets. GeneViewer is a genome browser that is able to dynamically filter and plot genomic elements by various data values. Because a database project could utilize both a traditional relational

database and a non-relational data store of Genome Feature Objects, these tools could be used to provide any genome database with high performance search and visualization of genomic data while still reaping the benefits of a relational database structure.

REFERENCES

1. Avery, O. T., Macleod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.* **79,** 137–158 (1944).
2. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
3. Sanger, F., Thompson, E. O. P. & Kitai, R. The amide groups of insulin. *Biochem. J* **59,** 509–518 (1955).
4. Sanger, F. Sequences, Sequences, And Sequences. *Annu. Rev. Biochem.* **57,** 1–28 (1988).
5. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).
6. Holley, R. W., Everett, G. A., Madison, J. T. & Zamir, A. NUCLEOTIDE SEQUENCES IN THE YEAST ALANINE TRANSFER RIBONUCLEIC ACID. *J. Biol. Chem.* **240,** 2122–2128 (1965).
7. Holley, R. W. *et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science* **147,** 1462–1465 (1965).
8. Nirenberg, M. *et al.* RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc. Natl. Acad. Sci. U. S. A.* **53,** 1161–1168 (1965).
9. Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237,** 82–88 (1972).
10. Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94,** 441–448 (1975).
11. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 560–564 (1977).
12. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265,** 687–695 (1977).
13. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome.

Nature **290,** 457–465 (1981).

14. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32,** 314–331 (1980).
15. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238 (1983).
16. Strauss, E. C., Kobori, J. A., Siu, G. & Hood, L. E. Specific-primer-directed DNA sequencing. *Anal. Biochem.* **154,** 353–360 (1986).
17. Rodefeld, M. D., Beau, S. L., Schuessler, R. B., Boineau, J. P. & Saffitz, J. E. Beta-adrenergic and muscarinic cholinergic receptor densities in the human sinoatrial node: identification of a high beta 2-adrenergic receptor density. *J. Cardiovasc. Electrophysiol.* **7,** 1039–1049 (1996).
18. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6,** 2601–2610 (1979).
19. Roach, J. C., Boysen, C., Wang, K. & Hood, L. Pairwise end sequencing: a unified approach to genomic mapping and sequencing. *Genomics* **26,** 345–353 (1995).
20. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269,** 496–512 (1995).
21. Goffeau, A. *et al.* Life with 6000 Genes. *Science* **274,** 546–567 (1996).
22. The C. elegans Sequencing Consortium & The C. elegans Sequencing Consortium. Genome Sequence of the Nematode C. elegans: A Platform for Investigating Biology. *Science* **282,** 2012–2018 (1998).
23. Adams, M. D. The Genome Sequence of Drosophila melanogaster. *Science* **287,** 2185–2195 (2000).
24. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–1351 (2001).
25. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* **281,** 363, 365 (1998).
26. Canard, B. & Sarfati, R. S. DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene* **148,** 1–6 (1994).
27. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45,** 81–94 (2008).
28. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

29. Ecker, J. R. *et al.* Genomics: ENCODE explained. *Nature* **489,** 52–55 (2012).
30. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–45 (2016).
31. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
32. Kanehisa, M., Fickett, J. W. & Goad, W. B. A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic Acids Res.* **12,** 149–158 (1984).
33. Dayhoff (Ed), M. O. & Silver Spring, Md. National Biomedical Research Foundation. *Atlas of Protein Sequence and Structure*. (1965).
34. Burks, C. *et al.* The GenBank nucleic acid sequence database. *Comput. Appl. Biosci.* **1,** 225–233 (1985).
35. Zuckerkandl, E. & Pauling, L. *Molecular Disease, Evolution, and Genic Heterogeneity*. (1962).
36. Kimura, M. *The Neutral Theory of Molecular Evolution*. (1983).
37. Weir, B. S. & Clark Cockerham, C. ESTIMATING F -STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution* **38,** 1358–1370 (1984).
38. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).
39. Miyata, T. & Yasunaga, T. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. *J. Mol. Evol.* **16,** 23–36 (1980).
40. Needleman, S. B. & Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. in *Molecular Biology* 453–463 (1989).
41. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).
42. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
43. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32,** 650–654 (2002).
44. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **28,** 15–18 (2000).
45. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res.* **39,** D28–31

(2011).
46. Tateno, Y. *et al.* DNA Data Bank of Japan (DDBJ) for genome scale research in life science. *Nucleic Acids Res.* **30,** 27–30 (2002).
47. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **41,** D8–D20 (2013).
48. Cherry, J. M. *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26,** 73–79 (1998).
49. Gelbart, W. M. *et al.* FlyBase: a Drosophila database. The FlyBase consortium. *Nucleic Acids Res.* **25,** 63–66 (1997).
50. Blake, J. A. *et al.* The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39,** D842–8 (2011).
51. Hampton, T. Cancer Genome Atlas. *JAMA* **296,** 1958 (2006).
52. Rigden, D. J. & Fernández, X. M. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **46,** D1–D7 (2018).
53. Kim, M. *et al.* GEneSTATION 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. *Nucleic Acids Res.* **44,** D908–16 (2016).
54. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
55. Phillips, J. B., Abbot, P. & Rokas, A. Is preterm birth a human-specific syndrome? *Evol Med Public Health* **2015,** 136–148 (2015).
56. Washburn, S. L. Tools and human evolution. *Sci. Am.* **203,** 63–75 (1960).
57. Rosenberg, K. & Trevathan, W. Bipedalism and human birth: The obstetrical dilemma revisited. *Evolutionary Anthropology: Issues, News, and Reviews* **4,** 161–168 (2005).
58. Martin, J. A., Hamilton, B. E., Osterman, M. J. K., Driscoll, A. K. & Drake, P. Births: Final Data for 2016. *Natl. Vital Stat. Rep.* **67,** 1–55 (2018).
59. Mungall, C. J., Emmert, D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23,** i337–46 (2007).

CHAPTER II

GEneSTATION 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes[1]

Mara Kim,[1] Brian A. Cooper,[1] Rohit Venkat,[1] Julie B. Phillips,[1] Haley R. Eidem,[1] Jibril Hirbo,[1] Sashank Nutakki,[1] Scott M. Williams,[2] Louis J. Muglia,[3] J. Anthony Capra,[1,4] Kenneth Petren,[5] Patrick Abbot,[1] Antonis Rokas,[1,4] and Kriston L. McGary[1]

[1]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA

[2]Department of Genetics, Geisel School of Medicine, Dartmouth College, Hanover, NH 03755, USA

[3]Center for Prevention of Preterm Birth, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA

[4]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37235, USA

[5]Department of Biological Sciences, University of Cincinnati, Cincinnati, OH 45221, USA

1This chapter was published in Nucleic Acids Res. 2016 Jan 4;44(D1):D908-16. doi: 10.1093/nar/gkv1137.

ABSTRACT

Mammalian gestation and pregnancy are fast evolving processes that involve the interaction of the fetal, maternal and paternal genomes. Version 1.0 of the GEneSTATION database (http://genestation.org) integrates diverse types of omics data across mammals to advance understanding of the genetic basis of gestation and pregnancy-associated phenotypes and to accelerate the translation of discoveries from model organisms to humans. GEneSTATION is built using tools from the Generic Model Organism Database project, including the biology-aware database CHADO, new tools for rapid data integration, and algorithms that streamline synthesis and user access. GEneSTATION contains curated life history information on pregnancy and reproduction from 23 high-quality mammalian genomes. For every human gene, GEneSTATION contains diverse evolutionary (e.g. gene age, population genetic and molecular evolutionary statistics), organismal (e.g. tissue-specific gene and protein expression, differential gene expression, disease phenotype), and molecular data types (e.g. Gene Ontology Annotation, protein interactions), as well as links to many general (e.g. Entrez, PubMed) and pregnancy disease-specific (e.g. PTBgene, dbPTB) databases. By facilitating the synthesis of diverse functional and evolutionary data in pregnancy-associated tissues and phenotypes and enabling their quick, intuitive, accurate and customized meta-analysis, GEneSTATION provides a novel platform for comprehensive investigation of the function and evolution of mammalian pregnancy.

INTRODUCTION

Placental mammals, which originated 160 million years ago, uniformly share a conserved set of reproductive traits related to embryonic development within a uterus and nutrient provisioning through a chorioallantoic placenta[1]. Paradoxically, this conservation of reproductive mode and function during mammalian evolution is starkly juxtaposed with the evolution of the placenta,

one of the most variable of all mammalian organs[2,3]. At present, there is no comprehensive explanation for the diversity of evolutionary tempos and modes exhibited by the processes associated with mammalian gestation and pregnancy. The consequences are important not only for our understanding of mammalian pregnancy[4], but also for major features of human evolution, such as the encephalization and bipedalism[5], and how natural selection has acted on and shaped human biology[6,7]. And clinically, complications of pregnancy in humans are a major cause of infant mortality around the world[8]; for example, complications stemming from birth before term (pre-term birth or PTB), defined in humans as birth before 37 completed weeks of gestation[9], are the leading cause of death in newborns and in children under the age of five[10,11].

Several funding agencies have recognized both the seriousness of pregnancy associated medical problems and the persistence of many unanswered questions about the process. Consequently, they are currently increasing their investments in the study of the biology and pathologies of pregnancy, which will lead to the generation of large amounts of diverse types of data in the next few years. Two notable examples are the NIH-sponsored Human Placenta Project[12], aimed to 'understand the role of the placenta in health and disease', and the March of Dimes-sponsored Prematurity Research Centers (http://prematurityresearch.org/), 'dedicated to solving the mysteries of premature birth'. Because PTB has a significant genetic component[13,14], there is general consensus that emerging molecular and genomic resources provide new opportunities to not only make fundamental advances in our understanding of the evolution and function of mammalian pregnancy[4,15–20], but to also make breakthroughs in treating its diseases[8,21–24].

At present, however, such advances are limited by the fact that such data and resources are dispersed either in many different journals' supplements or across several different databases, making synthesis of available information slow and costly, and hampering powerful system approaches that involve overlaying diverse data types and analyses in the treatment of

disease[25,26]. To facilitate this synthesis, we have developed GEneSTATION (http://genestation.org), a database that integrates diverse types of -omics data across mammals to advance understanding of the genetic basis of pregnancy-associated phenotypes and to accelerate the translation of discoveries from model organisms to humans. The database's name, GEneSTATION, is a compound word created by blending together ‘gene’ and ‘gestation’; it can be read as ‘gestation’ if the reader considers only the capitalized letters, or as ‘gene station’ if the reader considers all letters, and is intended to highlight the fact that this database is focused on synthesizing information on genes related to gestation.

GEneSTATION provides the data and tools to easily explore pregnancy from three complementary perspectives, evolutionary, organismal, and molecular, at three levels of synthesis. At the first level, individual gene pages integrate the evolutionary, organismal, and molecular perspectives in three easily accessible tabs, providing a comprehensive picture of the breadth of the data available for a single gene and introducing researchers to analyses and data that they have not previously considered. At the second level, individual analysis pages provide access to genome-wide information from a single perspective, such as natural selection in the human lineage, differential expression in complications of pregnancy, or protein-protein interactions among genes known to be involved in pregnancy. At the final level of synthesis, the Gene Set Analysis tool and the novel ‘SynTHy’ (Synthesis and Testing of Hypotheses) tool enable researchers to synthesize on-the-fly the many types of information available through the development and evaluation of testable hypotheses.

DATA SOURCES AND DATA ORGANIZATION

Organism life history data

GEneSTATION contains information on pregnancy- and reproduction-associated characteristics for every mammal genome present in the database: human (*Homo sapiens*), elephant (*Loxodonta africana*), chimpanzee (*Pan troglodytes*), cow (*Bos taurus*), macaque (*Macaca fascicularis*), cat (*Felis catus*), dog (*Canis lupus*), goat (*Capra hircus*), guinea pig (*Cavia porcellus*), horse (*Equus caballus*), mouse (*Mus musculus*), gibbon (*Nomascus leucogenys*), rat (*Rattus norvegicus*), baboon (*Papio anubis*), vole (*Microtus ochrogaster*), rabbit (*Oryctolagus cuniculus*), rhesus monkey (*Macaca mulatta*), sheep (*Ovis aries*), orangutan (*Pongo abelii*), gorilla (*Gorilla gorilla*), marmoset (*Callithrix jacchus*), wild boar (*Sus scrofa*), and platypus (*Ornithorhynchus anatinus*). Specifically, life history characteristics including mean gestation length, neonate development, placental structure and shape, litter size, interbirth interval, adult body mass, maximum longevity and the timing of neonatal brain growth, an interesting characteristic relevant to potential complications of early parturition[27], are provided for each species (for data sources see Materials and Methods).

Gene-specific data

In addition to the life history data for the 23 mammals, every gene in each mammalian genome has a page on GEneSTATION that depicts the available evolutionary, organismal and molecular knowledge for that gene, with data from each category reported in a separate tab (Figure 1). The juxtaposition of diverse data is designed to guide users toward a more comprehensive understanding of genes of interest and facilitate serendipitous construction of novel hypotheses. For example, a GEneSTATION user may look up a gene of interest with enriched expression in the placenta and quickly discover that this gene additionally: (i) is often differentially expressed in studies on preeclampsia, a complication of pregnancy characterized by high blood pressure

(both of these data types are reported in the ORGANISMAL tab), (ii) originated coincidentally with the placental mammals (reported in the EVOLUTIONARY tab) and (iii) interacts with known pregnancy related genes (reported in the MOLECULAR tab). Collectively, these associations would suggest that the gene would be a good candidate for further exploration.

The EVOLUTIONARY category contains a variety of population and evolutionary data on human genes, and in some instances (e.g. ancient selection, orthology) on genes from diverse mammals (see Materials and Methods). These include the strength of recent selection (measured by FST) and ancient selection (measured by dN/dS), a gene's estimated date and lineage of origin, the SNPs from every human gene, and mammalian orthology relationships.

The data in the ORGANISMAL category include the Online Mendelian Information in Man (OMIM) phenotypes for each human gene, if available, RNA and protein expression across many tissues from Protein Atlas, including several pregnancy related tissues, as well as all differentially expressed genes from 106 genome-wide comparisons from pregnancy studies across gestational tissues (including placenta, cervix, myometrium, decidua, chorion, amnion) and pathologies (including preeclampsia, intrauterine growth restriction, chorioamnionitis and spontaneous preterm birth) (see Materials and Methods).

The data in the MOLECULAR category include the gene annotation[28] information for human and seven other mammalian genomes, and the protein interactions, through STRING[29], for proteins from humans and 15 other species.

In addition to the data sets in the three categories listed above, GEneSTATION displays RefSeq summaries for genes as well as gene-specific links to a wide variety of general databases, e.g. Entrez Gene, PubMed, UniProt, TreeFam[30–33], where available. In addition, GEneSTATION contains links to two pregnancy disease-specific databases, dbPTB (http://ptbdb.cs.brown.edu/dbPTBv1.php) and PTBgene (http://ric.einstein.yu.edu/ptbgene/). Both databases are focused on human PTB; dbPTB contains the output from a computational

mining of the literature as well as of the KEGG and dbSNP databases to identify studies, pathways and variants associated with candidate disease-risk genes[34], whereas PTBgene represents the summary of the fewer than 100 genes that show genetic association with preterm birth[34,35].

DATA PRESENTATION

A number of general purpose or species-centric databases, e.g., Genecards, SGD, WormBase and FlyBASE[36–39], provide access to diverse data sets on individual genes. By design, such databases aim to present the breadth of all available data for a given gene (e.g. http://www.genecards.org/cgi-bin/carddisp.pl?gene=BRCA1), which means that pages of genes that have been extensively studied can become either cumbersome to navigate or saturated with large amounts of different types of data, potentially obscuring the biological interpretation of relationships among the various resources proffered. For ease of access, each major category on the gene page, e.g. evolutionary, is presented on a separate tab that is divided into subsections, e.g. Evolution in Mammals. These subsections are intended to expand and include additional related types of data in future versions of GEneSTATION.

Aided by its focus on a specific biological process, GEneSTATION was developed using state-of-the-art web frameworks to provide a clean visual layout that is easy to interpret and efficient to navigate efficiently (Figures 1–3). For faster access to the data, GEneSTATION is designed to be highly responsive to users and focuses on providing low latency interaction and feedback. We have implemented multiple custom-made visualizations to allow users to quickly grasp the various types of available data and analyses, both for individual genes and across the genome, such as creating interactive summary figures that chart the distributions of the underlying data or studies. For example, the gene expression page (Figure 2) shows the number of studies available by keyword (e.g. ‘myometrium’ or ‘spontaneous preterm birth’),

providing instantaneous and meaningful filters of the data, while simultaneously highlighting deficiencies in the number of publicly available data sets and identifying opportunities for meta-analyses, as recently described by Eidem et al.[40].

Custom-made visualizations are a key design feature of GEneSTATION pages. Examples include a density plot of each analysis in the SynTHy tool (Figure 3), which allows users to quickly select an appropriate cutoff based on the distribution of the values (http://www.genestation.org/SynTHy); the distribution of gene ages plot (http://www.genestation.org/analysis/gene/age), and the distribution of available pregnancy related expression studies by keyword plot (http://www.genestation.org/analysis/gene/expression). To support these custom visualizations, additional html/css/js libraries were included (see Materials and Methods).'

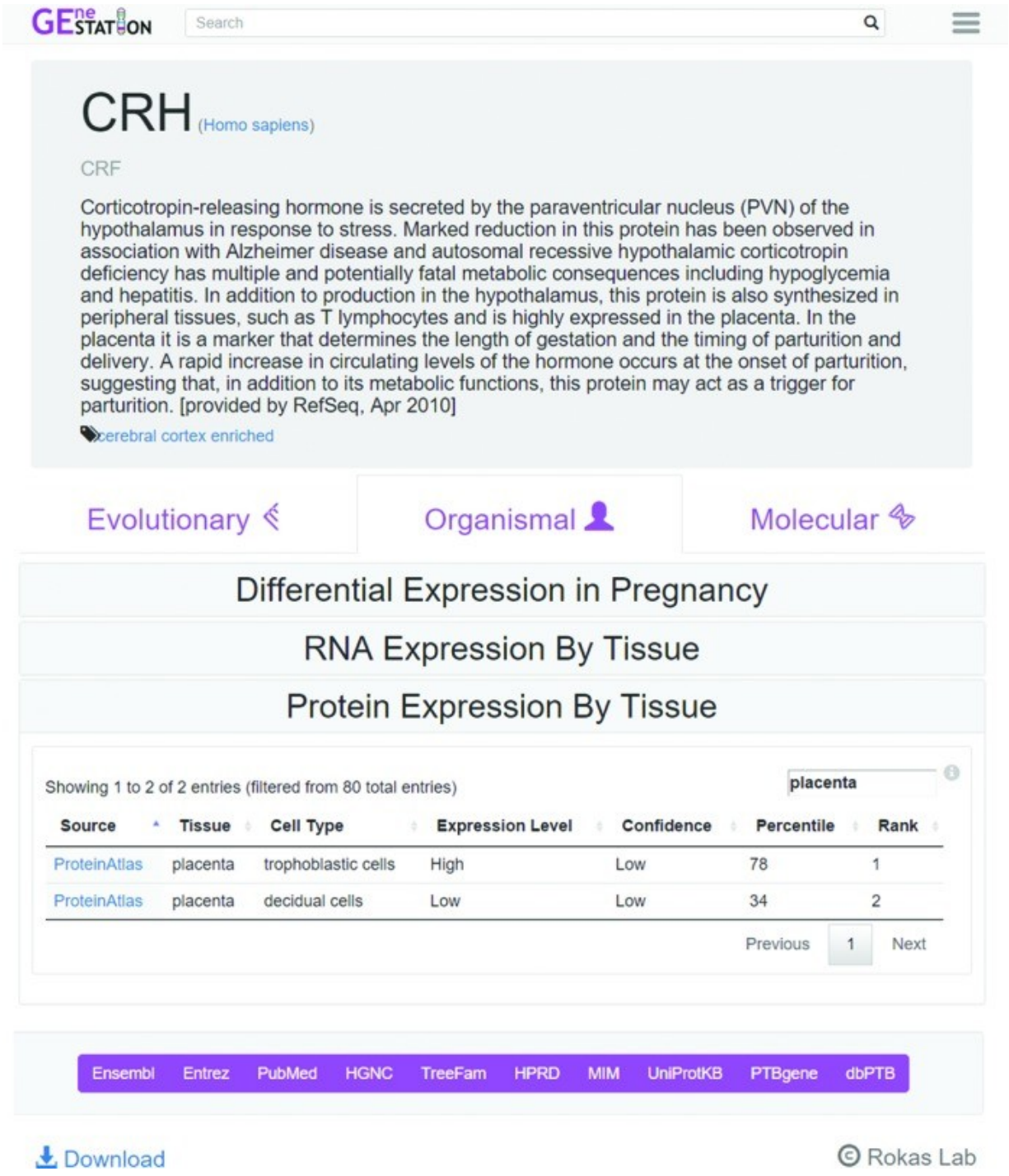

Figure 1. Screenshot of a typical GEneSTATION gene page. Each gene page includes a summary from RefSeq and data organized into three tabs, EVOLUTIONARY, ORGANISMAL and MOLECULAR. In this figure, the ORGANISMAL tab for the CRH gene is open and the Protein Expression by Tissue section is expanded. In this example, the table is filtered by the

search term ‘placenta’ and displays the expression levels of CRH protein by the two cell types annotated in this tissue. Links at the bottom of the page provide easy access to other relevant databases. The search bar, which provides instant search, is prominently visible on all pages. Additional analyses, information about GEneSTATION, forms to upload data, and frequently asked questions are accessible by clicking the triple line button on the top right.

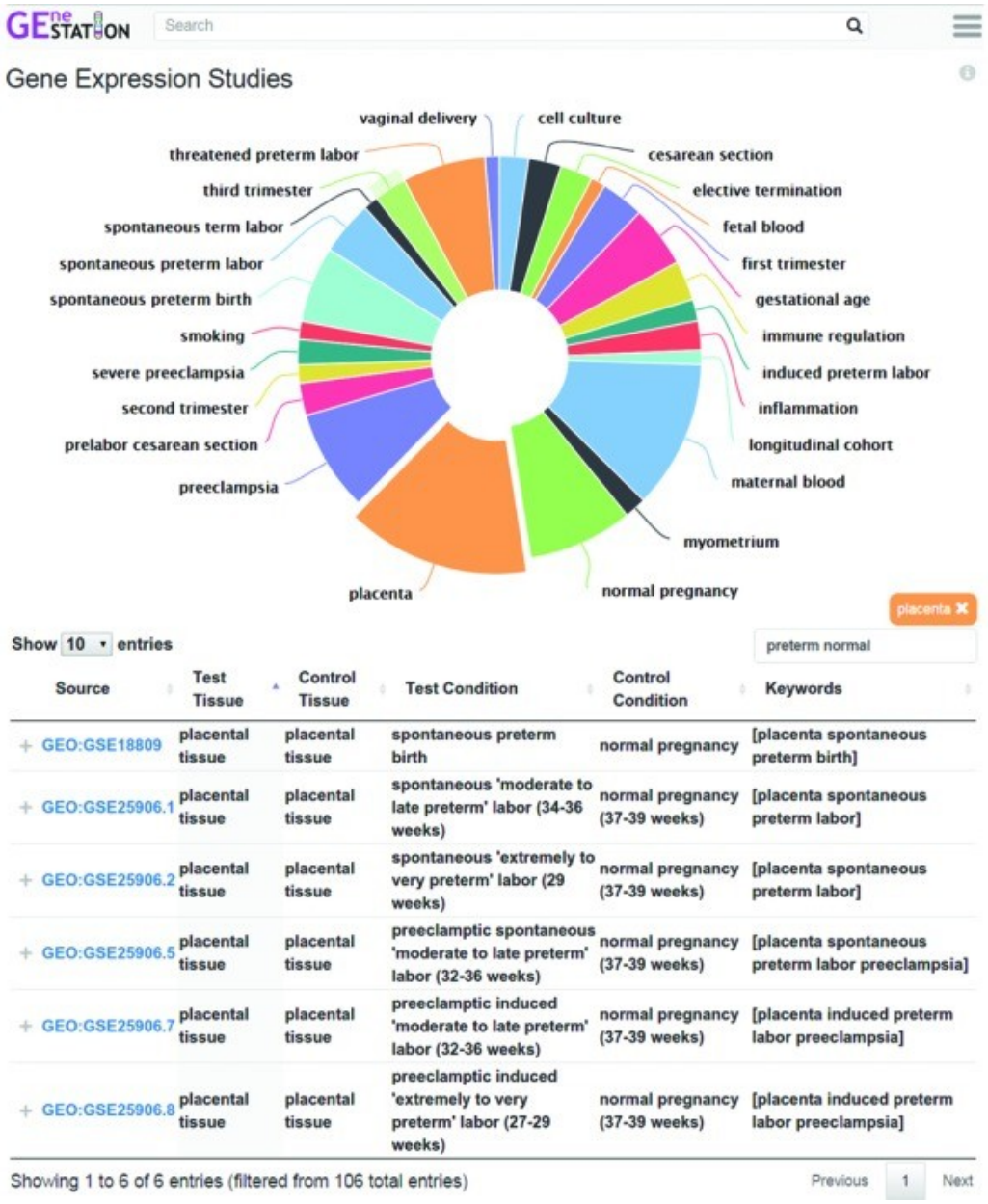

Figure 2. Screen shot of the Gene Expression Studies page. Each analysis page provides a summary figure to help users understand the scope of the data. The Gene Expression Studies page displays the most frequent keywords associated with available studies. The size of each segment in the pie chart is proportional to the number of studies with the keyword. In this case, the user has clicked on the placenta segment of the pie chart, which automatically filters the table for studies involving the placenta. In addition, the user has typed ‘preterm’ and ‘normal’ into the search box, which filters studies based on whether their

experimental and control condition descriptions contain these keywords. Each entry in the table provides a link to a page with additional details of the study, the tissue or cell line of the experiment, the test and control conditions, and relevant keywords. The page with study details also includes all genes reported in the study along with expression fold-change and significance.

Figure 3. Screen shot of the results of a complex query using the SynTHy tool. The SynTHy tool allows users to form and evaluate hypotheses rapidly, with instant visual

feedback guiding exploration. In this example, the user has created a rule group to include genes with significant gene expression differences in studies matching ‘spontaneous’ where the tissue is ‘placenta’. This rule group is visually represented by the blue circle in the Venn diagram. The user has also added a rule to include only genes that arose <160 million years ago (MYA), which is represented in the diagram by a green circle. The distribution of ages is presented to allow users to estimate how many genes are being filtered out. A slider is available to quickly select a cutoff. The final rule selects for genes near SNPs that have very strong differentiation between East Asian populations and European populations (FST ≥ 0.4), which is represented by the orange circle. The distribution to the right of the rule provides users with an estimate of how many genes are being removed by the filter. Users can interact with the Venn diagram to see the number of genes (47 in this example) in each segment along with the search query to find those genes (the black text box). Selecting segments in the Venn diagram, which are scaled to approximate the number of genes, or clicking the search button (in blue, bottom left) will take users to a search page where the genes matching the criteria are listed. The Gene Set Analysis button (in green, bottom left) submits the list of genes matching the intersection of all the rules to the Gene Set Analysis tool to find enrichment for any data stored in GEneSTATION. The Download button (in orange, bottom left) provides users with the selected list of genes and their associated data in JSON format. A screencast tutorial for SynTHy is available in the FAQ page and can be reached from the SynTHy Tutorial link below the search button and from the information icon.

SYNTHESIS AND ANALYSIS

The promise of GEneSTATION is that the rapid exploration of its diverse data types will allow users to generate a synthetic view of the genetic networks underlying pregnancy and its pathologies. Users with lists of genes obtained from experimental results, e.g. differential expression using RNA-seq, but short of fully developed hypotheses, can submit lists of candidate genes for enrichment analysis (http://www.genestation.org/analysis/gene/set) across the various data types (e.g. gene age, tissue expression, differential expression and methylation in disease, GO annotation, protein interactions) and examine statistically significant associations (see Materials and Methods). It has long been recognized that such interactive data exploration phases such as GEneSTATION provides are not only important in the analysis of complex data sets, but also in the formation of new hypotheses[41].

Alternatively, users may visit GEneSTATION with a specific hypothesis about genes involved in a particular process or pathology, or develop one while browsing through the gene pages. With GEneSTATION, finding candidate genes that test a hypothesis has been rendered intuitive and quick by the development of SynTHy (after Synthesize and Test Hypotheses; http://www.genestation.org/SynTHy), a novel tool that goes far beyond typical search tools by visualizing the distribution of the underlying data, giving immediate visual feedback and showing how the various components of a hypothesis impact the resulting list of genes. Rapid exploration of multiple variations on a hypothesis facilitates the development of an integrated view of the genetic relationships underlying the many different data types. For example, a user could ask whether genes with SNPs that have very different frequencies (high FST) in populations with high preterm birth rates versus populations with lower preterm birth rates[42] are also preferentially expressed in the placenta or arose in the mammalian ancestor. Any gene list results generated using SynTHy can be easily transferred to the gene set analysis tool for further refinement and exploration. SynTHy thus allows users to ‘find the question’ as readily as

to find the answers[41]. SynTHy is not intended to fully replace careful statistical analysis using original data but rather to synthesize disparate but high-quality data and analyses and facilitate rapid exploration.

DATA ACCESS

To facilitate specifically tailored statistical analyses, GEneSTATION makes all data available for easy download in JSON format using the download button on the bottom left of each gene page, analysis page and SynTHy result page.

SUMMARY AND FUTURE PERSPECTIVES

Understanding the complex functional landscape of pregnancy, how abnormalities of pregnancy arise, or how the biological mechanisms of gestation evolve and translate between species can be greatly augmented by the integration and synthesis of multiple types of experimental data, genomic data, and evolutionary analyses. Importantly, such genome-scale data sets are becoming more frequent and current funding priorities will only accelerate this trend. Consequently, populating GEneSTATION with additional high-quality evolutionary, organismal and molecular data sets is an active and ongoing process, with transcriptomic, proteomic, and imaging data being a high priority. In parallel, we are developing algorithms that will point users interested in a particular gene to other genes or biological processes with similar functional or evolutionary characteristics. Furthermore, GEneSTATION's integration of evolutionary and experimental data will support the development of algorithms that evaluate the likelihood that those specific biological systems or medical interventions that work in a model organism such as mouse or macaque will also work in human pregnancy.

In summary, GEneSTATION facilitates integrative analyses that draw from many types of data, providing a novel platform and paradigm for comprehensive understanding of pregnancy

across mammals. It is our hope that GEneSTATION's synthesis becomes a catalyst for the identification and evaluation of candidate genes by biologists interested in the function and evolution of mammalian pregnancy, as well as its complications. More generally, GEneSTATION's synthetic focus on a specific biological process has the potential to become a model for databases aimed at synthesizing the diverse types of biology's ‘big data’ for a wide variety of biological processes.

MATERIALS AND METHODS

Publically available data on GEneSTATION

Publically available data that did not need reanalysis or normalization (e.g. gene age, dN/dS, Protein Atlas) were added to GEneSTATION without modification. Details about these data are listed in the online methods page along with links to the original data source. In addition, the methods page describes in more detail data sets that may be difficult to interpret for non-specialists or that has potential caveats for interpretation. Small information icons on each analysis page or in the relevant subsection of gene pages provide links to both the methods page and the original data.

Sources of organism life history data

Mean gestation length and standard deviations were taken from primary accounts in the research literature focused on reproductive characteristics of each individual species. Neonate development state was either recorded directly from the literature or inferred using average litter size as a proxy[43]. For all species, placental type and shape were taken from Hradecký and Mossman 1987[1]. For non-primates, adult body mass was taken from the PanTHERIA database, as well as average litter sizes, interbirth intervals and maximum longevity for many species[44]. For primates, body mass data specifically reflect adult female body mass[44,45]. Finally, data on

the timing of brain growth across mammals were taken from[46].

Variant reanalysis

To provide consistent genome wide analyses of human genetic variation, variant call format (VCF) files, with coordinates lifted to genome build GRCh38, were downloaded from the 1000 Genomes Project FTP site (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/), representing variants for all 2504 unrelated individuals in the 1000 Genome Project Phase 3 cohort[47]. Variants were filtered to exclude non-SNP variants, fixed sites, and sites with uncalled or unphased genotypes. Variants were validated against dbSNP build 144[48] using the ValidateVariants tool in Genome Analysis Toolkit[49]. VCFtools[50] was used to calculate pairwise FST statistics[51].

Microarray reanalysis

We reanalyzed all microarray datasets that were downloaded from NCBI's Gene Expression Omnibus (GEO)[52] using the R package GEOquery. Ambiguous probes that map to multiple genes were discarded. For multiple probes mapping to a single gene, the probe with median significance value was reported. Pairwise differential expression statistics were computed using the eBayes algorithm in the limma package.

Database design

The data for GEneSTATION is stored in PostgreSQL, a highly reliable and durable open source database. For consistent integration of multiple types of data, we use a highly normalized and non-redundant biology focused schema, Chado[53], which is collaboratively developed by the Generic Model Organism Project[54]. We have written extensions to this schema so that GEneSTATION can handle the large numbers of genomes and diverse associated data (currently 3.4Tb) already loaded as well as those to be added in the future.

GEneSTATION's large datasets have required the development of high-performance custom tools for data loading, which are built using C++ with libpq, and allow loading of genome-wide datasets in seconds and SNP datasets (e.g. 1000 genomes) in minutes. Python with SQLAlchemy provides a flexible pipeline for loading the highly varied data sources in GEneSTATION, annotations, database cross-references, and controlled vocabularies. GEneSTATION loads both standard data formats, e.g. GFF, FSTA, OBO, and generic formats using JSON files for metadata and tab-delimited files for the data, which facilitates integration of diverse analysis pipelines.

GEneSTATION also uses Elasticsearch, a distributed, in-memory search engine for full-text search and as a store for precalculated, denormalized SQL queries that are computationally intensive. Elasticsearch allows GEneSTATION to respond to advanced queries from users, including boolean and full-text, with lower latency and higher throughput than is possible with PostgreSQL alone.

Web interface

The web interface to the GEneSTATION database is delivered by a high performance custom server written in Go, a new language developed at Google to power their web services. The server handles querying both PostgreSQL (using sqlx) and Elasticsearch (using elastic), parsing user input, and performing custom analyses on-the-fly (e.g., analysis of user submitted gene sets). The server provides an NCBI-like query language to query the Elasticsearch search engine and provides uniform access to all GEneSTATION data via a REST interface. The server caches most pages, reducing latency to much less than a second in most cases.

The foundation of the user interface is Bootstrap, an integrated library of html elements, css, and javascript, which allows consistent visual layouts even on mobile devices and provides tools for rich user interaction, e.g. tabs, tooltips, popovers, while supporting users on both

handheld devices (e.g. iPads and iPhones) and older browsers. Additional functionality was achieved with specialized html/css/js libraries, which include: autocompeter.js for immediate search results feedback and autocomplete; multiple libraries, selectize.js, autosize.js, d3.js, venn.js, and react.js where used to build the SynThy tool; multiple libraries, jquery.js, jquery.form.js, responsive-bootstrap-toolkit.js, jquery-highlighttextarea.js, jquery-hoverIntent.js and rPage.js, were used for simpler JavaScript development and richer user interaction.

Visualization

Data in charts and graphs are presented using highcharts.js, data in table formats are displayed with jquery.dataTables.js and the interactive organisms page uses jquery.mixitup.js.

Gene set analysis

P-values for gene set analyses are calculated using the cumulative hypergeometric distribution (similar to Fisher's Exact Test). The background is adjusted for each set to match the number of genes reported in each data set, either analysis or annotation.

ACKNOWLEDGEMENTS

We are grateful to the investigators of the March of Dimes Prematurity Research Center Ohio Collaborative, the March of Dimes leadership and boards of external reviewers, as well as to members of the Rokas lab for their constructive feedback during the construction and development of GEneSTATION. This work was conducted in part using the resources of the Advanced Computing Center for Research and Education at Vanderbilt University.

FUNDING

March of Dimes, as part of the March of Dimes Prematurity Research Center Ohio

Collaborative. Funding for open access charge: March of Dimes, as part of the March of Dimes Prematurity Research Center Ohio Collaborative (http://prematurityresearch.org/ohiocollaborative/).

Conflict of interest statement. None declared.

REFERENCES

1. Hradecký, P. & Mossman, H. W. Vertebrate Fetal Membranes: Comparative Ontogeny and Morphology; Evolution; Phylogenetic significance; Basic Functions; Research Opportunities. *The Journal of Zoo Animal Medicine* **18,** 55 (1987).
2. Carter, A. M. & Mess, A. Evolution of the placenta in eutherian mammals. *Placenta* **28,** 259–262 (2007).
3. Gundling, W. E., Jr & Wildman, D. E. A review of inter- and intraspecific variation in the eutherian placenta. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20140072 (2015).
4. Wildman, D. E. *et al.* Evolution of the mammalian placenta revealed by phylogenetic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 3203–3208 (2006).
5. Wittman, A. B. & Wall, L. L. The evolutionary origins of obstructed labor: bipedalism, encephalization, and the human obstetric dilemma. *Obstet. Gynecol. Surv.* **62,** 739–748 (2007).
6. Haig, D. Intimate relations: Evolutionary conflicts of pregnancy and childhood. in *Evolution in Health and Disease* 65–76 (2007).
7. Brown, E. A., Ruvolo, M. & Sabeti, P. C. Many ways to die, one way to arrive: how selection acts through pregnancy. *Trends Genet.* **29,** 585–592 (2013).
8. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
9. Spong, C. Y. Defining ‘term’ pregnancy: recommendations from the Defining ‘Term’ Pregnancy Workgroup. *JAMA* **309,** 2445–2446 (2013).
10. Beck, S. *et al.* The worldwide incidence of preterm birth: a systematic review of maternal mortality and morbidity. *Bull. World Health Organ.* **88,** 31–38 (2010).
11. Martin, J. A., Hamilton, B. E., Ventura, S. J., Osterman, M. J. K. & Mathews, T. J. Births:

final data for 2011. *Natl. Vital Stat. Rep.* **62,** 1–69, 72 (2013).

12. Guttmacher, A. E., Maddox, Y. T. & Spong, C. Y. The Human Placenta Project: placental structure, development, and function in real time. *Placenta* **35,** 303–304 (2014).
13. Treloar, S. A., Macones, G. A., Mitchell, L. E. & Martin, N. G. Genetic influences on premature parturition in an Australian twin sample. *Twin Res.* **3,** 80–82 (2000).
14. Allen, C. M. & Founds, S. A. Genetics and preterm birth. *J. Obstet. Gynecol. Neonatal Nurs.* **42,** 730–736 (2013).
15. Knox, K. & Baker, J. C. Genomic evolution of the placenta using co-option and duplication and divergence. *Genome Res.* **18,** 695–705 (2008).
16. Hannibal, R. L. *et al.* Copy number variation is a fundamental aspect of the placental genome. *PLoS Genet.* **10,** e1004290 (2014).
17. Elliot, M. G. & Crespi, B. J. Genetic recapitulation of human pre-eclampsia risk during convergent evolution of reduced placental invasiveness in eutherian mammals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370,** 20140069 (2015).
18. Gallant, J. R. *et al.* Nonhuman genetics. Genomic basis for the convergent evolution of electric organs. *Science* **344,** 1522–1525 (2014).
19. Lynch, V. J. *et al.* Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10,** 551–561 (2015).
20. Phillips, J. B., Abbot, P. & Rokas, A. Is preterm birth a human-specific syndrome? *Evol Med Public Health* **2015,** 136–148 (2015).
21. Chaudhari, B. P. *et al.* The genetics of birth timing: insights into a fundamental component of human development. *Clin. Genet.* **74,** 493–501 (2008).
22. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med.* **5,** 34 (2013).
23. Ouyang, Y., Mouillet, J.-F., Coyne, C. B. & Sadovsky, Y. Review: placenta-specific microRNAs in exosomes - good things come in nano-packages. *Placenta* **35 Suppl,** S69–73 (2014).
24. Kosova, G., Stephenson, M. D., Lynch, V. J. & Ober, C. Evolutionary forward genomics reveals novel insights into the genes and pathways dysregulated in recurrent early pregnancy loss. *Hum. Reprod.* **30,** 519–529 (2015).
25. Gracie, S. *et al.* An integrated systems biology approach to the study of preterm birth using '-omic' technology--a guideline for research. *BMC Pregnancy Childbirth* **11,** 71 (2011).

26. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29,** 150–159 (2013).
27. Rosenberg, K. & Trevathan, W. Birth, obstetrics and human evolution. *BJOG* **109,** 1199–1206 (2002).
28. Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucleic Acids Res.* **43,** D1049–56 (2015).
29. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–52 (2015).
30. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43,** D36–42 (2015).
31. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **43,** D6–17 (2015).
32. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–12 (2015).
33. Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M. & Bateman, A. TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res.* **42,** D922–5 (2014).
34. Uzun, A. *et al.* dbPTB: a database for preterm birth. *Database* **2012,** bar069 (2012).
35. Dolan, S. M. *et al.* Synopsis of preterm birth genetic association studies: the preterm birth genetics knowledge base (PTBGene). *Public Health Genomics* **13,** 514–523 (2010).
36. Safran, M. *et al.* GeneCards Version 3: the human gene integrator. *Database* **2010,** baq020 (2010).
37. Costanzo, M. C. *et al.* Saccharomyces genome database provides new regulation data. *Nucleic Acids Res.* **42,** D717–25 (2014).
38. Harris, T. W. *et al.* WormBase 2014: new views of curated biology. *Nucleic Acids Res.* **42,** D789–93 (2014).
39. dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **43,** D690–7 (2015).
40. Eidem, H. R., Ackerman, W. E., 4th, McGary, K. L., Abbot, P. & Rokas, A. Gestational tissue transcriptomics in term and preterm human pregnancies: a systematic review and meta-analysis. *BMC Med. Genomics* **8,** 27 (2015).
41. Tukey, J. W. We Need Both Exploratory and Confirmatory. *Am. Stat.* **34,** 23–25 (1980).
42. Patel, R. R., Steer, P., Doyle, P., Little, M. P. & Elliott, P. Does gestation vary by ethnic

group? A London-based study of over 122,000 pregnancies with spontaneous onset of labour. *Int. J. Epidemiol.* **33,** 107–113 (2004).

43. Müller, D. W. H. *et al.* Dichotomy of eutherian reproduction and metabolism. *Oikos* **121,** 102–115 (2011).
44. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90,** 2648–2648 (2009).
45. Smith, R. J. & Jungers, W. L. Body mass in comparative primatology. *J. Hum. Evol.* **32,** 523–559 (1997).
46. Dobbing, J. Vulnerable Periods in Developing Brain. in *Brain, Behaviour, and Iron in the Infant Diet* 1–17 (1990).
47. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
48. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
49. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).
50. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
51. Cockerham, C. C. & Weir, B. S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40,** 157–164 (1984).
52. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991–5 (2013).
53. Mungall, C. J., Emmert, D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23,** i337–46 (2007).
54. Stein, L. D. *et al.* The generic genome browser: a building block for a model organism system database. *Genome Res.* **12,** 1599–1610 (2002).

CHAPTER III

GEneSTATION 2.0: the integrative encyclopedia of evolutionary and functional data of human coding and non-coding elements for the study of pregnancy-associated diseases

ABSTRACT

GEneSTATION (http://genestation.org) is an integrative resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. The version 2.0 update has expanded the focus of the project to include evolutionary and functional data associated with non-coding elements, including enhancers and transposons. In order to support exploration of these topics, the underlying database has been overhauled from a SQL based schema to a ElasticSearch engine. This change has enabled the development of new and upgraded analysis and visualization tools for the encyclopedia that support the exploration of these additional datasets and the synthesis and test of new hypotheses.

INTRODUCTION

While viviparous reproduction has evolved multiple times within vertebrates, the evolution of mammalian pregnancy and parturition is uniquely characterized by early recognition, invasive placentation, and the evolution of the decidual stromal cell[1]. These changes have been associated with the rewiring of gene regulatory networks via ancient transposable elements[2]. The evolutionary mechanisms behind these processes are crucial to the understanding of human pregnancy[3]. Reproduction and parturition in humans is theorized to have evolved alongside recent developments along the human lineage, such as bipedalism and encephalization[4], and study of these topics has the potential to contribute to a greater understanding of our evolutionary history.

Pregnancy complications are one of the major unsolved human health problems today, with preterm-birth (PTB) remaining a leading cause of infant mortality[5]. Surviving infants are at greater risk for a multitude of long term health problems[6], and current treatments for PTB show little efficacy[7]. Until very recently, there were no robust genetic associations with increased risk for PTB[8–10], despite strong evidence of heritability demonstrated by epidemiological studies[11–14]. The difficulties with understanding the genomic underpinnings of PTB are believed to be due to its nature as a complex, multifactorial syndrome caused by subclinical dysregulation of any one of several potential pregnancy processes during the interaction of the maternal and fetal genomes in a diversity of tissues and environmental conditions[5].

The challenges unique to the study of pregnancy and PTB motivated the design and development of GEneSTATION 1.0, released in 2016[15]. The first release of the database boasted organism life history data[16–20] for 23 mammals and a page for every gene in these genomes. The database integrated several genome wide datasets at three levels of synthesis: evolutionary, organismal, and molecular. Evolutionary metrics included gene level metrics of recent[21] and ancient selection[22,23], estimates of gene age[24], mammalian orthologs and gene family membership[25]. Organismal data included Online Mendelian Information in Man phenotypes[26], RNA and protein expression from Protein Atlas[27], and 106 pregnancy related gene expression studies from the Gene Expression Omnibus[28]. In addition, the first version of the Synthesis and Test of Hypotheses (SynTHy) tool facilitated integrative analysis of gene-associated data.

The past two years have marked major developments in our understanding of the genomics of PTB. In 2017 Zhang et al. associated four loci with gestational duration and spontaneous preterm birth[29], and in 2018 Tan et al. associated three genomic regions with epigenetic changes in the adult genomes of twins born preterm[30]. These findings, along with our knowledge of the evolutionary history of human pregnancy, have highlighted the importance

of understanding the role played by non-coding regulatory elements and human genetic variation in pregnancy phenotypes.

GEneSTATION 2.0 marks an expansion in focus far beyond the gene-ome, with the inclusion of enhancer associations, repetitive element annotations, epigenetic data, and a greatly expanded library of evolutionary and functional analyses of sequence variation data. Furthermore, the integrative Synthesis and Test of Hypotheses (SynTHy) tool (Chapter V), has been updated to support these new genomic features and upgraded to natively perform statistical analyses in real time. To help explore the vast collection of additional data, we have developed the new GeneViewer tool (Chapter VI), a holistic genome browser that visualizes how the distribution of functional and evolutionary metrics in regions compare to the genomic background, as well as provide insights into the properties of non-coding elements within the context of the genomic region and background.

DATA SOURCES AND ORGANIZATION

Non-coding regulatory elements

GEneSTATION 2.0 has added FANTOM5[31] enhancer annotations, tissue/cell-type specific enhancer differential expression, and enhancer-TSS associations. In addition, repetitive element annotations from the Dfam repetitive element database[32] have been added to dataset.

Human Sequence Variation Data

The sequence variation data in GEneSTATION has been upgraded since the 1.0 release. In addition to statistics on recent selection[33] (F_{st}) and ancient selection (D_n/D_s), the encyclopedia now has statistics on nucleotide diversity (π) and tests of Hardy-Weinberg Equilibrium calculated from 1000 Genomes Project Phase 3 data[34]. In addition to these evolutionary metrics, we have incorporated functional association data from the GWAS Catalog[35], PheWas

Catalog[36], and significant eQTLs from the GTEx Project[37].

DATA PRESENTATION

Starting from the 1.0 release of GEneSTATION, the encyclopedia has taken a measured approach to data presentation. While comprehensive JSON dumps of the underlying data are available from each genomic feature, the presentation of excessive amounts of data can quickly become overwhelming for user. Following this philosophy, the browsable pages focus on presenting the data that would be most relevant to a hypothesis generation. The redesigned gene page (Figure 1) shows how the GeneViewer tool is used to summarize the regulatory landscape of a gene. Because most sequence variations will be functionally and evolutionarily neutral, by default the tool only displays variants that have demonstrated function from either the GWAS Catalog, PheWas Catalog, or GTEx project data. This allows an investigator to gain a holistic understanding of where the known functionally relevant variation exists on the gene.

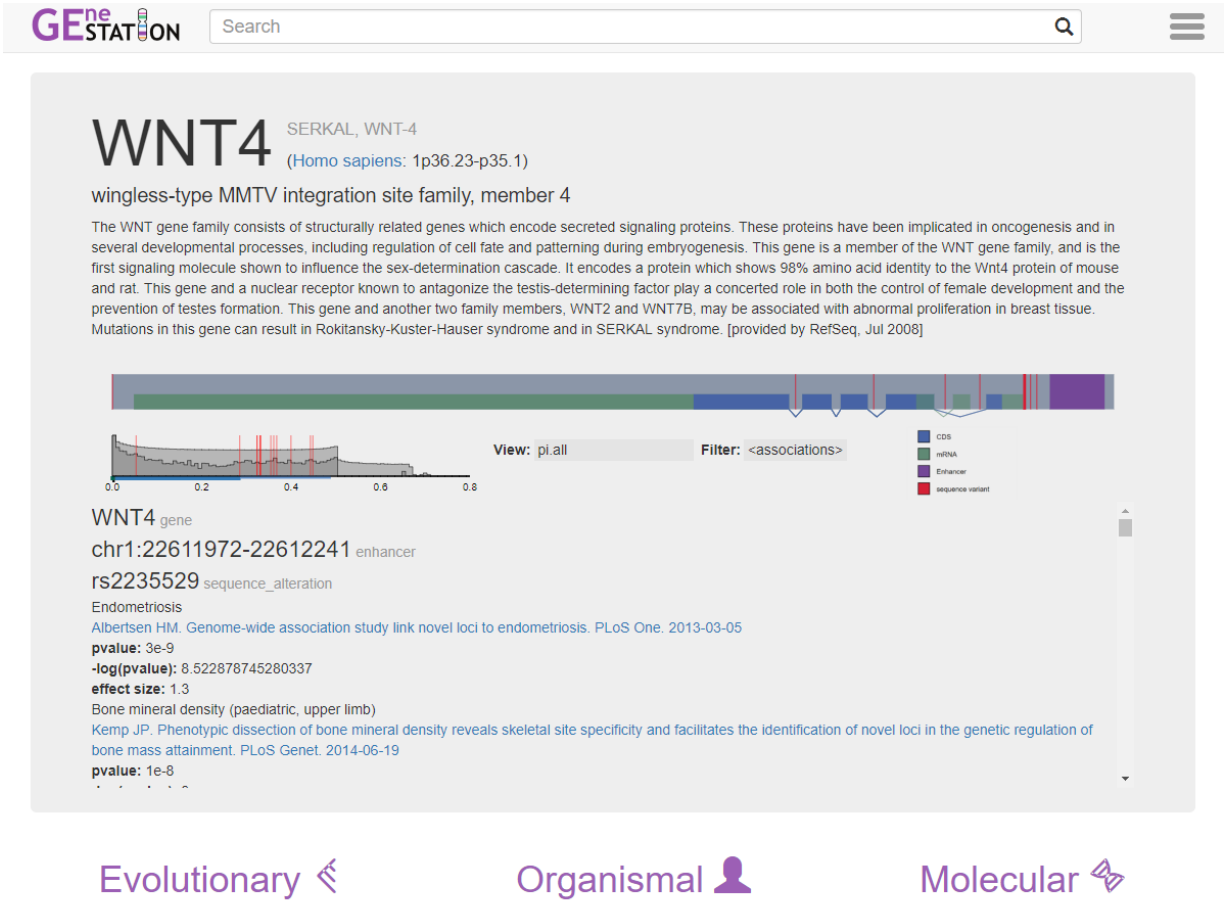

Figure 1. The updated 2.0 gene page featuring the GeneViewer. This new tool visualizes the genomic neighborhood of the gene and its associated regulatory elements: Blue - CDS; Green - mRNA; Purple - Enhancer; Red - sequence variant. Intergenic and intronic regions without regulatory elements are compressed to facilitate visualization of functional regions of the genome. By default, only variants that have functional associations are drawn on the genomic map and the histogram. The histogram shows the genomic background distribution of global nucleotide diversity (pi.all) of all variants in the genome (light grey) and display region (dark grey). Individual variants may be explored in the list below the GeneViewer.

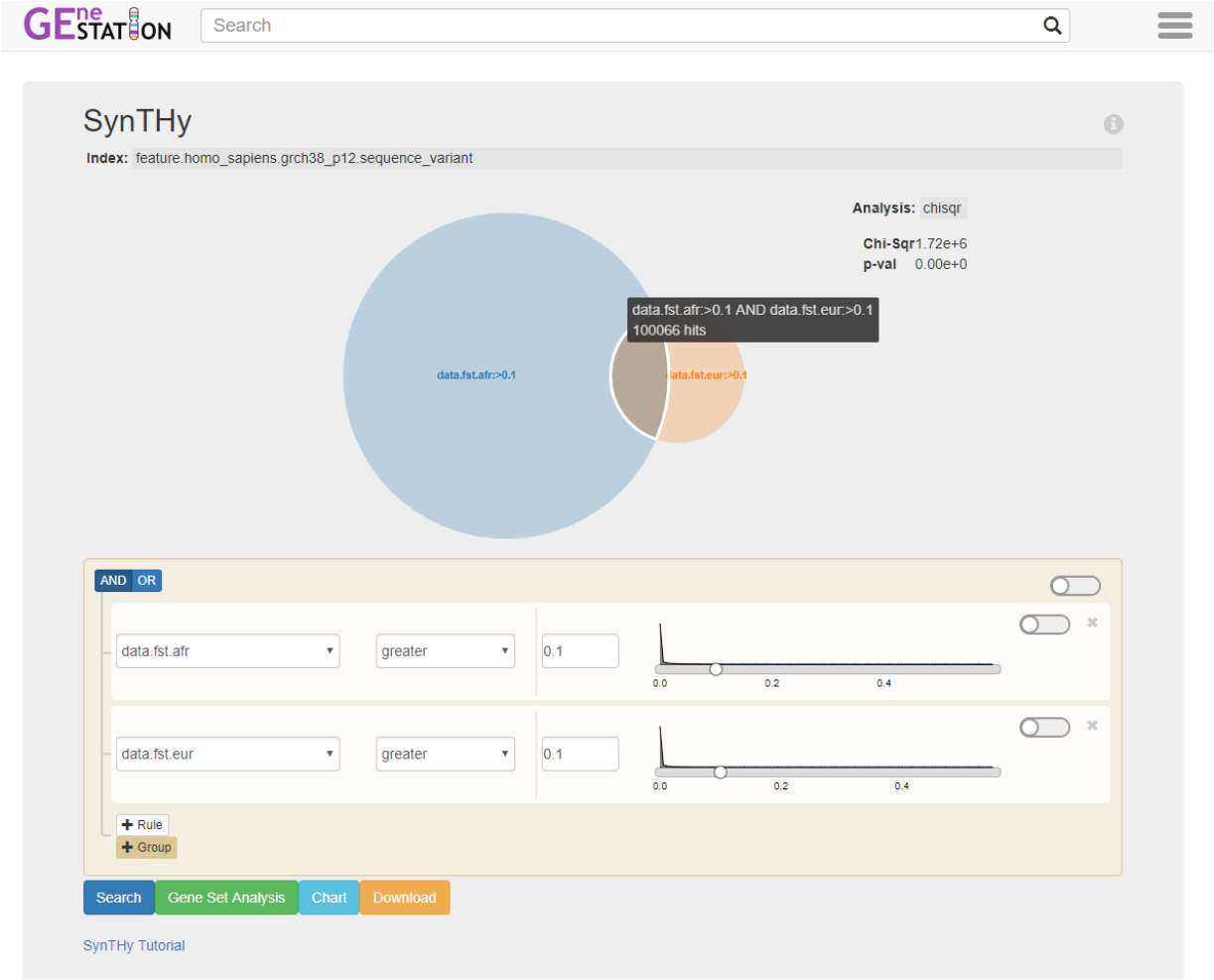

Figure 2. The updated SynTHy tool. In this example, the tool visualizes the intersection of the set of sequence variants with fixation index (F_{st}) greater than 0.1 in the African superpopulation and the European superpopulation. SynTHy automatically performs a Chi-squared Test of Independence for these two sets. The genomic background distribution of each measure is displayed in the rule filters at the bottom, illustrating the extreme skew towards 0 of this metric.

SYNTHESIS AND ANALYSIS

The Synthesis and Test of Hypotheses (SynTHy) tool was introduced in the 1.0 release of the encyclopedia. It facilitated the building of advanced search queries in an explorative manner by visualizing the genomic background distribution of numeric data fields and provided real-time visualization of the sets and intersections defined by compound queries in the form of a Venn diagram. In the 2.0 release, this tool has been updated to be able to perform search and analysis on the non-coding genomic elements in the database. In addition, the tool can perform additional statistical tests on the sets defined by the Venn diagram, such as chi-squared tests of independence and t-tests of dependent variables (Figure 2).

SUMMARY

Recent advances in the understanding of the evolution and regulation of human pregnancy underline the need for the consideration of genes within the context of their regulatory environment. The inclusion of non-coding elements in GEneSTATION 2.0 facilitates the exploration and investigation of a new dimension of genomic data, and the addition of functional and evolutionary sequence variation data allows researchers to perform novel synthetic analyses.

MATERIALS AND METHODS

Database Design

To facilitate the storage of non-coding data, we redesigned the backend database to utilize ElasticSearch, a JSON document store and search engine. The adoption of this new technology necessitated the design of a novel schema (Chapter IV) that could support the genomic data represented by the generic, ontological structure of the Generic Model Organism Database project's Chado schema[39], which had previously been utilized for GEneSTATION 1.0

Non-coding genetic elements and associations

Enhancer-TSS associations, differential tissue and cell-type enhancer expression were downloaded from FANTOM5[31] SlideShare (http://enhancer.binf.ku.dk/presets/). Non-redundant repetitive element annotations were downloaded from Dfam database of repetitive DNA families[32], version 2.0 data release (http://www.dfam.org/web_download/Release/Dfam_2.0/hg38_dfam.nrph.hits.gz). The genomic elements described by these data were loaded into ElasticSearch using the tools in the Genestation Search Engine Toolkit (Chapter IV).

Sequence Variation Data

Sequence variation data was compiled from the 1000 Genomes Project Phase 3 dataset[34] and dbSNP build 151. Evolutionary metrics were calculated from the 1000 Genomes data using . VCFtools[38]. Functional associations were downloaded from the GWAS Catalog v1.0 and PheWAS catalog v1.0 releases. Single tissue cis-eQTL data was downloaded from the GTEx Project version 7 data (https://gtexportal.org/home/datasets). These variants and their associations were loaded into ElasticSearch using the tools in the Genestation Search Engine Toolkit (Chapter IV).

REFERENCES

1. Wagner, G. P., Kin, K., Muglia, L. & Pavlicev, M. Evolution of mammalian pregnancy and the origin of the decidual stromal cell. *Int. J. Dev. Biol.* **58,** 117–126 (2014).
2. Lynch, V. J. *et al.* Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.* **10,** 551–561 (2015).
3. Brown, E. A., Ruvolo, M. & Sabeti, P. C. Many ways to die, one way to arrive: how selection

acts through pregnancy. *Trends Genet.* **29,** 585–592 (2013).
4. Wittman, A. B. & Wall, L. L. The evolutionary origins of obstructed labor: bipedalism, encephalization, and the human obstetric dilemma. *Obstet. Gynecol. Surv.* **62,** 739–748 (2007).
5. Romero, R., Dey, S. K. & Fisher, S. J. Preterm labor: one syndrome, many causes. *Science* **345,** 760–765 (2014).
6. Institute of Medicine (US) Committee on Understanding Premature Birth and Assuring Healthy Outcomes. *Preterm Birth: Causes, Consequences, and Prevention*. (National Academies Press (US), 2010).
7. Smith, V., Devane, D., Begley, C. M., Clarke, M. & Higgins, S. A systematic review and quality assessment of systematic reviews of randomised trials of interventions for preventing and treating preterm birth. *Eur. J. Obstet. Gynecol. Reprod. Biol.* **142,** 3–11 (2009).
8. Bezold, K. Y., Karjalainen, M. K., Hallman, M., Teramo, K. & Muglia, L. J. The genomics of preterm birth: from animal models to human studies. *Genome Med.* **5,** 34 (2013).
9. Swaggart, K. A., Pavlicev, M. & Muglia, L. J. Genomics of preterm birth. *Cold Spring Harb. Perspect. Med.* **5,** a023127 (2015).
10. Monangi, N. K., Brockway, H. M., House, M., Zhang, G. & Muglia, L. J. The genetics of preterm birth: Progress and promise. *Semin. Perinatol.* **39,** 574–583 (2015).
11. Goldenberg, R. L., Culhane, J. F., Iams, J. D. & Romero, R. Epidemiology and causes of preterm birth. *Lancet* **371,** 75–84 (2008).
12. Winkvist, A., Mogren, I. & Högberg, U. Familial patterns in birth characteristics: impact on individual and population risks. *Int. J. Epidemiol.* **27,** 248–254 (1998).
13. Porter, T. F., Fraser, A. M., Hunter, C. Y., Ward, R. H. & Varner, M. W. The risk of preterm birth across generations. *Obstet. Gynecol.* **90,** 63–67 (1997).
14. Boyd, H. A. *et al.* Maternal contributions to preterm delivery. *Am. J. Epidemiol.* **170,** 1358–1364 (2009).
15. Kim, M. *et al.* GEneSTATION 1.0: a synthetic resource of diverse evolutionary and functional genomic data for studying the evolution of pregnancy-associated tissues and phenotypes. *Nucleic Acids Res.* **44,** D908–16 (2016).
16. Jones, K. E. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. *Ecology* **90,** 2648–2648 (2009).
17. Smith, R. J. & Jungers, W. L. Body mass in comparative primatology. *J. Hum. Evol.* **32,**

523–559 (1997).
18. Dobbing, J. Vulnerable Periods in Developing Brain. in *Brain, Behaviour, and Iron in the Infant Diet* 1–17 (1990).
19. Hradecký, P. & Mossman, H. W. Vertebrate Fetal Membranes: Comparative Ontogeny and Morphology; Evolution; Phylogenetic significance; Basic Functions; Research Opportunities. *The Journal of Zoo Animal Medicine* **18,** 55 (1987).
20. Müller, D. W. H. *et al.* Dichotomy of eutherian reproduction and metabolism. *Oikos* **121,** 102–115 (2011).
21. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
22. Gayà-Vidal, M. & Albà, M. M. Uncovering adaptive evolution in the human lineage. *BMC Genomics* **15,** 599 (2014).
23. Plunkett, J. *et al.* An evolutionary genomic approach to identify genes involved in human birth timing. *PLoS Genet.* **7,** e1001365 (2011).
24. Capra, J. A., Williams, A. G. & Pollard, K. S. ProteinHistorian: tools for the comparative analysis of eukaryote protein origin. *PLoS Comput. Biol.* **8,** e1002567 (2012).
25. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34,** D572–80 (2006).
26. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43,** D789–98 (2015).
27. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347,** 1260419 (2015).
28. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res.* **41,** D991–5 (2013).
29. Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 1156–1167 (2017).
30. Tan, Q. *et al.* Epigenetic signature of preterm birth in adult twins. *Clin. Epigenetics* **10,** 87 (2018).
31. Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.* **16,** 22 (2015).
32. Hubley, R. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res.* **44,** D81–9 (2016).

33. Cockerham, C. C. & Weir, B. S. Covariances of relatives stemming from a population undergoing mixed self and random mating. *Biometrics* **40,** 157–164 (1984).
34. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
35. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
36. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31,** 1102–1110 (2013).
37. Carithers, L. J. *et al.* A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. *Biopreserv. Biobank.* **13,** 311–319 (2015).
38. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).
39. Mungall, C. J., Emmert, D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23,** i337–46 (2007).

CHAPTER IV

The Genomic Search Engine

ABSTRACT

The increase in the size of biological datasets has necessitated the adoption of new database technologies designed to address the challenges of “big data”. Previous published designs for biological databases have depended on relational database structures that perform poorly when handling datasets with millions of samples (eg. sequence variation data). This has necessitated the design and development of a novel data schema to handle genomic data in non-relational databases. The database schema and associated tools, collectively called the Genestation Search Engine Toolkit, provide a complete solution for managing a biological database in the ElasticSearch JSON document store and search engine.

INTRODUCTION

Database Management Systems

The first generation of database technologies arose with the development of disk based data storage systems which enabled non-sequential access to data in a manner not possible with earlier tape storage systems[1]. These “navigational databases” were the first database management systems and include the network model database, the Integrated Data Store (IDS) developed at General Electric[1,2], and the hierarchical model database, the IBM Information Management System.

Navigational database systems provided a standardized method of storing information; however, they required the programmer to manually explore the links within the stored data and most notably lacked the ability to efficiently search data fields. To address these issues, Edgar Codd proposed the “relational database”, a method of normalizing data and splitting it into tables that could be searched independently and reconstituted via relational algebra and tuple

relational calculus[3]. These concepts enabled the development of a new language for querying the database. Whereas previous *procedural languages* required the programmer to manually explore the structure of the database, this new *declarative language* would allow a programmer to simply specify the desired output from the database system and the task of searching and fetching this information would be handled by the machine. This concept would be most famously implemented by the Structured Query Language (SQL) developed by Chamberlin and Boyce at IBM[4].

Biological Databases

With the growing availability of sequence data following the elucidation of Fred Sanger's "first sequence" in 1955, multiple independent research groups realized the need for a method of organizing this new information. The first genomic databases began their development as private projects: Margaret Dayhoff's *Atlas of Protein Sequence and Structure*[5] and Walter Goad's Los Alamos Sequence Library, implemented on the FRAMIS relational database management system[6]. The Los Alamos Sequence Library would go on to become the Genbank nucleic acid sequence database in 1982[7].

The sequencing of "model organisms" during the 1990s would lead to the creation and spread of model organism databases, of which the first was the Flybase *Drosophila* database in 1997[8]. In the coming years, more such projects were completed including the *Saccharomyces* Genome Database[9], the WormBase Nematode Genome Database[10], and the Mouse Genome Database[11]. These four model organism database projects collaborated with the aim to provide a common set of tools that could be utilized by all model organism databases. This collaboration grew to include many model organism projects and resulted the creation of several projects and associated tools, including the Gene Ontology[12], a controlled vocabulary of core biological functions for genes, the Sequence Ontology[13], a controlled vocabulary for genome

annotations, and the Generic Model Organism Database (GMOD) Chado schema[14], a generic ontological schema to describe biological data in SQL.

Today, there are thousands of genomic databases, with the *Nucleic Acids Research* Molecular Biology Database Collection counting 1737 entries in its 2018 database issue[15], including databases focused on sequence data, structural data, variation data, and projects focused on specific diseases, organisms, and cells. The National Center for Biotechnology Information (NCBI) Entrez system supports search over 39 databases of biological data[16].

On the Need for Biological Analytical Processing Systems

Recent advances in the ability to assay non-coding and regulatory elements of the genome in a high-throughput fashion have greatly expanded the scale and scope of genomic data available in the literature. As the volume of biological information increases, the cost of reconstituting data that has been split up into a relational database structure ("joining") increases with proportionally the size of the data and the number of joins. Traditional relational database schemas, such as the GMOD project's Chado schema, do an excellent job at describing data in a manner that is compact and consistent. However, due to the highly normalized structure of the data, complex queries can become prohibitively costly to execute.

During the development of GEneSTATION 2.0 (Chapter III), I experienced difficulty with delivering satisfactory query performance using the Chado schema as we moved from querying genic data to non-coding data, such as SNPs. The increase in scope from thousands to millions of data points placed a growing burden on the SQL database. It became apparent that while a relational database was an excellent *online transaction processing system* it was imperative to create an efficient *online analytical processing system* to support web-based user interaction. This recent expansion of data has affected other large scale projects, including dbSNP, which announced their intention to deprecate their SQL database solution within the 2018 year (https://

ncbiinsights.ncbi.nlm.nih.gov/2017/07/07/dbsnp-redesign-supports-future-data-expansion/). In a 2016 evaluation of relational and non-relation datastores for the storage of biological data, Schulz et al. found that relational databases performed worse than noSQL solutions, and that the differences between the solutions grew wider with increasing data size[17].

The challenges that come with the need to store large datasets, such as sequence variation data, are an ongoing field of research in computer science today. Often called the problem of "big data", the unprecedented size of modern datasets are overwhelming traditional database designs. The reasons for this are manifold, but can be mainly attributed to the cost of reconstituting a document once it has been split into a relational structure. Furthermore, it is difficult for a relational database to take full advantage of recent advances in parallel computation. This has encouraged the development of new non-relational databases, such as ElasticSearch[18], Google's BigTable[19], and Amazon's SimpleDB[20], to handle datasets at this larger scale.

In this paper, I present a framework for storing genomic data in ElasticSearch, an open-source JSON document store and search engine with proven performance in large scale commercial applications. In addition, I provide tools to assist bioinformaticians in managing and interfacing with this new database solution. The database schema and associated tools are collectively called the Genestation Search Engine Toolkit.

A Search Framework for Genomic Data

The success of the Generic Model Organism Database Project's Chado schema demonstrates the value of establishing a standardized database schema. The creation of a common data framework has fostered the development of a rich ecosystem of tools. This tookit seeks to extend this ecosystem by providing a analytical search engine framework for web based applications and tools. A Chado database is not necessary to use the Genestation

Search Engine Toolkit, and it has been designed to supplement existing model organism databases with a scalable search solution in a schema agnostic manner.

The toolkit has been designed to mirror the original design principles behind the Chado schema to provide a solution that is generic, extensible, and available as open source. Furthermore, it aims to provide a common platform that can be utilized by an individual researcher with a few data files to a large database project encompassing an array of genomes. It accomplishes this by enforcing a shared concept of genomic features and locations while allowing each analysis to customize and define its own data structures.

METHODS

Genestation uses ElasticSearch for storage and search. It includes a python command line interface (CLI) to help interact with the ElasticSearch instance and create data analysis workflows. It also provides a javascript library for web-based tools and interaction. The python and javascript libraries are available via the Python Package Index and the Node Package Management system respectively. Source code for these packages are available on GitHub at www.github.com/genestation.

DESIGN

Genomic data is loaded via the Genestation CLI, which supports loading data from GFF, FASTA, VCF, JSON, and TSJV (tab-separated JSON values), via a Genomic Data Descriptor (GDD) file (see Appendix 3). Once the data is loaded into Elasticsearch, the CLI provides utilities for analysis and maintenance of the database. For a complete reference of the functionality of the CLI, see Appendix 2.

The CLI stores data in ElasticSearch as JSON documents. These documents are collected as indexes that may be queried independently or as a group. Fields in these

documents have distinct types (https://www.elastic.co/guide/en/elasticsearch/reference/current/mapping-types.html) which ElasticSearch requires to remain consistent within a single index whether they are defined with an explicit mapping or dynamically mapped during document upload. Notably, ElasticSearch does not distinguish between a singular data type and an array of that type.

Indexes

The Genestation data model recognizes 3 primary index types. The first is the Genome Index. This index always has the name `'genome'` and contains Genome Objects, which describe the organism and genome version. The genus, species, subspecies (if any), and genome version are used to construct a **genome identifier**. This identifier takes the form *`genome_species.version`* or *`genome_species_subspecies.version`*, where all alphabetic characters are converted to lowercase and periods and spaces are converted to underscores (eg. `'homo_sapiens.grch38_p12'` and `'canis_lupus_familiaris.3_1'`). Each genome identifier should refer to a unique genome within a genestation instance. The genome identifier is used to organize the other indexes in the database.

Genomic features are indexed into Feature Indexes. These are named in the form `feature.`*`genome_identifier.feature_type`*, where *`feature_type`* is a Sequence Ontology term (eg. `'feature.homo_sapiens.grch38_p12.sequence_variant'` and `'feature.canis_lupus_familiaris.3_1.gene'`). By prefixing all feature indexes with the constant string `"feature."`, these indexes are namespaced under this identifier and allows ElasticSearch to perform multi-index queries on all Feature Indexes simultaneously.

The final core index is the Meta Index, which stores general purpose index-level statistics and metadata. This index can be generated for any ElasticSearch index by the Genestation CLI and is named in the form `meta.`*`index_name`* (eg.

`'meta.feature.homo_sapiens.grch38_p12.sequence_variant'` and `'meta.feature.canis_lupus_familiaris.3_1.gene'`). These indexes store Field Objects, which describe fields in the associated index and can be used to store descriptive information about the index. The Genestation CLI should be used to (re)generates a Meta Index after loading new genomic data into a Feature Index (See Appendix 2). See Appendix 1 for a more complete description of the Genome Index, Feature Indexes, and Meta Indexes.

Genomic Data

Genomic features in Genestation are represented as Genome Feature Objects which are typed by Sequence Ontology terms. These objects describe the location (if any) of the feature and any analysis data (eg. expression level, odds-ratio) associated with that feature. In addition, these objects describe any ontological associations (eg. Gene Ontology function) and associations with other features. The Genome Feature Object structure is recursive and supports nested objects to represent features described in the General Feature Format (GFF).

Genome Feature Objects are stored in Feature Indexes (ie. an ElasticSearch index). Each index contains features of one single Sequence Ontology type (and any nested children) from a single genome. These feature indexes are grouped together into a genome and referenced by a Genome Object contained in the Genome Index. The Genome Index contains references to all the genomes that are available in the Genestation instance. For a complete description of the Genome Feature Object format, see Appendix 1. A high level summary of these object follows.

Locations: Genomic locations in Genestation use the BED format "0-start half-open" coordinate system (roughly analogous to Chado's interbase coordinate system) where locations are ranges with inclusive starts and exclusive ends. For example, a feature comprising the first four bases

of a region would have the location `start = 0` and `end = 4`. Locations are represented with the locrange ElasticSearch type, with in the form `{“gte”: start, “lt”: end}`. Discontiguous locations are represented with locrange arrays, and require no special treatment in ElasticSearch. This convention should be followed whether describing the primary location of a feature or describing secondary alignment or coordinates (eg. in sequence alignments and Hidden Markov Models). Complex locations including strand and phase information are discussed in further detail in Appendix 1.

Analysis Data: Analysis data is stored dynamically within the Genome Feature Object. Data integrity may be explicitly defined with an index level mapping; however, this is not strictly necessary as ElasticSearch automatically checks for type consistency within a field across each index. Analysis fields may be given detailed descriptions in an associated Meta Index (see Appendix 1).

The hierarchical nature of the JSON syntax allows each analysis to adopt its own data structure. This affords great flexibility to the search engine, but it requires the adoption of idioms to enforce consistent representations and allow code reuse. What follows are idioms that have been used in GEneSTATION to represent annotations and analysis results.

Publication Provenance: Publications may be described with a `pub` field. This may be present in the top level data object to represent feature discovery, or be nested as a subfield of an annotation to represent provenance of an association. If the `pub` field is of type string, it should be a complete citation in the MEDLINE/PubMed Citation Format (https://www.nlm.nih.gov/bsd/policy/cit_format.html). Alternatively, this field may be of type long, in which case it is interpreted as the PMID of the publication.

Ontology Association and Database Cross References: Ontological associations and database cross references must follow the format described by the GFF3 specification (https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md). These follow the format "DBTAG:ID" and are stored using the ElasticSearch keyword type. Hierarchical ontological queries (ie. is_a assertions) are supported using ElasticSearch's synonym token filter (https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-synonym-tokenfilter.html), which computes the ontological transitive closure of each vocabulary term at index time.

Numerical Data: Analyses that generate numerical data should use an appropriate numeric datatype (https://www.elastic.co/guide/en/elasticsearch/reference/current/number.html). Unlike the Chado schema, analysis results may label their output types explicitly (eg. `fpkm` instead of `normscore`) and including additional outputs (eg. standard deviation) does not require extension of the base schema. The meaning of each numerical field may be determined for each analysis, and additional context (eg. human readable descriptions, default sort order) may be stored in the `"description"` field of FIeld Objects in a Meta Index.

PERFORMANCE

The performance of querying genomic data in ElasticSearch (v5.6.5) using Genome Feature Objects was compared with the storage in PostgreSQL (v9.6.2) using the GMOD Chado schema. These tests were done with the datasets served by the current GEneSTATION build, using a prewarmed cache over 10 runs.

Table 1. Time to count the features on human chromosome 2 in seconds

	SQL	ElasticSearch
Genes	49.018	0.016
SNPs	50.101	0.020

These results show how query times may improve in real-world scenarios with the usage of noSQL database technologies. In each case, the unwarmed query took significantly more time, around 200 seconds for SQL and around 150 milliseconds for ElasticSearch.

The very similar runtime for the gene and SNP query in SQL can largely be explained by the underlying structure of the Chado schema, which combines all genomic feature types into one table. This forces the database to filter through records of all types in order to search for any given feature. In the Genome Feature Object model, each feature type is partitioned into its own index, which ensures that there are no performance penalties with increasing feature diversity in the database.

The large difference in magnitude in the runtime of the two systems is also a side effect of the performance penalty of data normalization in the SQL database. In the Chado schema, it is necessary to perform two additional search operations to determine whether a gene is located on a chromosome, while in the Genome Feature Object model this information can be accessed directly (Figure 1).

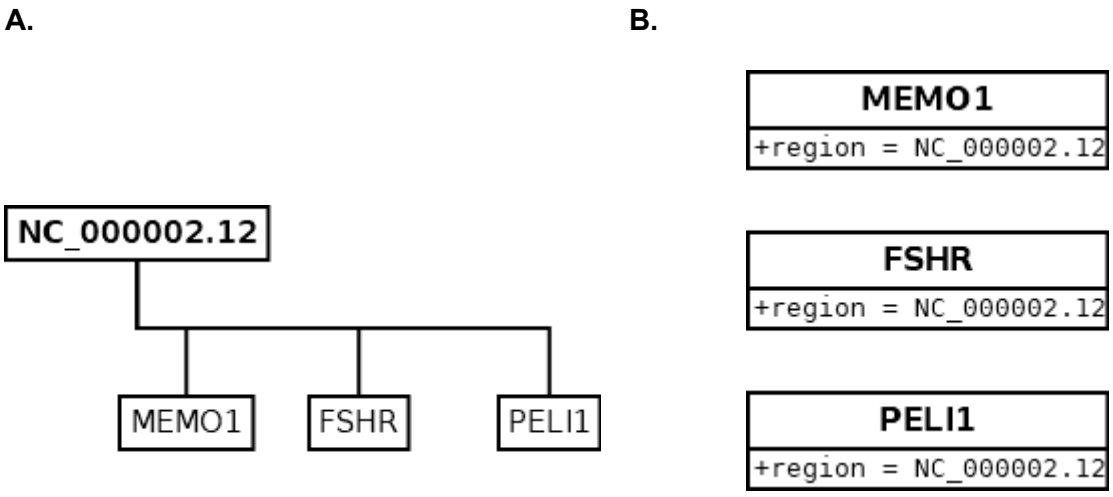

Figure 1. Feature location data models in relation and non-relational databases. A) The relational data model. Each box represents an entry in the feature table, while lines represent locational relationships. This model is space efficient and updating the chromosome name is efficient. However, determining the location of a gene requires 2 joins. First, the gene is joined to the feature location, and then the location is joined to the chromosome. B) The non-relational data model. Each gene is a document with the full location information. Data is duplicated and this model is space inefficient and is costly to update. However, the location of the gene is able to be queried directly without any joins.

CONCLUSIONS

The progressive enhancements of computational technology and the growing needs of bioinformaticians have gradually shifted the focus and needs of genomic tools. Early database systems prioritised the optimal utilization of storage space, and thus relational databases were designed to achieve extreme data compaction through data normalization.

As technology has progressed, the cost of both processors and storage solutions have massively plummeted. In addition, advancements in network computing have created new paradigms for parallel computing. This has allowed new database technologies (often collectively referred to as noSQL) to take advantage of this increase in performance to develop more distributed systems that undertake aggressive replication that sacrifices traditional data

normalization techniques to allow for arbitrarily wide horizontal scaling (ie. distributed computing systems). The advantages of these systems is the ability to handle increasing prodigious data sets while maintaining quick response times via massive parallelization.

While these systems have been heavily utilized within the commercial space, there has yet to be a standard for genomics and bioinformatics investigators to take advantage of these new tools. It is my hope that this new standard fosters the development of new tools for the investigation of biological questions, as well as ushering in a new wave of technological advancement.

REFERENCES

1. Bachman, C. W. The programmer as navigator. *Commun. ACM* **16,** 653–658 (1973).
2. Haigh, T. Charles W. Bachman: Database Software Pioneer. *IEEE Ann. Hist. Comput.* **33,** 70–80 (2011).
3. Codd, E. F. A relational model of data for large shared data banks. *Commun. ACM* **13,** 377–387 (1970).
4. Chamberlin, D. D. Early History of SQL. *IEEE Ann. Hist. Comput.* **34,** 78–82 (2012).
5. Dayhoff (Ed), M. O. & Silver Spring, Md. National Biomedical Research Foundation. *Atlas of Protein Sequence and Structure*. (1965).
6. Kanehisa, M., Fickett, J. W. & Goad, W. B. A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic Acids Res.* **12,** 149–158 (1984).
7. Burks, C. *et al.* The GenBank nucleic acid sequence database. *Comput. Appl. Biosci.* **1,** 225–233 (1985).
8. Gelbart, W. M. *et al.* FlyBase: a Drosophila database. The FlyBase consortium. *Nucleic Acids Res.* **25,** 63–66 (1997).
9. Cherry, J. M. *et al.* SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26,** 73–79 (1998).
10. Harris, T. W. *et al.* WormBase: a comprehensive resource for nematode research. *Nucleic Acids Res.* **38,** D463–7 (2010).

11. Blake, J. A. *et al.* The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res.* **39,** D842–8 (2011).
12. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25,** 25–29 (2000).
13. Eilbeck, K. *et al.* The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.* **6,** R44 (2005).
14. Mungall, C. J., Emmert, D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23,** i337–46 (2007).
15. Rigden, D. J. & Fernández, X. M. The 2018 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res.* **46,** D1–D7 (2018).
16. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46,** D8–D13 (2018).
17. Schulz, W. L., Nelson, B. G., Felker, D. K., Durant, T. J. S. & Torres, R. Evaluation of relational and NoSQL database architectures to manage genomic annotations. *J. Biomed. Inform.* **64,** 288–295 (2016).
18. Gormley, C. & Tong, Z. *Elasticsearch: The Definitive Guide*. ('O'Reilly Media, Inc.', 2015).
19. Chang, F. *et al.* Bigtable. *ACM Trans. Comput. Syst.* **26,** 1–26 (2008).
20. Sciore, E. SimpleDB. in *Proceedings of the 38th SIGCSE technical symposium on Computer science education - SIGCSE '07* (2007). doi:10.1145/1227310.1227498

CHAPTER V

SynTHy: Synthesis and Testing of Hypotheses

BACKGROUND

Genomic databases are tasked with organizing and presenting biological data to users in an interactive and intuitive manner. The diverse array of annotations and analyses that are compiled in an integrative database may span a variety of evolutionary and functional data types. To help investigators understand and leverage the information contained within these databases, it is important to provide a powerful and responsive search interface.

Traditionally this need is addressed with interactive query builders. These tools are useful because they teach users the kind of data that is available, allowing them to search in an exploratory manner. However, in general these tools do not provide users with an understanding of the nature of fields available in the database. This is important because often the distributions of these data values may be deviate from assumptions maintained by the investigator, for example normality. In addition, these tools often do not provide feedback on how compound queries perform in concert to sample information from the database.

SynTHy aims to augment the basic query builder with dynamic interactive visualizations that show the distribution of each data field while the user is searching. This allows investigators to maintain an intuitive understanding of the overall database during query construction. In addition, SynTHy provides a responsive Venn Diagram that reflects how complex searches sample from the database. This allows users to understand how different components of a multi-part query overlap in real-time. Together, these visualizations improve a researcher's understanding of the underlying data in order to gain greater insights and generate new hypotheses.

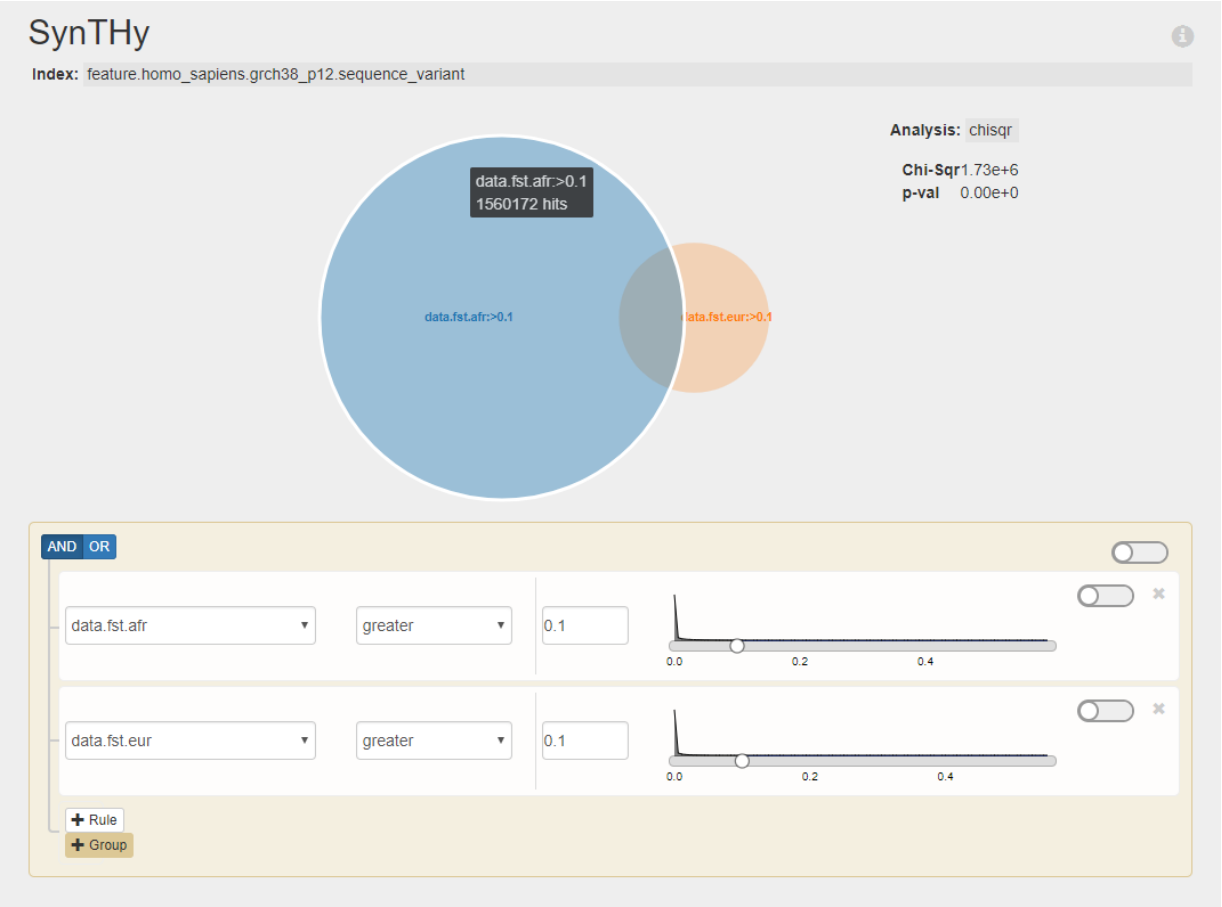

Figure 1. The SynTHy user interface. The index selector at the top indicates the organism and type of genomic feature currently being searched: sequence variants in the human genome. The interactive Venn diagram shows the relative size of each dataset: Blue, the set of variants with fixation index > 0.1 in the African superpopulation; Orange, the set of variants with fixation index > 0.1 in the European superpopulation. The analysis panel in the upper right reports statistical data about the sets, currently performing a Chi-square test of independence. The rule filters at the bottom are used to construct the data sets visualize the distribution of each data field, in this case showing the extreme skew of both datasets towards 0. Data taken from the 1000 Genomes project Phase 3 data[1].

IMPLEMENTATION

The Synthesis and Test of Hypotheses (SynTHy) tool uses the Genestation Search Engine Toolkit (Chapter IV) to store genomic data and generate statistical data. Statistical analyses are performed using ElasticSearch aggregations and the jStat JavaScript statistical library. The tool leverages this data to generate dynamic visualizations of genomic data and an advanced query builder using HTML and Javascript. As a web-based tool, it may be used in a standalone manner or within the context of an existing genomic encyclopedia as a database search widget. Source code is available at https://github.com/genestation/synthy.

RESULTS AND DISCUSSION

Many examples of search tools exist in model organism databases, such as the query builders in the Flybase *Drosophila* genome database[2] and the *Saccharomyces* Genome Database[3], and bioinformatics institutes, such as the National Center for Biotechnology Information's Entrez Gene[4]. All of these tools support interactive query construction and integrative search.

SynTHy is more than just an advanced query builder. It provides real-time, interactive visualizations of statistical information about each data field, including detailed histograms of numeric data fields and autocomplete suggestions for textual search fields, facilitating exploration of the data while in the process of formulating a search. In addition, as a query is constructed, a Venn diagram displays how each component of the search input and their subsets overlap within the database. The responsive diagrams help investigators gain an intuitive understanding of the properties of the data and their relationships to each other (Figure 1). With better knowledge of the underlying information, a user is able to synthesize new hypotheses and receive instant feedback from the database.

In addition to these dynamic visualizations, SynTHy is able to perform statistical tests using the rules defined by the query builder. Supported statistical tests include the chi-squared

test of independence between two sets[5], t-test of means of a dependent variable of two sets[6], Mann-Whitney U test of a dependent variable of two sets[7], paired t-test of means for two dependent variables of one set[8], Wilcoxon signed-rank test of two dependent variables of one set[9], and Analysis of Variance of a dependent variable of K sets[5]. With access to tools to assess the significance of the trends in the data, an investigator is able test integrative hypotheses about the relationships between different analyses in the database.

Extensibility

The behavior of SynTHy may be extended via JavaScript callbacks passed into the web widget. These callbacks can provide sets of genomic features and their data values to external scripts. This enables SynTHy to serve as an advanced search engine and query builder for an existing genomic database, as well as an entry to other data analysis and visualization tools.

CONCLUSION

Investigators can utilize the data visualization and analysis tools provided by SynTHy to synthesize and test hypotheses about the annotations and analyses contained within genomic databases. This tool facilitates gaining a holistic understanding of the data types associated with genomic features while enabling rigorous statistical analysis of queries in an interactive manner. Existing genome databases may utilize the extension functionality to use SynTHy as a search engine and query builder, while other tools can use SynTHy as the front-end of their own analysis and visualization tools.

REFERENCES

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

2. Gramates, L. S. *et al.* FlyBase at 25: looking to the future. *Nucleic Acids Res.* **45,** D663–D671 (2017).
3. Cherry, J. M. The Saccharomyces Genome Database: Advanced Searching Methods and Data Mining. *Cold Spring Harb. Protoc.* **2015,** db.prot088906 (2015).
4. Brown, G. R. *et al.* Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* **43,** D36–D42 (2014).
5. Fisher, R. A. *Statistical Methods For Research Workers*. (Genesis Publishing Pvt Ltd, 1925).
6. STUDENT. THE PROBABLE ERROR OF A MEAN. *Biometrika* **6,** 1–25 (1908).
7. Mann, H. B. & Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Stat.* **18,** 50–60 (1947).
8. Rubin, D. B. Matching to Remove Bias in Observational Studies. *Biometrics* **29,** 159 (1973).
9. Wilcoxon, F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* **1,** 80 (1945).

CHAPTER VI

GeneViewer: Integrative visualisation of the holistic gene

BACKGROUND

Since the advent of genomics, there has been an increasing availability of both genic and non-genic data, such as sequence variants, promoters, and enhancers. These additional genetic elements are important for the understanding the full regulatory context of a gene, so it is important to develop visualizations that enable researchers to gain greater insights into the function of a gene and its associated regulatory elements.

Traditional genome browsers take a comprehensive approach to rendering genomic elements. These tools are powerful because they allow users to see every element in the genome, but can be difficult to use due to how quickly their displays can overwhelm users with information. In addition, these tools often lack high level analyses and visualizations that allow an investigator to explore the genome in terms of regions, as opposed to single elements at a time.

GeneViewer aims to solve this problem by prioritizing genomic features according to their relevance. This allows a researcher to focus their attention on elements that are most interesting to their concerns and gain insights about these features in their genomic context. In addition, GeneViewer calculates summary statistics for data fields and shows how regions of the genome deviate from the genomic background. This allows for a better understanding of these regions and how they vary across the genome.

IMPLEMENTATION

GeneViewer uses the Genestation Search Engine Toolkit (Chapter IV) to store genomic data and generate statistical data. Statistical analyses are performed using ElasticSearch

aggregations (https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations.html) and the jStat JavaScript statistical library (http://jstat.github.io}. The tool leverages this data to generate dynamic visualizations of genomic data using HTML and Javascript. As a web-based tool, it may be used in a standalone manner or within the context of an existing genomic encyclopedia as a gene figure widget. Source code is available at https://github.com/genestation/gene-viewer.

Figure 1. GeneViewer showing the *WNT4* locus. Blue: CDS, Green: mRNA, Purple: Enhancer; Red: Sequence variant. Intronic and non-enhancer intergenic regions are compressed. Only sequence variants with functional associations are drawn in red. The histogram in the bottom left shows the distribution of global nucleotide diversity for all variants in across the whole genome (light grey) and the display region (dark grey), with red lines corresponding to the location of variants with functional associations. Data taken from 1000 Genomes project Phase 3 data[1], dbSNP build 151[2], GWAS catalog[3], and PheWAS catalog[4].

RESULTS AND DISCUSSION

Several genome browsers have been developed to visualize genomic elements and data, including the Ensembl genome browser[5], the UCSC genome browser[6], the NCBI Genome Data Viewer[7], Gbrowse[8], and Jbrowse[9]. These tools support the display and alignment of genomic elements.

GeneViewer does not aim to be a genome browser in the traditional sense that exhaustively displays every genomic element true to scale. Rather, it aims to display gene models, associated regulatory elements, and the overall landscape of sequence variation in a holistic manner by compressing intergenic and intronic regions to make it easier to see transitions between coding and non-coding regions. By default, GeneViewer only shows

sequence variants that have been associated with a functional outcome (Figure 1).

GeneViewer calculates the distribution of analytical values, such as nucleotide diversity and fixation index, for the genomic background. Users can select regions of the diagram to see how the distribution of values for that region compares against the background. Variants are plotted on this distribution so that investigators can track how a single variant behaves over multiple metrics (Figure 2).

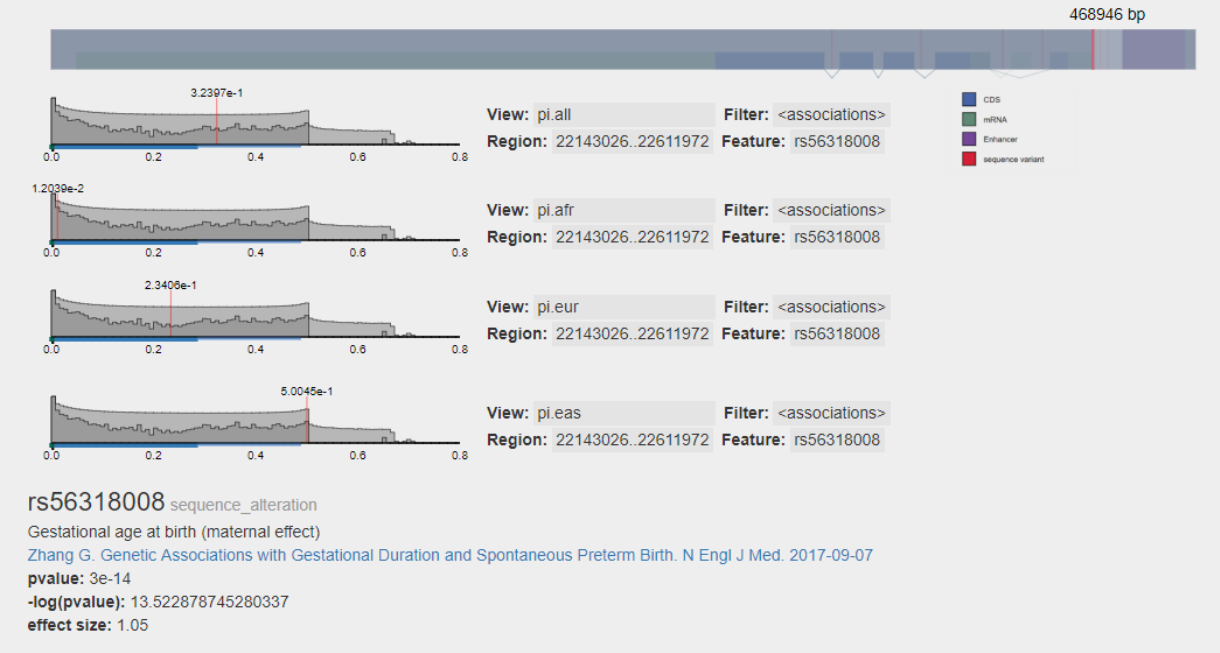

Figure 2. Examining human SNP rs56318008. In this example the nucleotide diversity of the SNP is shown in multiple contexts. The histogram shows the region highlighted at the top as the local background (dark grey) and the genomic background (light grey). The selected SNP shows much lower nucleotide diversity in the African superpopulation compared to the global background and the European and East Asian superpopulations. Data taken from 1000 Genomes project Phase 3 data[1] and Zhang 2017[10].

CONCLUSIONS

GeneViewer provides a way for researchers to gain a holistic understanding of genes and other genomic elements within the context of their regulatory environment. By rendering only the most informative lines of evidence, users are able to explore the data intuitively. GeneViewer

generates visualizations of the genome that are similar to traditional gene model diagrams and provides the ability to investigate how functional and evolutionary statistics of different regions compare to the genomic background.

REFERENCES

1. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
2. Sherry, S. T. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29,** 308–311 (2001).
3. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45,** D896–D901 (2017).
4. Denny, J. C. *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol.* **31,** 1102–1110 (2013).
5. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46,** D754–D761 (2018).
6. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
7. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **46,** D8–D13 (2018).
8. Stein, L. D. Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.* **14,** 162–171 (2013).
9. Buels, R. *et al.* JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17,** 66 (2016).
10. Zhang, G. *et al.* Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N. Engl. J. Med.* **377,** 1156–1167 (2017).

CHAPTER VII

Conclusion

Genes and their regulation form the basis of the functions, processes, and structure of all life. Since their initial characterization as discrete units of heredity by Gregor Mendel[1] in 1866 and rediscovery of his laws in in the early 1900s, their importance in the determination of the fates of biological organisms has been recognized by geneticists. The physical organization of genes on chromosomes was first described by Thomas Morgan and Alfred Sturtevant[2] in 1915. The discovery of DNA as the purveyor of genetic information in 1944[3] and the subsequent characterization of its double helical structure in 1953[4], enabled a new level of understanding of these heritable units. The development of DNA sequencing[5–8], the discovery of the ribosome[9] and RNA polymerase[10] and splicing[11,12] and proposal of the Central Dogma of Molecular Biology in 1958[13] began to paint a rich model for the way genes direct cellular and organismal activities. This model would continue to develop with the discovery transcriptional regulation via promoter regions[14] and enhancer elements[15–17].

This physical understanding of the organization of genes in the genome enabled the mapping of the locus of the Huntington's disease to the human chromosome 4 in 1983[18]. This discovery would be followed by the identification of the cystic fibrosis gene in 1989[19]. The elucidation of these loci has enabled a greater understanding of the mechanisms and pathophysiology of these diseases. In more recent times, the important biological questions have shifted to more complex disorders. In a National Institutes of Health bulletin released in 2009 (https://grants.nih.gov/grants/funding/challenge_award/high_priority_topics.pdf) detailing the highest priority challenge topics, the agency placed a focus on topics such as aging, mental health, and preterm birth. These diseases are complex, multifactorial syndromes that involve a wide array of processes and genetic loci, and they require the development of a sophisticated

understanding of a diverse set of functional and evolutionary analyses.

In Chapters 2 and 3, I described my work on one the development of an online integrative encyclopedia for the study of pregnancy and preterm birth, GEneSTATION. This database collected a diverse set of functional and evolutionary data associated with genes (v1.0) and regulatory elements (v2.0). In addition, it provides tools to perform synthetic analyses of genes and sequence variants in an interactive fashion. Available at http://www.genestation.org/, this database serves as a rich resource for the study of the function and evolution of human pregnancy and preterm birth.

In Chapter 4, I detailed the design of the Genestation Search Engine Toolkit, a novel schema for the storage of genomic data in ElasticSearch, a JSON document store and search engine. The development of this framework and the associated tools were motivated by the need to handle the ever increasing size of biological datasets that overwhelmed previous biological database technologies. Inspired by the design of the Generic Model Organism Database project's Chado schema[20], I followed a similar philosophy to describe a generic ontologically structured database. The framework and tools I have created have the potential to be applied to a wide variety of biological data and problems, and I believe that the Genestation Search Engine Toolkit will serve as an important resource for future genomic databases.

In Chapters 5, I describe the Synthesis and Test of Hypotheses (SynTHy) tool. SynTHy uses the Genestation Search Engine Toolkit to provide an interactive search interface written in HTML5 and JavaScript that serves as an advanced query builder that provides instant statistical feedback. With potential to be easily extended via callbacks, SynTHy can be used by existing genome databases to add a sophisticated search interface or used standalone to visualize the data in a database using the Genestation framework.

In Chapter 6, I describe the integrative holistic gene visualizer, GeneViewer. Using the Genestation Search Engine Toolkit, GeneViewer depicts genes and their associated regulatory

elements in a style similar to gene model diagrams and selectively displays sequence variants with evidence of functional implications. In addition, the tool visualizes the distributions of quantitative metrics, such as nucleotide diversity, within regions and how they compare to the genomic background.

The tools that I have created during the course of my dissertation research have only just begun to unlock new modes of exploration of the rich ecosystem of genomic data that has developed since the sequencing of the first genome in 1977. By enabling investigators to more readily explore and access the information contained within these resources, my technologies will facilitate more rapid discovery and dissemination of their knowledge. I have designed these works to be generic and extensible, with the ability to be utilized by other researchers for their own projects.

Future investigators should focus on expanding the functional and evolutionary data types collected by genomic encyclopedias, such as GEneSTATION. While my own work has collected a diverse set of genomic data, many analyses remain unintegrated, such as ChIP-seq data, 3D and 4D genome structural data, and many many more. The design of the Genome Feature Object used in the Genestation Search Engine Toolkit is generic and extensible, but the work of designing these extensions will be an important task for future users of the schema. As new biological assays and analyses are discovered, it will be important to develop of new analysis and visualization tools to explore and process this data. By publishing this work and open sourcing the software behind this project, I hope to provide a foundation for future investigators that will grow beyond the scope of my own work into a resource for the scientific community that facilitates a greater understanding of the mechanisms and processes of life.

REFERENCES

1. Mendel, G., Punnett, R. C. & Burndy Library. *Versuche über Pflanzen-Hybriden /*. (1866).
2. Morgan, T. H. *The mechanism of Mendelian heredity, by T. H. Morgan [et al.]*. (1915).
3. Avery, O. T., Macleod, C. M. & McCarty, M. STUDIES ON THE CHEMICAL NATURE OF THE SUBSTANCE INDUCING TRANSFORMATION OF PNEUMOCOCCAL TYPES : INDUCTION OF TRANSFORMATION BY A DESOXYRIBONUCLEIC ACID FRACTION ISOLATED FROM PNEUMOCOCCUS TYPE III. *J. Exp. Med.* **79,** 137–158 (1944).
4. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738 (1953).
5. Sanger, F., Thompson, E. O. P. & Kitai, R. The amide groups of insulin. *Biochem. J* **59,** 509–518 (1955).
6. Holley, R. W. *et al.* STRUCTURE OF A RIBONUCLEIC ACID. *Science* **147,** 1462–1465 (1965).
7. Min Jou, W., Haegeman, G., Ysebaert, M. & Fiers, W. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237,** 82–88 (1972).
8. Sanger, F. *et al.* Nucleotide sequence of bacteriophage φX174 DNA. *Nature* **265,** 687–695 (1977).
9. Palade, G. E. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.* **1,** 59–68 (1955).
10. Hurwitz, J. The discovery of RNA polymerase. *J. Biol. Chem.* **280,** 42477–42485 (2005).
11. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA. *Cell* **12,** 1–8 (1977).
12. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5’ terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 3171–3175 (1977).
13. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).
14. Lifton, R. P., Goldberg, M. L., Karp, R. W. & Hogness, D. S. The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harb. Symp. Quant. Biol.* **42 Pt 2,** 1047–1051 (1978).
15. Gillies, S. D., Morrison, S. L., Oi, V. T. & Tonegawa, S. A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell* **33,** 717–728 (1983).

16. Banerji, J., Olson, L. & Schaffner, W. A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33,** 729–740 (1983).
17. Mercola, M., Wang, X. F., Olsen, J. & Calame, K. Transcriptional enhancer elements in the mouse immunoglobulin heavy chain locus. *Science* **221,** 663–665 (1983).
18. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–238 (1983).
19. Riordan, J. R. *et al.* Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* **245,** 1066–1073 (1989).
20. Mungall, C. J., Emmert, D. B. & FlyBase Consortium. A Chado case study: an ontology-based modular schema for representing genome-associated biological information. *Bioinformatics* **23,** i337–46 (2007).

APPENDIX 1

Genome Feature Object: Representing Genomic Features in JSON document stores

1 Introduction

The Genome Feature Object is a method of representing genomic features in JSON document stores, such as ElasticSearch. The goal of this schema is to be able to represent diverse biological data in an extensible manner that is conducive to search and visualization. This specification is designed to integrate the information in Variant Call Format (VCF), Gene Feature Format (GFF), and Tab-separated JSON Values (TSJV) files into a single JSON document for each genomic feature.

2 Mapping

A Genome Feature Object is **mapped in ElasticSearch** as follows:

```
{
        “dynamic” : false,
        “properties” : {
                “genome” : {“type” : “keyword”},
                “ftype” : {“type” : “keyword”},
                “name” : {“type” : “keyword”},
                “dbxref” : {“type” : “keyword”},
                “region” : {“type” : “keyword”},
                “locrange” : {“type” : “long_range”},
                “start” : {“type”: “long”},
                “end” : {“type”: “long”},
                “loc” : {
                        “nested”: “true”,
                        “properties” : {
                                “start” : {“type”: “long”},
                                “end” : {“type”: “long”},
                                “strand” : {“type” : “byte”},
                                “phase” : {“type” : “byte”}
                        }
                },
                “child” : {GenomeFeatureObject},
                “association” : {GenomeFeatureObject},
                “data” : {“dynamic” : true},
                extensions...
        }
```

```
}
```

The top level of the document is defined as `“dynamic” : false`, which disallows dynamic mapping of unknown fields. The “data” property is defined as being explicitly dynamic, as this field is meant to generically integrate diverse data. Note that dynamically mapped fields in ElasticSearch are still checked for consistency within that field (ie. attempting to store a string in a field that was dynamically mapped as an integer field will raise a mapping exception).

2.1 genome

Required. The genome identifier the genomic feature. While this can be inferred by parsing the index it is included in the document to enable searching.

2.2 ftype

Required. The feature type(s) of the genomic feature, which must be a Sequence Ontology term. Used by visualization tools to determine draw styles. The top level `“ftype”` of each document in a single ElasticSearch index should be the same (ie. one index for genes, another for sequence_variations).

2.3 name

Required. The name(s) of the genomic feature (eg. gene name or rs number). If `“name”` is an array, the first name is the primary identifier and the rest are considered aliases. The primary identifier is used as the `“_id”` of the document in the ElasticSearch index and must be unique within that index.

2.4 dbxref

Optional. The dbxref(s) associated with this feature (eg. “GeneID:7124”,

“Ensembl:ENSG00000232810”). This field must follow the GFF3 specification for ontology associations and database cross references described here: **https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md**

2.5 region

Required. The identifier for the contig or chromosome on which this feature is located (eg. “1”, “2”, “X”, “Y”). It is important to be consistent when locating features to use the same srcfeature identifiers. For example, it is not recommended to use both “NC_000001.11” and “1” to refer to human chromosome 1 within a single ElasticSearch cluster. It is recommended to use chromosome names when possible, as opposed to accessions.

2.6 locrange

Required. The start and end of the feature in the genome using the “0-start half open” coordinate system (eg. the first base has start=0, end=1). Locrange is automatically generated by tools as a `long_range` type from `“start”` and `“end”` as a range object `{“gte”: start, “lt”: end}`, or a range object array.

2.7 loc

Optional. GFF style location entries, represented as an array of objects with start, end, strand, and phase. Note that this still uses the “0-start half open” coordinate system (eg. the first base has start=0, end=1).

2.8 child

Any children of the feature, represented as Genome Feature Objects. Should only be used to

represent true parent-child relationships (eg. transcripts of genes). Do not use this field to represent associations (eg. SNPs or enhancers to genes). Avoid deep nesting of children (eg. parent-child structures more than 3 levels deep) and keep the structure of child objects consistent within a single index. If there are children, there should be a fixed level of recursion which should be explicitly reflected in the mapping for the index. If it will be common to perform complex searches using the children, these features should be replicated into their own feature index.

2.9 parent

Any parents of this feature. This should be usually only be structured as a Genome Feature Object with only “genome”, “ftype”, and “name”. Should only be used to represent true parent-child relationships. Commonly utilized when creating a redundant search index for children when they will be the targets of search queries (eg. an index for transcripts).

2.10 association

Any associations this feature has to other features. At minimum, this should be structured as a Genome Feature Object with at least “genome”, “ftype”, and “name”. The “data” subfield may be used to quantify the confidence/strength of the association.

2.11 data

Any data values associated with the feature. Each subfield is considered to be completely independent and may define its own data structure. Genome Annotation Table (GAT) files may include mapping directives that are included to describe the data structure of the generated JSON documents. The data object should be consistent across all documents in one ElasticSearch index (allowing for null values and missing fields).

See the documentation for `genestation.py` for how these objects are generated (Appendix 2).

2.12 Extension: variant

Sequence variants from a VCF file store allelic data in `“variant”`.

```
{
        ...
        “variant” : {
                “nested” : true,
                “properties” : {
                        “base” : {“type” : “text”},
                        “is_ref” : {“type” : “boolean”},
                        “data” : {“dynamic” : true}
                }
        }
}
```

2.12.1 variant

Any variants that this feature represents. This is a nested document with the following fields:

2.12.1.1 base

Required. The base(s) of this variant.

2.12.1.2 is_ref

A flag that indicates whether this variant matches the reference sequence.

2.12.1.3 data

Similar to the document level data object, this object integrates diverse allele specific data.

3 Indexing

3.1 Genome Index

The Genome Index contains descriptions for every genome available within the ElasticSearch cluster. This index is always called `‘genome’` and is the entry point for tools that interact with GeneStation. Genomes are described via Genome Objects that have the following structure.

```
{
        “dynamic” : false,
        “properties” : {
                “genus” : {“type” : “keyword”},
                “species” : {“type” : “keyword”},
                “subspecies” : {“type” : “keyword”},
                “version” : {“type” : “keyword”},
                “common_name” : {“type” : “keyword”},
                “dbxref” : {“type” : “keyword”},
                “data” : {“dynamic” : true}
        }
}
```

Each genome described in the index is uniquely identified by the genus, species, subspecies (nullable), and version. This takes the form `genus_species.version` or `genus_species_subspecies.version`. All alphabetic characters are converted to lowercase, with spaces and periods (eg. in version numbers) replaced with underscores. For example, the human genome release GRCh38.p12 would have the prefix `“homo_sapiens.grch38_p12”`, and the domesticated dog genome release 3.1 would have the prefix `“canis_lupis_familiaris.3_1”`. The genome identifier is used to organize the other indexes in the database.

3.1.1 genus

Required. The genus of the sequenced organism.

3.1.2 species

Required. The species of the sequenced organism.

3.1.3 subspecies

Optional. The subspecies identifier of the sequenced organism, if any.

3.1.4 name

Optional. The common name(s) of this organism. If `"name"` is an array, the first name is the primary name and the rest are considered aliases.

3.1.5 dbxref

Optional. The dbxref(s) associated with this organism or genome following the format described in the GFF3 specification (eg. "RefSeq:GCF_000001405.38", "taxon:9606").

3.2 Feature Index

Feature Indexes are named in the form `feature.genome_identifier.feature_type`, where feature_type is a Sequence Ontology term. For example, an index for human genes might have the identifier `feature.homo_sapiens.grch38_p12.gene`. This hierarchical prefixing structure namespaces all feature indexes and allows for use of the ElasticSearch MultiIndex API (**https://www.elastic.co/guide/en/elasticsearch/reference/current/multi-index.html**) to query all features (regardless of type) of a specific genome.

A single feature index contains Genome Feature Objects of a single type and its children. The children should only be represented within the child field and should be redundantly indexed in

their own index if they will be the targets of search queries (eg. A database supporting searches on mRNAs and genes).

3.3 Meta Index

Meta Indexes store general purpose index-level statistics and metadata. This index can be generated for any ElasticSearch index by the Genestation CLI and is named in the form `meta.index_name`. For example, a Meta Index for human genes might have the identifier *meta.feature.homo_sapiens.grch38_p12.gene*. These indexes store Field Meta Objects which describe fields in the associated index and have the following structure:

```
{
        "field": {"type": "keyword"},
        "description": {"type": "text"},
        "type": {"type": "keyword"},
        "stats": {FieldStatsMetaObject},
}
```

Note that most of these fields will be automatically generated by the Genestation CLI and the only manually populated field is the `"description"` text field. The Meta Index is generally used to precalculate genome wide statistics on Feature Indexes, but may be used to describe any index in ElasticSearch.

3.3.1 field

Automatically generated. The dot-notated name of the field profiled by the Field Meta Object.

3.3.2 description

Optional. A human readable description of the field being profiled.

3.3.3 type

Automatically generated. The field data type as mapped in ElasticSearch

3.3.4 stats

Automatically generated. This field is only calculated for numeric field data types. The Field Stats Meta Object has the following structure:

```
}
        "count": {"type": "long"},
        "min": {"type": "double"},
        "max": {"type": "double"},
        "sum": {"type": "double"},
        "sum_of_squares": {"type": "double"},
        "variance": {"type": "double"},
        "std_deviation": {"type": "double"},
        "std_deviation_bounds": {
                "properties": {
                        "upper": {"type": "double"},
                        "lower": {"type": "double"},
                }
        },
        "percentiles": {
                "properties": {
                        "1.0": {"type": "double"},
                        "5.0": {"type": "double"},
                        "25.0": {"type": "double"},
                        "50.0": {"type": "double"},
                        "75.0": {"type": "double"},
                        "95.0": {"type": "double"},
                        "99.0": {"type": "double"},
                }
        },
        "histogram": {
                "properties": {
                        "key": {"type": "keyword"},
                        "from": {"type": "double"},
                        "to": {"type": "double"},
                        "doc_count": {"type": "long"},
                }
        }
}
```

The ElasticSearch Extended Stats aggregation

(https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-metrics-

extendedstats-aggregation.html) is used to populate the following fields:

- **count**
- **min**
- **max**
- **sum**
- **sum_of_squares**
- **variance**
- **std_deviation**

- **std_deviation_bounds**

The ElasticSearch Percentiles aggregation (https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-metrics-percentile-aggregation.html) is used to calculate the value of the **percentiles** field.

The ElasticSearch Histogram aggregation (https://www.elastic.co/guide/en/elasticsearch/reference/current/search-aggregations-bucket-histogram-aggregation.html) is used to calate the value of the **histogram** field. By default, this will use 100 equally sized buckets over the range of the field value with unbounded first and last buckets.

3.4 Custom Indexes

If it is necessary to create custom indexes, for example a tool or script, these indexes should be created with an explicit namespace prefix. This will ensure that any created indexes will not interfere will other tools.

APPENDIX 2

Genestation Command Line Interface

1 Synopsis

genestation [--version] [--host HOST] COMMAND ARGS

1.1 --host HOST

Set the hostname of the ElasticSearch node. Defaults to `localhost:9200`

2 Commands

2.1 init

genestation init

Initialize a Genestation instance. Creates the Genome Index, loads index templates for Feature Indexes and Meta Indexes, and loads search templates for creating Meta Indexes.

2.2 load

genestation load descriptor [...descriptor]

Load genomic data referenced in the descriptor files. The format of these files is described in Appendix 3.

2.3 genome

genestation genome [**show** identifier]

Without any arguments, list all genomes

2.3.1 genome show

Show information about the specified genome, the associated Feature Indexes, and the number of documents they contain

2.4 index

genestation index [**[show|make-meta]** index]

Without any arguments, list all indexes and the number of documents they contain

2.4.1 index show

Show ElasticSearch metadata about the index

2.4.2 index make-meta

Create or update the Meta Index associated with the index

2.5 get

genestation get index id

Get the document in the specified index with matching id

2.6 search

genestation search index query

Search the specified index using the query in ElasticSearch QueryString Syntax

(https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-query-string-query.html#query-string-syntax)

APPENDIX 3

Genomic Data Descriptor JSON
A configuration file for loading genomic data into ElasticSearch

INTRODUCTION

The Genomic Data Descriptor JSON (GDD JSON) is a JSON file that describes and references genomic data files for the Genestation Command Line Interface (CLI). GDD files should have the file type suffix `*.gdd.json`. A GDD file may contain one GDD object, or a list of GDD objects which contain the keys described below. Genomic data for a given genome may be contained in any number of files and loaded separately. For example, one might have a GDD for a GFF containing genic data and another GDD for a VCF containing SNPs and other polymorphisms.

To understand how the Genestation CLI stores genomic data in ElasticSearch, see the documentation for the Genome Feature Object in Appendix 1.

GENOMIC METADATA KEYS

These keys are used to construct the **genome identifier**. This identifier takes the form *genome_species.version* or *genome_species_subspecies.version*, where all alphabetic characters are converted to lowercase and periods and spaces are converted to underscores (eg. `‘homo_sapiens.grch38_p12’` and `‘canis_lupus_familiaris.3_1’`). Each **genome identifier** should refer to a unique genome within a genestation instance.

- **genus:** The taxonomic genus of the organism (eg. **Homo** in Homo sapiens).
- **species:** The taxonomic species of the organism (eg. **sapiens** in Homo sapiens).
- **subspecies:** OPTIONAL. The taxonomic subspecies classification of the organism, if any (eg. **familiaris** in Canis lupus familiaris).
- **version:** The genome assembly version string (eg. GRch38.p12).

GENOMIC DATA KEYS

These keys describe various data formats that can be handled by the Genestation CLI. The following data keys are all optional, but may contain required subkeys if present. If none of these keys are present, the only result from loading the GDD will be the creation of the Genome Meta Object in the Genome Index.

It is important for chromosomal identifiers to be consistent across all genomic data files for a given genome in order to ensure that location based queries perform correctly. The Genestation CLI is capable of performing simple transformations to aid in the integration of diverse data sources, but it is recommended that complex transformations be performed in external data pipelines.

gff

A string or object. If it is a string, a relative file reference relative to a GFF file with genomic feature data. Multiple GFF files may be specified as a string array or object array. If this is an object, it has the following subkeys:

- **file**: A string or string array of file references to GFF file(s) relative to the GDD JSON.
- **ftype**: OPTIONAL. The genomic feature type(s) to load from the GFF file(s). Must be a Sequence Ontology term. Defaults to `'gene'`. This feature will be loaded into ElasticSearch. Its children will be nested inside the `'child'` field.
- **seqid_alias**: OPTIONAL. An object of strings to strings which maps SEQIDs in the GFF to another alias. Useful for converting chromosomal accessions to names (eg. NC_000001.11 to 1 in the human genome).
- **alias_attr**: OPTIONAL. The attribute key that should be used to populate the `'alias'` field for this feature. Defaults to `'Alias'`.
- **dbxref_attr**: OPTIONAL. The attribute key that should be used to populate the `'dbxref'` field for this feature. Defaults to `'Dbxref'`.
- **data_attr:** OPTIONAL. An object of strings to strings which maps attribute keys to dot-notated field names in the `'data'` field. An attribute key must be present in this object to be loaded into the `'data'` field.

fasta

A string or object. If it is a string, a relative file reference to a FASTA file with genomic sequence data. Multiple FASTA files may be specified as a string array. If this is an object, it has the following subkeys:

- **file**: A string or string array of file references to FASTA file(s) relative to the GDD JSON.
- **ftype**: The sequence feature type. Must be a Sequence Ontology term.
- **defline_part**: OPTIONAL. An integer which specifies a portion of the defline to use as

the unique name of the feature after splitting on `'|'` (pipe), zero indexed. Useful for extracting accessions from deflines with extra information. If this option is missing the entire defline is used.

- **defline_alias**: OPTIONAL. An object of strings to strings which maps deflines in the FASTA to another alias. If a **defline_part** has been specified, only the specified part will be used in this mapping process. Useful for converting chromosomal accessions to names (eg. NC_000001.11 to 1 in the human genome).

vcf

A string or object. If it is a string, a relative file reference to a VCF file with genomic feature data. Multiple VCF files may be specified as a string array. If this is an object, it has the following subkeys:

- **file**: A string or string array of file references to VCF file(s) relative to the GDD JSON.
- **chrom_alias**: OPTIONAL. An object of strings to strings which maps CHROM identifiers in the VCF to another alias. Useful for converting chromosomal accessions to names (eg. NC_000001.11 to 1 in the human genome). It is important for chromosomal identifiers to be consistent across all genomic data files in the GDD.
- **info_key_alias**: OPTIONAL. An object of strings to strings which maps INFO keys in the VCF to another alias. Useful for renaming and nesting data in the INFO column into a form more suitable for ElasticSearch (eg. nesting all FST calculations within an fst.* key).

tsjv

A string or object. If it is a string, a relative file reference to a TSJV file with genomic feature data. Multiple TSJV files may be specified as a string array. If this is an object, it has the following subkeys:

- **file**: A string or string array of file references to TSJV file(s) relative to the GDD JSON.
- **ftype:** OPTIONAL. The sequence feature type of the features named in **feature_col**. Must be a Sequence Ontology term. Defaults to `'gene'`.
- **feature_col**: OPTIONAL. String. The column name which contains the feature name. Defaults to `'feature'`.
- **start_col**: OPTIONAL. String. The column name which contains the feature start. Defaults to `null` (no location data in file).
- **end_col**: OPTIONAL. String. The column name which contains the feature end.

Defaults to `null` (no location data in file).

- **region_col**: OPTIONAL. String. The column name which contains the feature region. Defaults to `null` (no location data in file).
- **data_cols**: OPTIONAL. String array. The columns which contain feature data. Defaults to `null` (no data will be loaded).
- **data_mapping**: OPTIONAL. Object. ElasticSearch mapping for the data columns. May be a partial mapping. Defaults to `null` (no data will be loaded).
- **association_cols**: OPTIONAL. String array. The columns which contain the associated feature names. Defaults to `'association'`.
- **association_data_cols**: OPTIONAL. String array. The columns which contain feature association data. Defaults to `null` (no association data will be loaded).
- **association_data_mapping**: OPTIONAL. Object. ElasticSearch mapping for the association data columns. May be a partial mapping. Defaults to `null` (no data will be loaded).
- **association_genome**: OPTIONAL. String. The genome identifier of the associated genomic feature. Defaults to the current genome.
- **association_genome_col**: OPTIONAL. String. The column which contains the genome identifier of the associated genomic feature. Overrides **association_genome**. Defaults to `null`.
- **association_ftype**: OPTIONAL. String. The feature type of the associated genomic feature. Defaults to `'gene'`.
- **association_ftype_col**: OPTIONAL. String. The column which contains the feature type of the associated genomic feature. Overrides **association_ftype**. Defaults to `null`.