in NGS experiments is driving a need for automation and sustainable laboratories can easily lose track of the many changes to their based analyses generate new sequence and feature annotations and large datasets that they have created. Bioinformatic and bench expertise is molecular biology can have difficulty managing the assemblies (Farrer et al., 2009) that have made it possible for individual laboratories to create draft quality genome assemblies in very short timescales. Projects describing fragmented but valuable assemblies (Farrer et al., 2009; Kemen et al., 2011; Raffaele et al., 2010) have provided many new insights. Laboratories whose main expertise is molecular biology can have difficulty managing the large datasets that they have created. Bioinformatic and bench based analyses generate new sequence and feature annotations and laboratories can easily lose track of the many changes to their sequence and annotated features. Thus, the amount of data generated in NGS experiments is driving a need for automation and sustainable storage solutions for assembly and feature data. We have created an application called Gee Fu that stores genomic sequence and feature information and provides a platform to allow groups with diverse levels of bioinformatics and computing skills to easily store, track versions, edit and visualize feature data and keep track of versions, share and visualize basic sequence data. There are many genome browsers and annotation editors, including GBrowse and Apollo (Lewis et al., 2002). Gee Fu extends and integrates into this ecosystem by providing a lightweight database version system that can import from and export to other annotation systems and genome browsers if needed. It provides access and utility to bioinformaticians who want to access data programatically and others who prefer a graphical interface in one central repository that allows changes made by one worker to be reflected to others instantly. Gee Fu has been useful in our next-generation sequence based assembly and annotation efforts (Kemen et al., 2011).

2 IMPLEMENTATION

Gee Fu is an application built on Ruby-on-Rails, a rapid, open-source web application development framework based on the Ruby language (http://rubyonrails.org) and is centred around a straightforward relational database schema for feature and sequence data. In this schema, the organism is the root object and may have many genome versions and reference sequences. Each sequence object has many feature objects which are described as in GFF3 and parsed into the database using the BioRuby GFF3 class. Sequence read feature information from high-throughput alignments to reference sequence is stored in BAM format (Li et al., 2009) accessed using samtools-ruby. Features also belong to experiment objects which describe the study in which features are created. Features may be edited and old versions are stored in a predecessors table, providing a feature history. As with other systems, features are tied to a reference sequence and do not automatically transfer to a new sequence, allowing for complex rearrangements of the sequence between versions. Even for large assemblies from plants or mammals, system requirements are light, enough storage for the sequence and features in the RDBMS and for any BAM files is required. Memory requirements are low and 4 GB RAM will suffice.

3 DATA ACCESS

Input: organism, genome (including sequence), experiment and feature data may each be added to the database from FASTA and GFF3 format files. This is easily accomplished using the forms in the web-interface, with Rails rake tasks that can be executed from the command line.
API of Gee Fu can easily be extended to send data to other services, making it suitable for creating a fast genome analysis.

Web-service: Gee Fu implements a straightforward but useful API for retrieval of data via GET requests. Gee Fu can return data on all the model classes in the database in XML or JSON format. Feature data can be returned as lists of objects for local analysis or as a summary, e.g. of sequence read coverage in a specified genomic interval.

Integration with other Annotation Editors and Genome Browsers: Gee Fu is able to import and export features in GFF3, a common feature format used by many other annotation editing programs, like Apollo. It is possible to export features from an experiment and edit them externally, then load them back in with edits being automatically updated and versioned. GFF3 is compatible with popular browsers and databases such as GBrowse. Gee Fu has methods to respond to requests made by the freely available AnnoJ viewport, making it suitable for creating a fast genome browser with a few simple configuration steps. The web-service API of Gee Fu can easily be extended to send data to other services such as GBrowse for feature rendering.

4 DISCUSSION

Gee Fu helps a growing problem in genomics: the storage and sharing of in-development assemblies and feature data within small groups of collaborators. Lack of specialist data management skills in groups using NGS technologies mean that versions of an assembly can proliferate, and become out of sync with each other causing at best confusion and time spent remedying the situation, at worst loss of data or inaccuracy. Gee Fu helps to centralize assembly and feature storage and provides easy data access to workers with varying degrees of experience with databases and computational methods. As a Rails application Gee Fu benefits from the functionality of ActiveRecord, the ORM layer in Rails, and all database information can be accessed programmatically in an object-oriented fashion facilitating high-throughput analyses with custom scripts. Gee Fu uses our samtools-ruby gem, a Ruby wrapper around SAMtools (Li et al., 2009) that wraps all the API functions allowing BAM files to be used alongside the conventional database. Gee Fu allows for data sharing over the web with a light web-service API ideal for scripts and websites avoiding bulk data exchange, making collaboration on a growing project simple. Gee Fu’s database schema is open and lightweight, allowing for extension via the Rails framework which enjoys a large, supportive expert user-base, meaning the application is customisable to the needs of individual groups. Additional analysis tools can easily be integrated. Freely available viewport software such as AnnoJ and GBrowse mean that Gee Fu can be the data store for a easily set up and fast genome browser.

ACKNOWLEDGEMENT

Thanks to Raoul Bonnal for gem help and Julian Tonti-Filippini for AnnoJ help.

Funding: Gatsby Charitable Foundation (to D.M.); Biotechnology and Biological Sciences Research Council (to R.R.G. and M.C.).

Conflict of Interest: none declared.

REFERENCES


