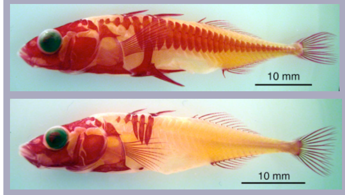
Managing Next Generation Sequence Data with GMOD

Dave Clements[1], Scott Cain[2], Paul Hohenlohe[3], Nicholas Stiffler[3], Paul Etter[3], Eric Johnson[3], William Cresko[3]
[1]National Evolutionary Synthesis Center, Durham, NC, USA, [2]Ontario Institute for Cancer Research, Toronto, ON, Canada, [3]University of Oregon, Eugene, OR, USA

Abstract

Next generation sequencing is flooding many organizations with enormous amounts of genomic data. A single machine can now produce three billion base pair reads (the size of the human genome) every three days. Components from the GMOD Project (http://gmod.org) can help visualize, manage and annotate this deluge of data. We describe how to use GMOD tools to gain new insights with high-throughput sequence data.

Population Genomics in Sticklebacks Using Illumina Sequenced RAD Tags

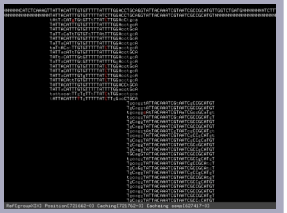


Marine (armored) and freshwater (unarmored) threespine sticklebacks.

We illustrate the potential of GMOD for displaying and analyzing genomic data from natural populations with a sample dataset from the threespine stickleback fish (*Gasterosteus aculeatus*). This species exhibits parallel patterns of morphological evolution in repeated colonization events from marine to freshwater habitats. To investigate the genetic basis of these evolutionary patterns, we generated homologous genomic sequence data for 16 individuals – eight each from a marine and a freshwater population – using the Illumina-based RAD sequencing technique[1]. The RAD (Restriction-site Associated DNA) technique isolates fragments of genomic DNA lying on either side of restriction enzyme recognition sites, which are found primarily at homologous locations across the genome in related individuals.

Using four-nucleotide barcodes ligated onto the genomic fragments to distinguish among individuals, we used an Illumina Genome Analyzer II to sequence 30bp of genomic sequence in each direction from each RAD site. A single run generated an average of ~200x coverage of the 60bp region at each of 28,000+ RAD sites.

We used Maq, a short read alignment program available from SourceForge, to align the small-read sequences to the existing stickleback reference genome, and identified putative single-nucleotide polymorphisms (SNPs). However, some nucleotide differences may be the result of sequencing error rather than actual SNPs. Therefore we developed a maximum-likelihood statistical approach for estimating the sequencing error rate and then assessing the most likely genotype at each site for each individual where possible.

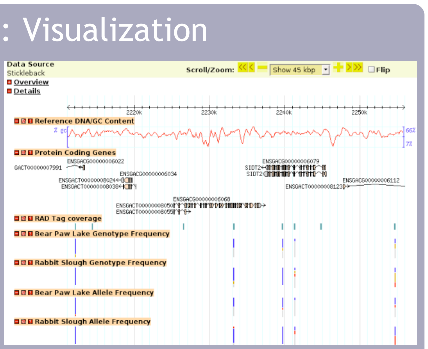


RAD sequence fragments aligned to the stickleback genome (top), visualized with Maqview. Sequence reads point in each direction from a restriction enzyme recognition site. Red nucleotides indicate putative SNPs, where more than one nucleotide was detected at a site across individuals.

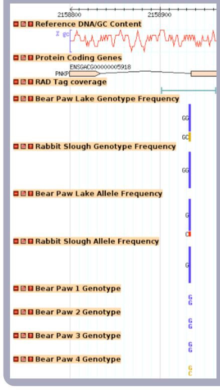
[1] Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA. 2008. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3(10):e3376.

GBrowse: Visualization

We used GMOD's GBrowse genome viewer to visualize our results in the context of the reference assembly and Ensembl gene predictions. Allele and genotype frequencies are shown for combined and individual populations and genotypes for each individual. RAD tag coverage (where we looked for SNPs) has a track as well.



45 kbp stickleback region, showing RAD tag coverage and population SNP frequencies



Additional information such as the exact position of SNPs and RAD tags, exact allele and genotype counts, and SNP call confidence scores are available in popups and linked pages.

Each track can be turned on or off, and can be configured and reordered for custom views. GBrowse can display any type of data that can be associated with a genomic region. GBrowse is designed to work with assembled and unassembled genomes.

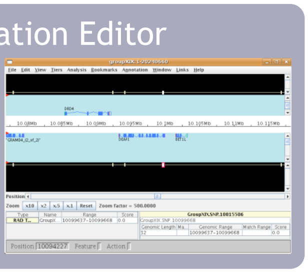
Detailed view of ~165 bp region, showing allele and genotype frequencies per population, and individual genotypes for a SNP within the *PNKP* gene.

Chado: Data Integration & Analysis

Chado is GMOD's modular database schema for managing biological data. Chado integrates genomic data with many other types of biological data (phenotypes, mapping, stocks, microarrays and targeted expression, publications, ontologies, …) into one database. It allows you to ask arbitrarily complex questions across biological data types using the (standard) SQL query language. Chado also backs many web sites, from ParameciumDB to FlyBase.

Apollo: Genome Annotation Editor

Apollo, GMOD's genome editor, is used to manually annotate genomic sequences. Apollo supports adding new annotations and refining computational annotations. It is used in several community annotation efforts, and by full-time curators as well.



GMOD is a collection of open source software components for managing, visualizing and annotating biological, mainly genomic, data. GMOD is also a community of people and organizations who support and use those tools. In addition to GBrowse, Chado, and Apollo, GMOD provides tools for comparative genomics, community annotation, web site generation, … See http://gmod.org for more.

This work is funded by NIH grants to Ian Holmes at UC Berkeley, James Hu at Texas A&M, and by NIH and NSF grants to William Cresko at Oregon.