Bioinformatics Pipeline for Selecting Anonymous Genetic Markers

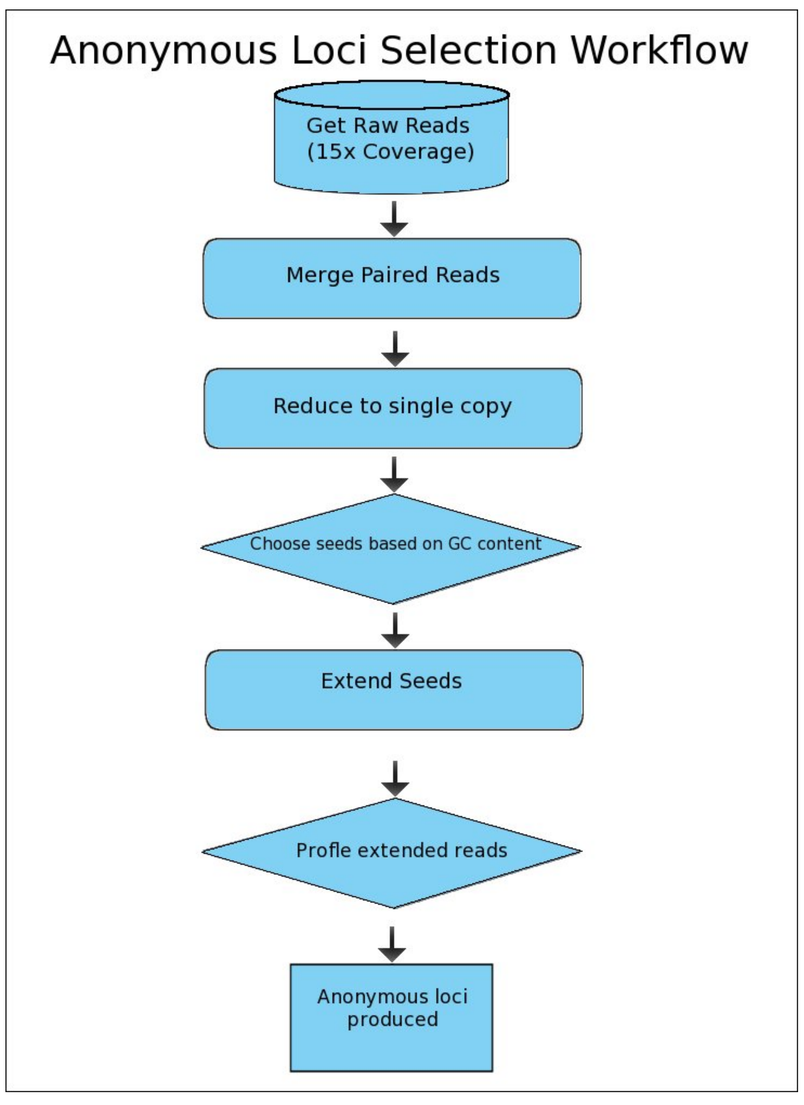
A. Rossi and M. Conry
Advisor: Dr. Alan Lemmon
Florida State University

Abstract Phylogenomic studies of non-model organisms often rely on genetic markers provided by well-known model organisms. This can lead to biased, improper tree estimation because only conserved genes are common across distantly related species. A new approach, made possible using hybrid enrichment techniques, would involve de novo selection of anonymous loci from low-coverage genome data. These regions would be more representative of the genome as a whole, and thus would likely evolve at a higher mutation rate. Such high mutation rates are preferable for more precise phylogenomic estimations because branch lengths can be more accurately estimated on shallow scales. Herein, we present a workflow for gathering suitable anonymous regions and discuss the careful considerations made when selecting anonymous loci.

Introduction

Illumina technology is currently the most successful and widely adopted next-generation sequencing technology.[1] Illumina can be used to sequence single-end or paired-end reads. Paired-end reads are two ends of the same DNA molecule where one end is sequenced, flipped around, and the other end is sequenced on the opposite strand. In this pipeline, paired-end reads are bioinformatically manipulated in order to select anonymous loci for probe design. The resulting probes capture select portions of DNA for sequencing in shallow scale studies of non-model organisms (e.g., between species of anoles). Anonymous loci selection is based on hybridization affinity, copy number, locus length, and similarity to other read sequences (read-seq similarity). Hybridization affinity is the affinity of a probe to a target site. This affinity increases with an increase in the percent conservation of the loci. There is a trade off here, however, since the more diversity a probe has, the more diversity can be captured using the probe. Hybridization affinity also increases with the percentage of nitrogenous bases in the molecule that are guanine or cytosine (GC content). The GC pair is bound by three hydrogen bonds whereas the adenine-thymine pair has only two hydrogen bonds. The number of hydrogen bonds along with stacking interactions increase the stability of the DNA.[2] More variation in GC content leads to more variable loci. The next two selectors are copy number and length. The duplication of genes and loss of genes over time leads to a lack of available single copy genes to choose from. Instead, genes with a low copy number are selected. Genes with low copy number and a long length in base pairs (bp) are of particular value because tree resolution improves with locus length. However, there is a trade off between the number of loci that can be selected and the length of each locus (e.g., 400 loci of 2000bp in length would be equivalent to 2000 loci of 400bp length). Finally, the proportion of loci with read-seq similarity is examined. Too high of a proportion could result in targeting uninformative loci for sequencing.

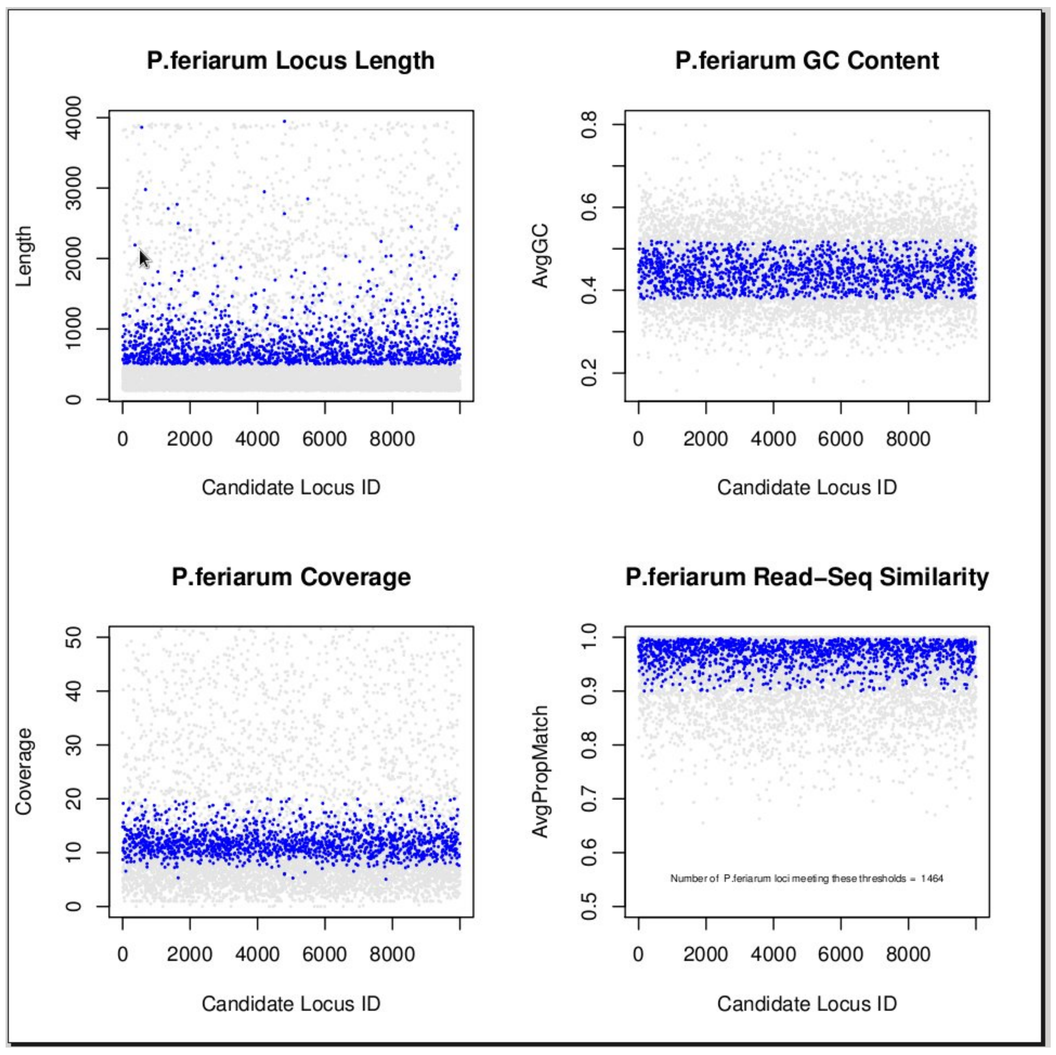
Pipeline



The pipeline starts with a whole genome or raw reads. Transcriptome data is not as informative. The data usually consists of raw reads of about 15x coverage. Larger genomes require more lanes in the Illumina sequencer to get the same coverage, which increases the cost of sequencing. Once the data is obtained, paired-end reads are merged. The Merge algorithm takes the paired-end reads and merges them. Two files, a forward read file and a reverse read file, are read. The quality scores are compared to each other and to the probability that a base call is incorrect. This is calculated using the binomial probability of k matches in n overlapping bases: $f(k) = \binom{n}{k} p^k (1-p)^{n-k}$ where n choose k is a combination of n base pairs taken k at a time. The result is a file of merged reads. The remaining unmerged reads are also printed out to files (forward and reverse). This reduces the size and increases the quality of the data. Next, reads are reduced to only those with low copy number. Then seeds are chosen based on GC content. An extending algorithm is then used to extend seeds. Given a list of seeds, the right side of each seed (forward orientation) is checked to see if it matches the beginning of a read. Next, the reverse compliment is checked against the end of a read. Extended reads are then profiled and selected on the basis of copy number, coverage, GC content, and lengths. Presence in related species is then determined.

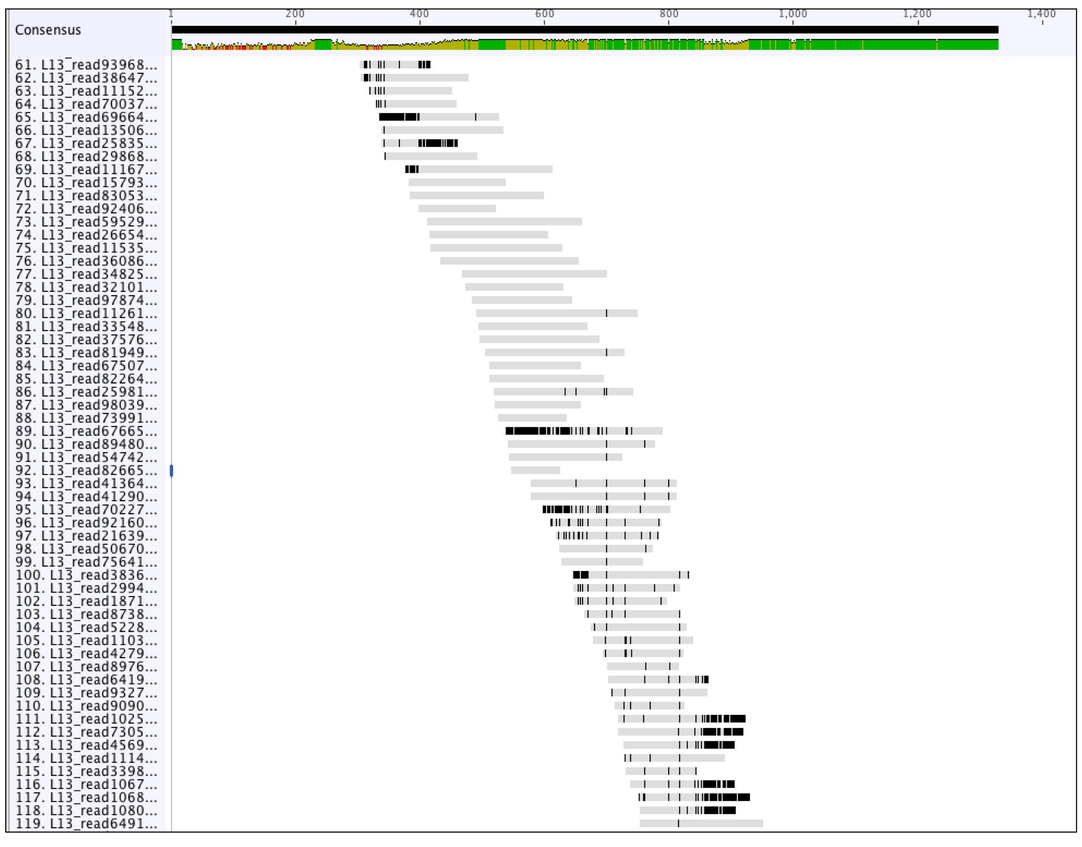
Example 1

Various criteria for selecting anonymous loci for use as genetic markers in probe kits. Loci are individually judged by length, GC content, sequence coverage, and read-seq similarity. Acceptable loci (blue) must be within a certain threshold specific to the given criteria.



Example 2

Manual inspection of a candidate locus viewed as a sequence alignment of assembled reads using Geneious.[3] In this case, the assembly has numerous heterozygous sites and sequencing errors. This assembly would likely be rejected by our criteria for selecting suitable anonymous genetic markers.



References

- DNA Sequencing, *www.illumina.com* N.p., n.d. Web. 20 Mar. 2014
- Yakovchuk P, Protozanova E, Frank-Kamenetskii MD 2006, *Base-stacking and base-pairing contributions into thermal stability of the DNA double helix*, Nucleic Acids Res. 34 (2): 56474.
- Geneious (version 7.1.3) created by Biomatters. Available from www.geneious.com