A Pipeline for Markers Selection Using Restriction Site Associated DNA Sequencing (Radseq)

Msc.Ms Hanan Begali Bonn university, Germany.

Abstract

The discovery and assessment of genetic variants for Next Generation Sequencing (NGS), including Restriction site Associated DNA sequencing (RADSeq), is an important task in bioinformatics and comparative genetics. The genetic variants can be single-nucleotide polymorphisms (SNPs), insertions and deletions (Indels) when compared to a reference genome. Usually, the short reads are aligned to a reference genome at first using NGS alignment software, such as the Burrows- Wheeler Aligner (BWA). The alignment is usually stored into a BAM file, a binary format of standard SAM (Sequence Alignment/Map) protocol. Then analysis software, such as Genome analysis Toolkit (GATK) or SAMTools, together with scripts written in R programming language, could provide an efficient solution for calling variants.

In this project, we focus on RADSeq-based marker selection for Arabidopsis thaliana. RADSeq consists of short reads which do not cover the whole reference genome. In order to obtain four call-sets of SNPs as output in Variant Call Format (VCF), SNPs have been called by GATK or SAMTools. Then VCF files have been visualized by Integrative Genomics Viewer (IGV) software. We found that the visualization of SNPs and Indels has been very helpful and has provided us with valuable insights on marker selection. We found that applying Chi-Square test for all target genotypes, which are homozygous reference 0/0, heterozygous variants 0/1 and homozygous variants 1/1, to test Hardy-Weinberg Equilibrium (HWE) in order to reduce false positive rate significantly. We show that our pipeline is efficient in RADSeq-based marker selection.

Image

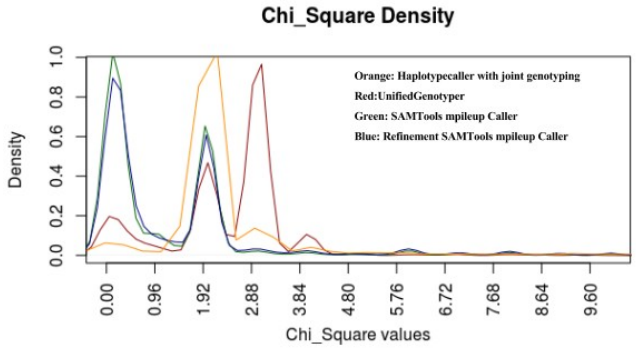


Figure 9: Density distribution for critical value of Chi-Square test for each SNPs.

Recent Publications

1. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of radseq for ecological and evolutionary genomics. Nature Reviews Genetics 17: 81-92.
2. Ishii K, Kazama Y, Hirano T, Hamada M, Ono Y, et al. (2016) Amap: A pipeline for whole-genome mutation detection in arabidopsis thaliana. Genes and genetic systems 91: 229-233.
3. De Summa S, Malerba G, Pinto R, Mori A, Mijatovic V, et al. (2017) Gatk hard filtering: tunable parameters to improve variant calling for next generation sequencing targeted gene panel data. BMC bioinformatics 18: 119.
4. Marrano A, Birolo G, Prazzoli ML, Lorenzi S, Valle G, et al. (2017) Snp-discovery by rad-sequencing in a germplasm collection of wild and cultivated grapevines (v. vinifera l.). PloS one 12: e0170655.
5. McKenna A, Matthew H, Eric B, Andrey S, Kristian C, et al. (2016) The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. Genome res 20: 1297-1303.

Photograph



Biography

Msc. Bioinofrmatics Ms.Hanan Begali comes from Yemen, has experiences on biological data analysis as well as math tutor of seven years. She has worked on both computational and bioinformatics criteria. She did her research in pipelines analysis of genetic variants discovery using NGS,RADSeq which included computational part. She had generated biological network model of macrophages, innate immunity using BEL. Furthermore, She was a assistant for analysis genetics and somatic mutations related to cancer in pathology institute, Dusseldorf, as well as worked as mathematics tutor. Also, she has worked on data analysis based on bio-informatics tools as well as she was working on microbiological department which is based on microscopic analysis, immunology and biochemistry tests on different samples for human. She prefers team working, international community and independent and friendly.

Email: hanan_albegali@yahoo.com

Notes/Comments: Place/Date: Duesselorf,Germany/24.Dec.2018