A program for annotating and predicting the effects of single nucleotide polymorphismsm SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w^{1118}; iso-2; iso-3

Dreycey Albin (da39)

April 7, 2019

Scientific question

The scientific question and goal for this paper focuses on *making a program that is able to find genomic differences between the genomes of variants.* This program allows evaluating different annotated genomic regions, such as intronic, translated, upstream, downstream, and splice site regions. In addition, this allows for the program to find codon gain, losses, and the effects of mutations on the resulting coding region. SnpEff allows for the genomic differences of wide range of different species to be evaluated, including 320 genome version of multiple species. The program also allows the user to add their own reference genomes to the system, as long as the user supplies a fasta file and an annotation file for constructing the database (which in turn creates the interval forest breifly explained below).

Hypothesis tested

The hypothesis tested in this research is that *this program can uniquely identify genomic differences between species variants.* The authors test this hypothesis by evaluating genomic differences between two variants of Drosophila melanogaster; strain w^{1118}; iso-2; iso-3 and strain y; cn bw sp. By evaluating the genomic differences of the two strains, the authors were able to convincingly show that the program has the ability to identify strain level differences in the annotated genomes for both of the strains.

Methods/Results

The methods and results for this paper can be described two fold: (1) How they evaluated they evaluated the strain level differences, and the results for the analysis; and (2) the algorithm the authors use to map different protein coding regions between the strains for the evaluation.

0.1 Strain Level Analysis of Drosphila

Using the mutations found within the annoted coding sequences, SnpEff is able to find which mutations that generate new start codons, generate new stop codons or loss of stops codons, and cause synonymous and non-synonyomous codon changes (abbreviated as S and N, respectively). Considering the generation of start codons, the authors found that the predominant proteins with gained start codons have gene ontology (GO) terms corresponding to proteins that are likely to have a wide range of intra and interspecies diversity. These incluse immunogenic proteins and developemental proteins, such as immunoglobulins and Hox proteins. Most interstingly, the authors hypothesize that the 5' UTR may contain potential regions that allow for a diversity of proteins, in which a diverse group of start sites may take advantage. The stop-gains and losses were differentially categorized by the authors, according to what segments of the gene are conserved across both strains, which allows for a heuristic in categorizing the implied importance of the gene effected. This included four different categories, where, in general, the higher the category indicated more conservation and implied higher importance. The authors found many stop-gains and several stop losses in both strains. However, it should be noted that many of these coding regions are not well known, and the ancestor to both strains is also unkown, so a stop-gain in one species is categorized as a stop-loss in the other, and vice-versa. Lastly, considering the synonymous and non-synonymous mutations, the usual metric is to compare the ratio (N/S). The higher this ratio, the higher the number of amino acid changes in the resulting proteins. The authors found that the non-normalized ratio for the strain w^{1118} compared to the reference strain is roughly 0.28. They also observed that this ratio was lower in the middle of chromosome arms, indicating that essential genes are typically in these regions, as compared to the centromeric and telomeric regions of the chromosomes. Most of the 356,660 SNPs found in the analysis can be verified by capilary sequencing, thereby giving confidence to the SNPs and analysis garnished by SnpEff.

0.2 Interval Forest Data Structure

They achieve this feat by implementing a data structure which they call an interval forest, which contains a hast to different interval trees, which in turn map an input interval, A, to a interval B in the tree, in which A is completely contained within B. This allows for a reduced time complexity algorithm for mapping genomic sections between varients (at O(log (m+n)), where m=number of intervals in the tree and n=number of intervals in the tree node). This data structure is fundamental to the work in this paper, as it allows for the coding regions to be efficiently mapped from the query genome to the reference genome in a time efficient manner.

Key implications of the results

The key implications from the paper, like the methods and results, are the analysis between Drosphila strains and the program used for the analysis, SnpEff. The authors implemented a fast coding sequence mapper between genomes using a data structure, called an interval forest, which uses a hash map to different interval trees. This allows for a reduced time complexity for interval mapping, or in this case, mapping annoted coding regions. After mapping, the program is able to evaluate the biologically significant changes between the coding regions. Using this program, the authors compared coding regions from the strain w^{1118}; iso-2; iso-3 to the reference strain y; cn bw sp. This gave rise to 356,660 SNPs that elucidated many differences in the resulting coding region output between the two strains.