JUST AGRICULTURE
multidisciplinary e-Newsletter

Variant Calling Identification in Biological Data

Bulbul Ahmed*, Nalini Kanta Choudhury

*Deptt. Of Bioinformatics
Indian Agricultural Research Institute, Library Avenue, PUSA, New Delhi-110012

ARTICLE ID: 44

Introduction

Deoxyribonucleic acids (DNAs) are the basic structure of development, functioning and reproduction in most of the living organisms. Maximum portion of DNA is non-coding which do not serve for protein sequences, but directly or indirectly helps in protein coding DNAs. A gene refers to the unit of a DNA that carries the instructions for making protein. Genes are actively transcribed and involved in production of huge complex substances. These products may be either RNAs or proteins. Different types of gene are *housekeeping gene or constitutive gene*, *Luxury gene and Pseudo-gene.*

Typically genes contain coding and noncoding region. Noncoding DNA sequences are components of an organism's DNA that do not encode protein sequences but helps in translation of coding portion into proteins. Advanced sequencing technology leads to the discovery of rare polymorphisms in an efficient manner. Objective behind studying exome region is to identify rare variants that carry complex traits. Different algorithms have been developed for finding the functional variants. Basic calling strategies with high precision can be used to call genotypes at high coverage but for regions with low coverage, genotype calling is still a challenge as it leads to more errors and missing data. Depending on the capture of information from sequence data across individuals and genomic positions, variant calling algorithms fall into three major categories:

A brief discussion of steps involved in variant calling:

A. **Quality Control**: Sequence from Illumina or other sequencing platforms are used to check the quality of the data in Fastq file format using FastQC tool. After this, bad quality data is removed by using the Trimmomatic, SRAToolkit or cutadapt. Overall quality of variant call sets included in the four call sets were assessed by calculating the transition-to-transversion ratio (Ts/Tv). A Ts/Tv > 2 is expected for inter-genic sites. This is typically much higher in coding regions due to purifying selection.

B. **Alignment:** Trimmed good quality reads are maped to reference genome, to get the desired gene sequences. Various tools available for alignement are Trinity, Picard, Valvet (de novo based) and TopHat (reference based).

C. **Local Realignment around Indels and Variant Calling**: Local realignment helps in correcting the alignment reads which are wrongly or falsely aligned using Genome Analysis Toolkit (GATK). GATK call HaplotypeCaller that identifies all the variant in the aligned reads in a variant calling file (VCF format).

Three different variant calling algorithms are:

1. *Individual-based single marker caller (IBC):* An individual-based prior, assuming that each allele has a probability θ = 0.001 being different from the reference was considered by IBC. A uniform prior probabilities for transitions and transversions ratio (Ts/Tv) was assigned for variant sites. The model assigned the most likely genotype when the posterior probability reached a threshold of 99% while genotypes with lower posterior probability were marked as missing. glfSingle (http://genome.sph.umich.edu/wiki/GlfSingle) was used to call genotypes (Yancy., 2012).

2. *Population-based single marker caller (PBC):* PBC uses a two-step procedure to call variants (Li *et al.*, 2011).

i. Considering that at least one read carries a non-reference allele, the model applied a population genetic prior which estimated probability of the site being polymorphic as a function of sample size, with per base pair heterozygosity of θ = 0.001 under the stationary neutral model (Watterson *et al.*, 1975).

ii. PBC estimates the population allele frequency '*f*' per polymorphism site which assume that it is a bi-allelic site in Hardy-Weinberg equilibrium.

3. LD-aware caller (LDC): This updates the genotype of each individual at each marker using a Hidden Markov Model derived from the haplotype based-model that is used in the imputation software MACH (Watterson *et al.*, 1975). This algorithm begins with randomly phased haplotypes for each individual. The algorithm compares one sequenced sample with a randomly picked subset of haplotypes per iteration. Based on the similarity of the sample haplotype to the reference haplotypes, it updates and imputes missing each genotype.

Variant Annotation: It can be done through identifying the SNPs using SNP annotation tools. Here all the biological information are extracted using SnPEff annotation tool (Cingolani *et al.*, 2012). It can identify the core relation between the variations in sequence level with respect to phenotype (Landrum *et al.*, 2013). Further it can be validated through clinical trials.

Where to use?

A. ***Individual-based single marker callers (IBC):*** These assign genotypes based on align reads from a single individual at a single position and is largely applied to exone sequencing data of high depth.

B. ***The population-based single marker callers (PBC):*** Here allele frequencies and polymorphism is determined from reads per position from all samples jointly. Estimated allele frequencies then call genotypes by using per individual read data. It is usually used in low-pass sequencing studies.

C. ***Linkage disequilibrium (LD):*** This makes use of information related to LD across large flanking each variants base that are identified by IBC or PBC.

Importance of Identification of Variants calling: Identification of variants give details about the variations in the individual's genome and underlying reasons for different diseases and phenotype changes in an organism. Variants analysis plays a crucial role in genome-wide association studies, more precisely in gene studies which are responsible for disease development. Moreover, SNP based arrays such as axiom array help in the improvement of diseases, crop yield and resistance to varieties of stress. Computationally SNP annotation helps in prediction of deleterious effects of SNPs and their role in diseases in living organisms.

Reference:

Li, Y., Sidore, C., Kang, H., Boehnke, M. and Abecasis, G. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* 2011;**21**:940–51.

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., and Ruden, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80-92.

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., and Maglott, D. R. 2013. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*, *42*(D1), D980-D985.

Watterson, G.A., On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*.1975;**7(2):**256–76.

Yancy., 2012. Alternative mutation model with uniform (uninformative) prior for transition to transversion ratio. Available at http: http://genome.sph.umich.edu/wiki/GlfSingle.