Certified Specialist Programme in Next-Generation Sequencing

Variant Detection and Genotyping

Variant detection and genotyping are key components of next-generation sequencing (NGS) data analysis. These processes allow researchers to identify genetic variations in samples being studied and determine the genotype of each sample at specific genetic loci. Here, we will explain some of the key terms and vocabulary related to variant detection and genotyping.

1. Variant: A variant is a genetic sequence that differs from a reference sequence. Variants can be classified as single nucleotide variants (SNVs), insertions and deletions (indels), structural variants (SVs), or copy number variants (CNVs).
2. Genotype: A genotype is the genetic makeup of an individual or sample at a specific genetic locus. Genotypes can be homozygous or heterozygous. A homozygous genotype means that both copies of the gene are identical, while a heterozygous genotype means that the two copies of the gene differ.
3. Reference genome: A reference genome is a genome sequence that is widely accepted and used as a standard for comparison. In NGS data analysis, reads from a sample are aligned to the reference genome to identify variations.
4. Alignment: Alignment is the process of comparing the sequence of short reads from an NGS experiment to a reference genome to identify regions of similarity. Alignment software uses algorithms to find the best matches between the reads and the reference genome.
5. Variant calling: Variant calling is the process of identifying genetic variants in a sample based on the alignment of short reads to a reference genome. Variant callers use statistical methods to determine the likelihood that a variant is present in the sample.
6. Quality score: A quality score is a measure of the confidence that a variant caller has in the accuracy of a variant call. Quality scores are typically reported as Phred scores, which are log-transformed probabilities.
7. Alternate allele: An alternate allele is a genetic sequence that differs from the reference allele at a specific genetic locus.
8. Read depth: Read depth is the number of times a specific nucleotide or region is sequenced in a sample. High read depth increases the confidence in variant calling.
9. Mapping quality: Mapping quality is a measure of the confidence that a read aligns to the correct location in the reference genome.
10. Base quality: Base quality is a measure of the confidence that a specific base call is correct.
11. Filtering: Filtering is the process of removing false-positive variant calls from a variant call set. Filters can be based on various criteria, such as read depth, mapping quality, base quality, and genotype quality.
12. Hard filtering: Hard filtering is the process of removing variant calls that meet or exceed predefined thresholds for specific criteria.
13. Soft filtering: Soft filtering is the process of adjusting the quality scores of variant calls based on various criteria.
14. VCF file: A Variant Call Format (VCF) file is a standard file format used to store genetic variant data. VCF files contain information about the location, type, and quality of each variant call.

15. Genotyping: Genotyping is the process of determining the genotype of a sample at specific genetic loci. Genotyping can be performed using various methods, including NGS, microarrays, or PCR.
16. Haplotype: A haplotype is a set of linked genetic variants that are inherited together on the same chromosome.
17. Linkage disequilibrium: Linkage disequilibrium is the non-random association of genetic variants on the same chromosome.
18. Imputation: Imputation is the process of inferring missing genotype data based on patterns of linkage disequilibrium and haplotype structure.
19. Population genetics: Population genetics is the study of genetic variation within and between populations.
20. Phylogenetics: Phylogenetics is the study of evolutionary relationships among organisms based on genetic data.

Now, let's look at some practical applications and challenges related to variant detection and genotyping.

One practical application of variant detection and genotyping is identifying genetic mutations associated with disease. For example, researchers can use NGS to sequence the entire genome or exome (the protein-coding regions of the genome) of individuals with a particular disease and compare their genetic data to that of healthy individuals. By identifying genetic variants that are unique to the diseased individuals, researchers can gain insights into the underlying genetic causes of the disease.

Another application is pharmacogenomics, which is the study of how genetic variation affects drug response. By identifying genetic variants that are associated with drug response, researchers can develop personalized treatment plans that are tailored to an individual's genetic makeup.

Challenges in variant detection and genotyping include dealing with false-positive and false-negative variant calls, as well as accurately genotyping low-frequency variants. False-positive variant calls can occur due to sequencing errors, alignment errors, or other factors. False-negative variant calls can occur due to low read depth or poor sequence quality. Accurately genotyping low-frequency variants can be challenging due to their low frequency and the potential for sequencing errors.

To address these challenges, researchers can use various strategies, such as increasing read depth, improving alignment algorithms, and developing more accurate variant callers. Additionally, researchers can use statistical methods to estimate the false-positive and false-negative rates of variant callers and adjust their analysis pipelines accordingly.

In conclusion, variant detection and genotyping are critical components of NGS data analysis. By identifying genetic variants and determining genotypes, researchers can gain insights into genetic variation within and between populations, as well as the genetic causes of disease. While there are challenges associated with variant detection and genotyping, researchers can use various strategies to improve the accuracy and reliability of their analysis pipelines.