* **what is allelic depth and genotype depth in bcftools mpileup command?**

In the context of the bcftools mpileup command, **allelic depth** and **genotype depth** refer to different types of depth information for variant calling and genotyping analysis.

1. **Allelic Depth (AD):**
   * **Definition:** Allelic depth refers to the number of reads that support each allele at a specific genomic position. It is a measure of how many reads contain the reference allele and how many contain an alternative allele (variant).
   * **Usage in bcftools:** For each variant, AD is output as part of the information in the VCF (Variant Call Format) file, with a split count for reference and alternate alleles (e.g., AD=20,5 would mean 20 reads support the reference allele and 5 reads support the alternate allele).
   * **Format in VCF:**
   * AD=<num\_ref\_reads>,<num\_alt\_reads>

Where:

* + - <num\_ref\_reads> is the number of reads supporting the reference allele.
    - <num\_alt\_reads> is the number of reads supporting the alternate allele.

1. **Genotype Depth (DP):**
   * **Definition:** Genotype depth refers to the total depth of reads that support a particular genotype. This is the total number of reads that support the called genotype at a specific position, regardless of whether they support the reference or alternate allele.
   * **Usage in bcftools:** It represents the sum of the allelic depths for both the reference and alternate alleles. It is reported as a single number indicating the total depth for the genotype (e.g., DP=25 means there were 25 reads covering the position).
   * **Format in VCF:**
   * DP=<total\_depth>

Where:

* + - <total\_depth> is the total number of reads supporting the called genotype (reference + alternate).

**Example in a VCF file:**

If you run bcftools mpileup and have a variant at a certain position, the VCF record might look like this:

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1

chr1 123456 . A G 100 PASS DP=30;AD=20,10 GT:DP:AD 0/1:30:20,10

Here:

* **DP=30** indicates the total number of reads supporting the genotype.
* **AD=20,10** indicates 20 reads support the reference allele (A), and 10 reads support the alternate allele (G).
* **GT=0/1** indicates the genotype is heterozygous with one copy of the reference allele (A) and one copy of the alternate allele (G).

**Summary:**

* **Allelic depth (AD)** gives the read count for each allele separately.
* **Genotype depth (DP)** gives the total read count supporting the genotype.
* **Strand bias?**

Strand bias occurs when the reads supporting a particular allele (reference or alternative) are disproportionately from one strand of the DNA (forward or reverse strand). Ideally, you want to see balanced representation from both strands to ensure that the variant is real and not an artifact.

**Strand Bias in VCF Format:**

To specifically track strand bias in VCF format, there are additional fields you can look for. The bcftools mpileup command can calculate a strand bias metric, usually indicated in the **INFO** field of the VCF output as SB (strand bias).

* **SB (Strand Bias) Field in VCF**: The SB field reports a statistical test for strand bias, which measures the ratio of reference and alternate allele support from each strand. The result will typically be a ratio or a p-value indicating if the variant might be biased toward one strand.

**Example with Strand Bias (SB) in VCF:**

Suppose you run bcftools mpileup and get the following VCF record for a variant:

#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT SAMPLE1

chr1 123456 . A G 100 PASS DP=30;AD=20,10;SB=1.2 GT:DP:AD:SB 0/1:30:20,10:1.2

Here:

* **DP=30** indicates 30 reads covering the position.
* **AD=20,10** means 20 reads support the reference allele (A), and 10 support the alternate allele (G).
* **SB=1.2** might indicate the strand bias ratio (depending on the specific method or tool used), which reflects the relative evidence from the forward and reverse strands.

**Identifying Strand Bias:**

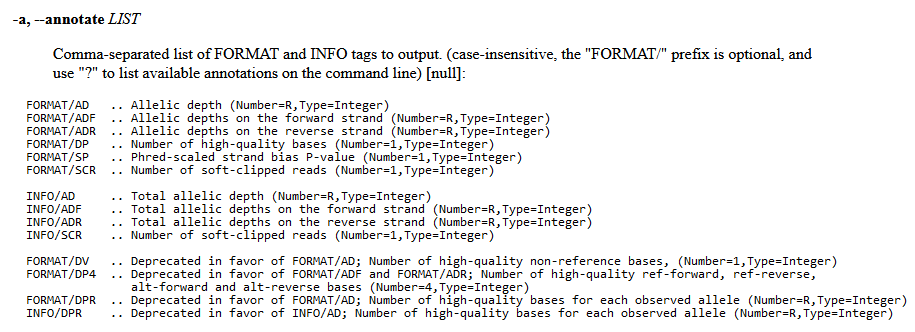
In practice, you can look at the ratio of reference vs. alternate allele counts from the two strands (forward vs. reverse). If the SB value is significantly skewed (e.g., a large ratio between forward and reverse strand support), it may suggest strand bias.

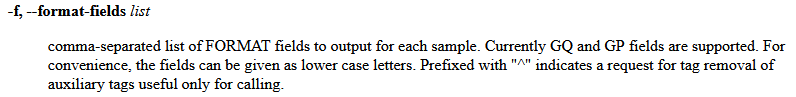
* A **low SB ratio** (closer to 1) means little to no bias between strands.
* A **high SB ratio** could indicate that the variant is supported more strongly by one strand, which might be a sign of a sequencing artifact or bias.

**How to Handle Strand Bias:**

When you detect strand bias in a VCF file, you may:

1. **Flag variants with high strand bias** for further investigation.
2. **Apply filters** to exclude variants with significant strand bias, especially if you suspect it’s due to sequencing errors.
3. **Examine the sequencing process** to understand whether certain regions or techniques could cause bias (e.g., PCR amplification).

mpileup:

call: