

# Scipio Documentation

## Contents

1. What is the difference between `--min_score` and `--min_coverage`?
2. What is the difference between `--max_assemble_size` and `--min_dna_coverage`?
3. How far does Scipio search for unmatched query sequence in the terminal regions?

## Difference between `--min_score` and `--min_coverage`?

The `--min_coverage` parameter is used to exclude all hits that by chance show some homology of short stretches (e.g. if your query sequence contains regions of low complexity there is some chance that you might find several "homologous" regions in the target genome). The `--min_coverage` parameter is applied first to the refined Blat hits (for more details look at the activity diagrams). Imagine the case that just two stretches of the query sequence by chance map with more or less mismatches to one of the target contigs, while a long part inbetween these stretches is not found. Such a "hit" is very unlikely to be part of the correct result, although the matched part might be identical between query and target resulting in a high score. Nevertheless, there might really be a gap in the assembled sequence (e.g. between the contigs part in a supercontig or chromosome assembly) and therefore the default value for `--min_coverage` is 60 (60 percent of the maximal possible query sequence must be covered on every (partial) contig). To the thus filtered hits, the `--min_score` parameter is applied next. The `--min_score` parameter is used to filter perfect ( $\Rightarrow$  complete) or nearly perfect results from those that either cover just a short stretch of the query sequence or that are too unrelated, meaning too many mismatches. The default value of 0.3 means, that at least about 30 percent of the query sequence must be found on a single contig (the exact formula for the score is: matches minus mismatches, divided by query length.). If the gene is spread on several contigs, the minimal score applies to the best-scoring partial hit. In the subsequent assembly of partial hits all other hits will be added to this hit regardless of their score.

## Difference between `--max_assemble_size` and `--min_dna_coverage`?

Max. Assemble Size defines the maximum size of intron parts at target boundaries as absolute number of nucleotides (default: 75000 nucleotides). If an intron would have to be created between two partial hits across two contigs, that exceeds the given size, the two hits cannot appear together as parts of one composed hit. Min. DNA Coverage is a possibility to define the minimum for the mapping of query sequence to the target (explicitly: hit length divided by intron length). By default, the `--min_dna_coverage` is set to 0 and would thus allow even single amino acids found on contigs to be included in the assembled gene. The `--max_assemble_size` parameter is useful for assemblies with large contigs, while the `--min_dna_coverage` parameter might be useful for assemblies with short contigs.

## Search in terminal regions?

If terminal exons are very short or divergent they will not be identified by Blat. Scipio tries to find the unmatched query using direct pattern search methods in the terminal regions defined by the `--max_assemble_size` parameter (default: 75000 nucleotides). The sequences to be searched for are limited in length by the `--gap_to_close` parameter, they must be identical between query and target, and the found hits must result in accepted splice sites for the new intron. E.g., an N-terminal methionine will only be identified if the intron to exon 2 has the standard splice sites GT---AG or GC---AG.