The largest discrepancy occurs in the number of stop codons between the coding sequences (separate genes) and the whole genome sequence, as shown in the figure below, the difference is close to 800. In the whole genome, the total length is 29,904 bases, of which 29,264 bases belong to coding regions, and the remaining 640 bases are non-coding regions. The non-coding region can form approximately 213 codons (640 ÷ 3 ≈ 213). The codons in these non-coding regions are randomly distributed and may contain many stop codons (TAA, TAG, TGA). These additional stop codons won't appear in individual gene coding sequences because they don't participate in protein production. From the whole genome sequence, we can print the positions of the first 10 stop codons: [25, 67, 112, 130, 154, 262, 403, 433, 457, 523], but the first stop codon in individual gene coding sequences is 21289, The did not appear primarily because the whole genome sequence includes large non-coding regions, which don't appear in individual gene coding sequences but contain randomly distributed stop codons. Another reason is that coding regions are always multiples of 3 (as each codon consists of 3 nucleotides), while non-coding regions don't have this restriction and can be of any length. When we read the entire genome in groups of three, the reading frame shifts, and sequences that are actually start codons might be misinterpreted as not being start codons. That's why the first ATG appears at position 901 instead of 266