

width = 221, height = 100, channel = 7

输入: 染色体 chr, 位置 k, 突变类型 (snp 或 in 或 del), 突变 alt

1, 取到覆盖该位点的所有 reads (深度 depth), 对每条 read 做循环

- 1> 若该条 read 没有突变信息或比对质量 read_q < 10 或该位点碱基质量 < 10, 则 continue
- 2> 获取该 read 的起点 start_point, 碱基序列 base, 碱基质量序列 base_q, 对应的 ref 序列, 正负链信息, cigar 信息。
- 3> 若 cigar 信息中存在 I, D, 则对 base 和 base_q 做修正:
 - 1> 对于 M 段, 保持不变
 - 2> 对于 D 段, base 用 "-" 表示, base_q 为 0
 - 3> 对于 I 段, base 用对应小写字母表示
- 4> 若 cigar 信息中存在 I, 则对 ref 序列做修正, 举例说明: base 为 ATGCatgGCC, 而 ref 为 ATGCGCC, 则将 ref 变为 ATGC***GCC, 为了方便后续一一对应做比较
- 5> 对于[pic_start, pic_end)中的每个点 p 做循环来计算该点值 int poi[7], 初始化均为 0
 - 1> 若 p - start_point < 0 或 > base 长度 - 1 (该点未被 read 覆盖), 则 continue
 - 2> poi[0]根据 base[p]的碱基, A 为 250, G 为 180, T 为 100, C 为 30, N 为 0, "-" 为 0, 小写字母对应的 I 区域为 42
 - 3> poi[1]根据 base_q[p], 和 40 取 min 后归一化到[0, 254]
 - 4> poi[2]根据 read_q, 和 60 取 min 后归一化到[0, 254]
 - 5> poi[3]根据正负链信息, 正链 70, 负链 240
 - 6> poi[4]根据 read 是否支持该 alt, 支持为 254*0.2 取整, 不支持为 254
 - 7> poi[5]根据 base[p]和 ref 是否一致, 一致为 254, 不一致为 254*0.6 取整
 - 8> poi[6]根据 cigar 信息, M 和 I 区域用当前片段长度表示, D 区域用 0 表示, 举例说明: 2M1D3M 对应为[2, 2, 0, 3, 3, 3], 2M3I4M 为[2, 2, 3, 3, 3, 4, 4, 4, 4]

根据事先读取的该区域 ref 片段, 根据通道 1 为碱基数值, 通道 2356 均为 254, 通道 4 为 70, 通道 7 为 0, 这样重复 5 行, 后将 reads 数值排在后面, 因而 depth = depth + 5 这样, 得到一个[depth, width, channel]的像素矩阵。

- 2, 1> 若 depth > height, 则从 depth 行中任意选择 height 行
- 2> 若 depth == height, 则保持不变
- 3> 若 1 < depth < height, 则用 0 补全至 height 行
- 4> 若 depth <= 1, 则放弃该位点

最终矩阵规格为[height, width, channel]

- 3, 因为矩阵中存在 insert 元素, 所以目前矩阵中每一列并没有和 ref 一一对应, 所以要调整矩阵来将所有的 insert 区域 (poi[0]==42) 空开, 空开部分用 0 补齐, 同时舍弃向右移动超出宽度的部分, 简化举例: 原矩阵和调整后的矩阵分别为

[[1. 1. 42. 42. 1. 1.]	[[1. 1. 42. 42. 1. 1.]
[1. 1. 42. 42. 1. 1.]	[1. 1. 42. 42. 1. 1.]
[1. 1. 42. 42. 1. 1.]	[1. 1. 42. 42. 1. 1.]
[1. 1. 2. 3. 1. 1.]	[1. 1. 0. 0. 2. 3.]
[1. 1. 2. 3. 1. 1.]	[1. 1. 0. 0. 2. 3.]
[1. 1. 2. 3. 1. 1.]	[1. 1. 0. 0. 2. 3.]