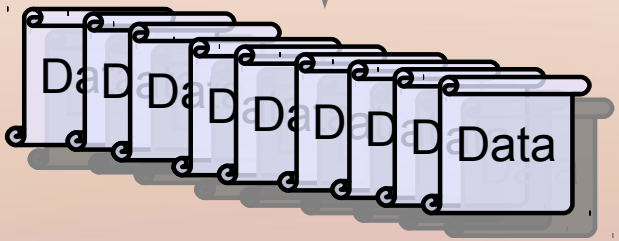
Efficient comparison of sets of intervals with NC-lists

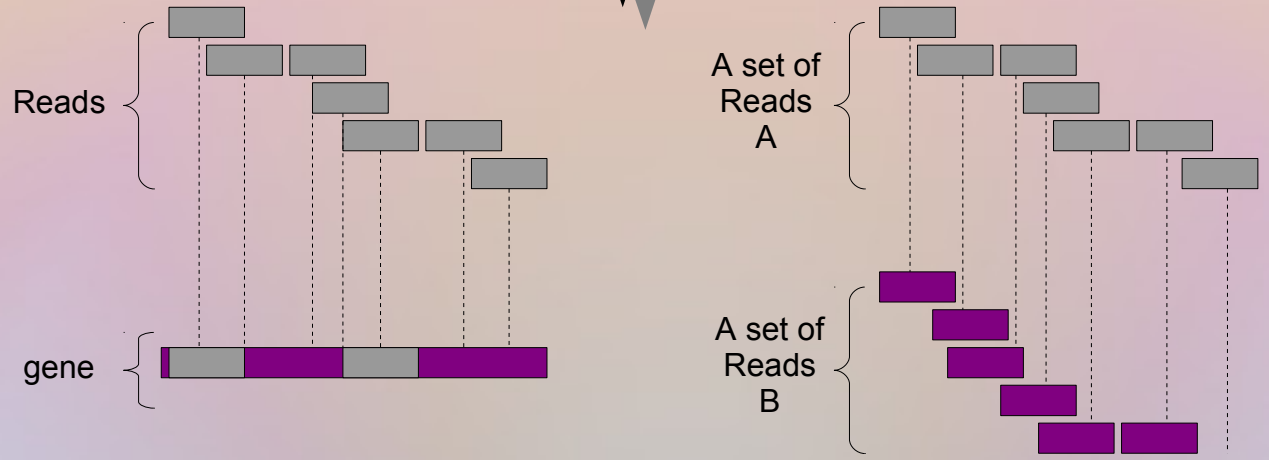
Yufei Luo, Matthias Zytnicki, Hadi Quesneville
URGI INRA Versailles

- Introduction
- Methods
- Results

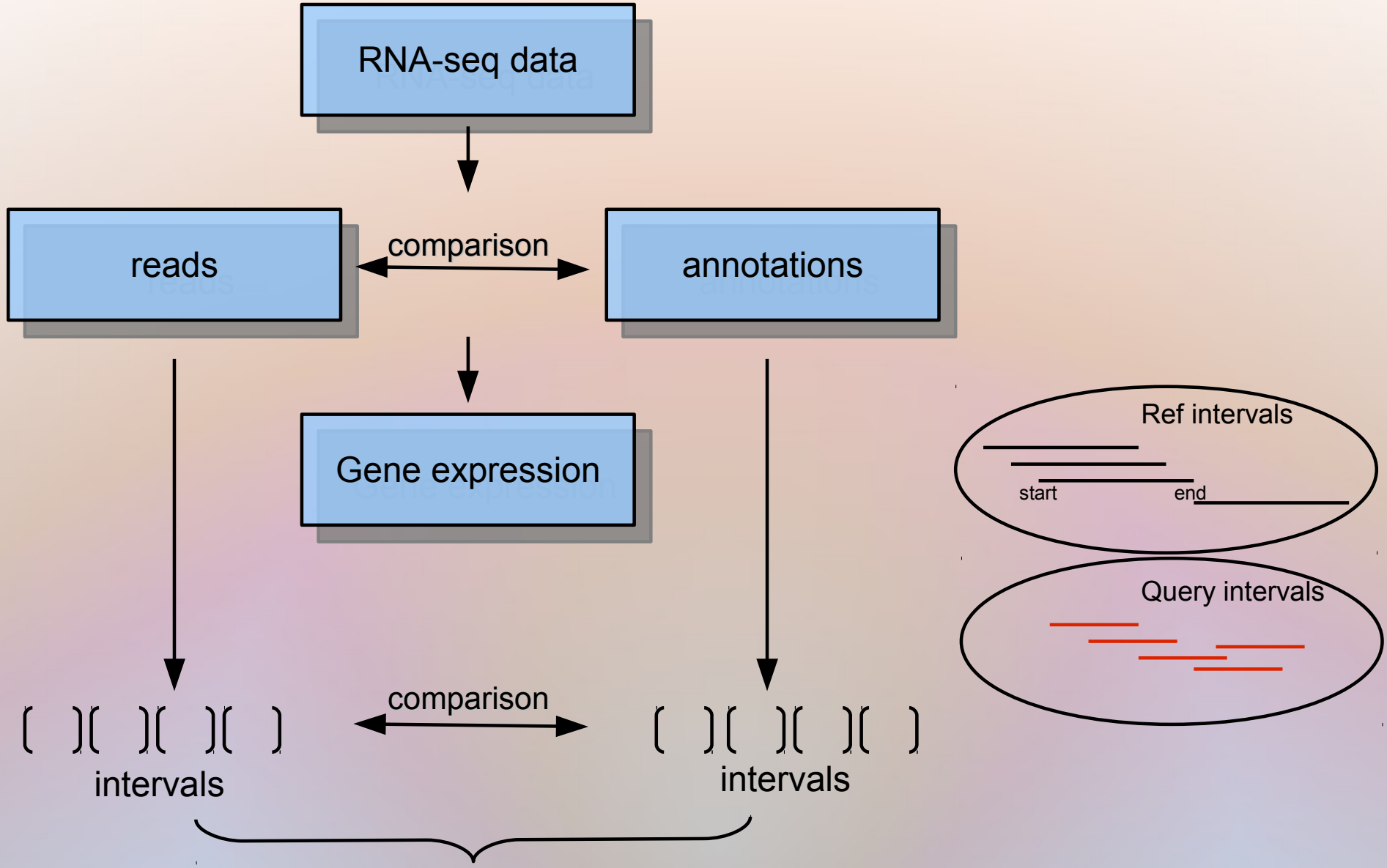
High-throughput sequencing



Genomic coordinates



Are there **optimized algorithms** for the fast comparison of large amount of sequenced data?

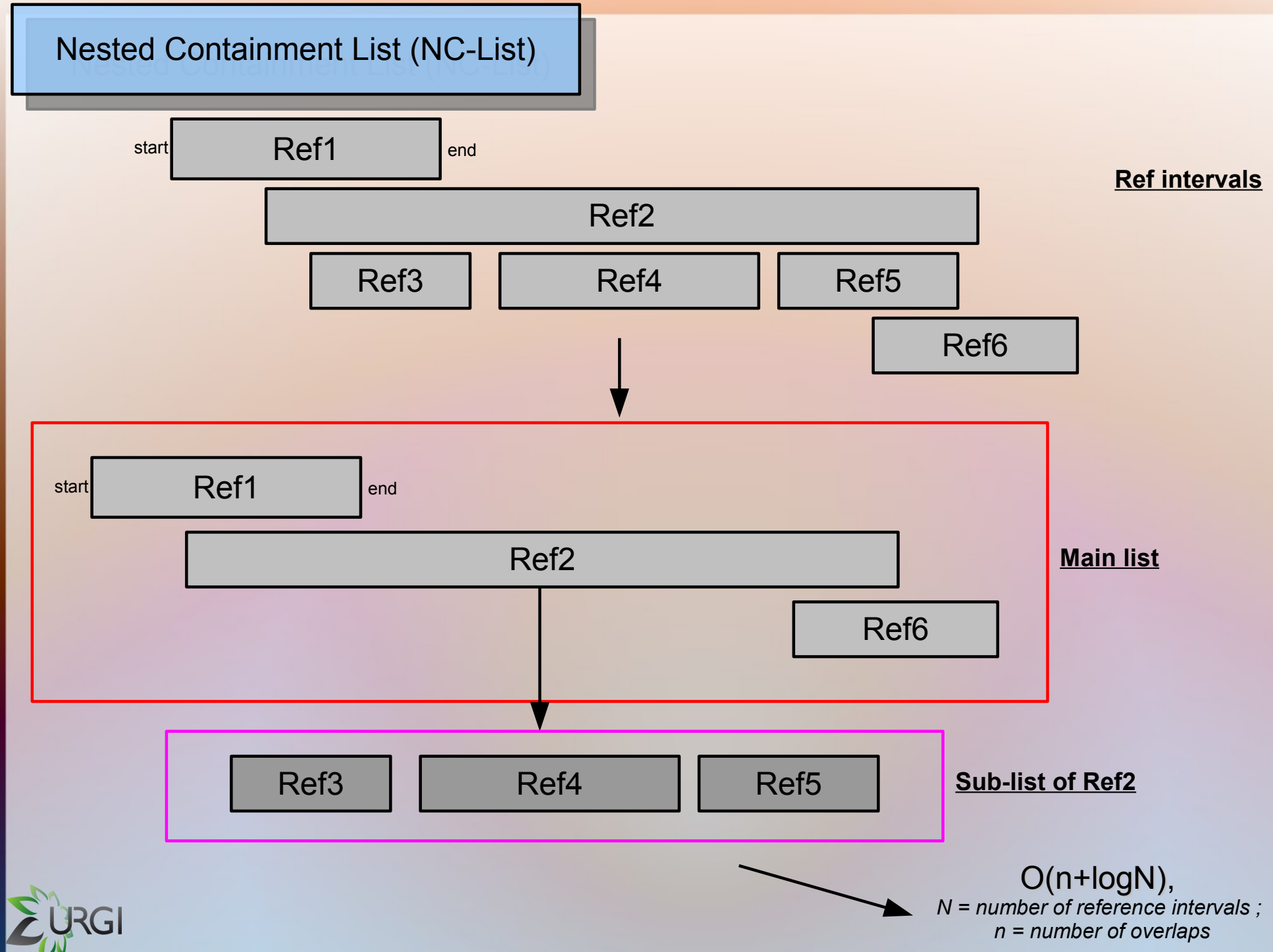


NC-List → O(n+logN), → Is it true?

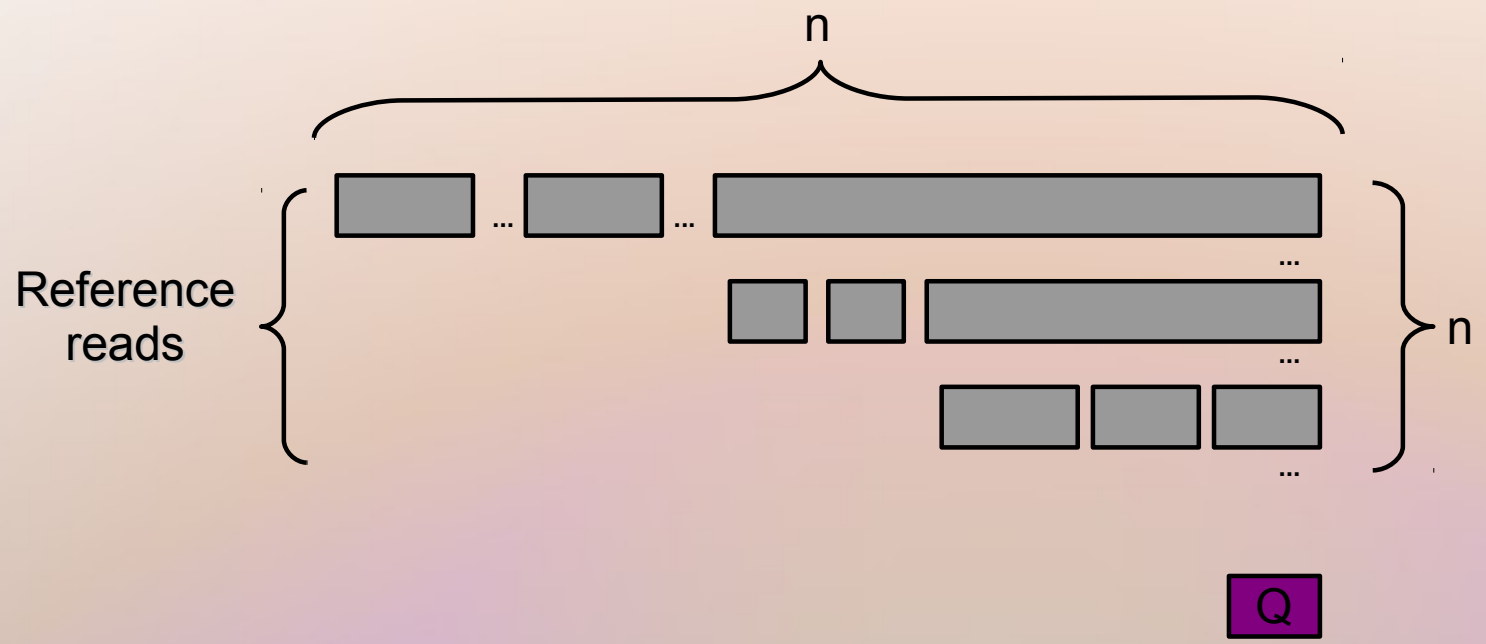
Alexander V. Alekseyenko and Christopher J.Lee. Bioinformatics, 2007

N = number of reference intervals ;
n = number of overlaps

Nested Containment List (NC-List)



Example:

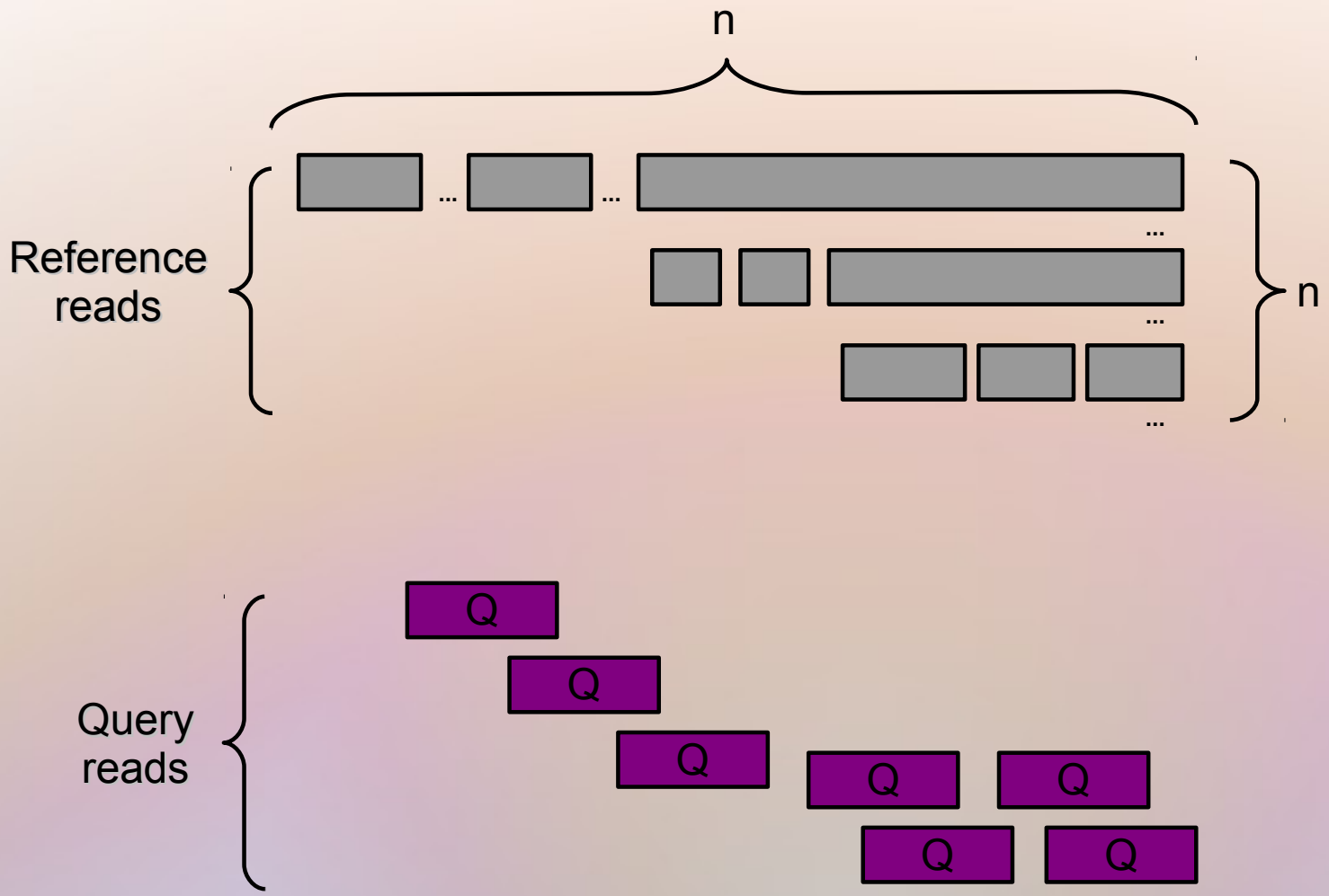


#Reference reads = n^2; #Overlaps = n

Complexity = O(log#Reference reads + #Overlaps)

= O(log(n^2) + n) = O(2logn+n)

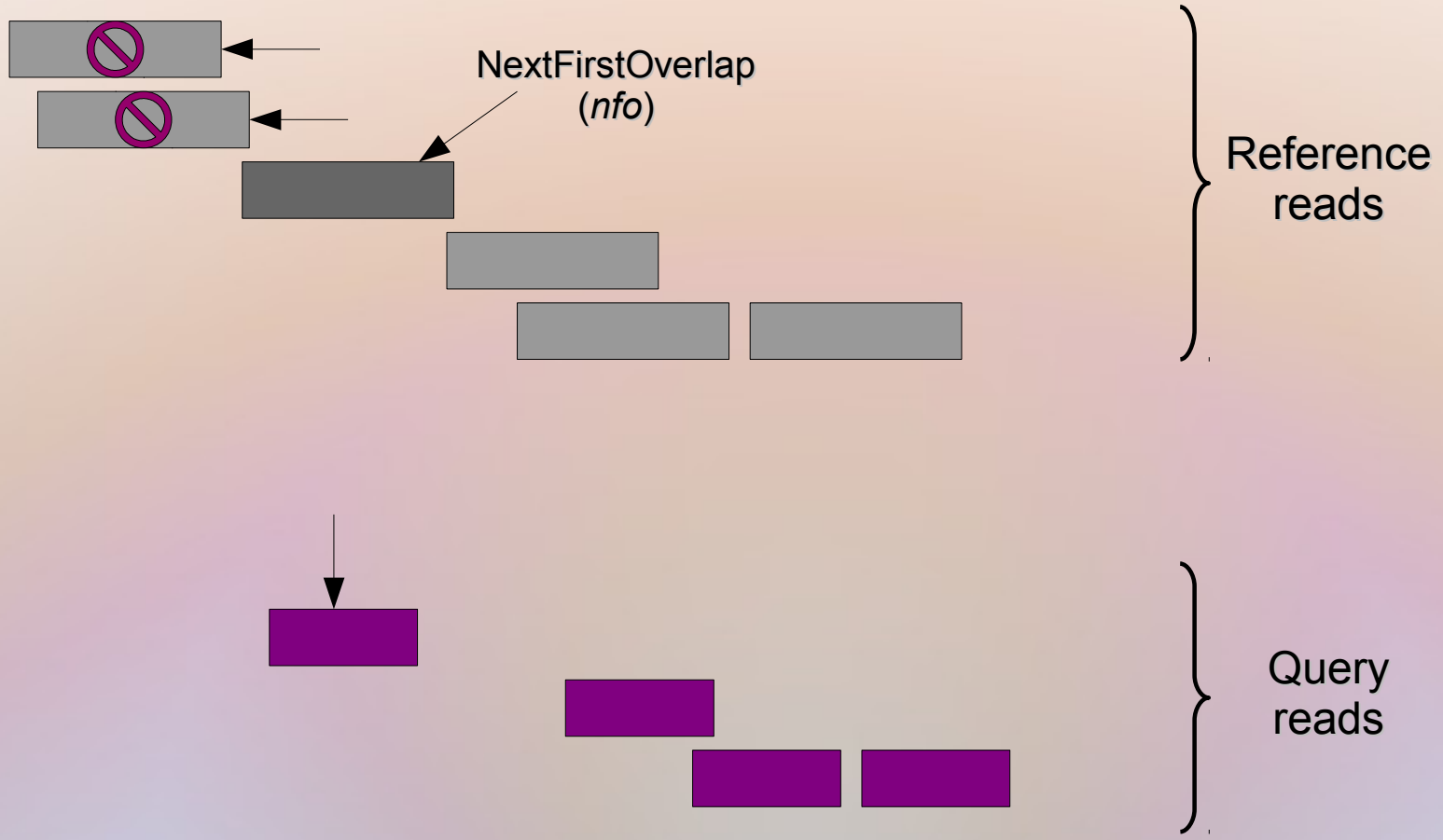
~ O(nlogn) >> O(logn + n)



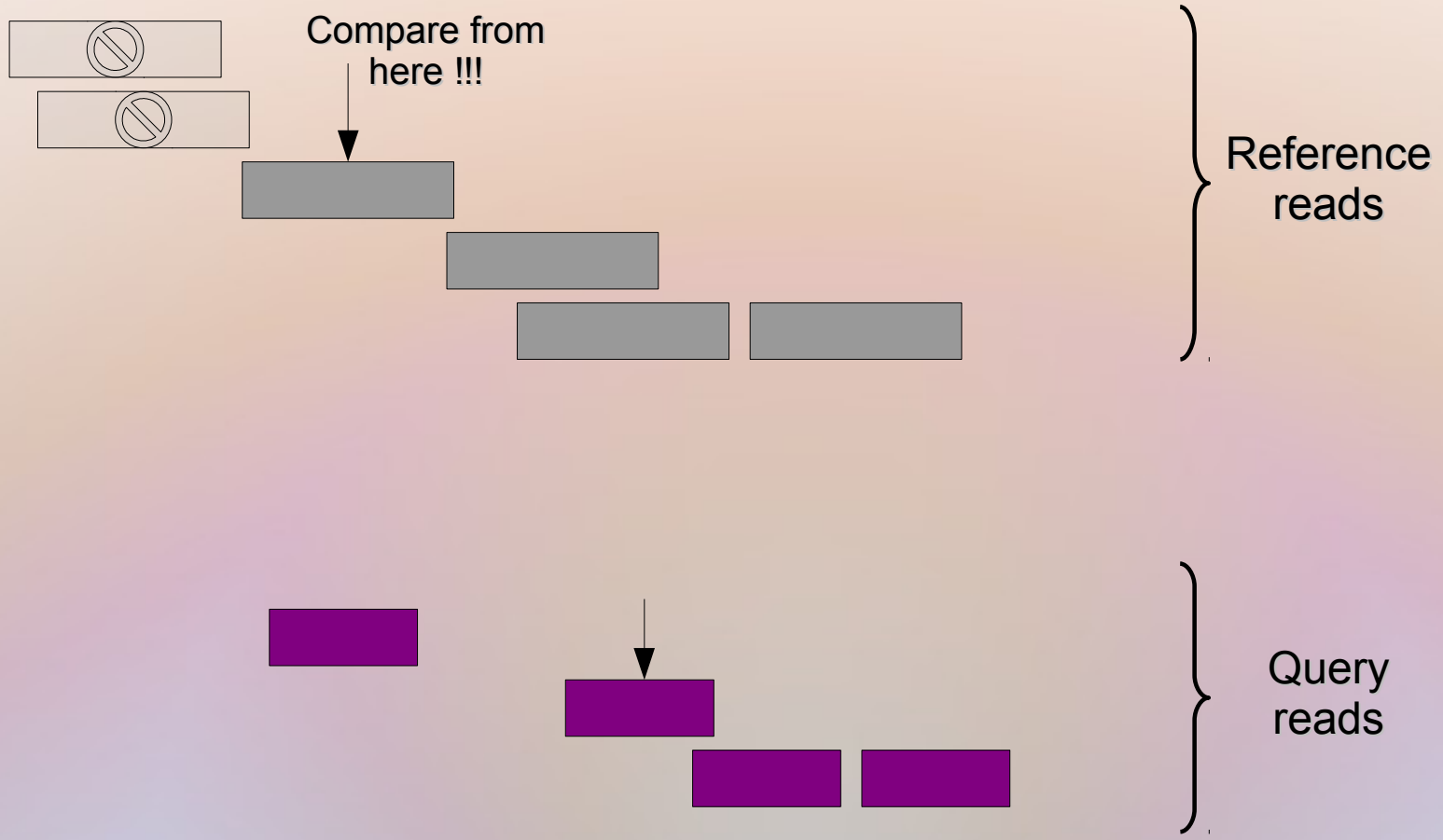
A new algorithm?

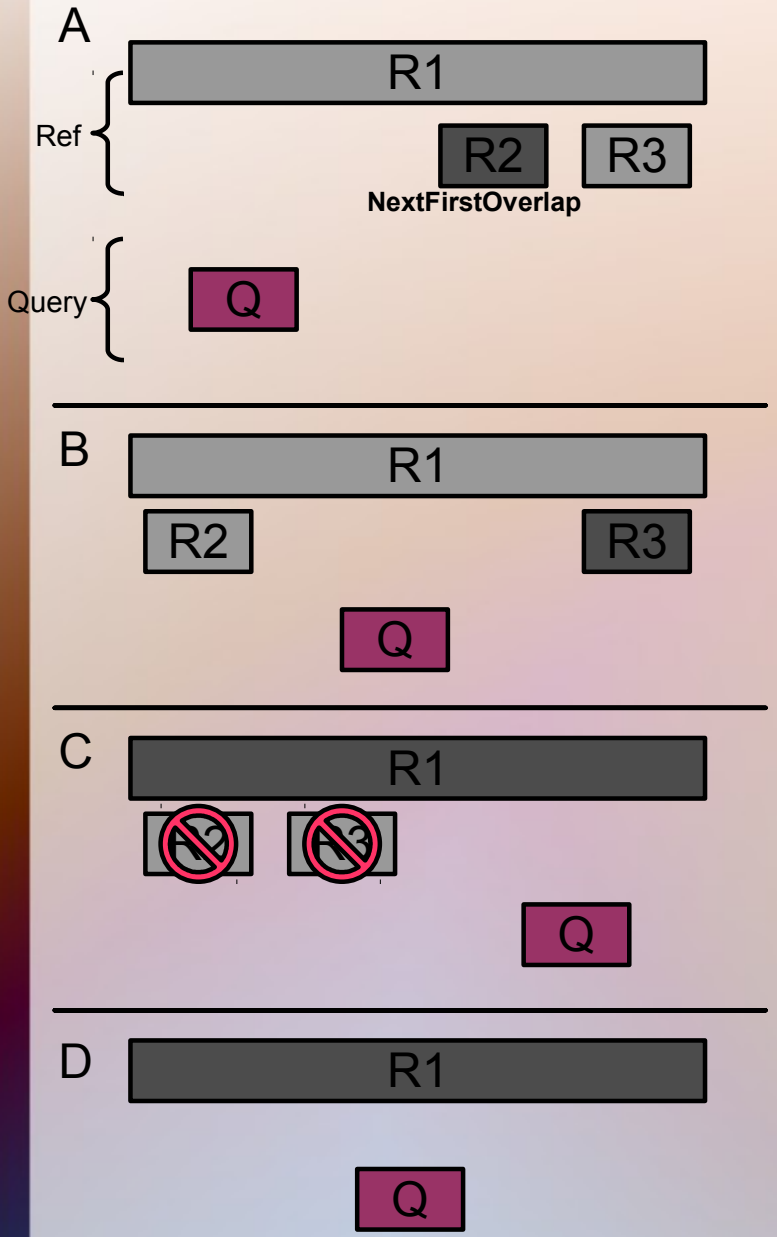
New idea

New idea



New idea



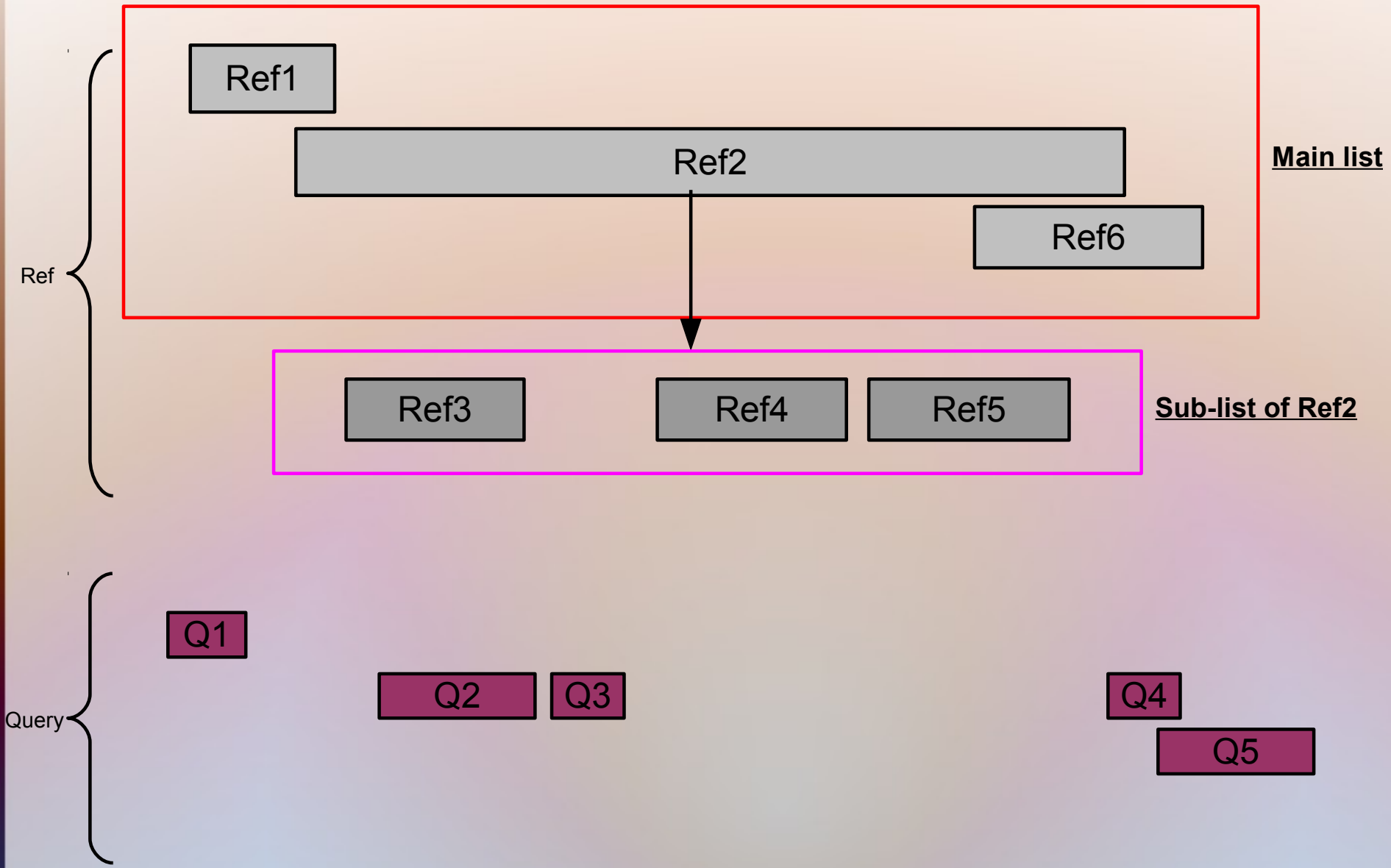


Function `simpleFindOverlap`

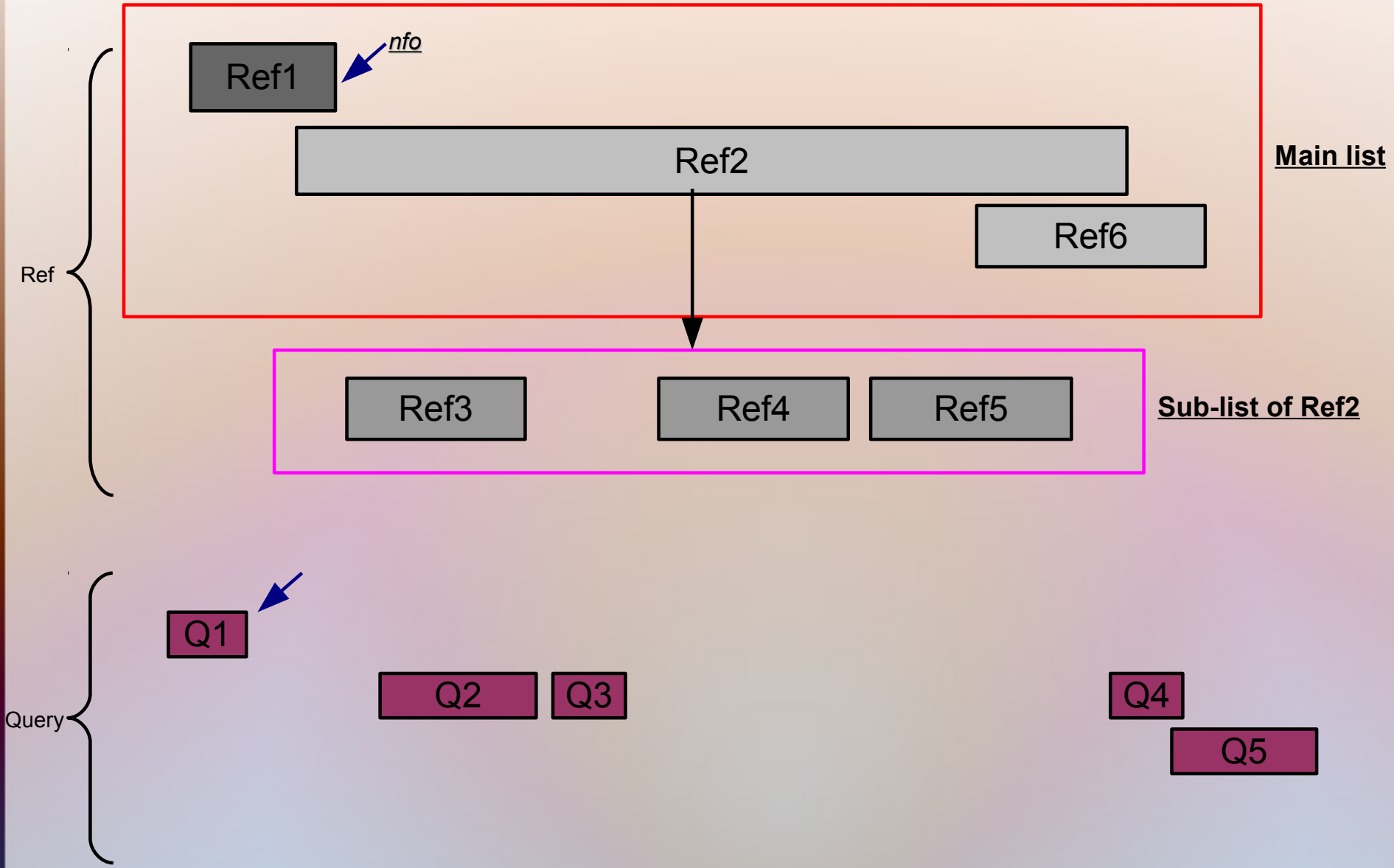
```
1  nfo ← R[0];
2  foreach q in Q.sorted() do
3      r ← nfo;
4      nfo ← None;
5      while true do
6          if r < q then r ← r.next;          // On the left side
7          else if r <> q then                // Overlaps with
8              if nfo = None then nfo ← r;
9              M.add(r);
10             r ← r.firstChild or r.next;
11         else break;
```

!! Compare firstly with all parents of *nfo for the next Query.*

Example:

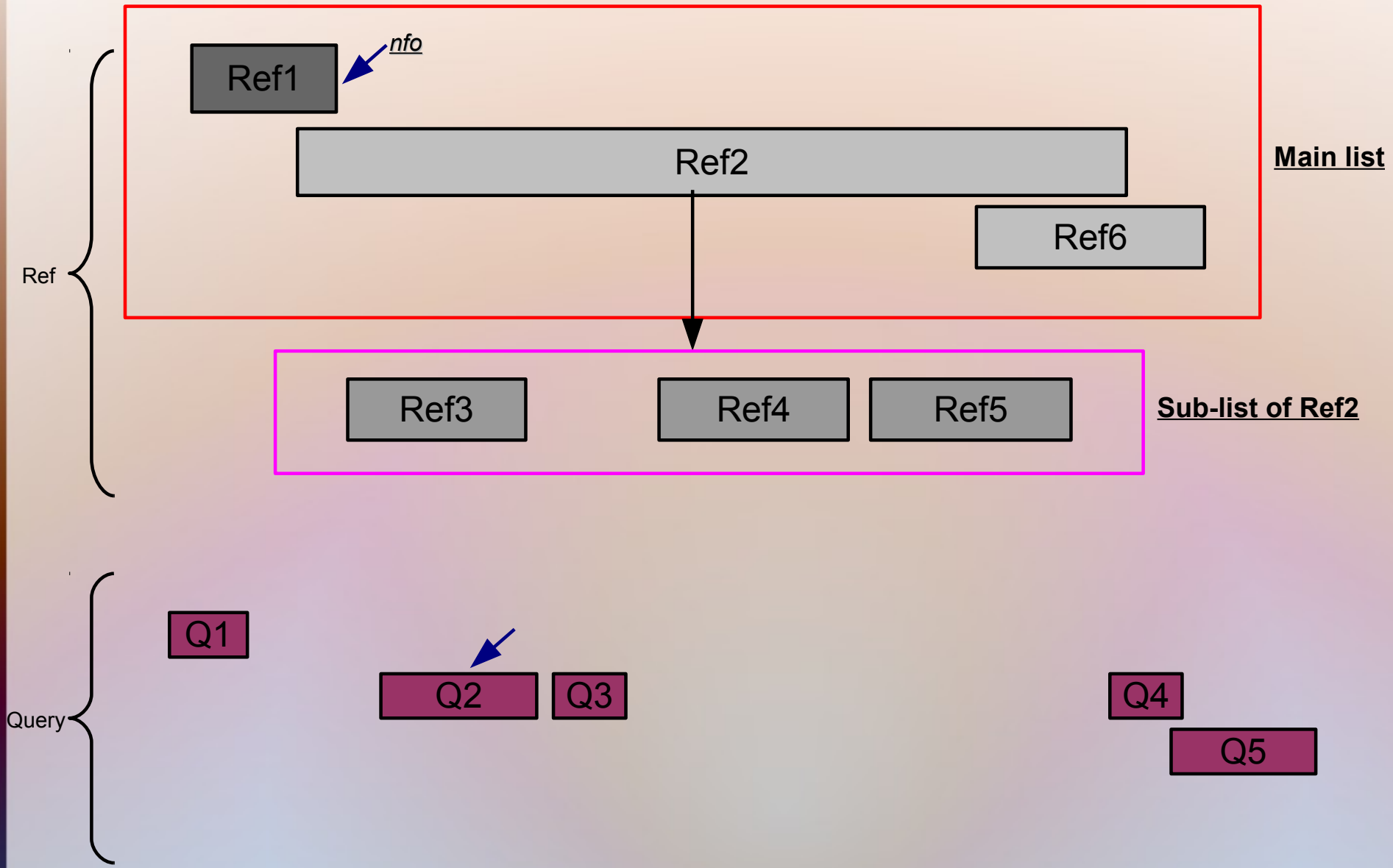

Ref
Ref1
Ref2
Ref6
Main list
Ref3
Ref4
Ref5
Sub-list of Ref2
Query
Q1
Q2
Q3
Q4
Q5

Example:



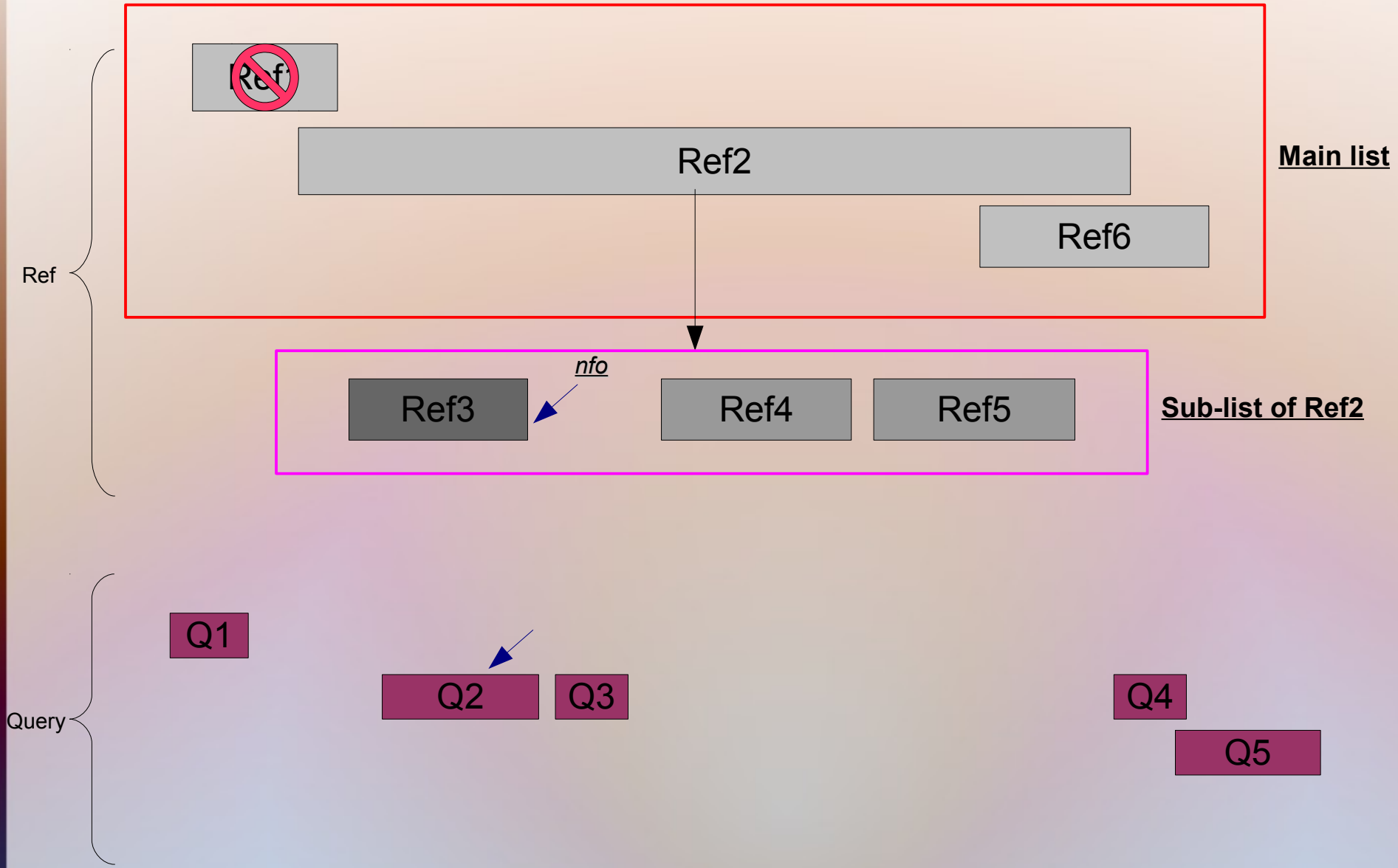
Q1 is overlap with: Ref1

Example:



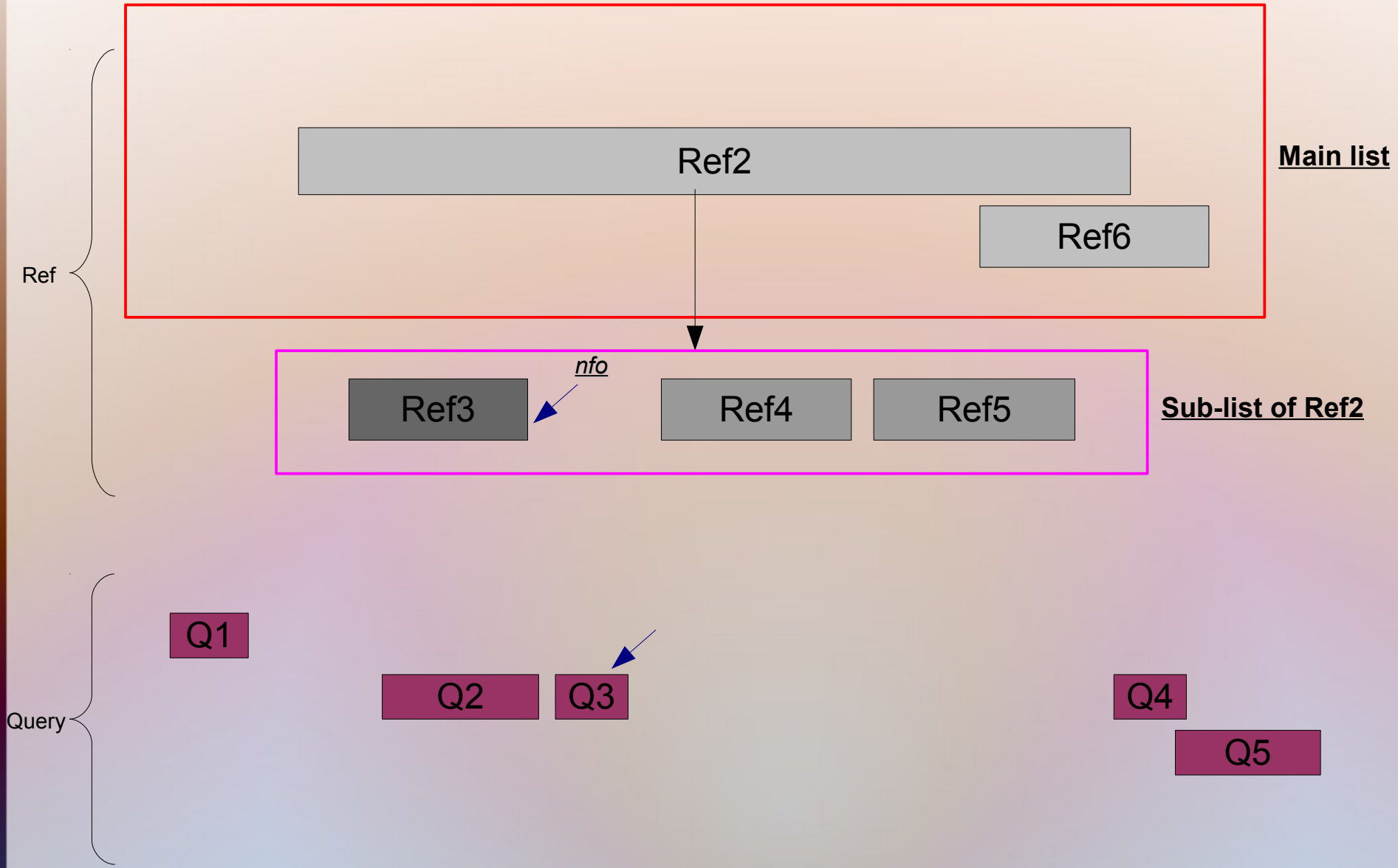
Q1 is overlap with: Ref1

Example:



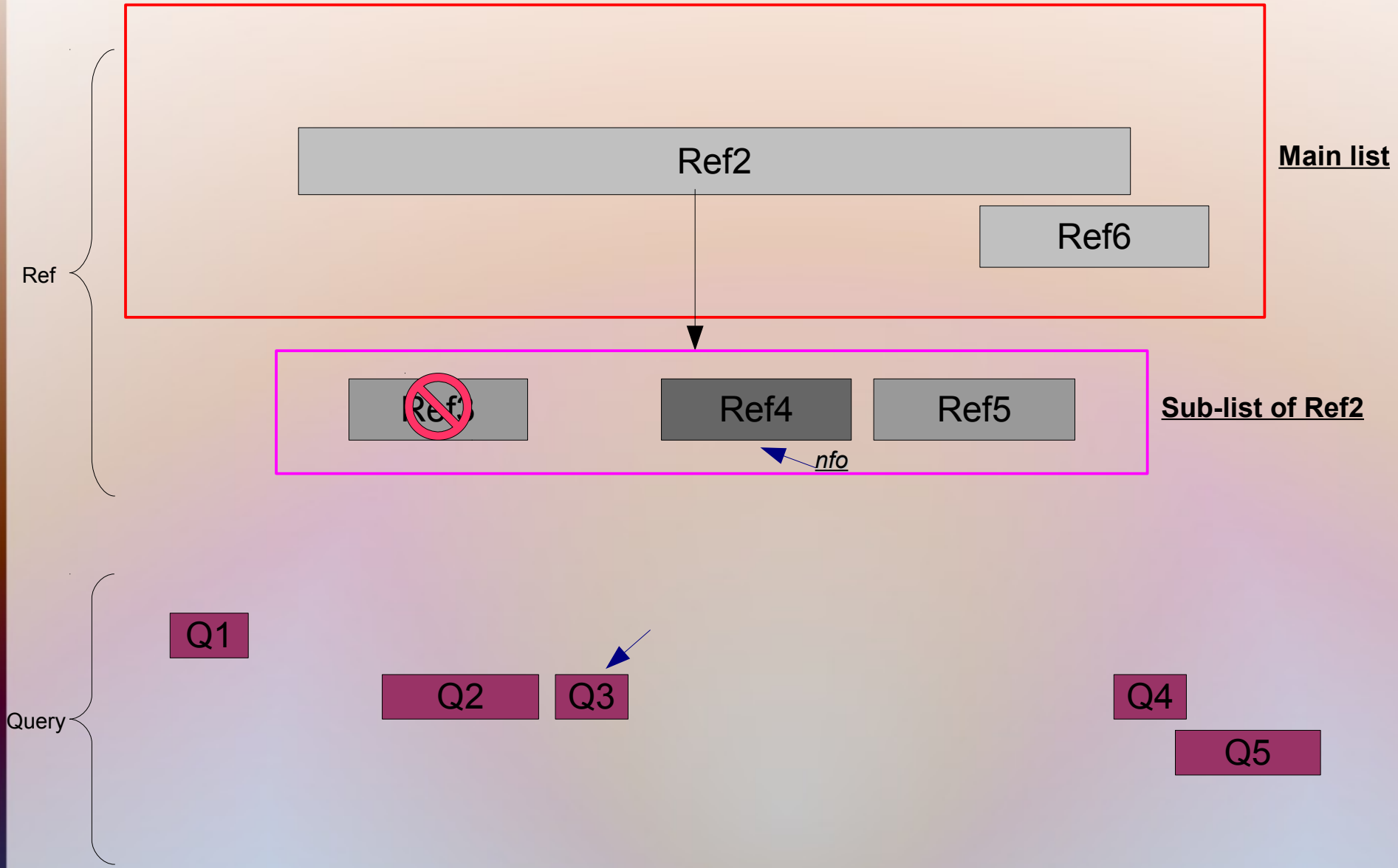
Q2 is overlap with: Ref2, Ref3

Example:



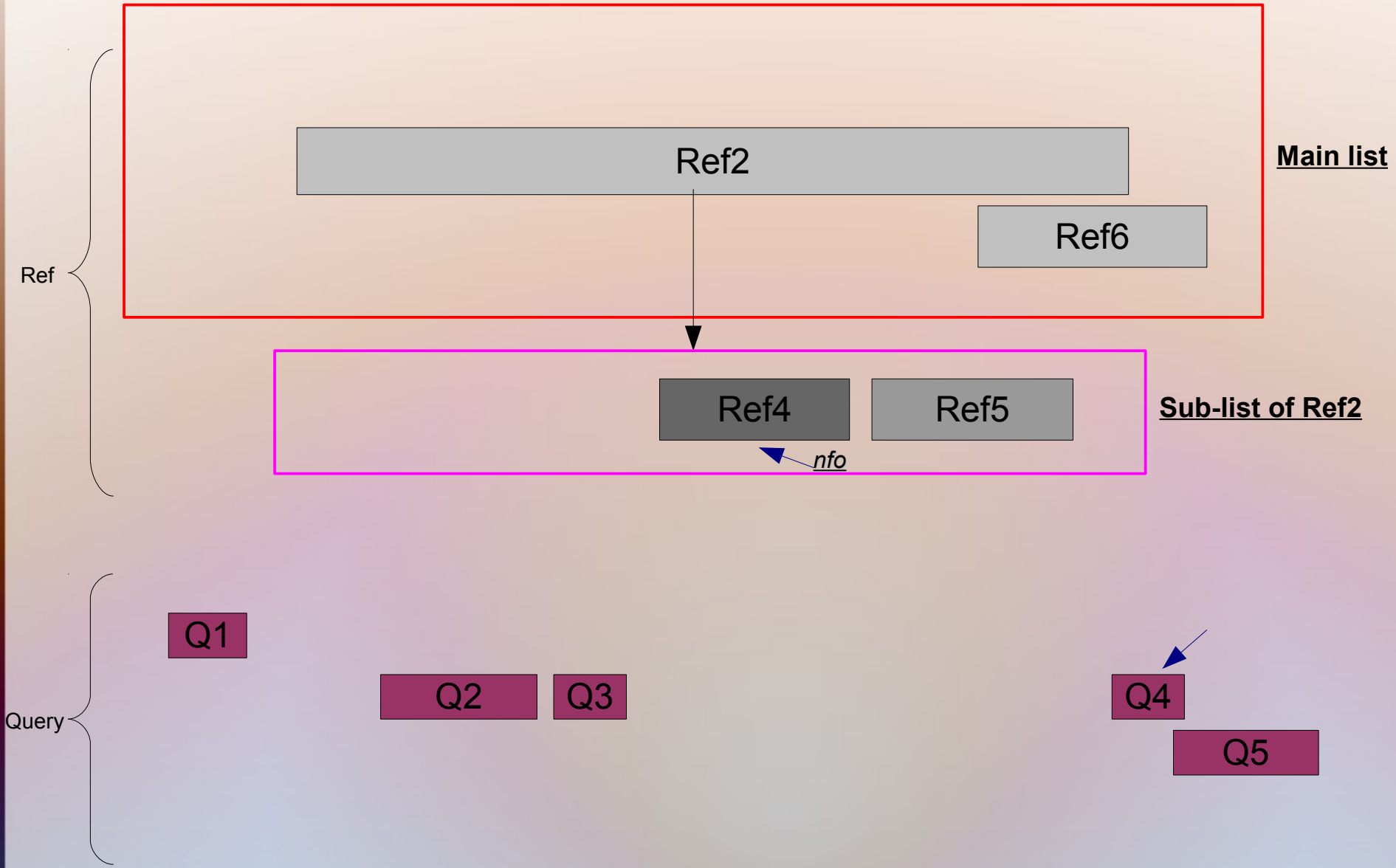
Q2 is overlap with: Ref2, Ref3

Example:



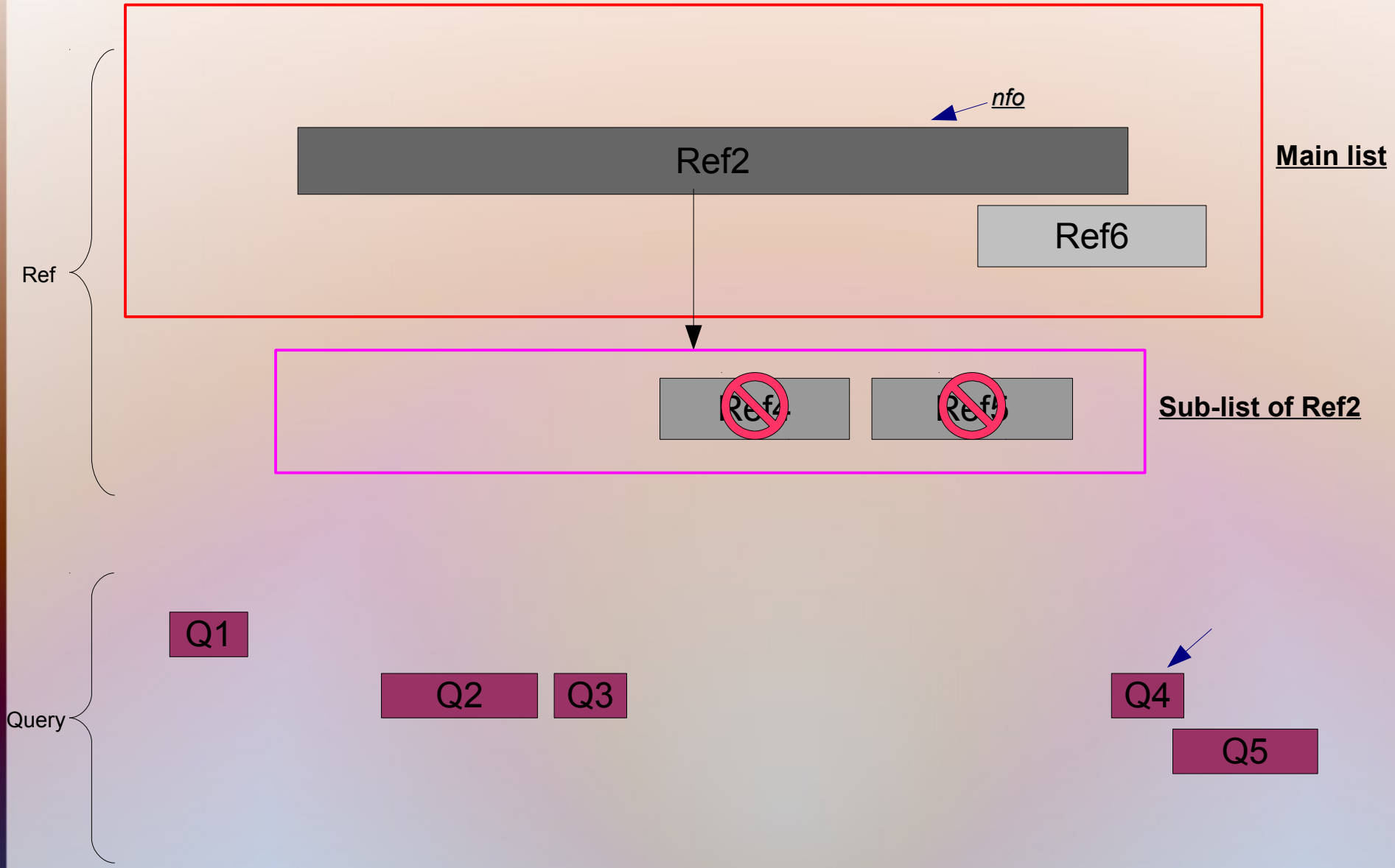
Q3 is overlap with: Ref2

Example:



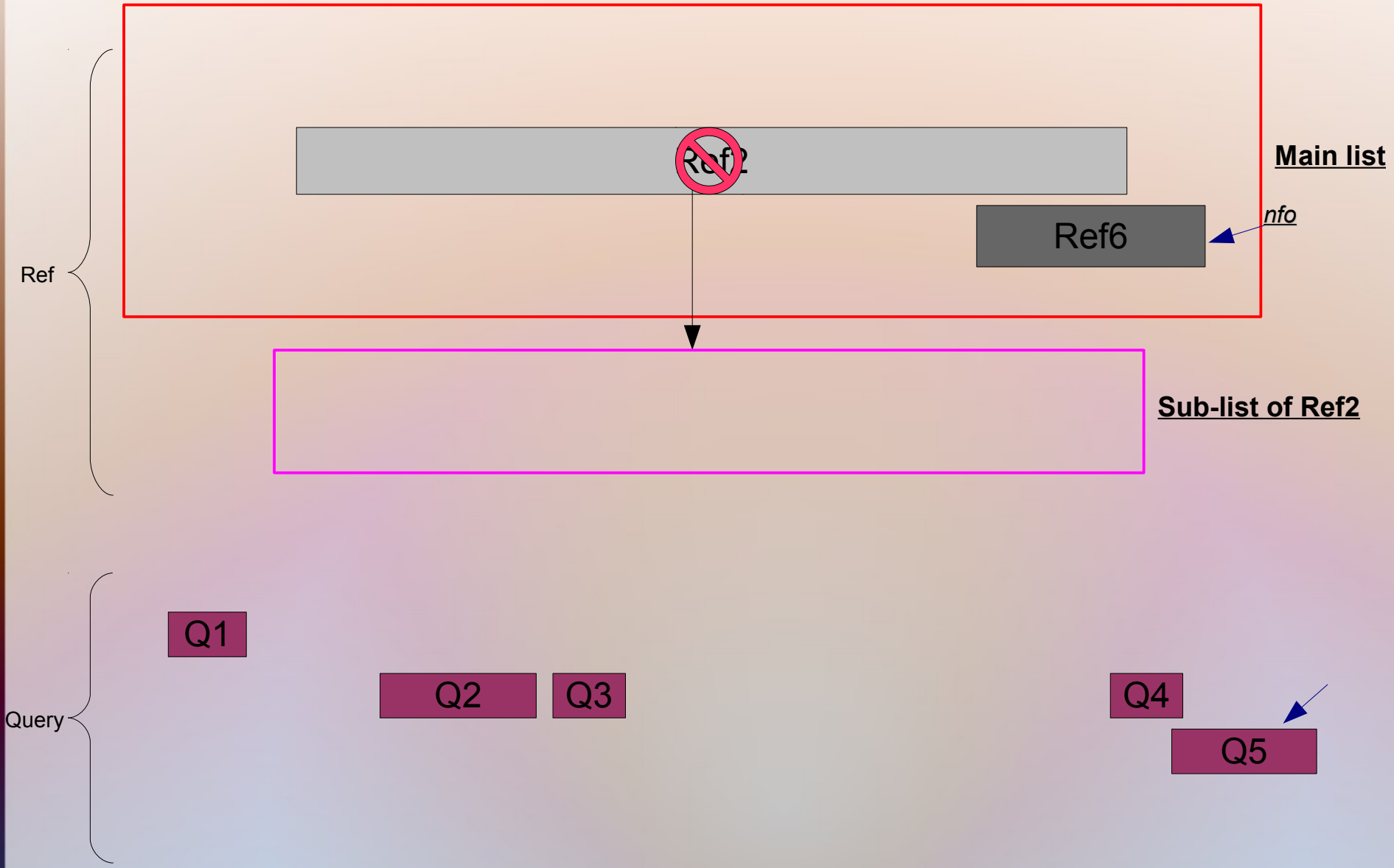
Q3 is overlap with: Ref2

Example:



Q4 is overlap with: Ref2, Ref6

Example:



Q5 is overlap with: Ref6

The complexity:

O(#R + #Q + #M)

Where R is the set of reference intervals;
Q is the set of query intervals;
M is the set of reference intervals overlapping with the query intervals.

Results

Compare with binning algorithm

Organism	Reference: RefSeq genes	Query: reads	Number of overlaps	Results
Drosophila melanogaster	18453	24274 454 technology, unique mapping	37761	Our algorithm: 10s
				Binning: 29s
		292085 454 technology, Non-unique mapping	467815	Our algorithm: 1m41s
				Binning: 4m27s
		3013398 Illumina Unique mapping	3716626	Our algorithm: 15m18s
				Binning: 46m45s

perspectives

Comparaison with other tools
Implementation in S-MART[1] and Galaxy

[1] *Matthias Zytnicki and Hadi Quesneville, PloS ONE, 2011*

acknowledgment

APLIBIO

eBIO

INRA