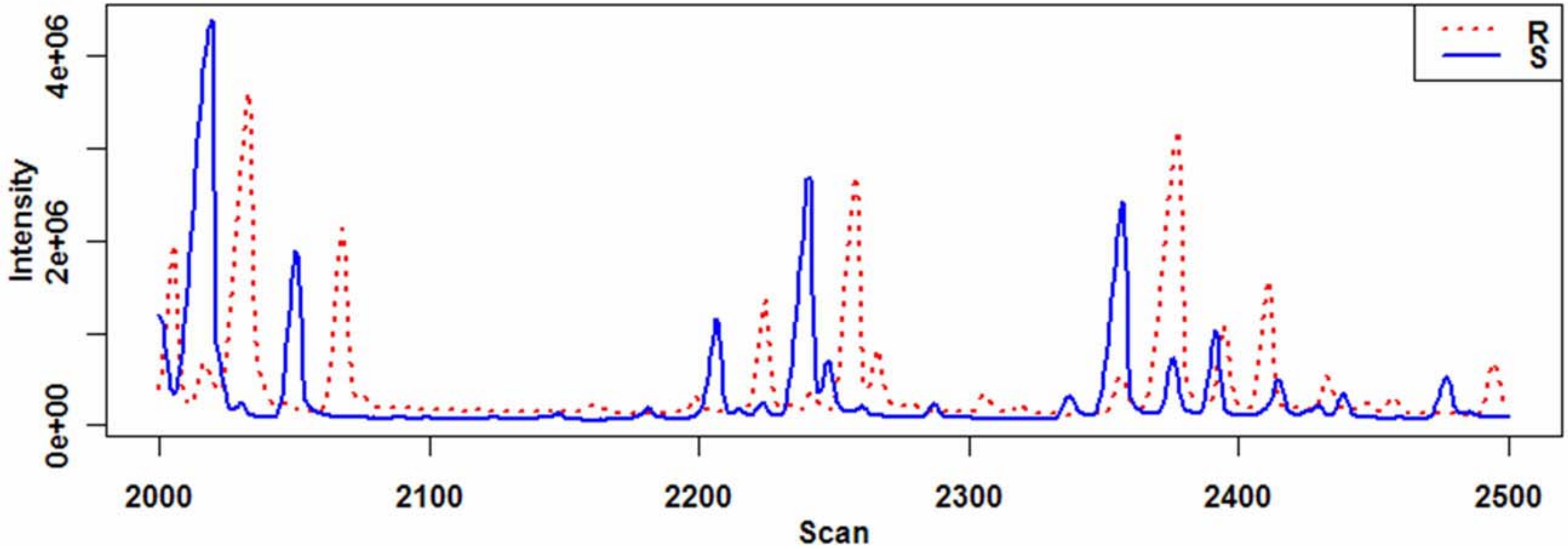
washAlign: a GC-MS Data Alignment Tool Using Iterative Block-Shifting of Peak Retention Times Based on Mass-Spectral Data

Minho Chae

UALR/UAMS Joint Graduate Program in Bioinformatics

GC-MS

- Powerful technique used in metabolomics study
- Identification is based on a retention time (*RT*) and a mass spectrum – build library
- Significant nonlinear inter-run variance in *RT*
  - Big hurdle for multi-dimensional analysis, i.e., MCR-ALS or PARAFAC
  - 2-way (*RT* space & *mz* space) data analysis more common

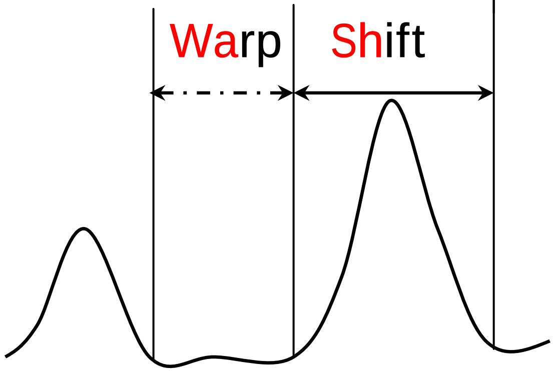


Alignment Methods

- COW (Correlation Optimized Warping) – *Nielson et al.*
  - Pairwise, difficult to find optimal input parameters (*N*, *S*)
  - Distortion of peak areas
- XCMS – *Smith et al.*
  - Statistical approach based on feature detection; median position of well behaved *peak-groups*
  - Better alignment result
- Why need one more?
  - Output more suitable to multi-dimensional analysis
    - Precise alignment
    - Little distortion of peak areas
  - Easier visual inspection

washAlign

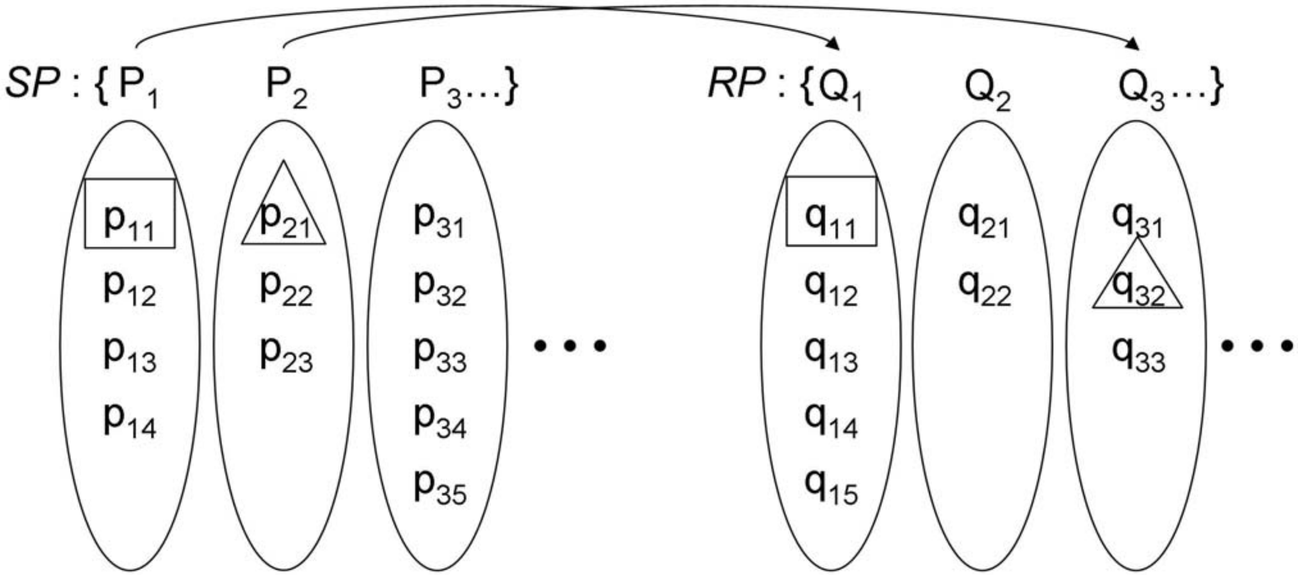
- Little peak distortion
  - Warping only non-peak regions while shifting peak regions
  - Possible distortion only in non-peak regions



- Precise
  - Feature detection (TIC & EIC)
  - Retention time & **mass spectral** information
  - Iterative peak matching: more likely ones matched first

washAlign

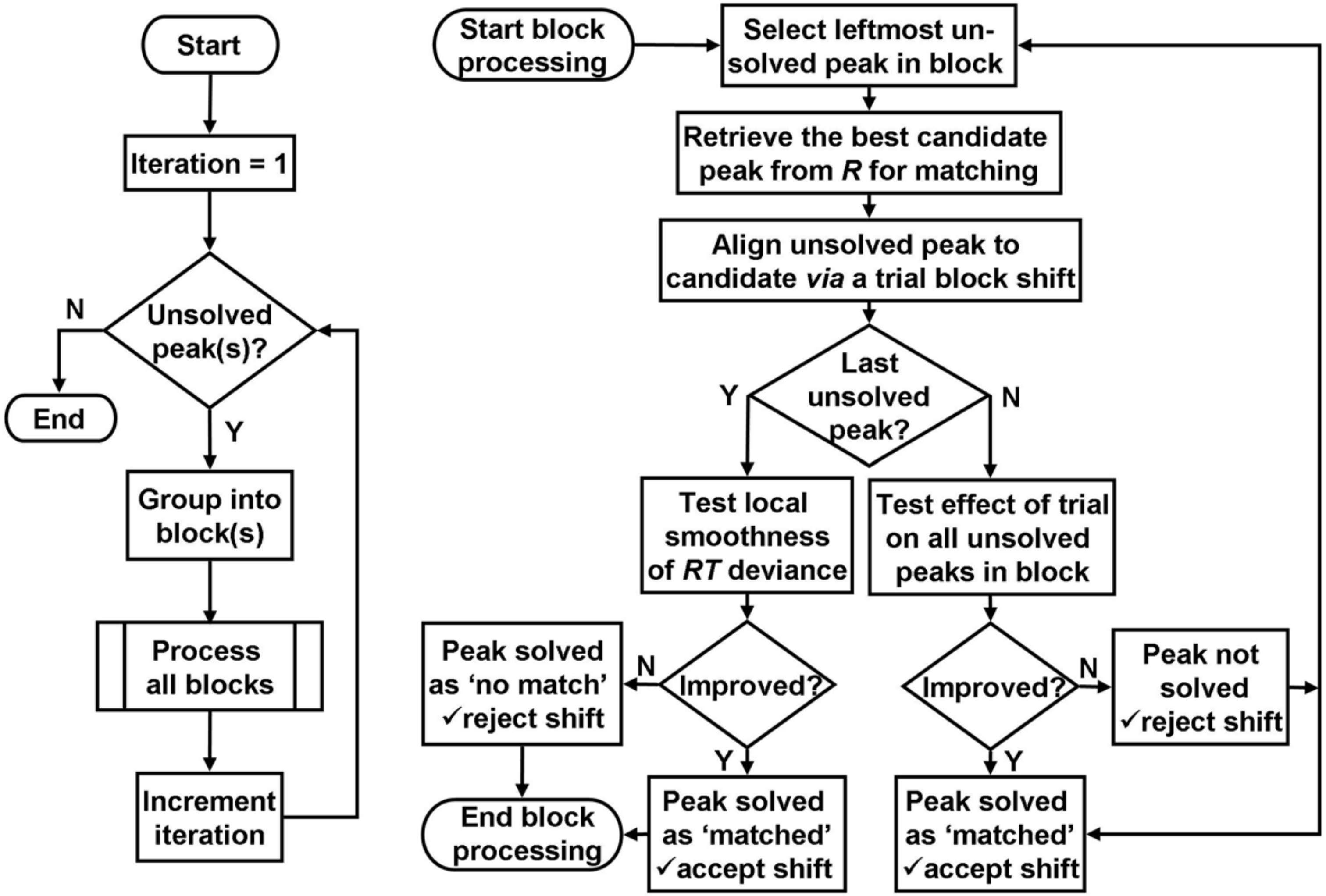
- Pairwise: Sample (*S*) and reference (*R*)
  - Dynamic reference peaks
- Steps:
  - Peak selections → peak matching → waSh
  - Peak matching (TIC vs TIC and EIC vs EIC)
    - Retention time, correlation of mass spectrum, simulation of subsequent peaks



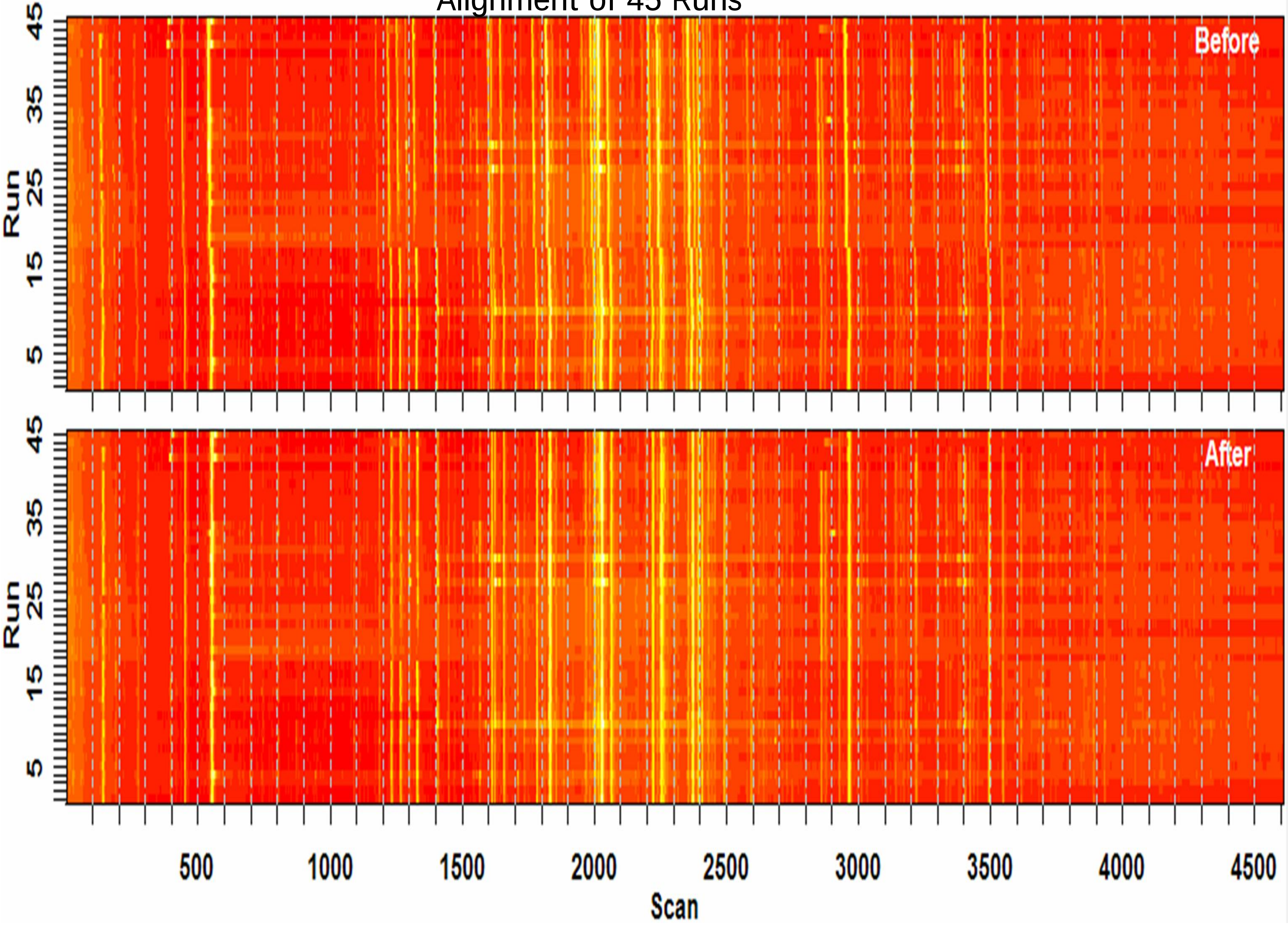
Terms Defined

- Every peak in *S* has a status
  - *Unsolved* : initial, will be tried to find a match
  - *Solved* : decision made on matching, no further trial
    - *Matched*
    - *No-match* found
- *Block*
  - Group of neighboring unsolved peaks
  - All peaks belong to one block, initially, will be broken
  - Smallest block: one peak

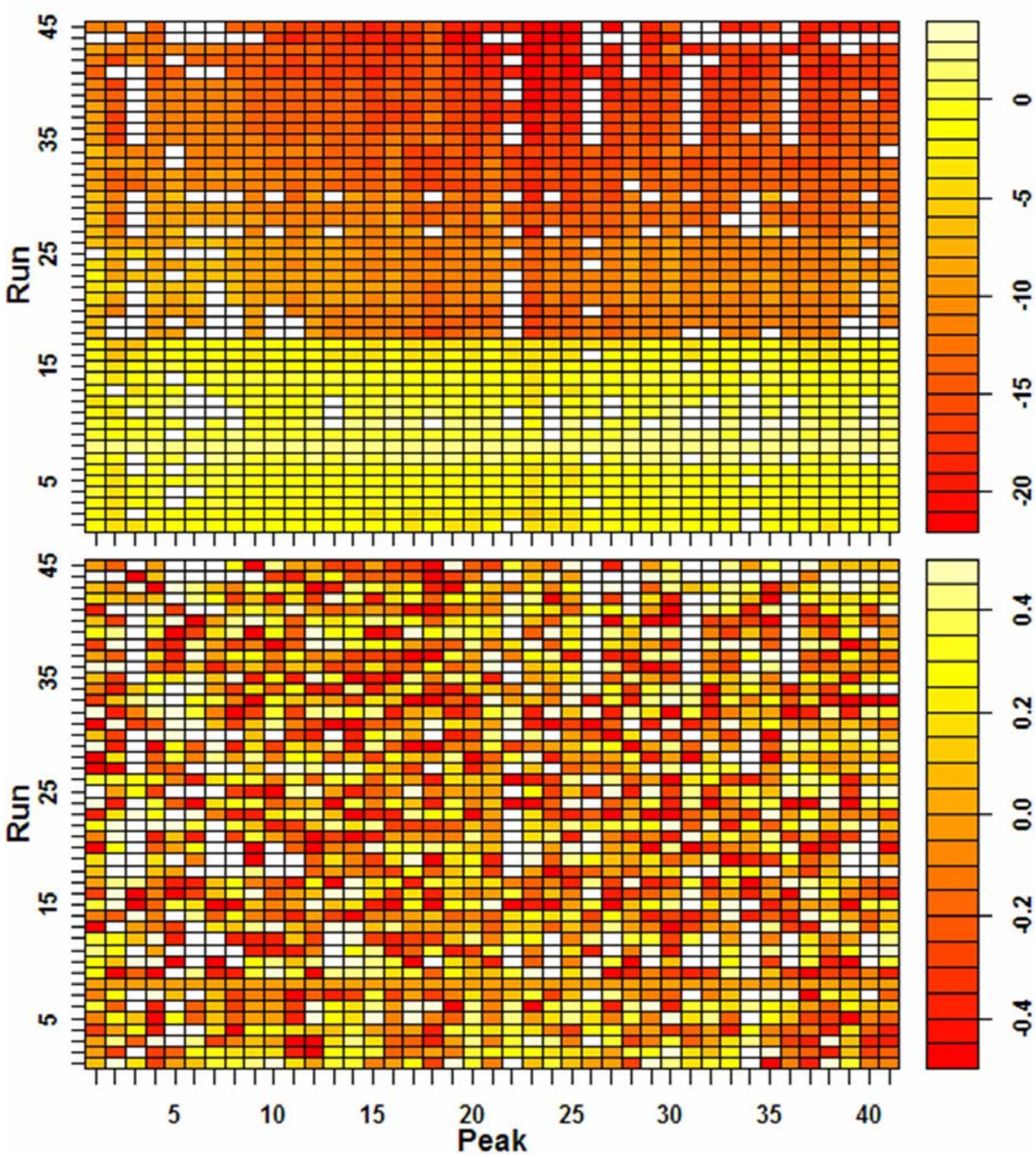
Iterative Peak Matching


Start
Iteration = 1
Unsolved peak(s)?
N
End
Y
Group into block(s)
Process all blocks
Increment iteration
Start block processing
Select leftmost unsolved peak in block
Retrieve the best candidate peak from *R* for matching
Align unsolved peak to candidate *via* a trial block shift
Last unsolved peak?
Y
N
Test local smoothness of *RT* deviance
Test effect of trial on all unsolved peaks in block
Improved?
N
Y
Improved?
N
Y
Peak solved as ‘no match’ ✓reject shift
End block processing
Peak solved as ‘matched’ ✓accept shift
Peak solved as ‘matched’ ✓accept shift
Peak not solved ✓reject shift

Alignment of 45 Runs

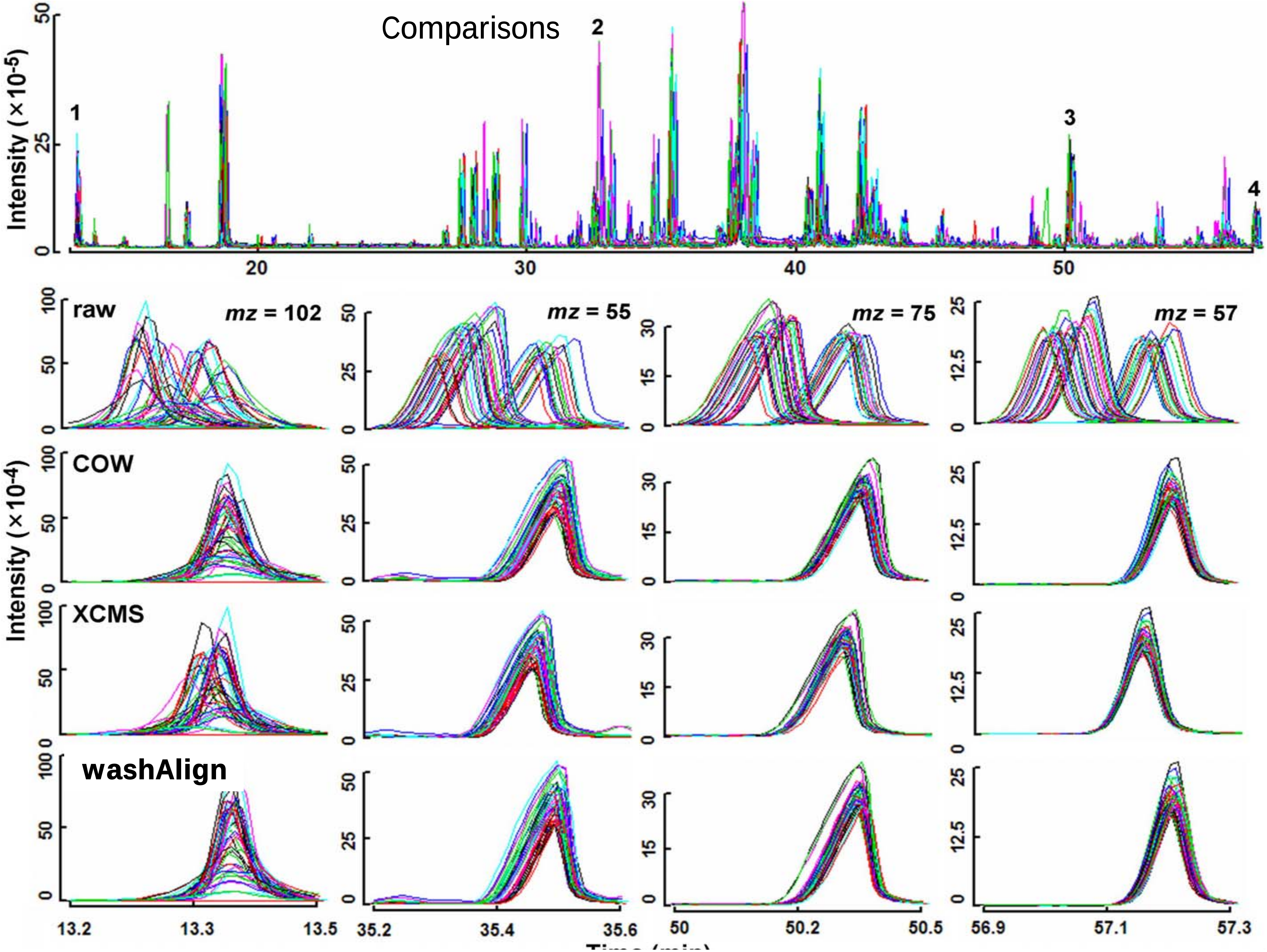


Deviations before and after



Max deviation:
22 scans → less than 1 scan !

Comparisons



Comparison (Cont’d)

Peak integration errors* caused by three alignment methods

	1	***2***	***3***	4
COW area %error ± SD	8.7 ± 5.2	4.7 ± 3.8	3.0 ± 2.4	4.5 ± 3.2
XCMS area %error ± SD	0.17 ± 00.14	1.29 ± 0.91	0.50 ± 0.89	0.11 ± 0.10
washAlgin area %error ± SD	0.000 ± 0.00	0.002 ± 0.01	0.18 ± 0.80	0.000 ± 0.00
washAlign *vs*. COW (*t*-test P val.)	$<10^{-10}$	$<10^{-10}$	$<10^{-10}$	$<10^{-10}$
washAlign *vs*. XCMS(*t*-test P val.)	$<10^{-10}$	$<10^{-10}$	0.08	$<10^{-10}$

*area %error = 100% × $(\text{area}_{\text{aligned}} – \text{area}_{\text{raw}}) / \text{area}_{\text{raw}}$

Demo

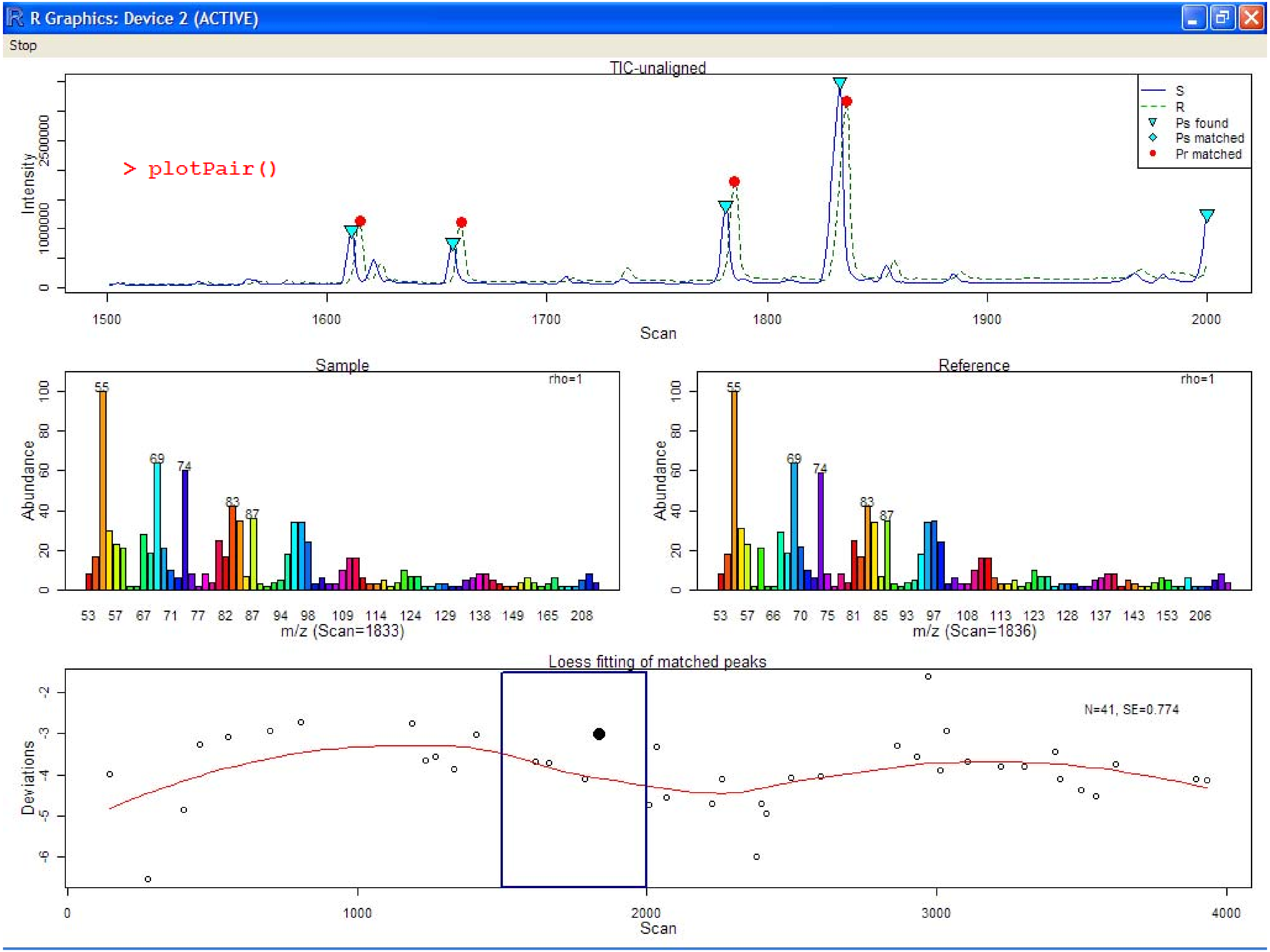
```
> alignResult <- alignOneSample(sFileNo=1, rFileNo=8)
Aligning sample no = 1 ...
Finding TIC peaks for Sam. ...Done!
Finding EIC peaks for Sam. .....................................................Done!
Finding EIC peaks for Ref. ...Done!
Aligning iteratively...Done!
WaShing Tic...Done!

Alignment Summary:
Ref. file: D:\Devs\Align\Worm_GCMS\01040704.CDF
Sam. file: D:\Devs\Align\Worm_GCMS\01030704.CDF
41 out of  44 peaks are matched in 4 iterations.
Mean shift of peaks: 3.97561
Mean peak mass correlation (before): 0.8856085
Mean peak mass correlation (after): 0.99646
Time(s): 65.555
> alignResult
   sTop rTop  mz shift    preCor   postCor
1   139  143  53     4 0.9877841 0.9977636
2   269  276 155     7 0.4524669 0.9816376
3   394  399 192     5 0.4281637 0.9852801
4   450  453   8     3 0.9714484 0.9999439
5   549  552  68     3 0.7278806 0.9996624
6   697  700 155     3 0.8455646 0.9953103
7   802  805 169     3 0.5728226 0.9786473
8  1185 1187 136     3 0.8530645 0.9907060
9  1231 1235 192     4 0.9446699 0.9991874
10 1265 1269 194     4 0.9499426 0.9988961
11 1329 1333 308     4 0.9571284 0.9997278
```

…

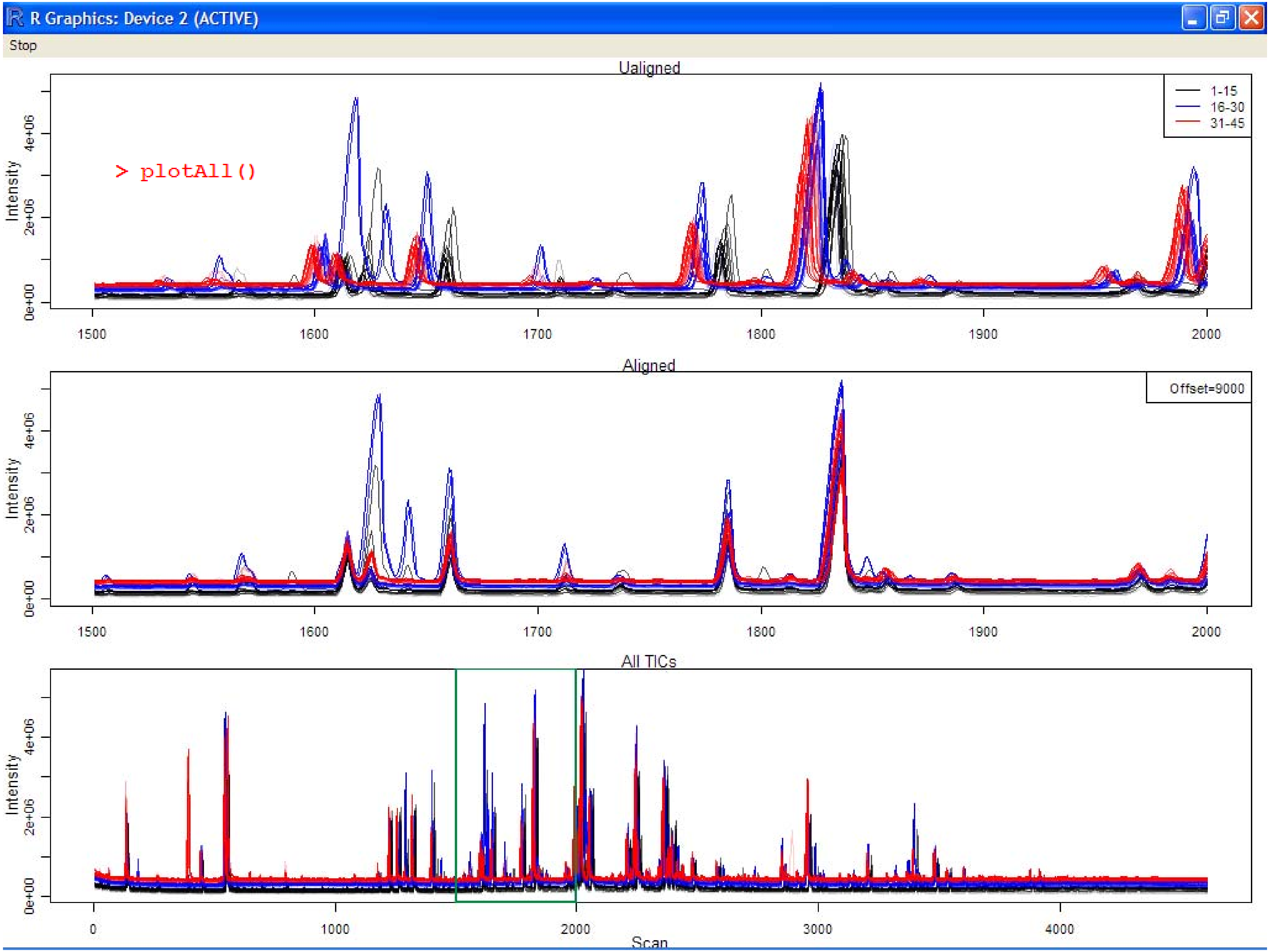
```
> alignedEics <- washAllEics()
WaShing all EICs............................
> alignedEics
        [,1]        [,2]        [,3]
  [1,]    10    10.00000    10.00000
  [2,] 17362 17237.79259 16744.81481
  [3,] 34610 34400.40000 34619.77778
  [4,]  4905  4408.17037  4304.57037
  [5,]   452   333.61481   429.71852
```

…



```
> plotPair()
```

R Graphics: Device 2 (ACTIVE)
Stop
TIC-unaligned
Intensity
0
1000000
2500000
Scan
1500
1600
1700
1800
1900
2000
S
R
Ps found
Ps matched
Pr matched
Sample
rho=1
Abundance
0
20
40
60
80
100
55
69
74
83
87
53 57 67 71 77 82 87 94 98 109 114 124 129 138 149 165 208
m/z (Scan=1833)
Reference
rho=1
Abundance
0
20
40
60
80
100
55
69
74
83
87
53 57 66 70 75 81 85 93 97 108 113 123 128 137 143 153 206
m/z (Scan=1836)
Loess fitting of matched peaks
N=41, SE=0.774
Deviations
-2
-3
-4
-5
-6
Scan
0
1000
2000
3000
4000



R Graphics: Device 2 (ACTIVE)
Stop
Ualigned
1-15
16-30
31-45
Intensity
0e+00
2e+06
4e+06
1500
1600
1700
1800
1900
2000
Aligned
Offset=9000
Intensity
0e+00
2e+06
4e+06
1500
1600
1700
1800
1900
2000
All TICs
Intensity
0e+00
2e+06
4e+06
0
1000
2000
3000
4000
Scan

```
> plotAll()
```

Summary

- washAlign
  - Precise alignment with minimal peak distortion
  - Interactive visual checking
- Plans
  - Improved packaging: S4 conversion
    - Maintenance
    - Easy use
  - Speed, i.e., peak detections
- More information
  - Chae M, Shmookler Reis RJ, Thaden JJ:
    BMC Bioinformatics 2008, 9(Suppl 9):S15

Acknowledgement

Supported by NIH #P20 RR-16460

Dr. Robert Reis

Dr. John Thaden

Dr. Steven Jennings

Dr. Chan-Hee Jo

Dr. Lulu Xu

Bill Starrett

R developers and users!