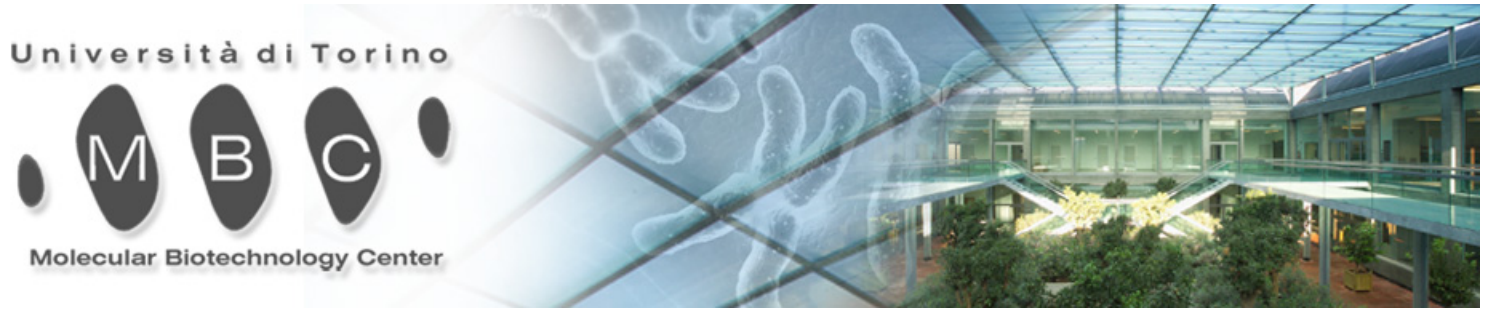


Chimera

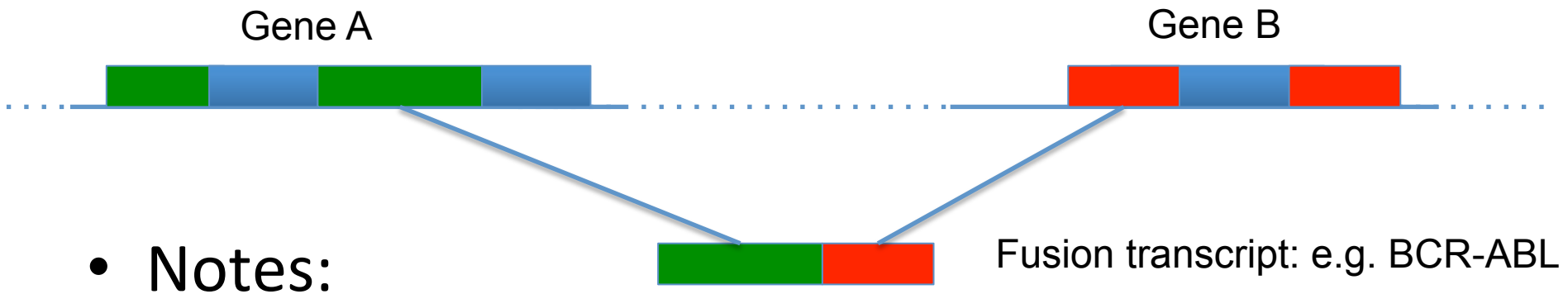
A package for secondary analysis of fusion products

Raffaele A Calogero

raffaele.calogero@unito.it

Transcription-induced chimeras

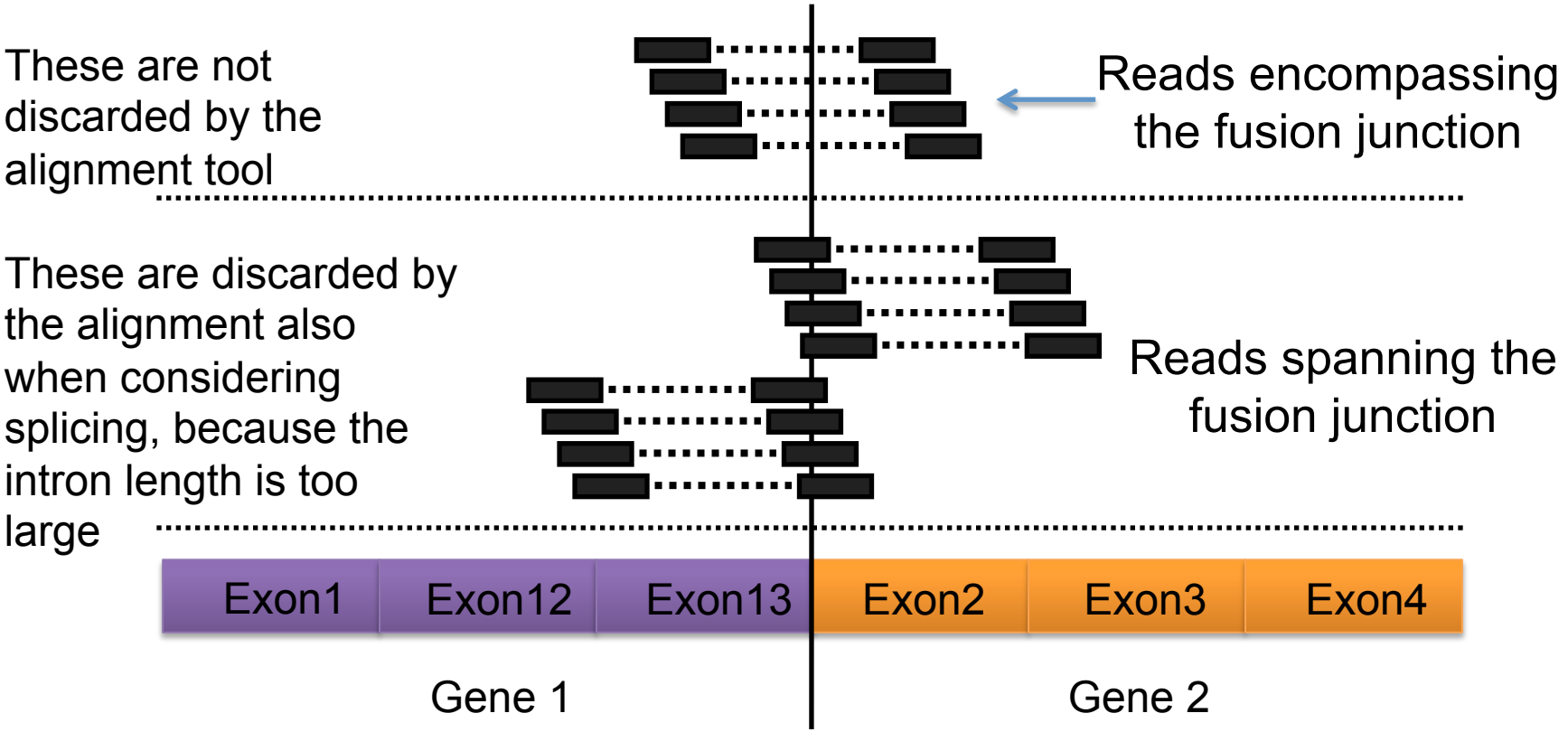
- RNA-seq has the potential to discover genes created by complex chromosomal rearrangements:
  - 'Fusion' genes formed by the breakage and re-joining of two different chromosomes have repeatedly been implicated in the development of cancer.



- Notes:
  - Fusion may not happen at exon boundaries
  - Non-canonical junctions must be considered

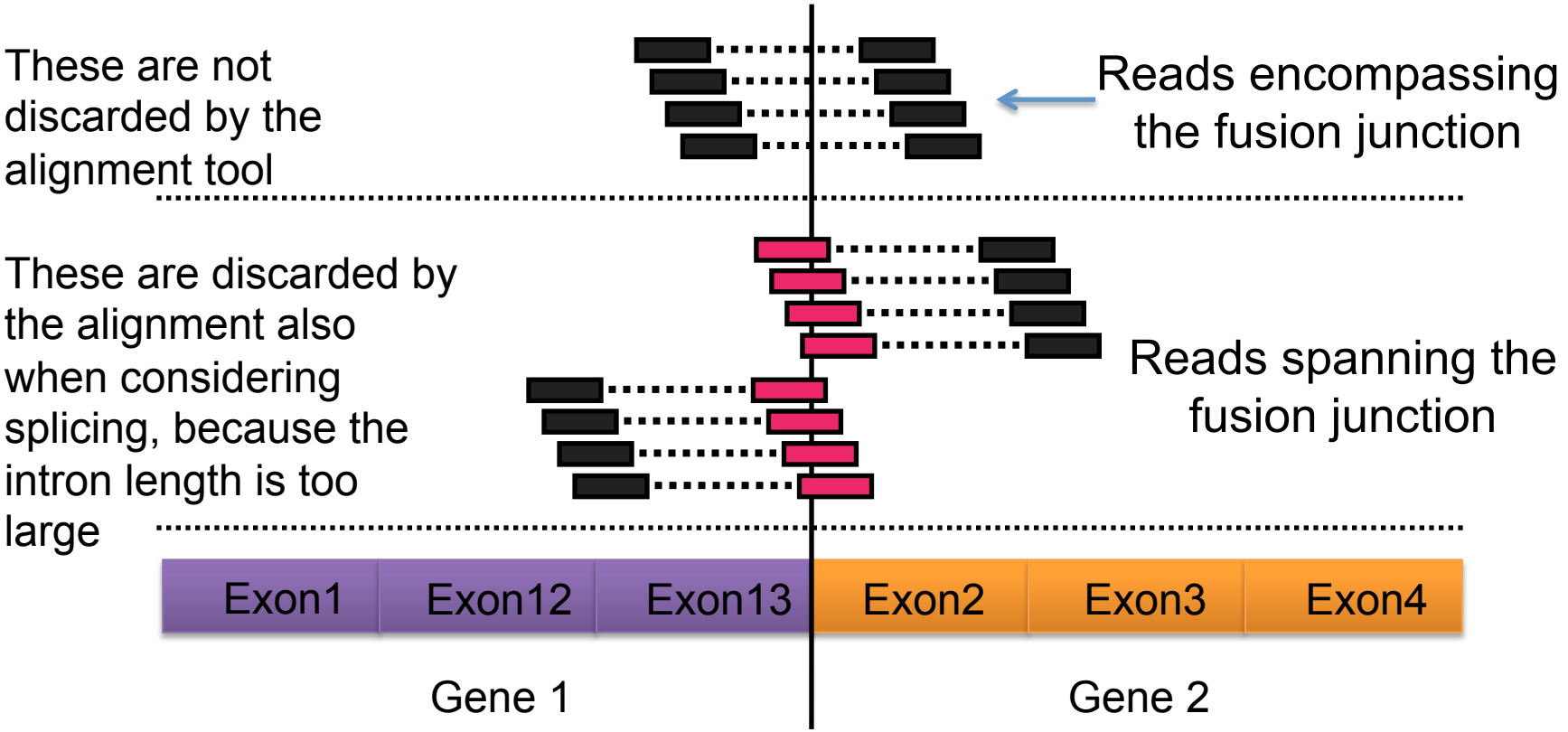
Chimeric Transcript Detection

Approach proposed by Maher et al., 2009



Chimeric Transcript Detection

Approach proposed by Maher et al., 2009



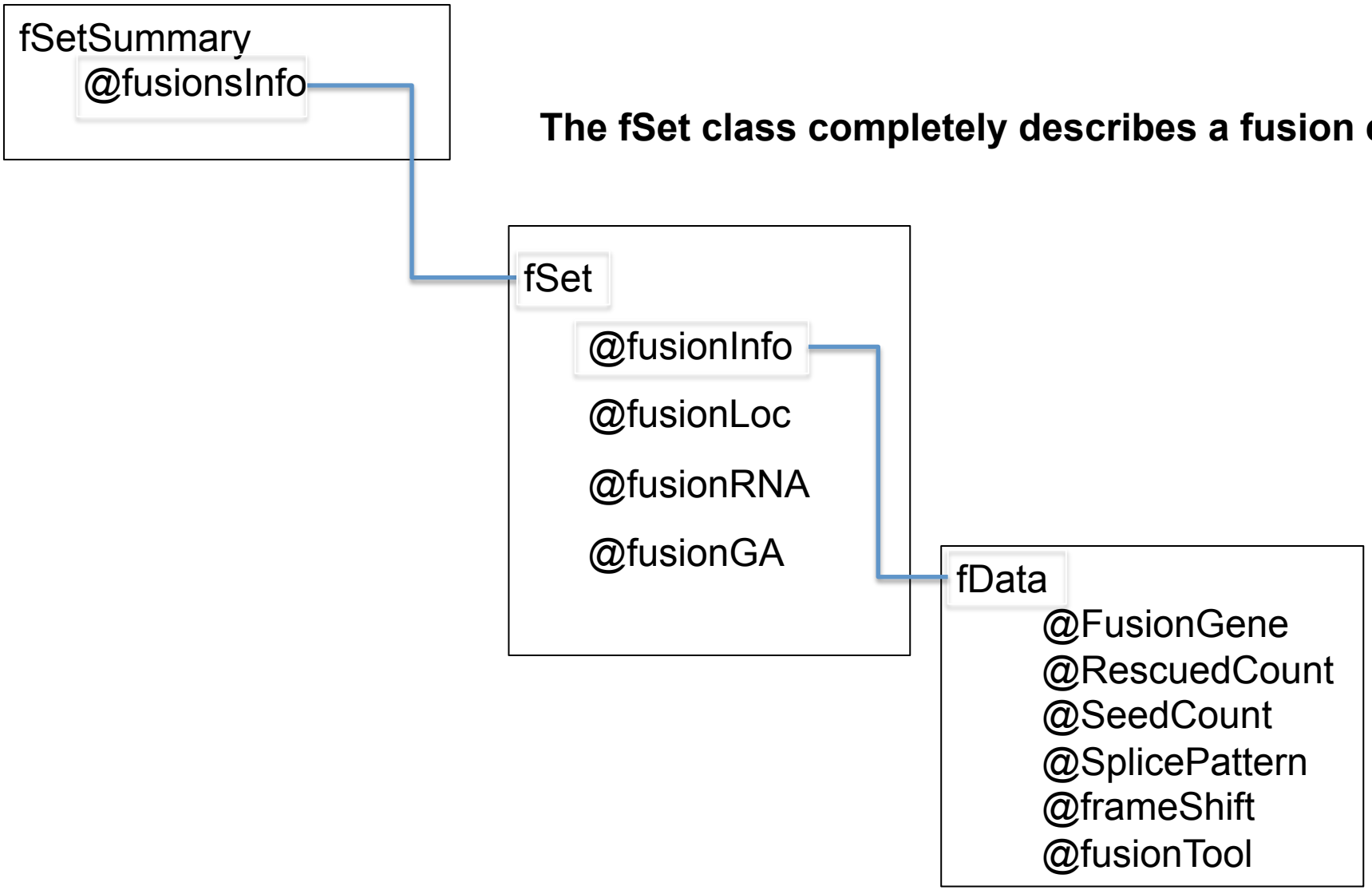
Fusion detection tools

- Recently have been presented a quite large number of fusion detection tools:
  - MapSplice (2010)
  - FusionHunter (2011)
  - Defuse (2011)
  - FusionMap (2011)
  - TopHat-fusion (2011)
  - FusionFinder (2012)
  - Bellerophontes (2012)
  - ShortFuse (2011)
  - ChimeraScan (2011)
  - EricScript (2012)
  - FusionCatcher (2012)

Chimera package

- On the basis of our knowledge no tools are available to manipulate the output generated by fusion finders.
- Each tool produces its own type of output.
- Output generated by fusion finders do not follow any standard structure.
- **chimera** is a package for downstream processing of fusion events.

Classes



The fSet class completely describes a fusion event

```
> showClass("fData")
Class "fData" [package "chimera”]
Slots:
Name:          fusionTool    UniqueCuttingPositionCount          SeedCount
Class:          character    numeric                             numeric

Name:          RescuedCount          SplicePattern          FusionGene
Class:          numeric               character              character

Name:          frameShift
Class:          character
```

```
An object of class "fData"
Slot "fusionTool":
[1] "FusionMap"

Slot "UniqueCuttingPositionCount":
[1] 14

Slot "SeedCount":
[1] 2

Slot "RescuedCount":
[1] 18

Slot "SplicePattern":
[1] "CT-AC"

Slot "FusionGene":
[1] "uc002xtx.4->uc002xto.3,uc002xtr.3,uc002xtq.3,uc010ghv.1"

Slot "frameShift":
[1] "0->1"
```

```
> showClass("fSet")
Class "fSet" [package "chimera”]
Slots:
Name:       fusionInfo       fusionLoc       fusionRNA       fusionGA
Class:       fData       GRangesList     DNAStringSet    GappedAlignments
```

```
Slot "fusionLoc":
GRangesList of length 2:
$gene1
GRanges with 1 range and 5 elementMetadata cols:
      seqnames               ranges strand |   KnownGene                                  KnownTranscript
         <Rle>            <IRanges>  <Rle> | <character>                                      <character>
  [1]    chr20 [46365656, 46365686]      + |       SULF2 uc002xto.3,uc002xtr.3,uc002xtq.3,uc010ghv.1
      KnownExonNumber KnownTranscriptStrand        FusionJunctionSequence
          <character>          <character>                    <character>
  [1]         3,3,3,3                 ---- GCCGGGTCTTGTTCATCACCTGCATGGAAC

$gene2
GRanges with 1 range and 5 elementMetadata cols:
      seqnames               ranges strand | KnownGene KnownTranscript KnownExonNumber
  [1]    chr20 [47538547, 47538577]      - |   ARFGEF2      uc002xtx.4               1
      KnownTranscriptStrand         FusionJunctionSequence
  [1]                     + cgagcgccacctggcaggccctgcgcagct
```

> showClass("fSetSummary")

Class "fSetSummary" [package "chimera"]
Slots:
Name: fusionsInfo
Class: list

Import functions

1. fmImport: FusionMap
2. fhImport: FusionHunter
3. ffImport: FusionFinder
4. dfImport: deFuse
5. msImport: MapSplice
6. bfImport: bellerophontes
7. thfImport: TopHat-fusion

- csImport: ChimeraScan
- sfImport: ShortFuse

Unique identifier for the fusion

- Each tool uses a different annotation resource.
- To obtain a unique identifier for the fusion.
  - In the import procedure:
    - chromosome location for the fused genes are associated to their gene symbols using the chromosome coordinates available in org.Hs.eg.db
  - Fusion identifier is in the format:
    - Symbol1:Symbol2
  - In case chomosomal location does not associate to a know gene the formats for the fusion are:
    - Symbol1:CHR:acceptorStart-acceptorEnd
    - CHR:donorStart-donorEnd:Symbol2
    - CHR:donorStart-donorEnd:CHR:acceptorStart-acceptorEnd

Methods fSetSummary

- show(fSetSummary)
- fset(fSetSummary, num)
- fusionsInfo(fSetSummary)
- fusionsGA(fSetSummary)
- supportingReads(fSetSummary)
- fusionName(fSetSummary)
- fusionJ(fSetSummary, fusion.name)
- extractFusion(fSetSummary, fusion.name)
- subsetSummary(fSetSummary, n)
- filterSummary(fSetSummary, type=c("supporting.reads","fusion.names"), query)

Methods fSet

- Extracting information:
  - fusionData (fSet)
  - fusionGRL (fSet)
  - fusionRNA (fSet)
  - fusionGA(fSet)

- Adding information:
  - addRNA(fSet, rna)
  - addGA(fSet, bam)

functions

- Functions modifying fSet object:
  - chimeraSeqs:

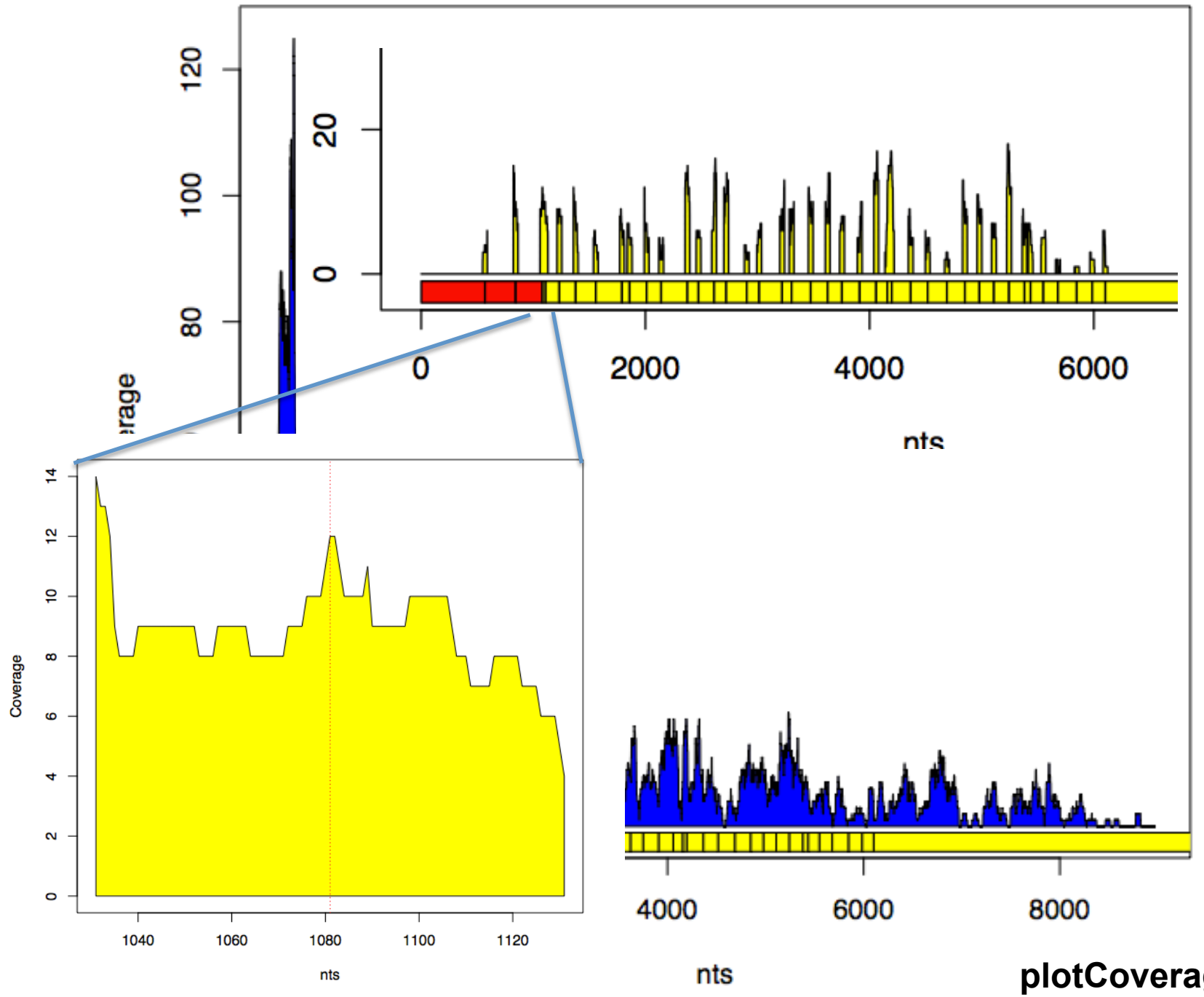
```
Slot "fusionRNA":
  A DNAStringSet instance of length 7
    width seq                                                                names
[1]  9998 CACTTGAGCCGAAGTTAATTTCTCGGGGAGTTCTCGG...TTGATCAGGTGGTACATCAATAAAATTTTAAAAAGTA ENST00000359930:E...
[2]  9815 GGGCCATTTCTGGACAACAGCTGCTATTTTCACTTGA...TTGATCAGGTGGTACATCAATAAAATTTTAAAAAGTA ENST00000484875:E...
[3]  9238 CTCGGGCGCGCACAGGCAGCTCGGTTTGCCCTGCGAT...TTGATCAGGTGGTACATCAATAAAATTTTAAAAAGTA ENST00000467815:E...
```

  - tophatRun:

```
Slot "fusionGA":
GappedAlignments with 11131 alignments and 0 elementMetadata cols:
                                seqnames strand       cigar    qwidth     start       end     width
                                   <Rle>  <Rle> <character> <integer> <integer> <integer> <integer>
     [1] ENST00000359930:ENST00000371917      +         50M        50        59       108        50
     [2] ENST00000359930:ENST00000371917      +         50M        50        90       139        50
     [3] ENST00000359930:ENST00000371917      -         50M        50       132       181        50
```

functions

- Functions describing a fusion given a fSet object:
  - plotCoverage
  - fusionPeptides

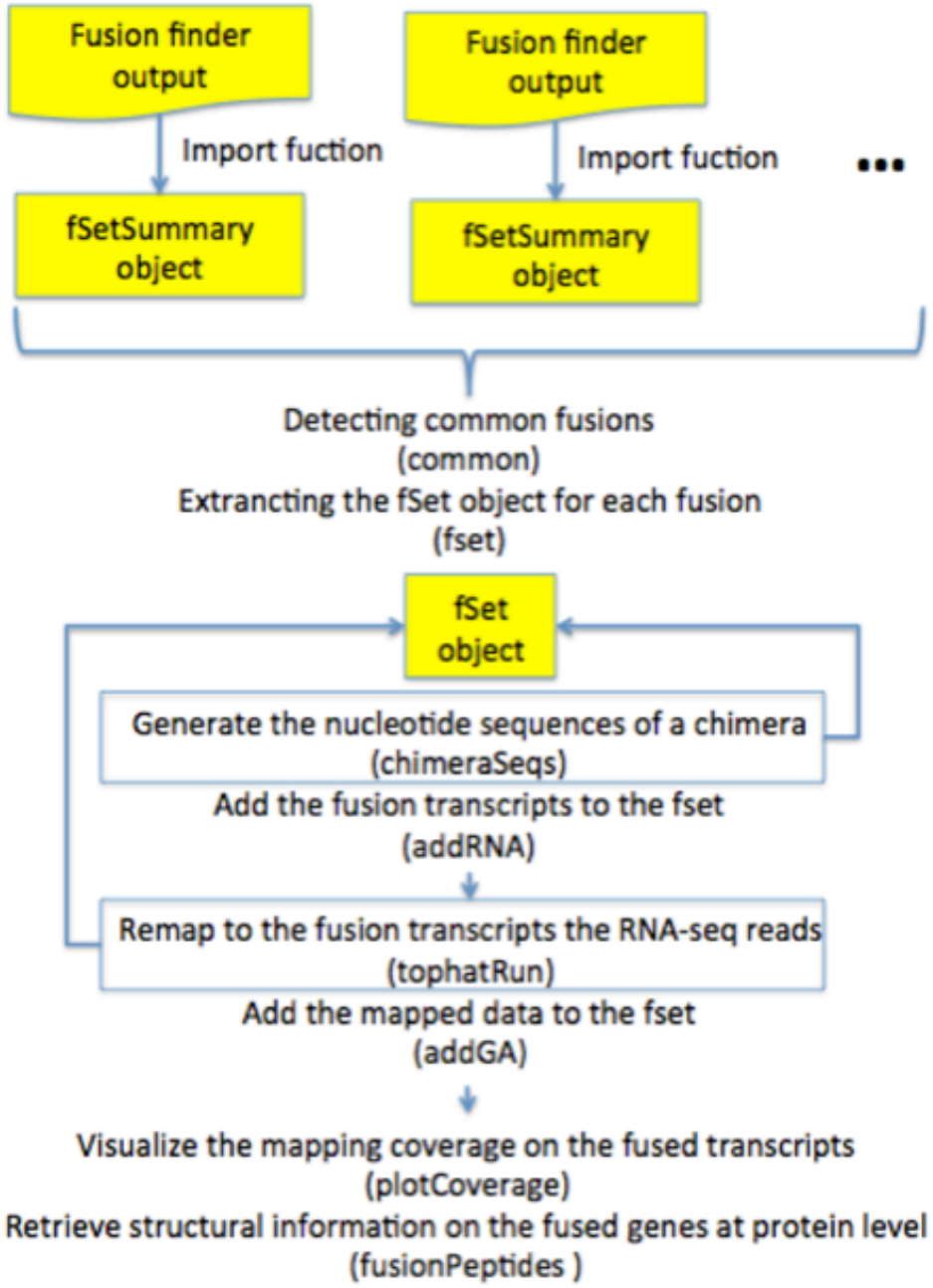


plotCoverage output

fusionPeptides output

$transcript1
  206-letter "DNAString" instance
seq: ATGGGCCCCCCGAGCCTCGTGCTGTGCTTGCTGTCCGCAACTGTGT...
$pep1
  68-letter "AAString" instance
seq: MGPPSLVLCLLSATVFSLLGGSSAFLSHHRLKGRFQRDRRNIR…
$frame.pep1
[1] 3
…
$validation.seq
[1]
"CCCCCCGAGCCTCGTGCTGTGCTTGCTGTCCGCAACTGTGTTCTCCCTGCTGG
GTGGAAGCTCGGCCTTCCTGTCGCACCACCGCCTGAAAGGCAGGTTTCAGAGG
GACCGCAGGAACATCCGCCCCAAC...TCGCATACGGGCACATCACTGGCAACGC
CCCTGACAG”
$junction.ga
GappedAlignments with 5 alignments and 0 elementMetadata cols:

	seqnames	strand	cigar	qwidth	start	end
[1]	ENST00000467815:ENST00000371917	-	50M	50	162	211
[2]	ENST00000467815:ENST00000371917	+	50M	50	181	230
[3]	ENST00000467815:ENST00000371917	+	50M	50	182	231
[4]	ENST00000467815:ENST00000371917	-	50M	50	187	236
[5]	ENST00000467815:ENST00000371917	-	50M	50	190	239

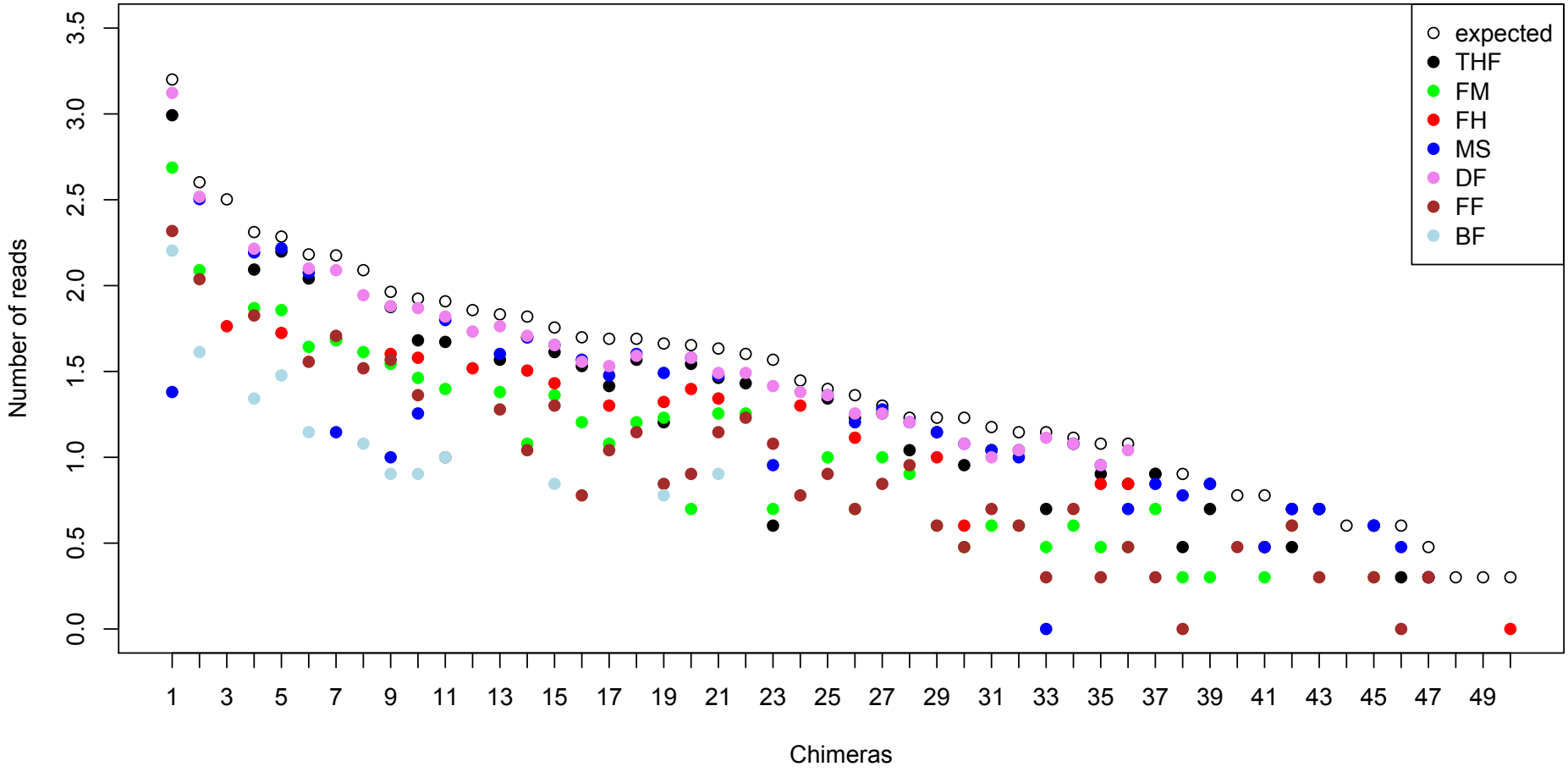

Fusion finder output
Fusion finder output
Import fuction
Import fuction
fSetSummary object
fSetSummary object
Detecting common fusions
(common)
Extrancting the fSet object for each fusion
(fset)
fSet object
Generate the nucleotide sequences of a chimera
(chimeraSeqs)
Add the fusion transcripts to the fset
(addRNA)
Remap to the fusion transcripts the RNA-seq reads
(tophatRun)
Add the mapped data to the fset
(addGA)
Visualize the mapping coverage on the fused transcripts
(plotCoverage)
Retrieve structural information on the fused genes at protein level
(fusionPeptides)

Why developing chimera?

- We are involved in a project on Leukemia biomarkers detection and we need to efficiently identify fusion products.
- Questions?
  - There is a fusion finder tool characterized by the highest sensitivity and specificity?
  - Are false positive critical?

Testing sensitivity of fusion detection tools

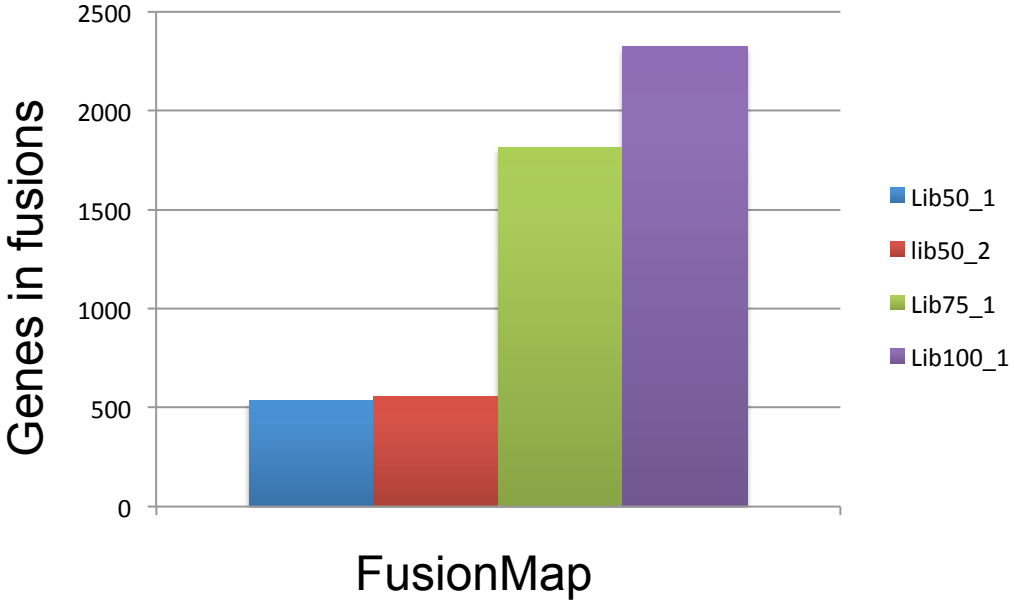
- FusionMap developers provide a synthetic dataset of simulated paired-end RNA-Seq reads (~60,000 pairs of reads, 75nt, fragment size=158bp).
- 50 fusions are represented with a range of supporting pairs going from by 9 to 8852.

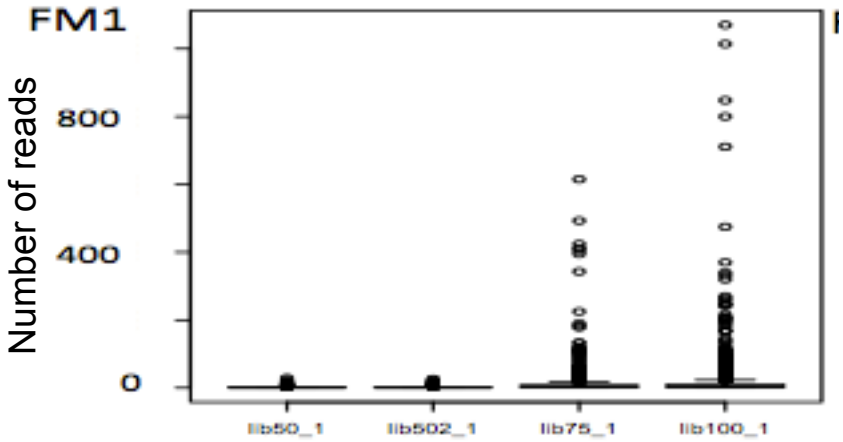


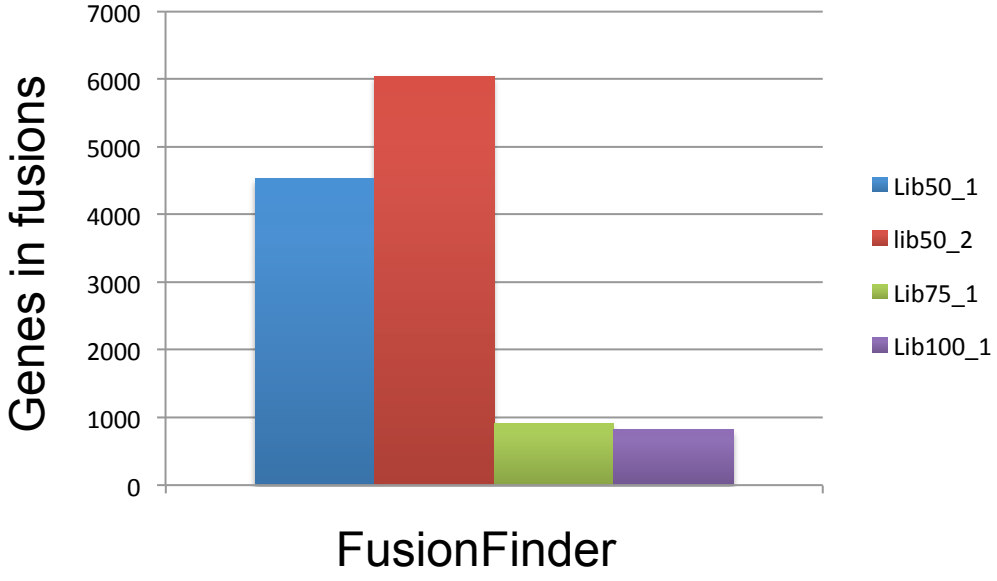
- **deFuse** shows the best correlation with the expected reads but as FusionHunters looses nearly all the fusions supported by less than 18 reads

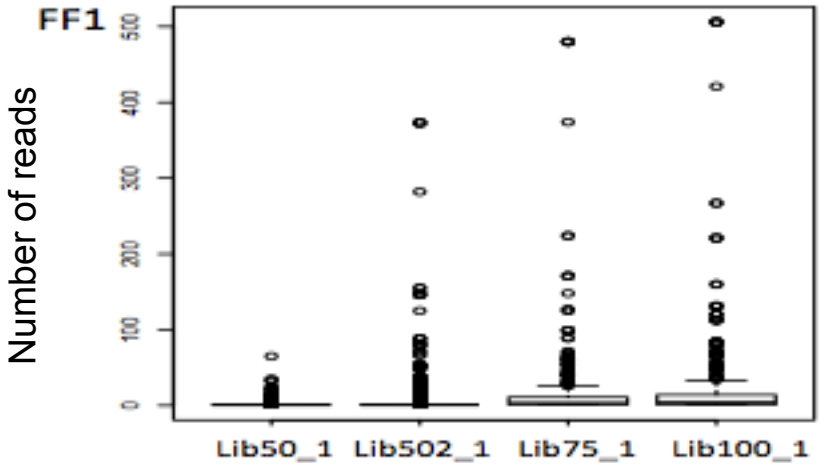
Testing specificy of fusion detection tools

- A fusion-free data set (lib100.fa) was made of 70 million 100 paired-end reads (BEERS software).
- 70 million reads the quality scores derived from two 2 x 100 nts paired-end read experiment run in our laboratory to generate lib100_1 and lib100_2 fastq files.
- From the 100 paired-end reads a set we generated:
  - 2 x 75 nts (lib75_1 and lib75_2)
  - 2 x 50 nts paired-end reads (lib50_1 and lib50_2)

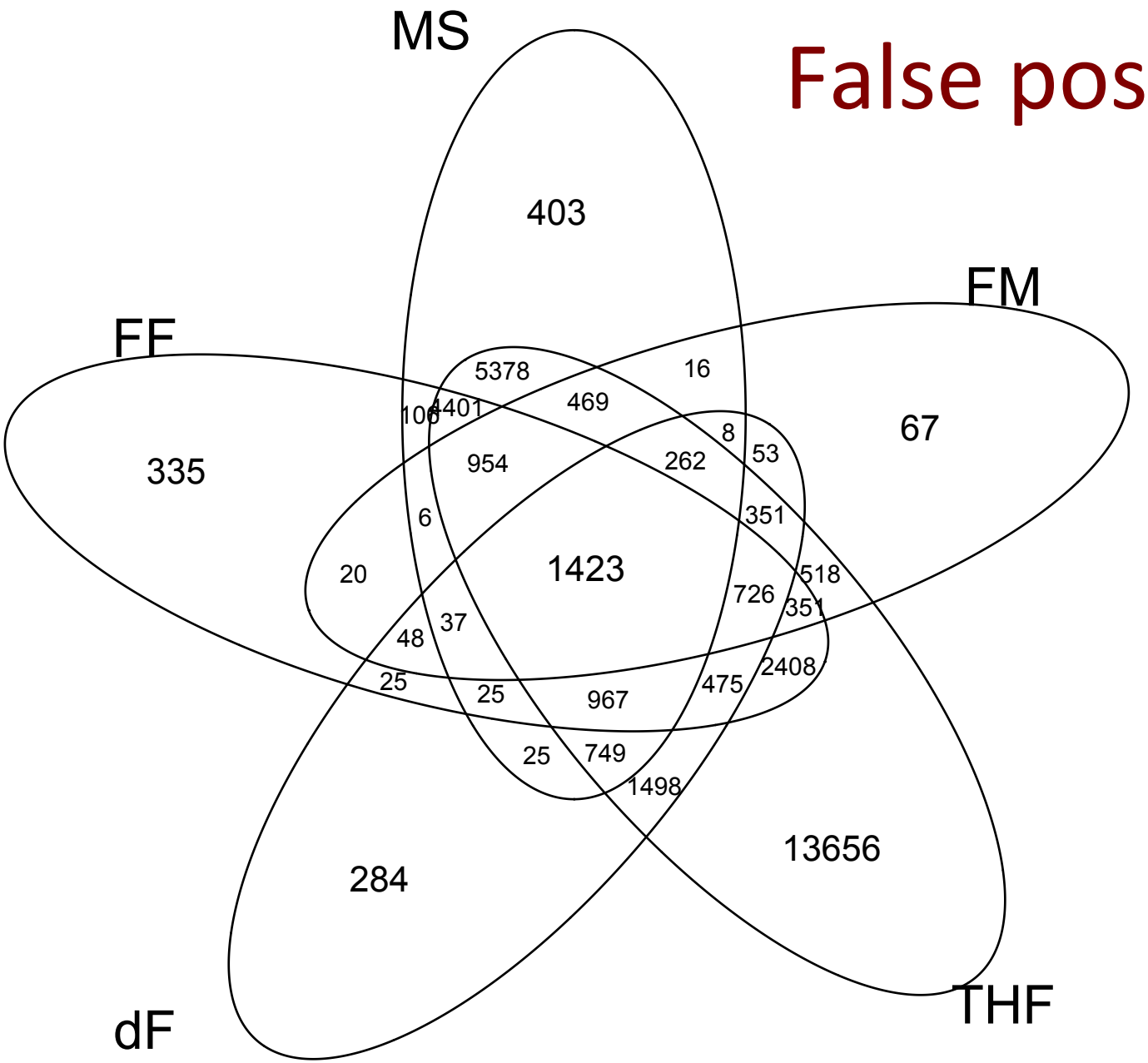

Genes in fusions
2500
2000
1500
1000
500
0
FusionMap
Lib50_1
lib50_2
Lib75_1
Lib100_1

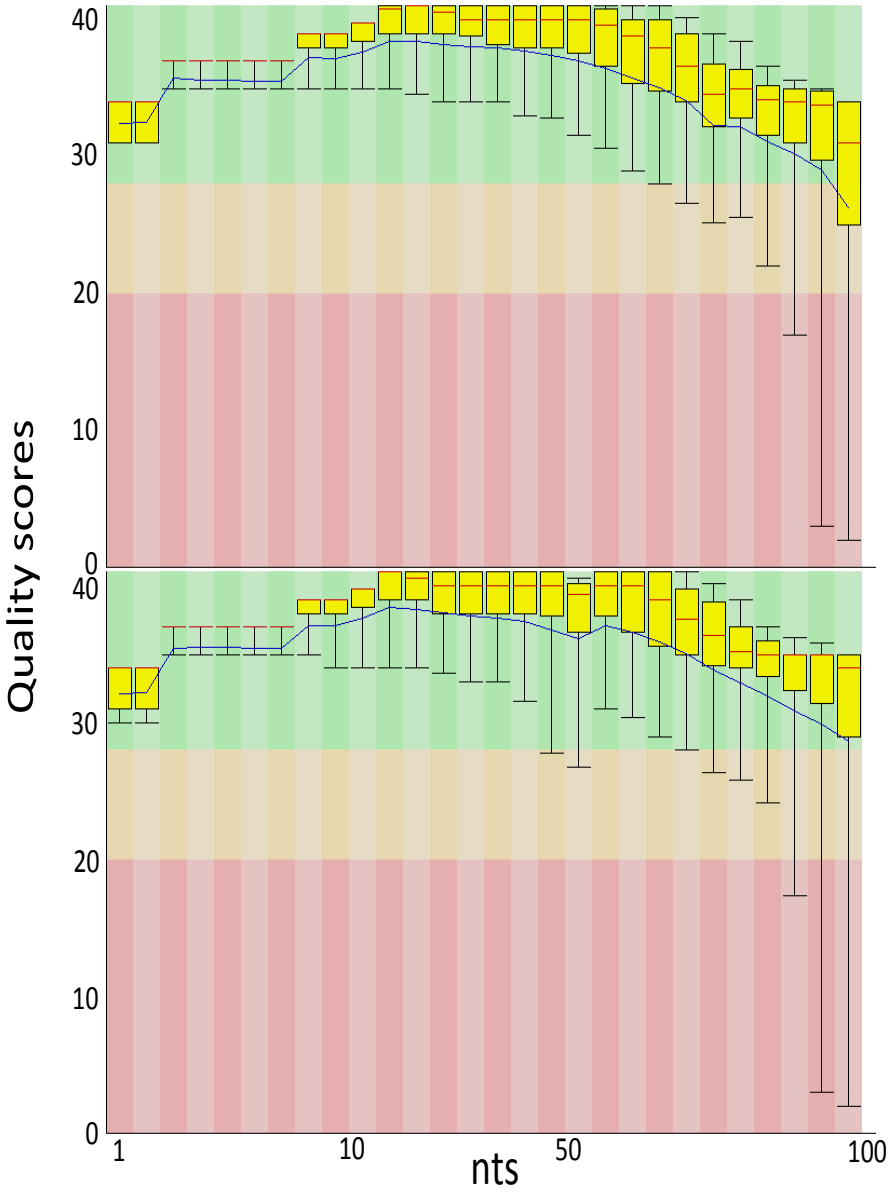

FM1
Number of reads
800
400
0
lib50_1
lib502_1
lib75_1
lib100_1

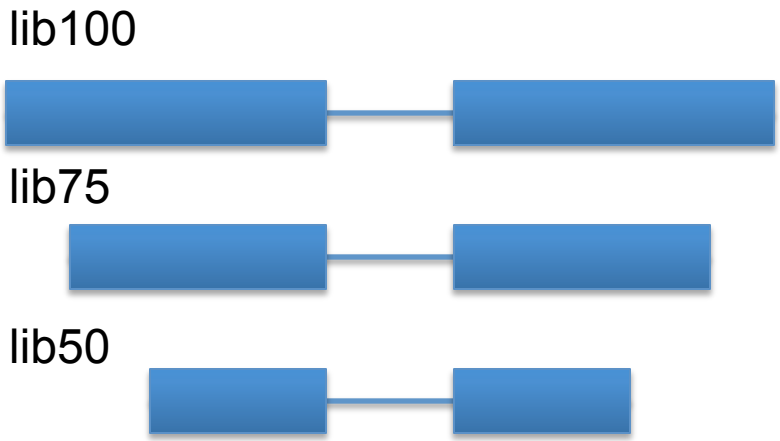

Genes in fusions
7000
6000
5000
4000
3000
2000
1000
0
FusionFinder
Lib50_1
lib50_2
Lib75_1
Lib100_1


FF1
Number of reads
500
400
300
200
100
0
Lib50_1
Lib502_1
Lib75_1
Lib100_1

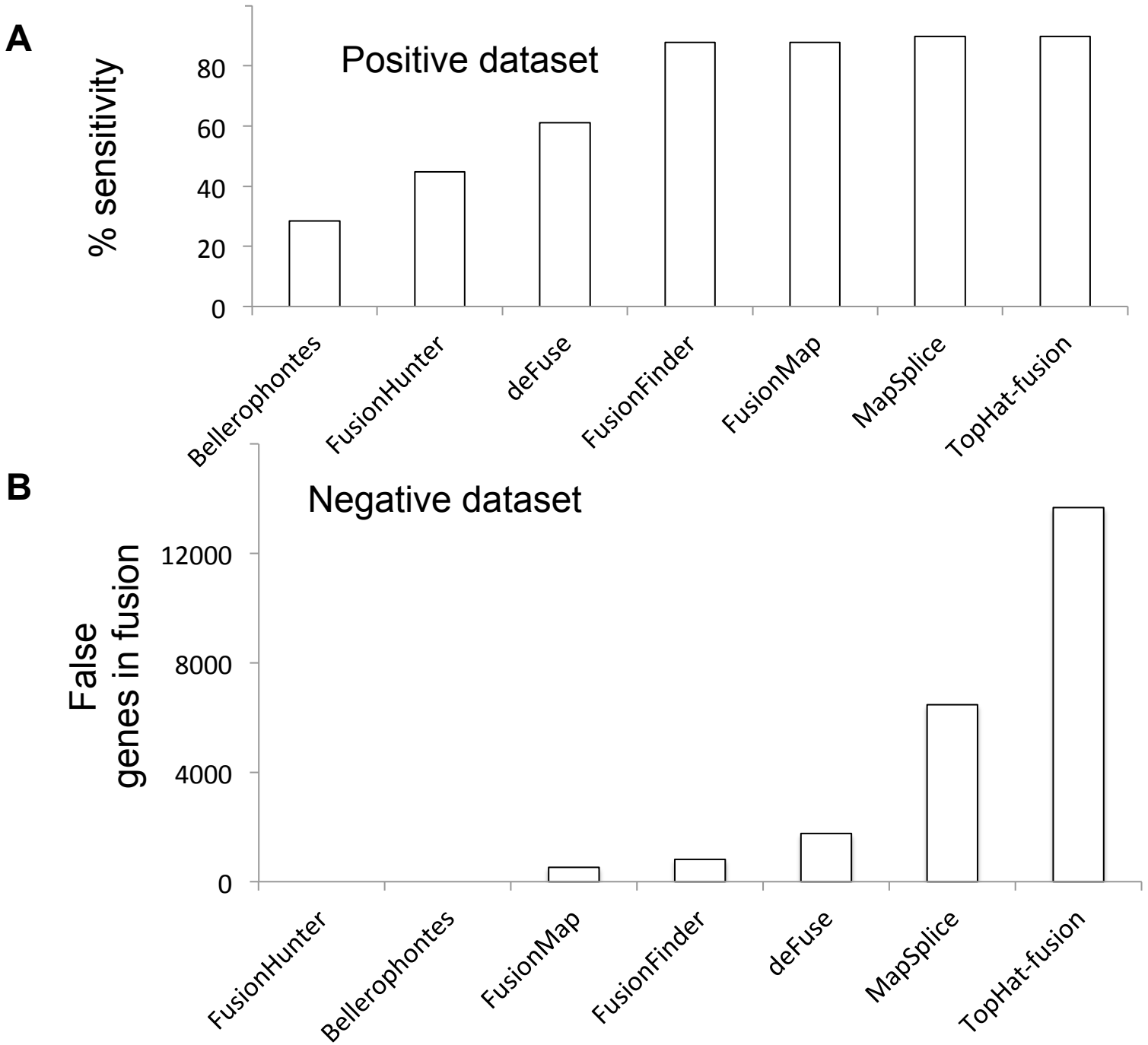
False positives

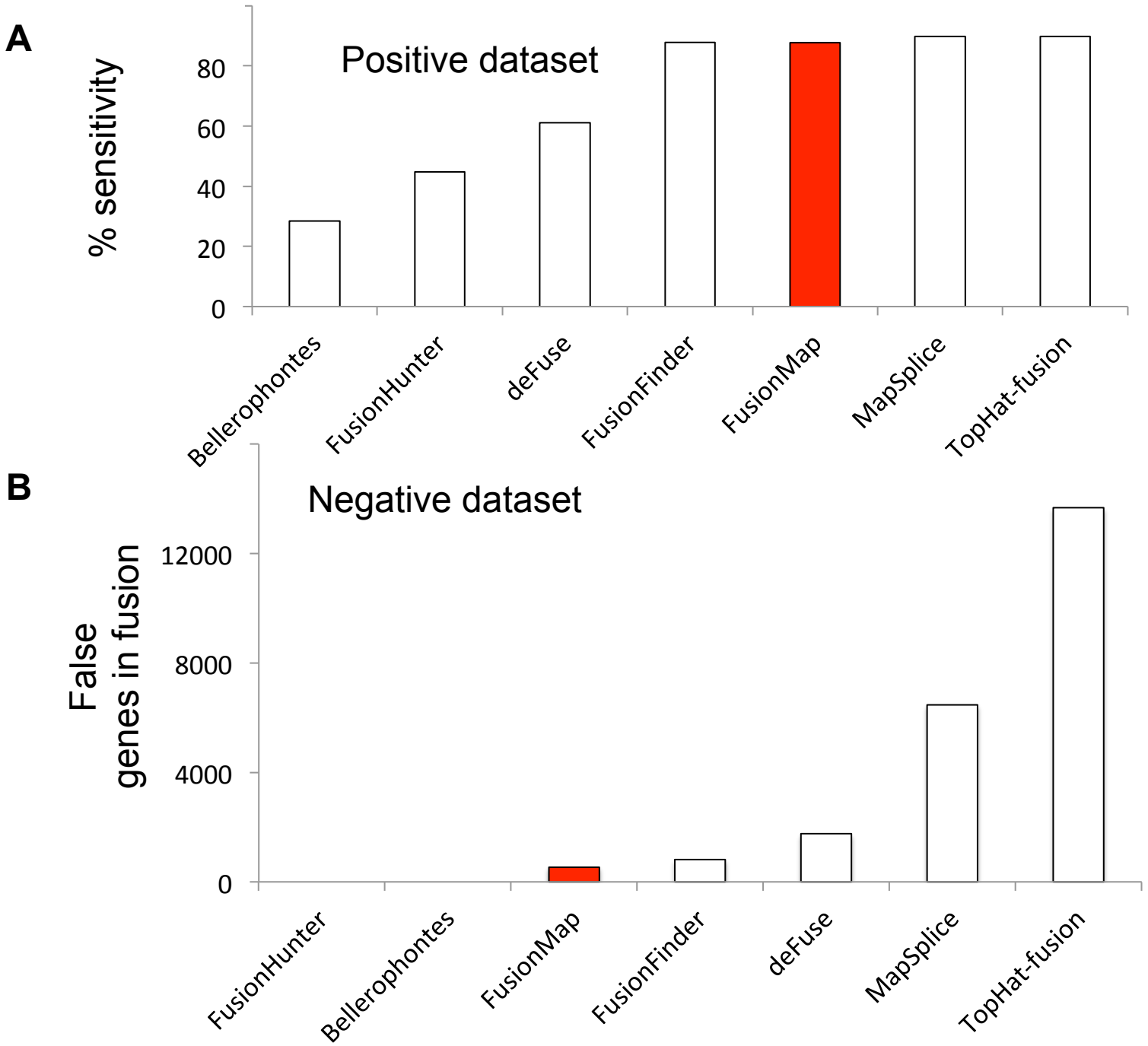



Quality scores
40
30
20
10
0
40
30
20
10
0
1
10
50
100
nts


lib100
lib75
lib50

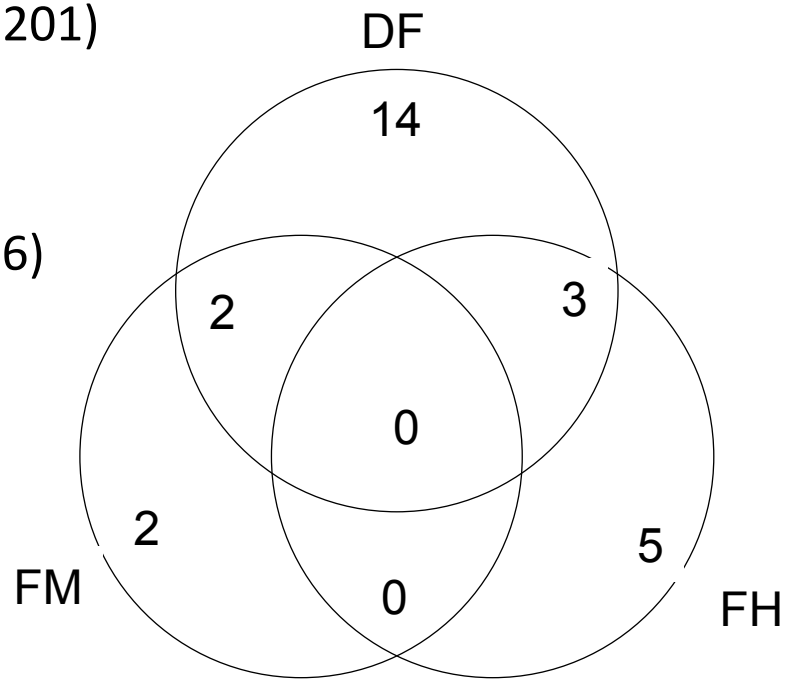

lib100
lib75
lib50


A
Positive dataset
% sensitivity
0
20
40
60
80
Bellerophontes
FusionHunter
deFuse
FusionFinder
FusionMap
MapSplice
TopHat-fusion
B
Negative dataset
False genes in fusion
0
4000
8000
12000
FusionHunter
Bellerophontes
FusionMap
FusionFinder
deFuse
MapSplice
TopHat-fusion


A
Positive dataset
% sensitivity
0
20
40
60
80
Bellerophontes
FusionHunter
deFuse
FusionFinder
FusionMap
MapSplice
TopHat-fusion
B
Negative dataset
False genes in fusion
0
4000
8000
12000
FusionHunter
Bellerophontes
FusionMap
FusionFinder
deFuse
MapSplice
TopHat-fusion

Edgren et al. Genome Biology 2011, 12:R6

- In Edgren paper a total of 27 fusions, validated experimentally in four breast cancer cell lines (MCF-7, KPL-4, SK-BR-3, BT-474) are described.
  - TopHat-fusion on-going 19 out 27 (301928)
  - FusionFinder detects 13 out of 27 (2201)
  - deFuse detects 19 out of 27 (915)
  - FusionMap detects 4 out of 27 (69)
  - FusionHunter detects 8 out of 27 (26)
  - Bellerophontes on-going
  - MapSplice on-going



False positive

- An important issue is the reduction of false positive.
- To maximize the true fusion detection in pathological samples results from different tools need to be combined.
- Question:
  - A collection of fusions detected in normal tissues might be useful to remove non-pathological chimera?

FusionMap example

- FusionMap detects 69 fusions in Edgren's dataset:
  - Only 4 are part of the 27 valided fusion.
- BodyMap 2.0 was used as source of fusions of normal tissues.
- BodyMap 2.0:
  - 16 normal human tissues sequenced PE 50 nts
  - 70 milions reads each tissue.
  - 299 fusions detected by FM in the 16 tissues
- Filtering out "normal fusion" from "pathological fusions":
  - 69 ➔ 53 (77%)
- We are collecting RNA-seq from normal data samples to enlarge the collection of "normal fusions" to be used as filtering instrument.
  - Fusions detected in normal tissues will be organized in an experimental package

Conclusions

- The tested fusion detection tools are far to be efficient.
- Results are affected by various extent by read length and quality
- Specificity issue cannot be solved by a simple intersection of results generated by different methods.
- We continue testing new tools…

Matteo Carrara

Susanna Donatelli
Francesca Cordero
Marco Beccuti

Thank you!

raffaele.calogero@unito.it

Bioinformatics wp

grant 306242