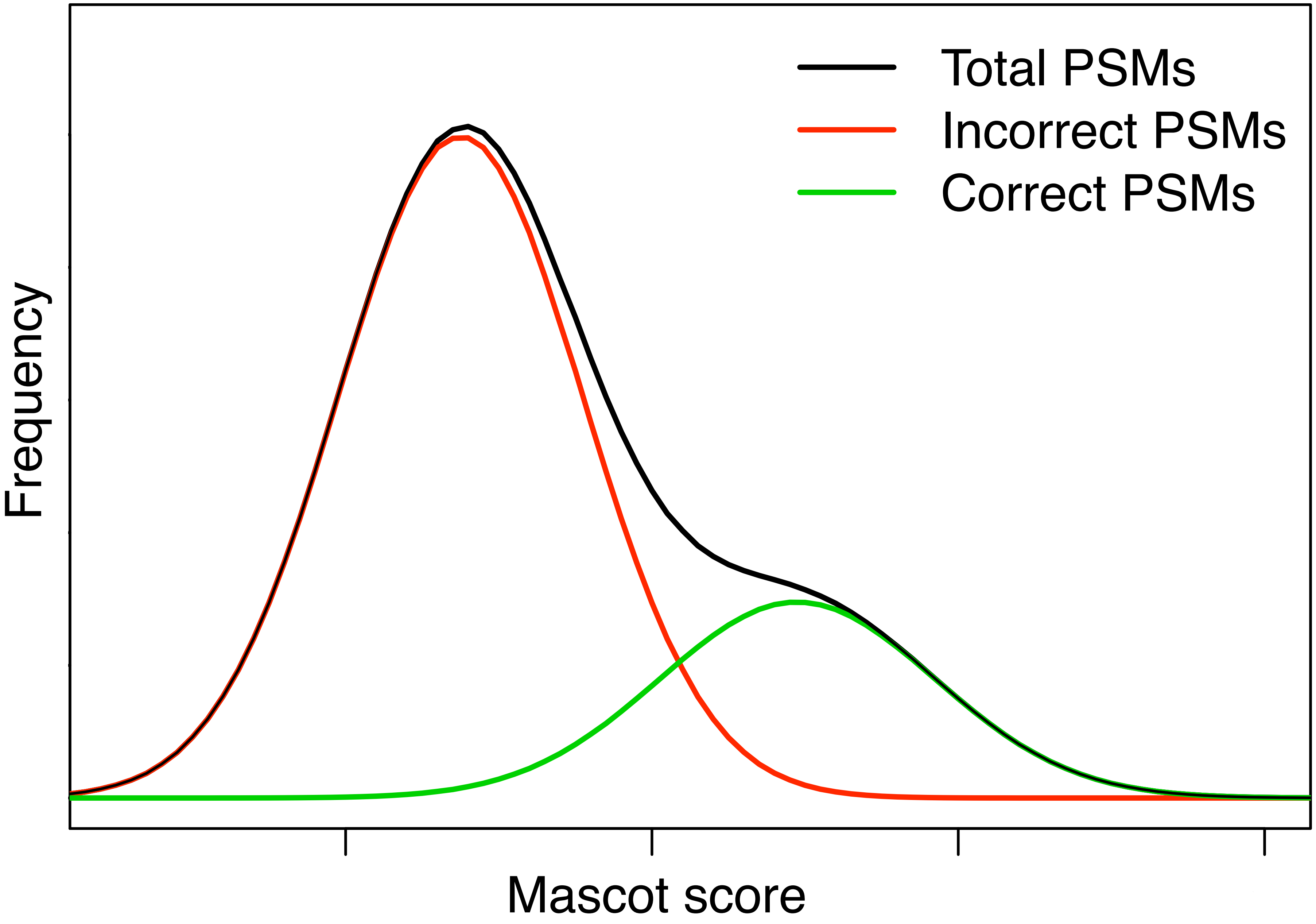
Sensitive and accurate peptide identification with Mascot Percolator

Markus Brosch
mb8@sanger.ac.uk

wellcome trust sanger institute

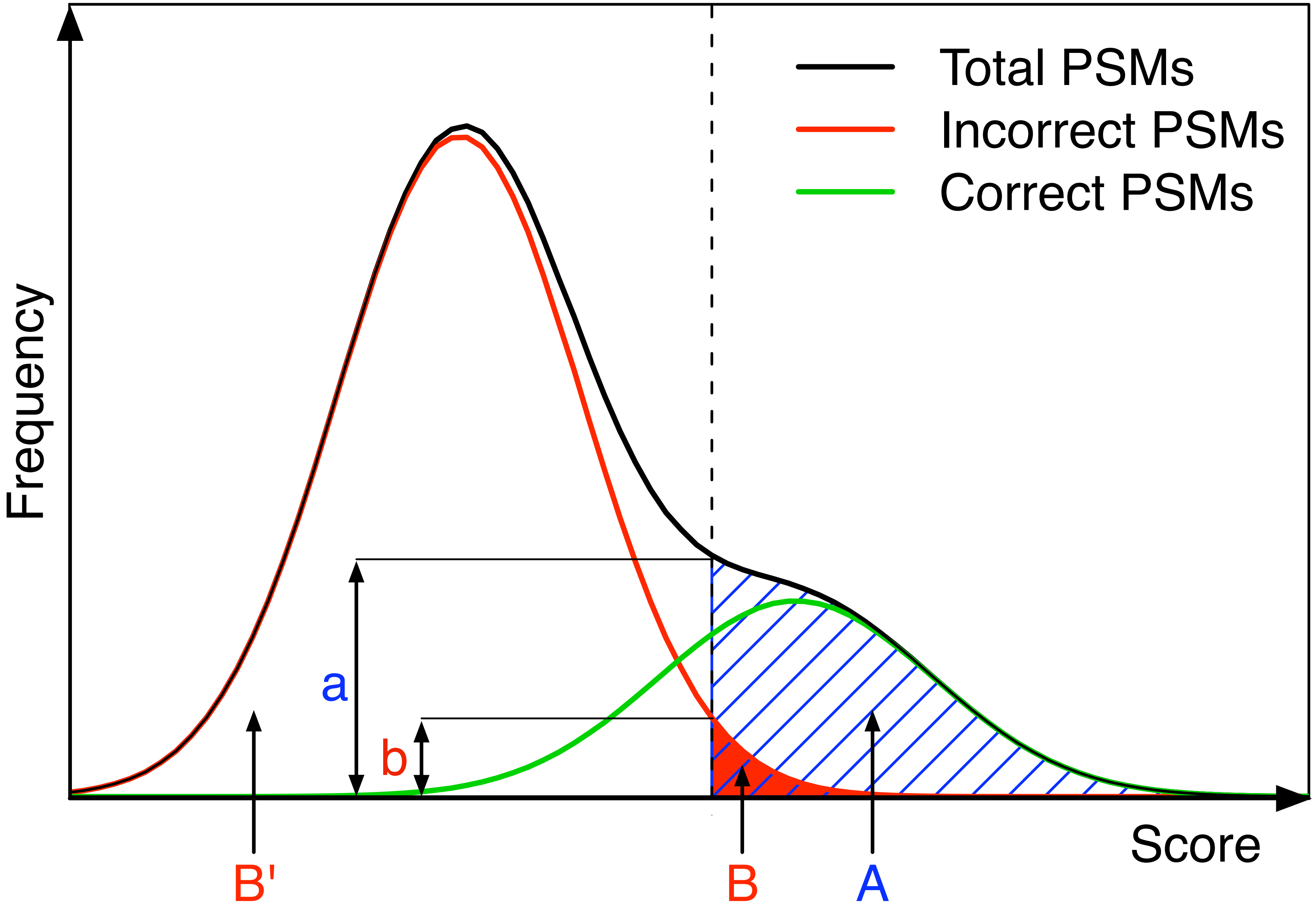
UNIVERSITY OF CAMBRIDGE

Terminology: FPR, FDR, PEP



L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, *J Proteome Res* 7, **29** (2008).
L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, J Proteome Res 7, **40** (2008).
M. Brosch, J. Choudhary, in Scoring and validation of tandem MS peptide identification methods, Eds. (Humana Press, 2009).

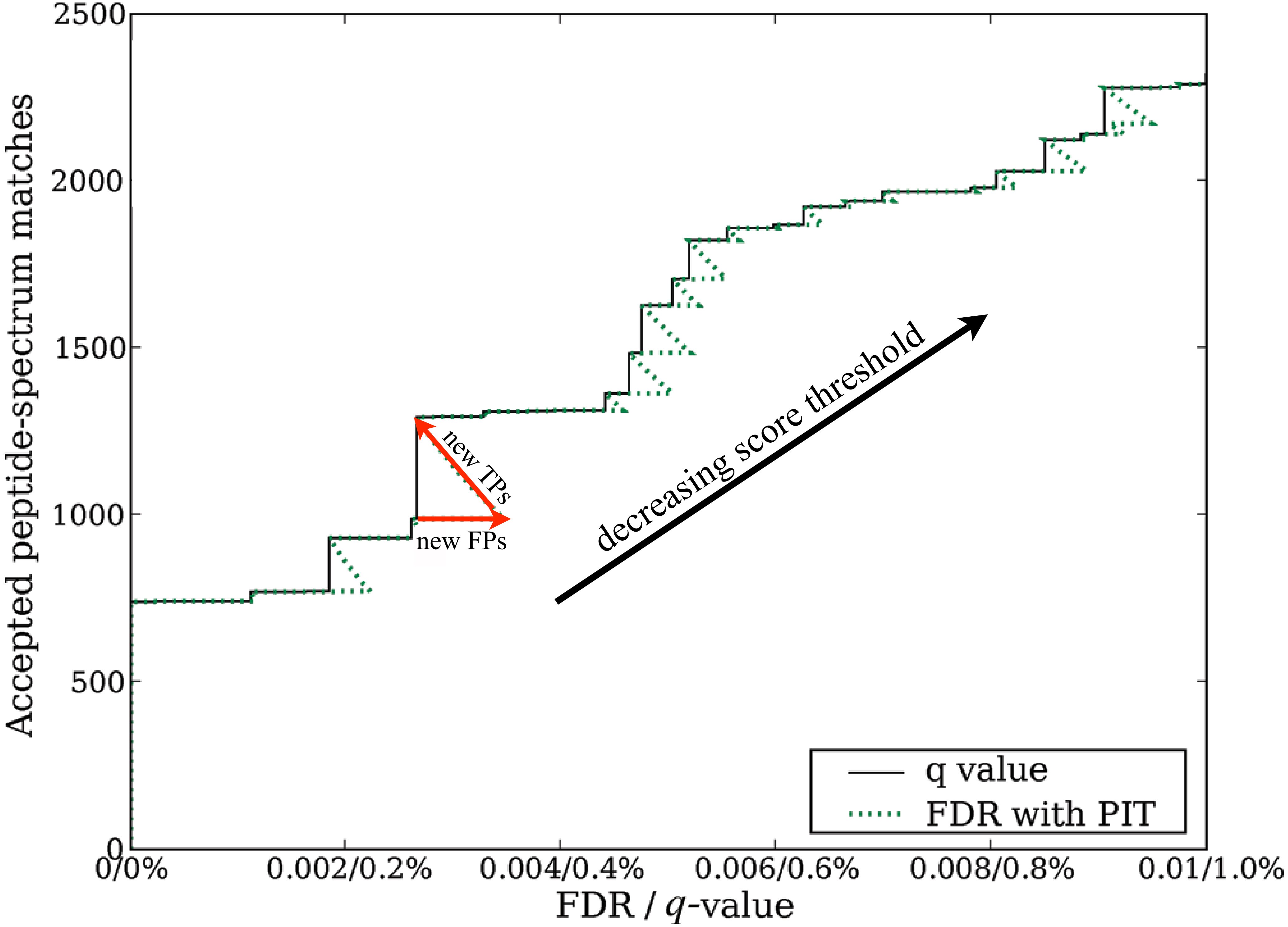
Terminology: FDR & PEP



~~FPR = B / (B'+B)~~ $\quad \mathrm{FDR} = B / A = (\sum_{i=1}^{A} \mathrm{PEP}_i) / A \quad \mathrm{PEP} = b / a$

L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, *J Proteome Res* 7, **29** (2008).
L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, J Proteome Res 7, **40** (2008).
M. Brosch, J. Choudhary, in Scoring and validation of tandem MS peptide identification methods, Eds. (Humana Press, 2009).

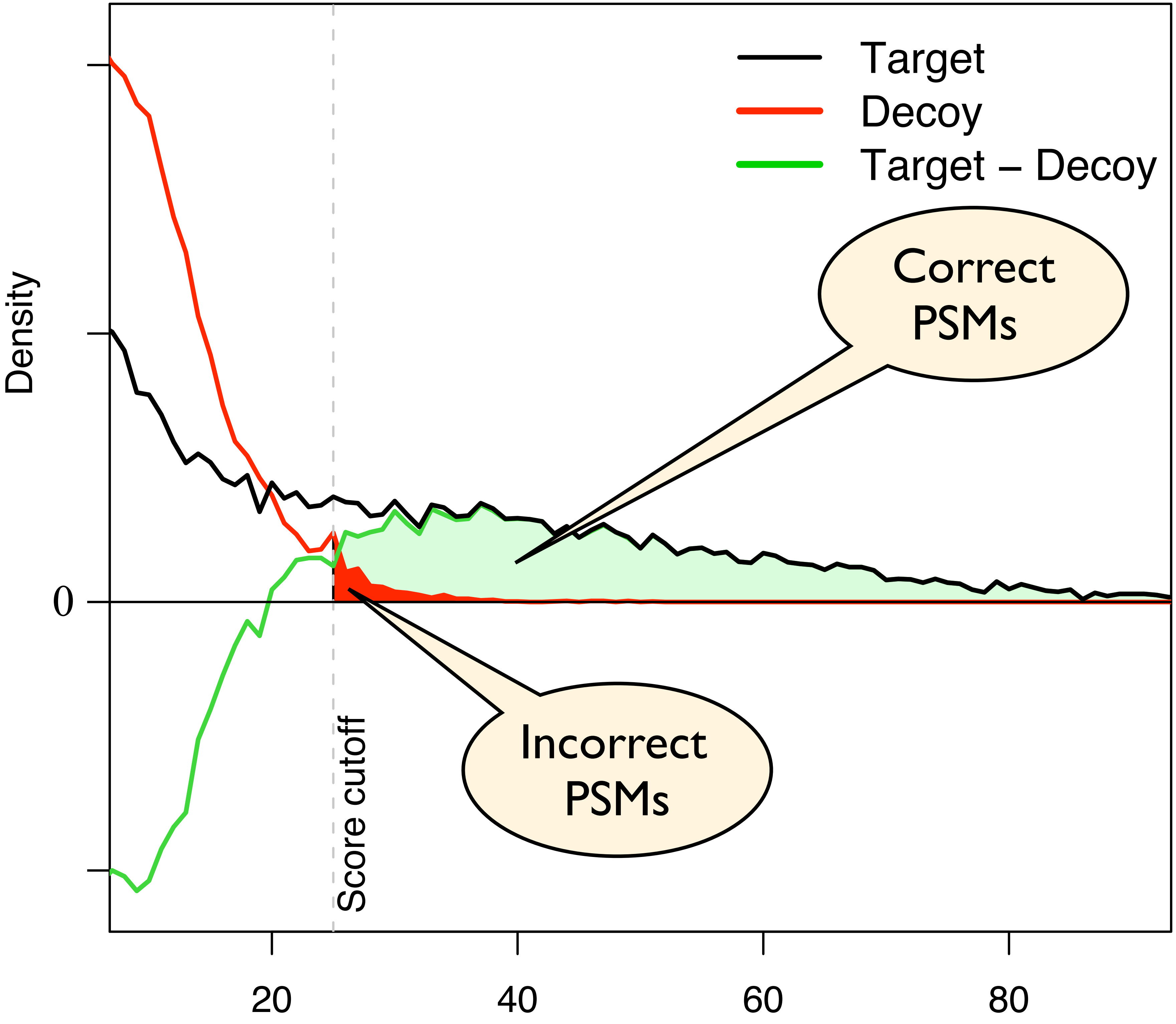
Terminology: FDR vs q-value



J. D. Storey, R. Tibshirani, Proc Natl Acad Sci U S A 100, 9440 (2003).

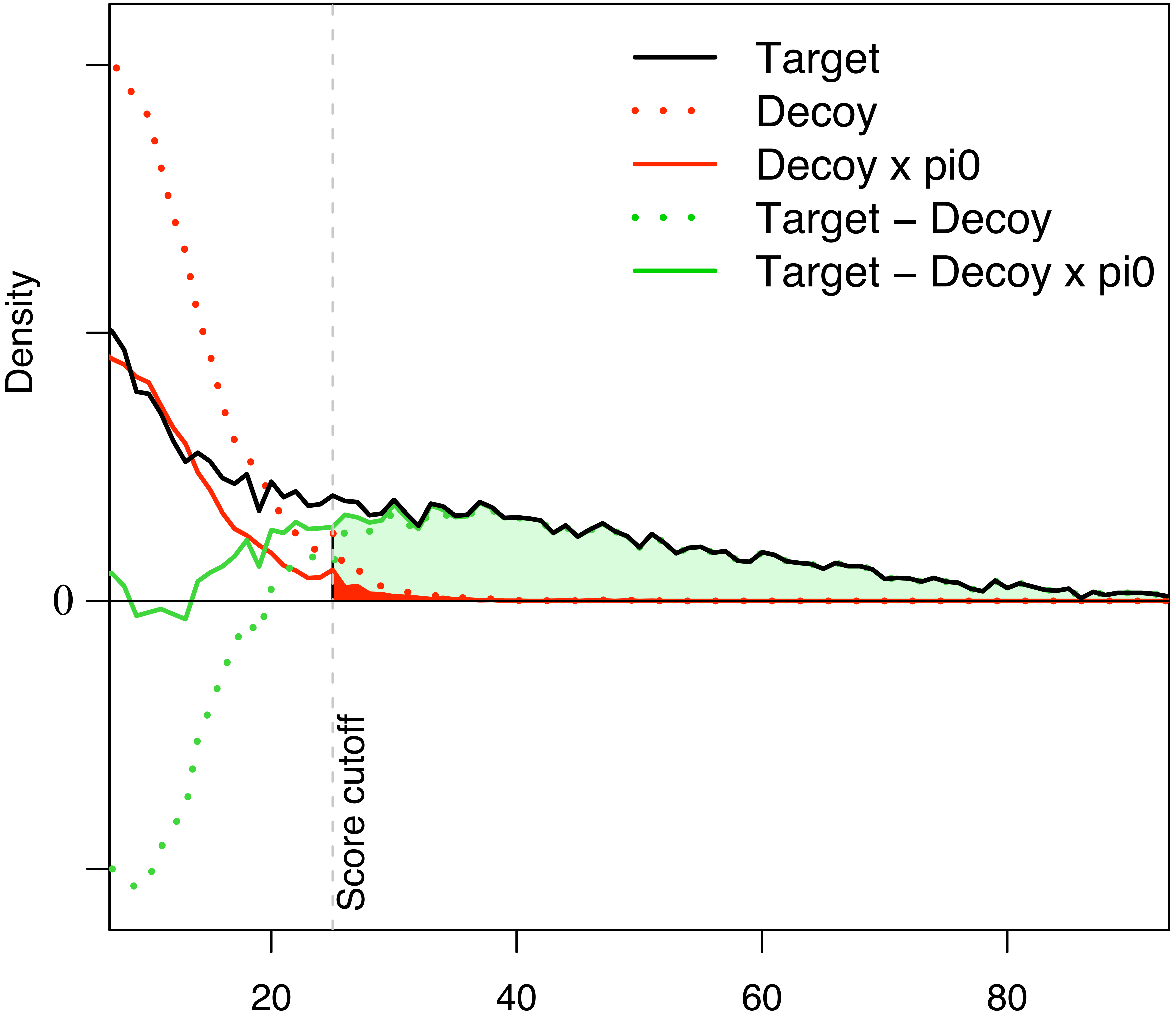
Figure: L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, *J Proteome Res* **7**, 29 (2008).

Target / Decoy database searching



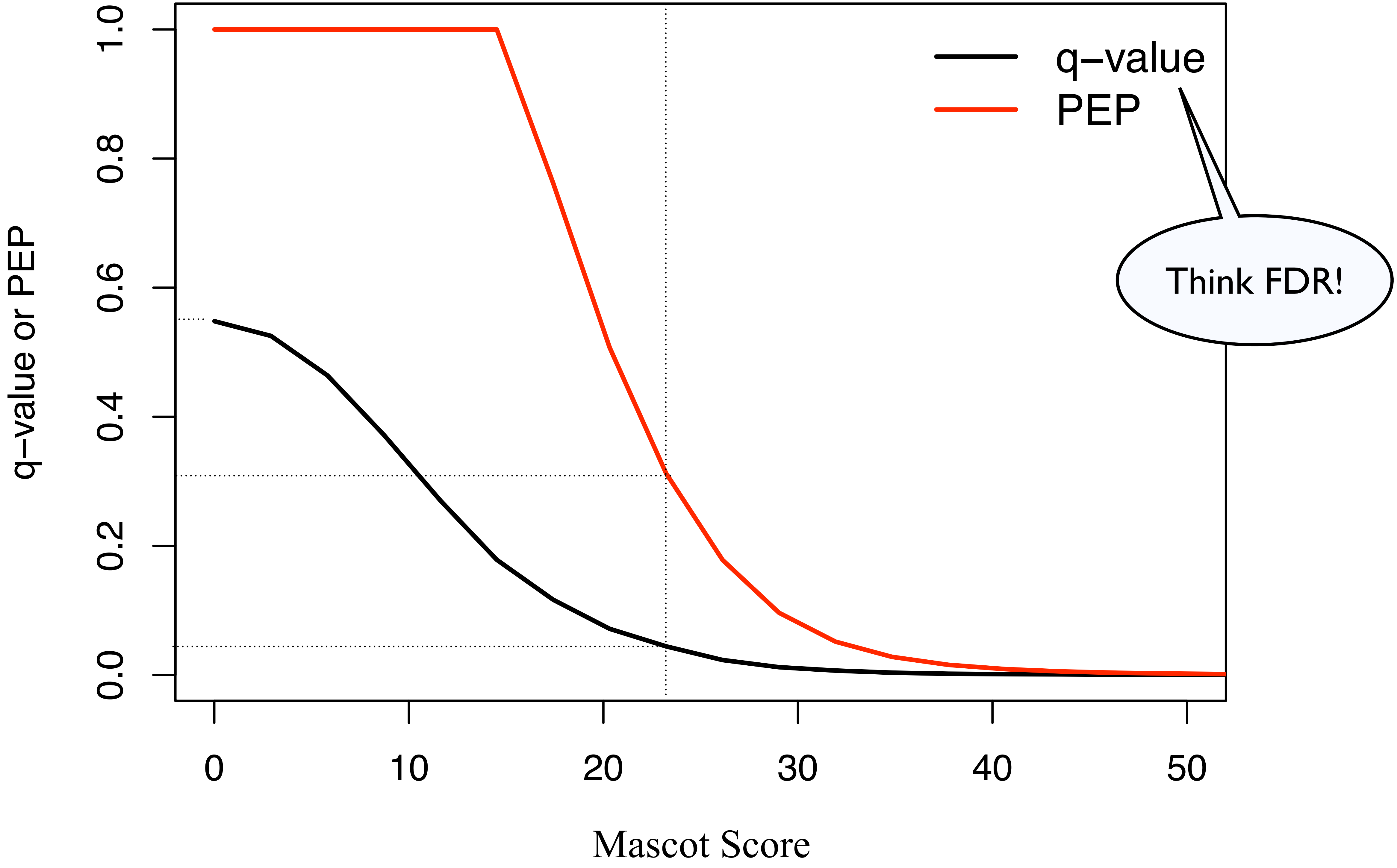
Target/Decoy concept: R. E. Moore, M. K. Young, T. D. Lee, J Am Soc Mass Spectrom 13, 378 (2002).

Target / Decoy database searching



L. Käll, J. D. Storey, M. J. MacCoss, W. S. Noble, *J Proteome Res* **7**, 29 (2008).

Accurate q-values (FDR) and PEPs



qvality software package: L. Käll, J. D. Storey, W. S. Noble, *Bioinformatics* **24**, i42 (2008).

Mascot score

$$\text{Mascot score} = -10\log_{10}(P)$$

$$P = 10^{-(\text{Mascot score} / 10)}$$

Probability that match is random

Example:
A 1% probability that the peptide spectrum match is a random event would translate into a Mascot score of 20.

‘Theoretical’

MIT

Mascot Identity Threshold

$$\text{MIT} = -10\log_{10}(P)$$

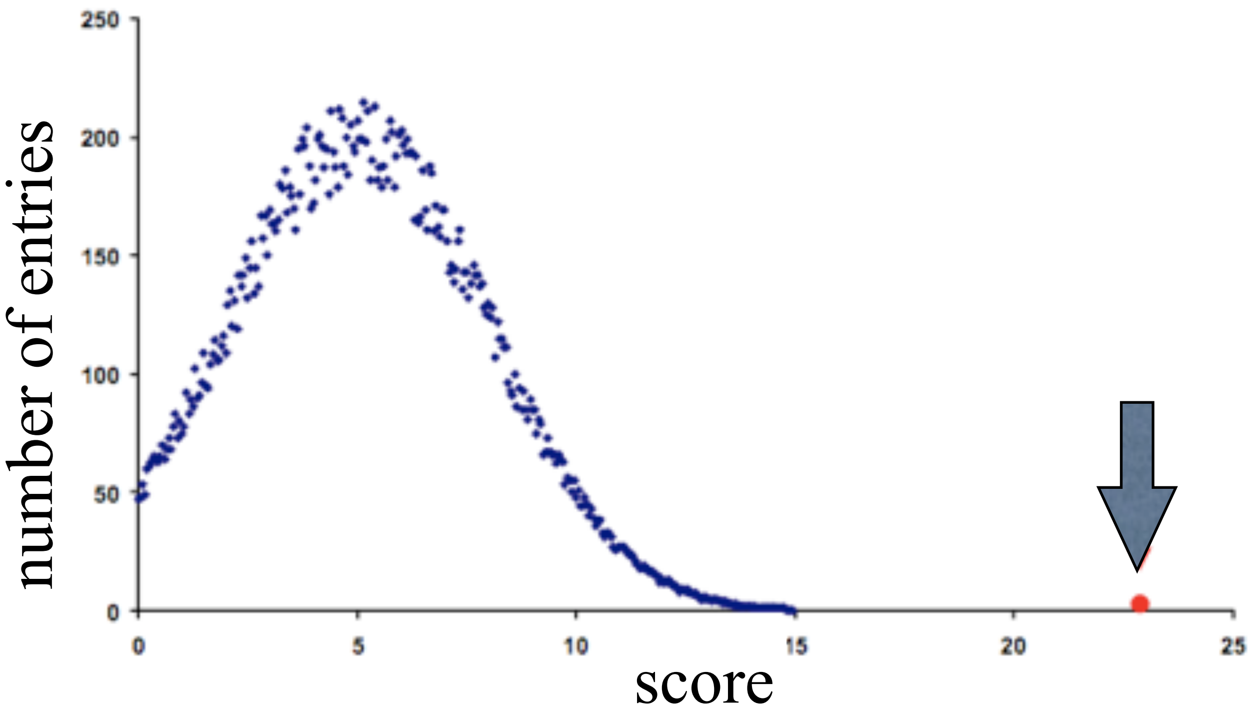
Example:
If there are 5000 precursor matches, a 1 in a 20 chance of getting a false positive match is a probability of
$P = 1 / (20 \times 5000 \times n)$

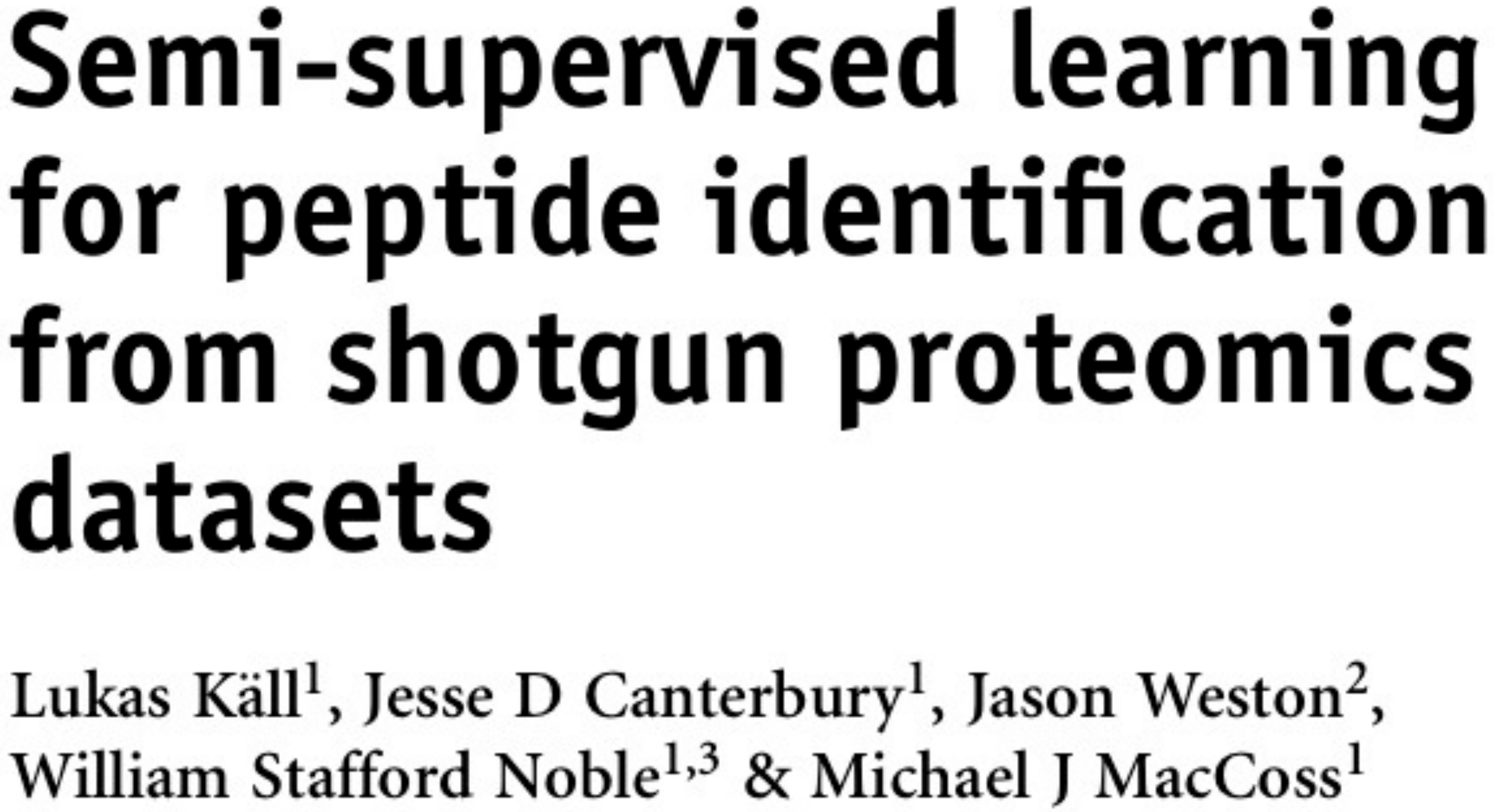
http://www.matrixscience.com/pdf/optimization.pdf
http://www.matrixscience.com/pdf/2005WKSHP4.pdf

‘Empirical’

MHT

Mascot Homology Threshold





Semi-supervised learning for peptide identification from shotgun proteomics datasets

Lukas Käll[1], Jesse D Canterbury[1], Jason Weston[2], William Stafford Noble[1,3] & Michael J MacCoss[1]

- Good news: convincing method
- Bad news: only available for Sequest
- Good news: command line interface with generic input and output formats, so we can extend it for use with Mascot

Support Vector Machine

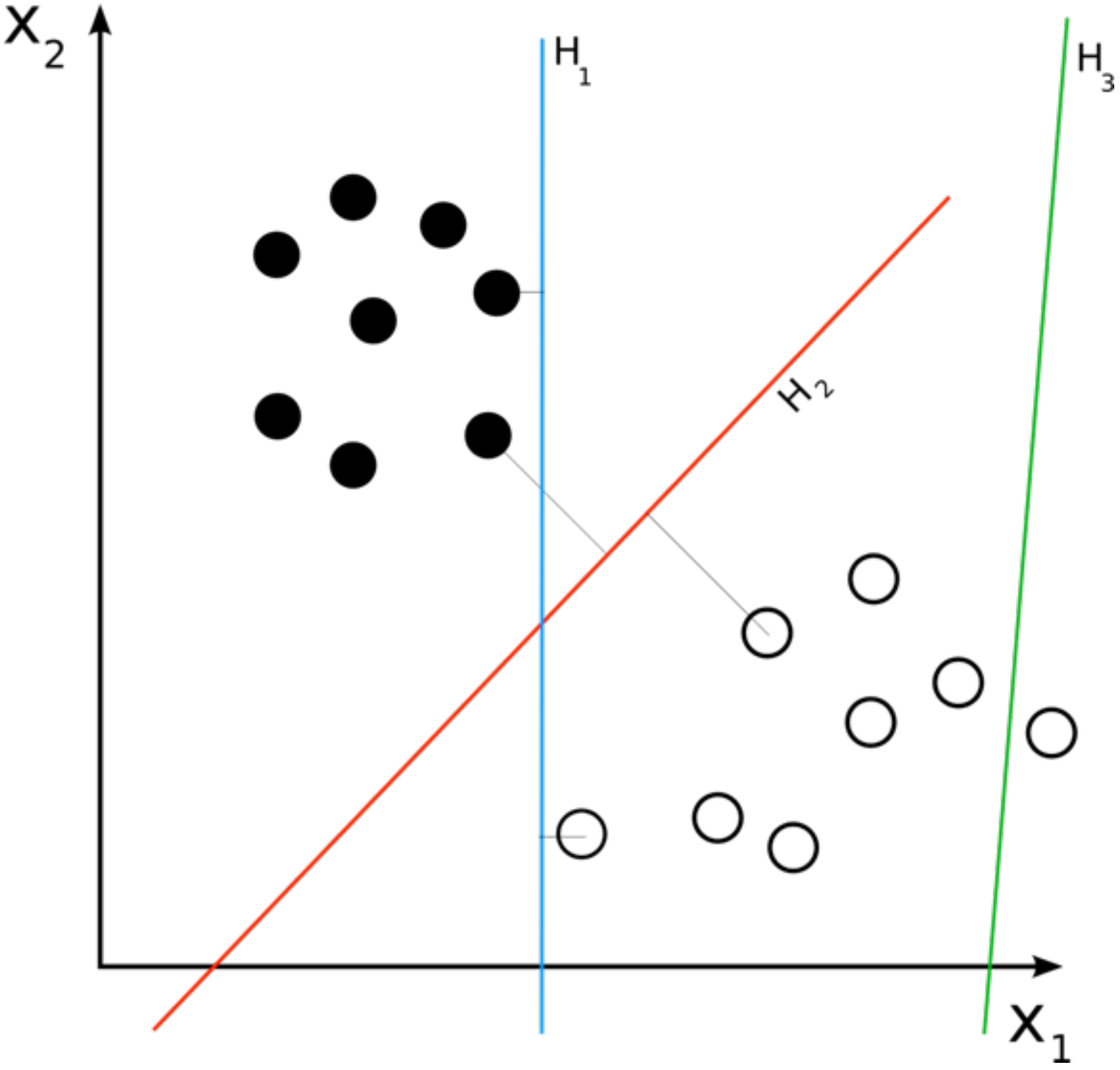
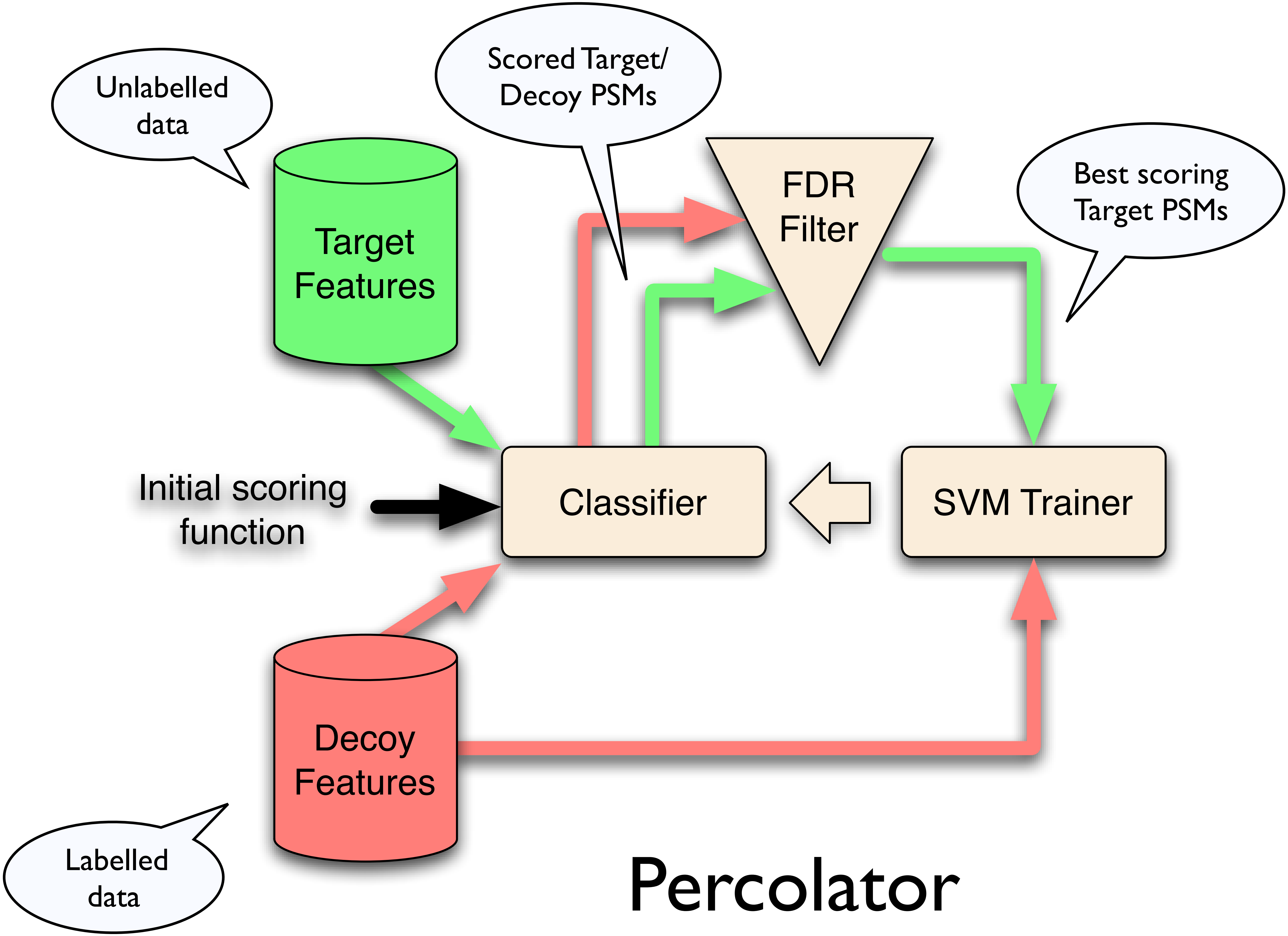


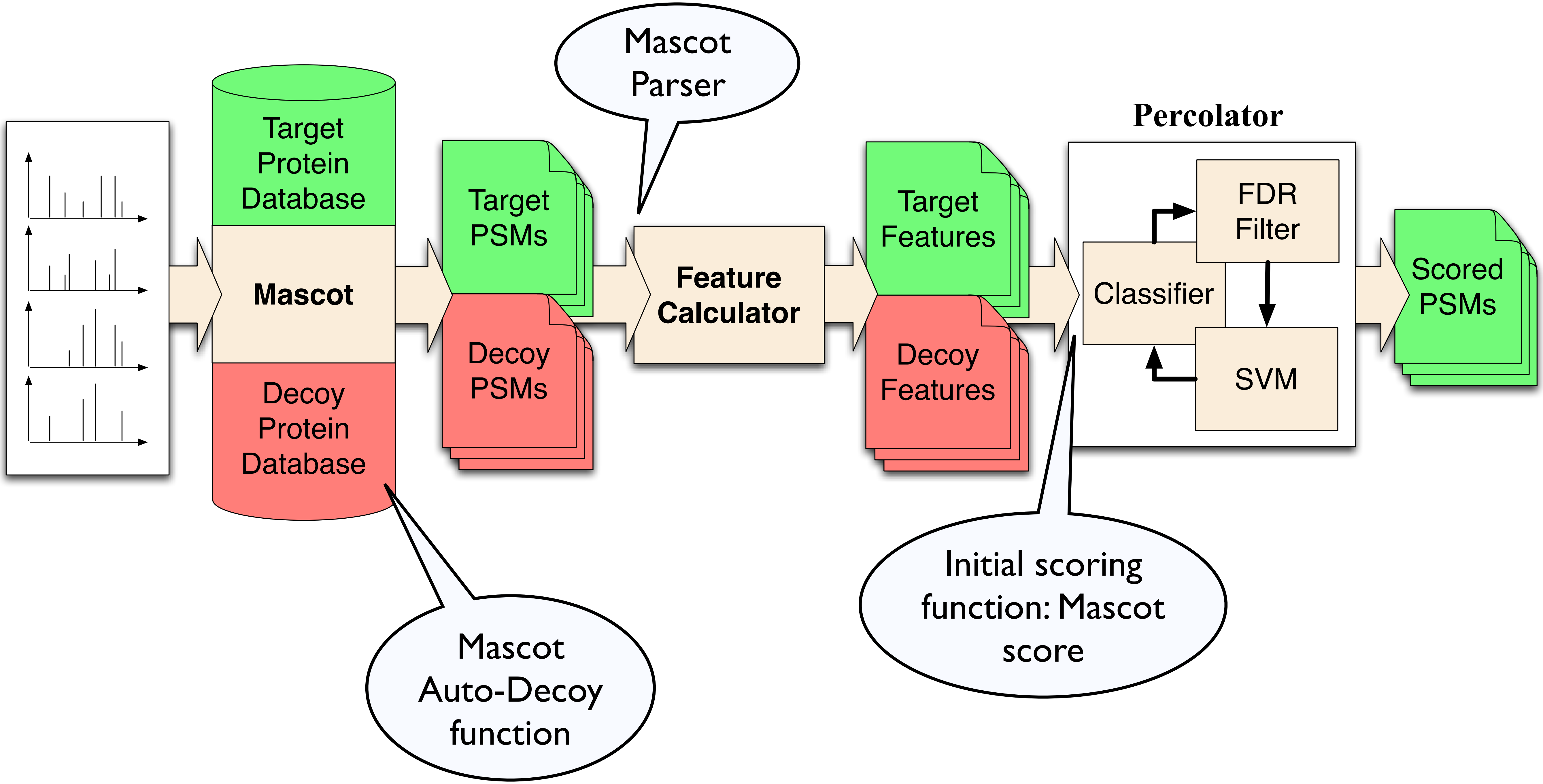
Figure: http://en.wikipedia.org/wiki/Support_vector_machine



Percolator

Käll, L., Canterbury, J. D., Weston, J., Noble, W. S., & MacCoss, M. J. (2007). *Nature Methods*

Mascot Percolator



M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

Mascot Percolator Features

(a) Peptide Matching Scores (Mascot score, Peptide score [1])

(b) Peptide properties (mass, charge, mc, var mods)

(c) Delta mass, absolute delta mass, delta mass accounting for incorrect peak detection (^{13}C)

(d) Fragment delta mass, absolute fragment delta mass

(e) Total intensity, matched intensity, relative matched intensity

(f) Fraction of ions matched (per ion series)

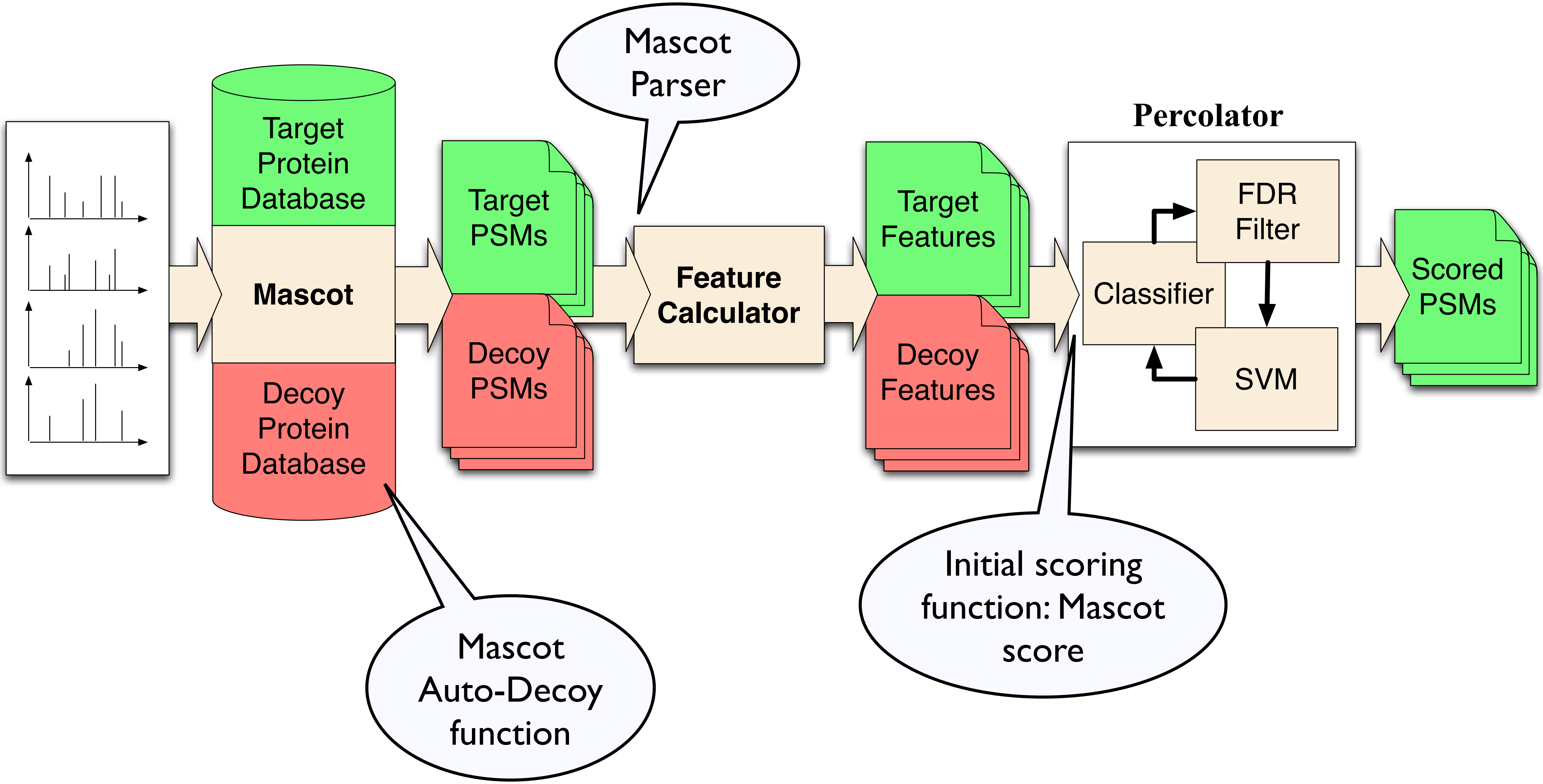
(g) Sequence coverage (per ion series)

(h) Intensity matched (per ion series)

(i) Retention time where available (see talk by Lukas Käll, WOE 3:10 pm)

[1] S. A. Beausoleil, J. Villen, S. A. Gerber, J. Rush, S. P. Gygi, *Nat Biotechnol* **24**, 1285 (2006).

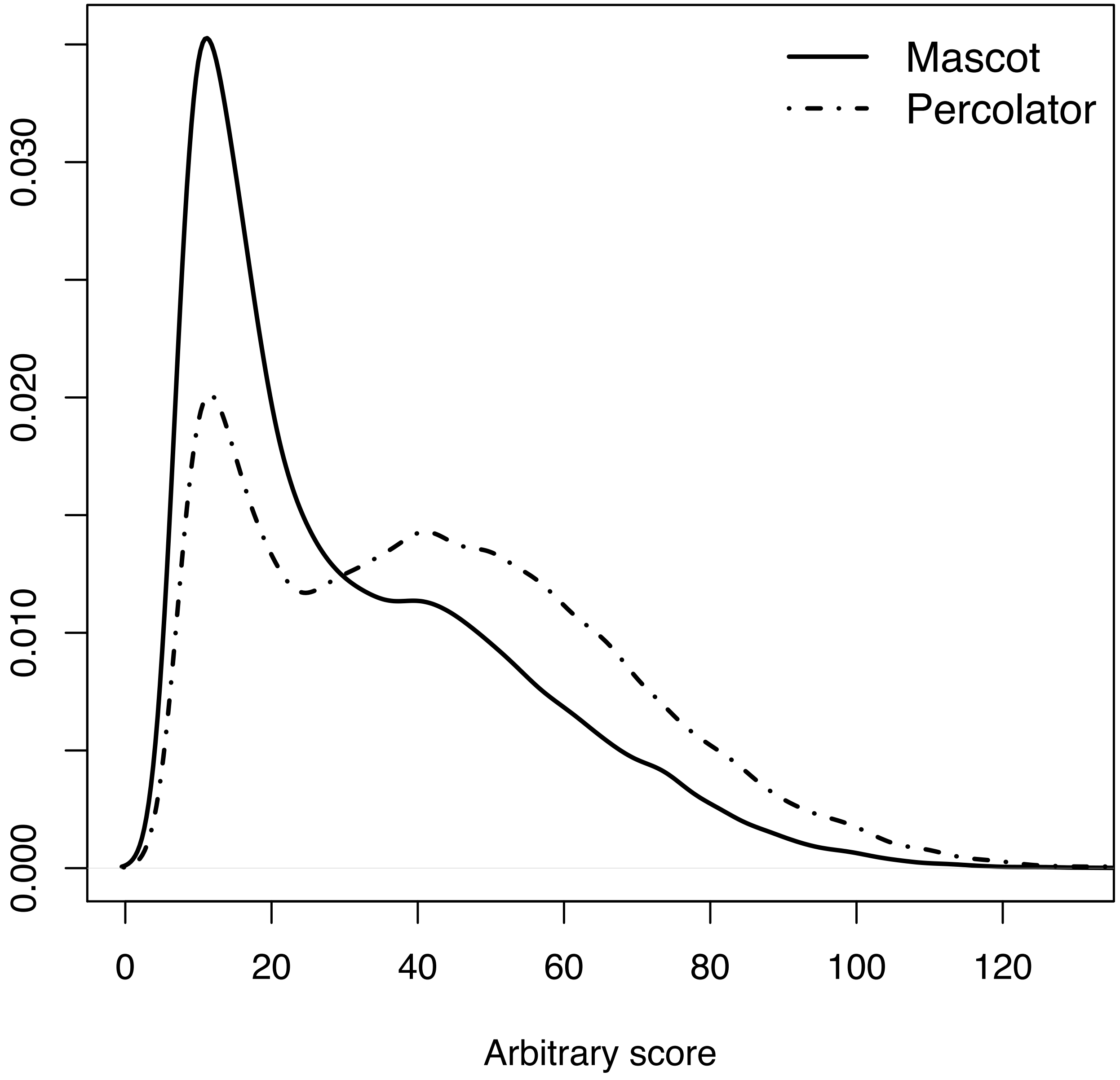
Mascot Percolator



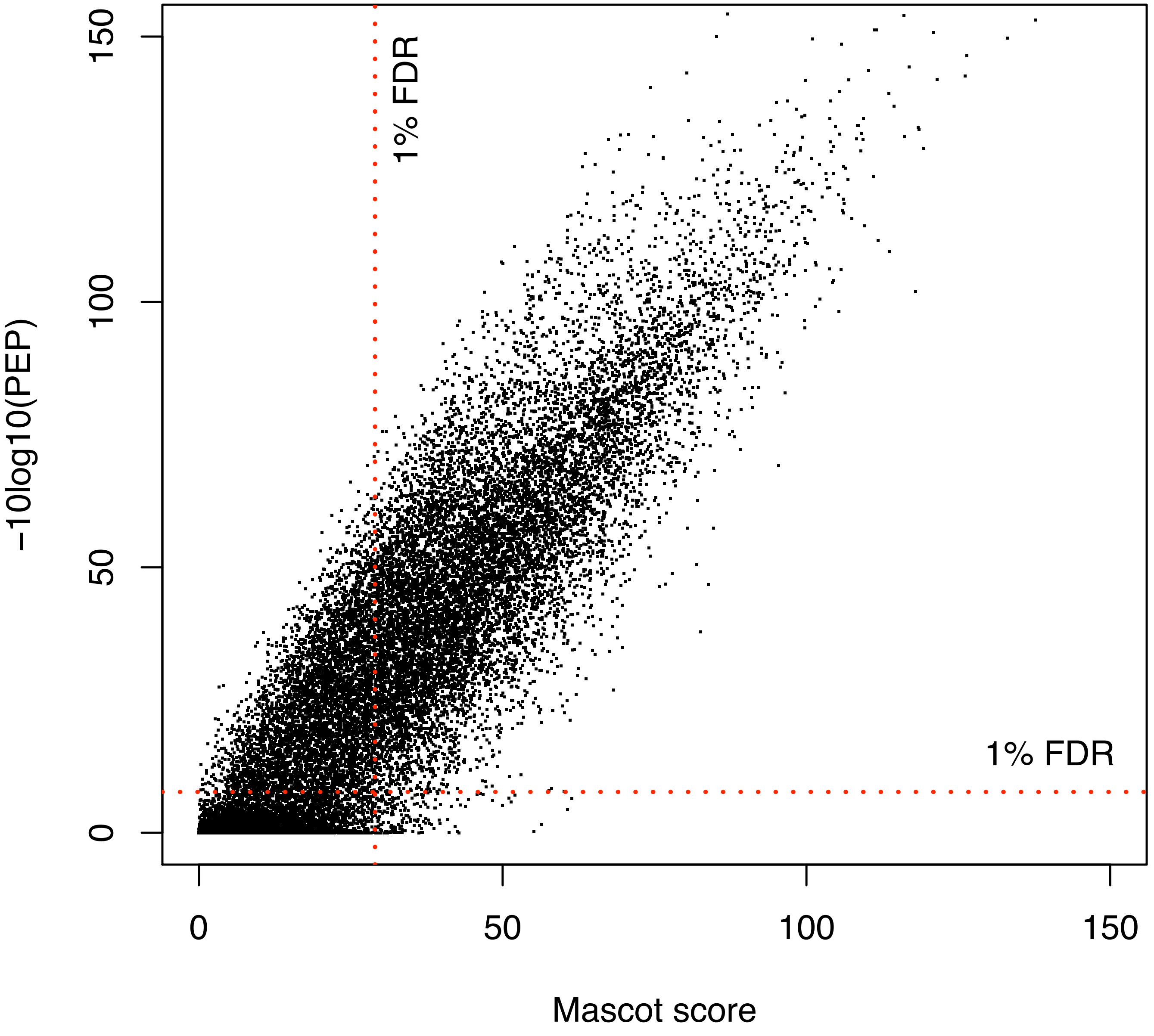
M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

Mascot score vs Percolator PEPs

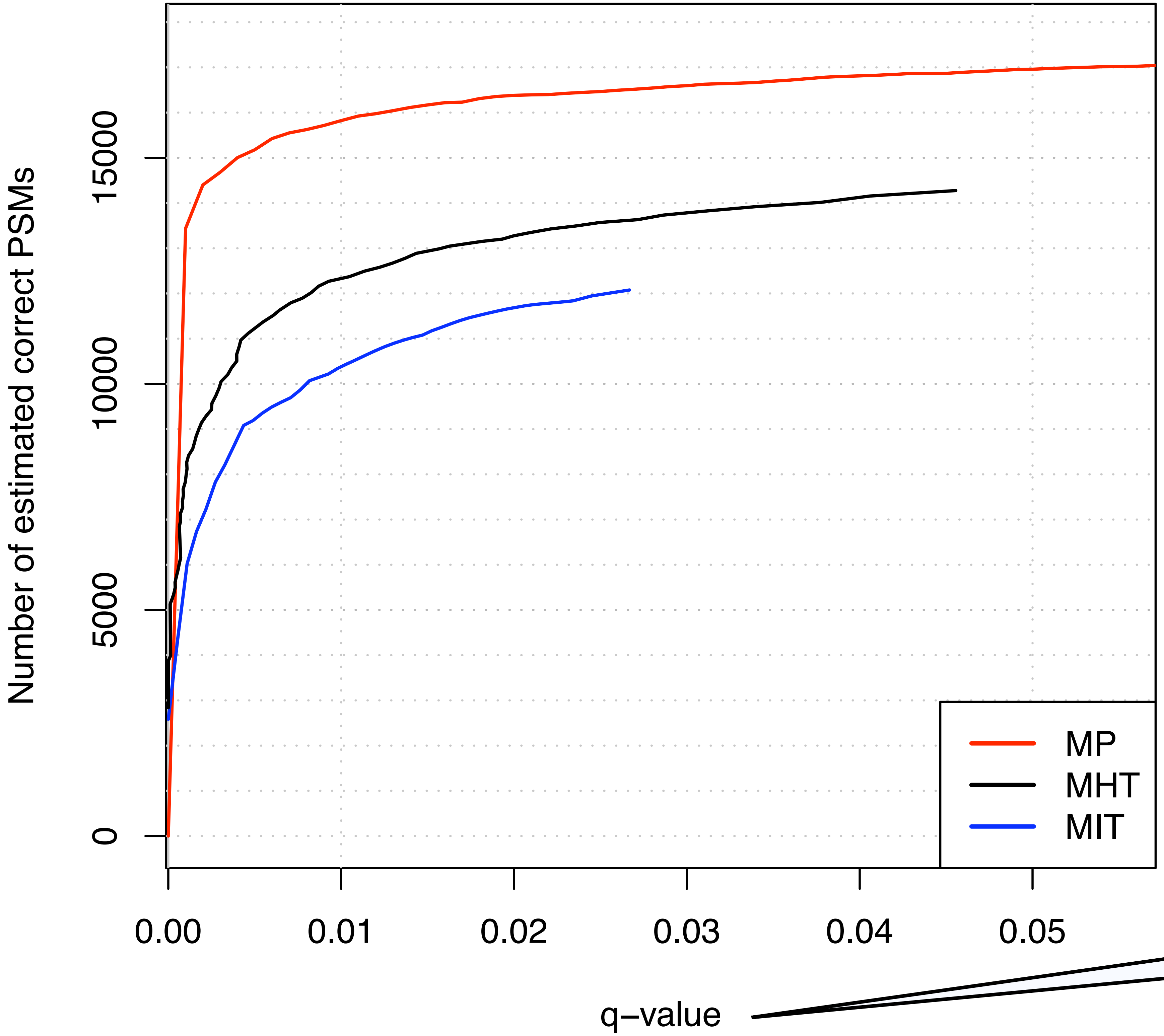
Mascot vs Percolator score density



Mascot score vs PEP



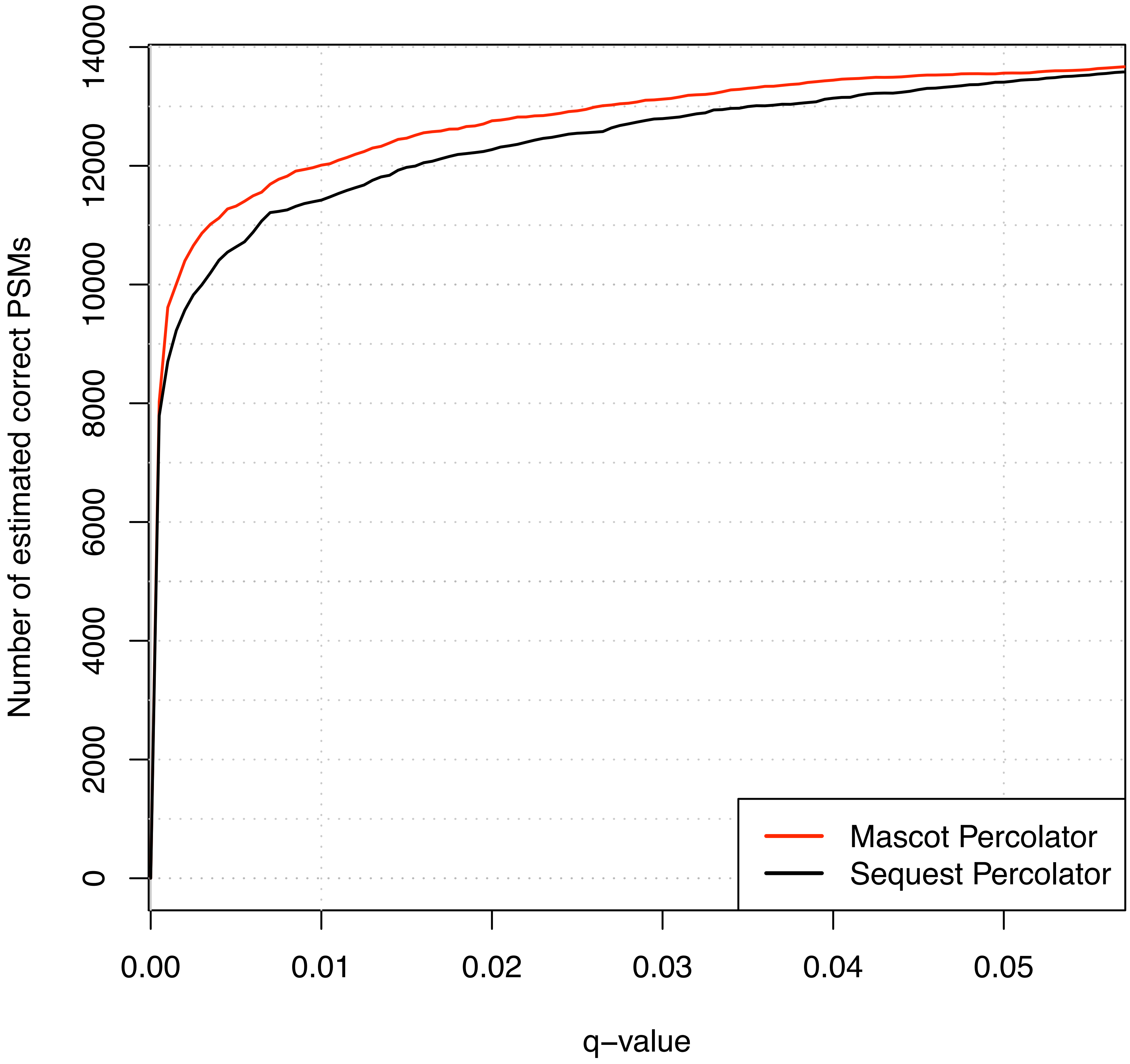
ROC comparison



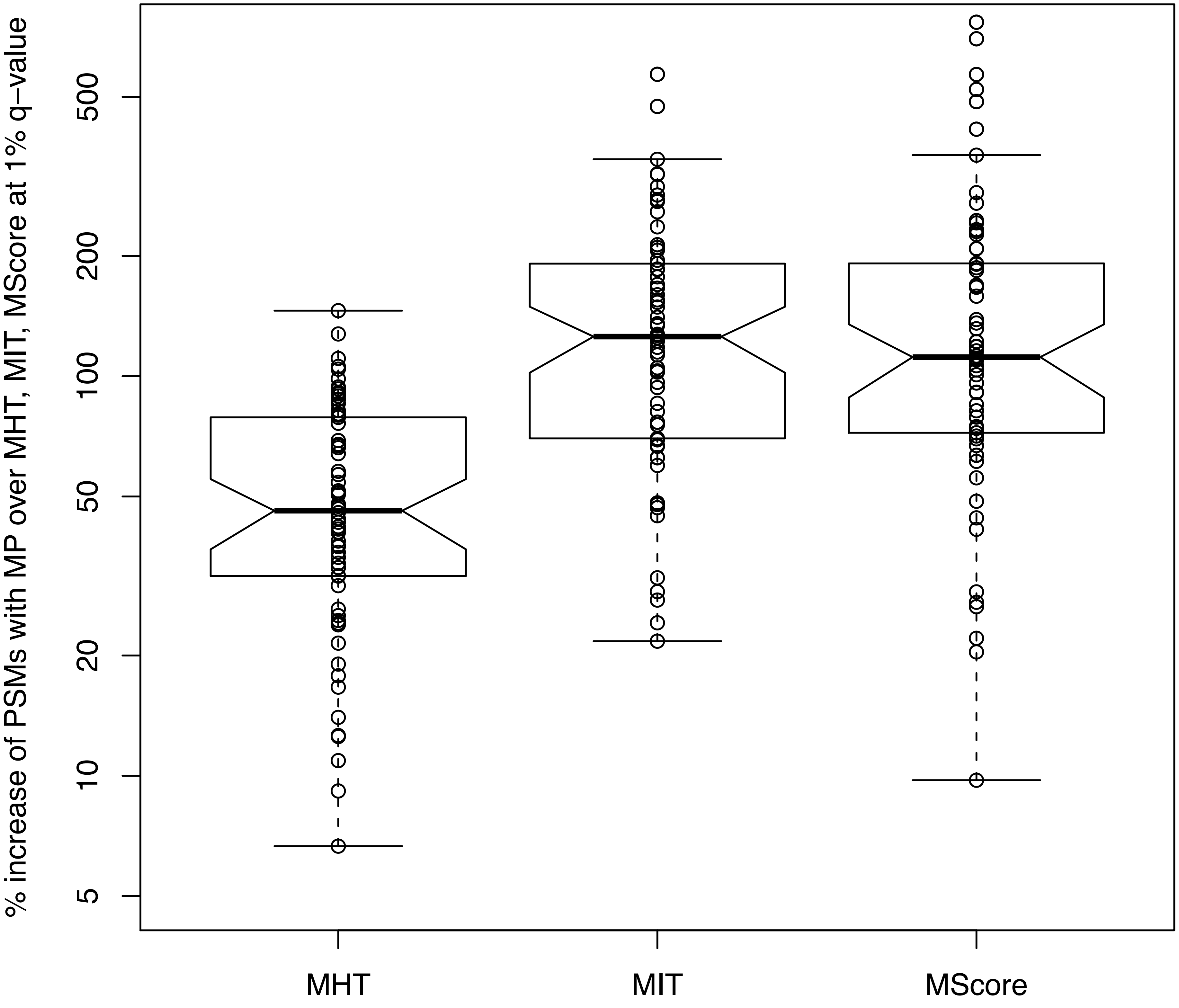
Think FDR

M. Brosch, L. Yu, T. Hubbard, J. Choudhary, *J Proteome Res* (2009).

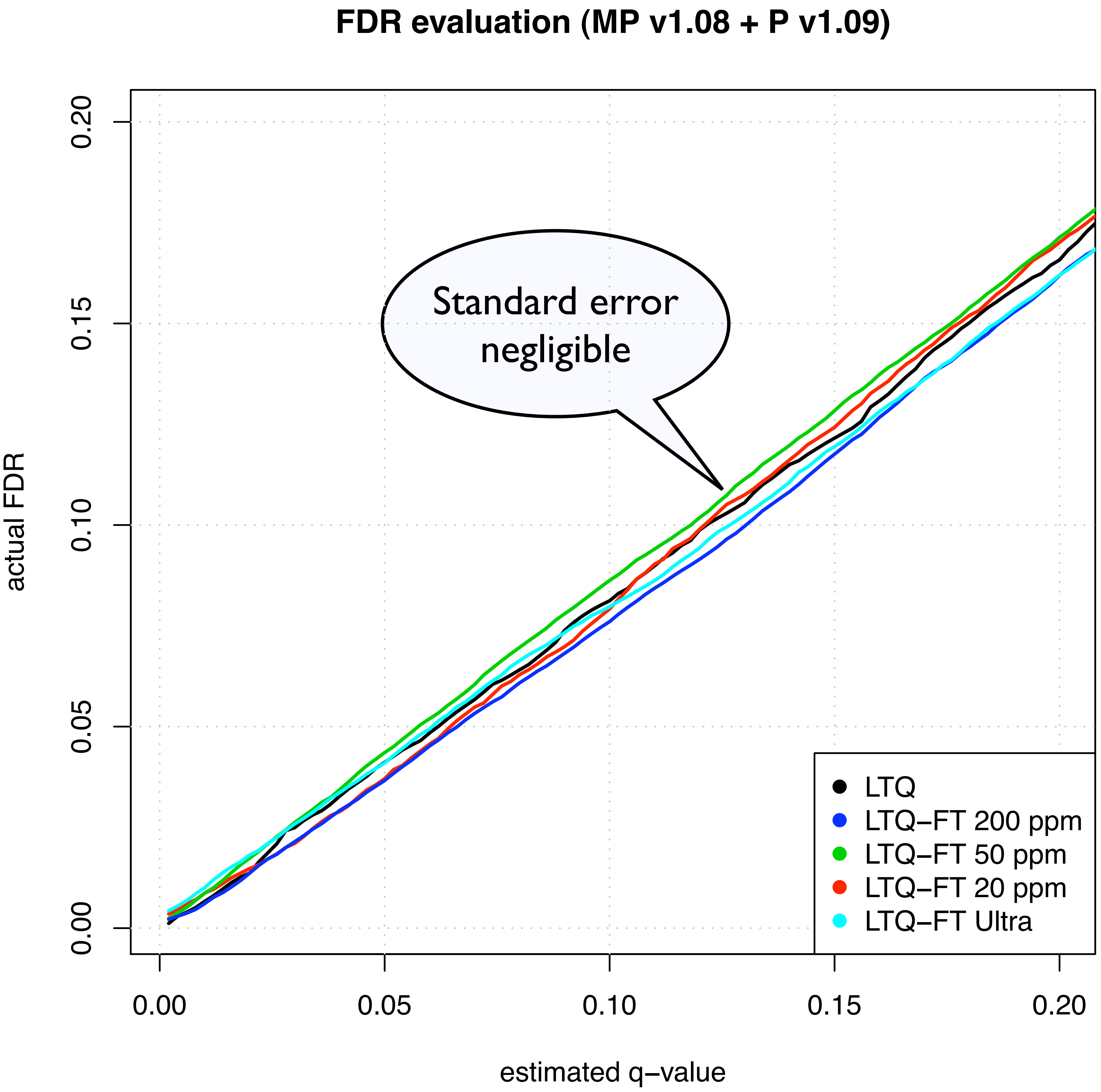
Mascot vs Sequest Percolator



All of Peptide Atlas Mouse



Evaluate q-value accuracy



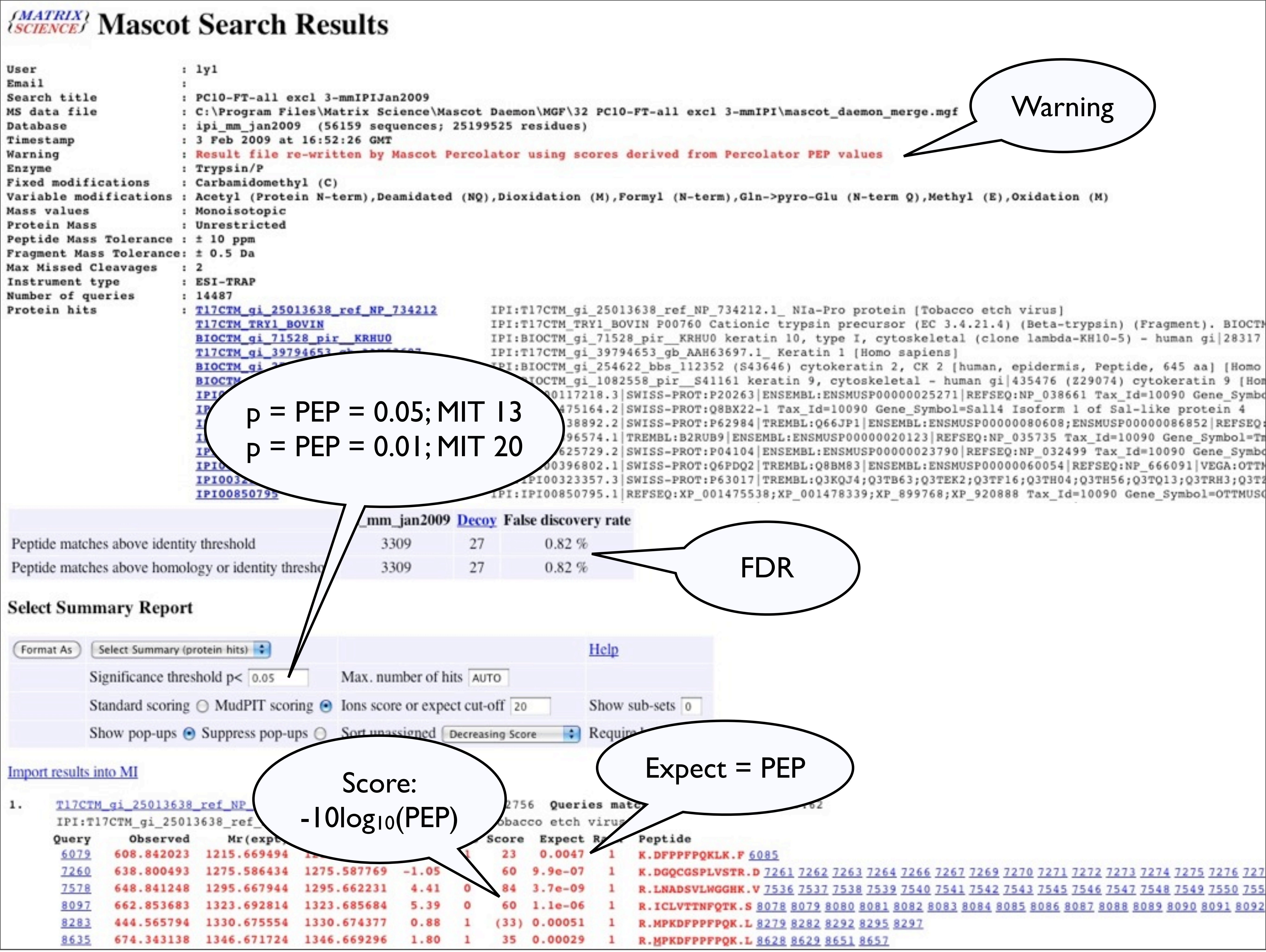
Mascot Percolator software package

- Simple command line interface:

```
java -cp MascotPercolator.jar cli.MascotPercolator
     -target 11083
     -decoy 11084
     -out 11083-11084
     -rankdelta 1
     -newDat
```

Mascot log IDs

- Tested with Mascot 2.2; limited experience with versions <= 2.1
- Runtime: 10-150 spectra per second
- Memory requirements typically 1-2 GB for up to 100k spectra. Largest Mascot search processed was 350k spectra which required 6GB of memory.



Mascot Search Results

```
User                   : lyl
Email                  :
Search title           : PC10-FT-all excl 3-mmIPIJan2009
MS data file           : C:\Program Files\Matrix Science\Mascot Daemon\MGF\32 PC10-FT-all excl 3-mmIPI\mascot_daemon_merge.mgf
Database               : ipi_mm_jan2009  (56159 sequences; 25199525 residues)
Timestamp              : 3 Feb 2009 at 16:52:26 GMT
Warning                : Result file re-written by Mascot Percolator using scores derived from Percolator PEP values
Enzyme                 : Trypsin/P
Fixed modifications    : Carbamidomethyl (C)
Variable modifications : Acetyl (Protein N-term),Deamidated (NQ),Dioxidation (M),Formyl (N-term),Gln->pyro-Glu (N-term Q),Methyl (E),Oxidation (M)
Mass values            : Monoisotopic
Protein Mass           : Unrestricted
Peptide Mass Tolerance : ± 10 ppm
Fragment Mass Tolerance: ± 0.5 Da
Max Missed Cleavages   : 2
Instrument type        : ESI-TRAP
Number of queries      : 14487
Protein hits           : T17CTM_gi_25013638_ref_NP_734212       IPI:T17CTM_gi_25013638_ref_NP_734212.1_ NIa-Pro protein [Tobacco etch virus]
                         T17CTM_TRY1_BOVIN                      IPI:T17CTM_TRY1_BOVIN P00760 Cationic trypsin precursor (EC 3.4.21.4) (Beta-trypsin) (Fragment). BIOCTM
                         BIOCTM_gi_71528_pir__KRHU0             IPI:BIOCTM_gi_71528_pir__KRHU0 keratin 10, type I, cytoskeletal (clone lambda-KH10-5) - human gi|28317
                         T17CTM_gi_39794653                     IPI:T17CTM_gi_39794653_gb_AAH63697.1_ Keratin 1 [Homo sapiens]
                         BIOCTM_gi                              PI:BIOCTM_gi_254622_bbs_112352 (S43646) cytokeratin 2, CK 2 [human, epidermis, Peptide, 645 aa] [Homo
                         BIOCTM                                 IOCTM_gi_1082558_pir__S41161 keratin 9, cytoskeletal - human gi|435476 (Z29074) cytokeratin 9 [Hom
                         IPI                                    00117218.3|SWISS-PROT:P20263|ENSEMBL:ENSMUSP00000025271|REFSEQ:NP_038661 Tax_Id=10090 Gene_Symbo
                         IP                                     475164.2|SWISS-PROT:Q8BX22-1 Tax_Id=10090 Gene_Symbol=Sall4 Isoform 1 of Sal-like protein 4
                         I                                      38892.2|SWISS-PROT:P62984|TREMBL:Q66JP1|ENSEMBL:ENSMUSP00000080608;ENSMUSP00000086852|REFSEQ:
                         I                                      96574.1|TREMBL:B2RUB9|ENSEMBL:ENSMUSP00000020123|REFSEQ:NP_035735 Tax_Id=10090 Gene_Symbol=Tm
                         IP                                     625729.2|SWISS-PROT:P04104|ENSEMBL:ENSMUSP00000023790|REFSEQ:NP_032499 Tax_Id=10090 Gene_Symbo
                         IPI0                                   00396802.1|SWISS-PROT:Q6PDQ2|TREMBL:Q8BM83|ENSEMBL:ENSMUSP00000060054|REFSEQ:NP_666091|VEGA:OTTM
                         IPI0032                                PI00323357.3|SWISS-PROT:P63017|TREMBL:Q3KQJ4;Q3TB63;Q3TEK2;Q3TF16;Q3TH04;Q3TH56;Q3TQ13;Q3TRH3;Q3T
                         IPI00850795                            IPI:IPI00850795.1|REFSEQ:XP_001475538;XP_001478339;XP_899768;XP_920888 Tax_Id=10090 Gene_Symbol=OTTMUS
```

	_mm_jan2009	Decoy	False discovery rate
Peptide matches above identity threshold	3309	27	0.82 %
Peptide matches above homology or identity thresho	3309	27	0.82 %

Select Summary Report

Format As	Select Summary (protein hits)		Help
	Significance threshold p< 0.05	Max. number of hits AUTO	
	Standard scoring ○ MudPIT scoring ●	Ions score or expect cut-off 20	Show sub-sets 0
	Show pop-ups ● Suppress pop-ups ○	Sort unassigned Decreasing Score	Requi

Import results into MI

```
1.   T17CTM_gi_25013638_ref_NP_                        2756   Queries matc                 62
     IPI:T17CTM_gi_25013638_ref_                       obacco etch virus
   Query    Observed    Mr(expt                             Score  Expect  Rank  Peptide
   6079   608.842023  1215.669494  12                     1    23  0.0047   1    K.DFPPFPQKLK.F 6085
   7260   638.800493  1275.586434  1275.587769  -1.05          60  9.9e-07  1    K.DGQCGSPLVSTR.D 7261 7262 7263 7264 7266 7267 7269 7270 7271 7272 7273 7274 7275 7276 727
   7578   648.841248  1295.667944  1295.662231   4.41     0    84  3.7e-09  1    R.LNADSVLWGGHK.V 7536 7537 7538 7539 7540 7541 7542 7543 7545 7546 7547 7548 7549 7550 755
   8097   662.853683  1323.692814  1323.685684   5.39     0    60  1.1e-06  1    R.ICLVTTNFQTK.S 8078 8079 8080 8081 8082 8083 8084 8085 8086 8087 8088 8089 8090 8091 8092
   8283   444.565794  1330.675554  1330.674377   0.88     1  (33) 0.00051  1    R.MPKDFPPFPQK.L 8279 8282 8292 8295 8297
   8635   674.343138  1346.671724  1346.669296   1.80     1    35  0.00029  1    R.MPKDFPPFPQK.L 8628 8629 8651 8657
```

Warning

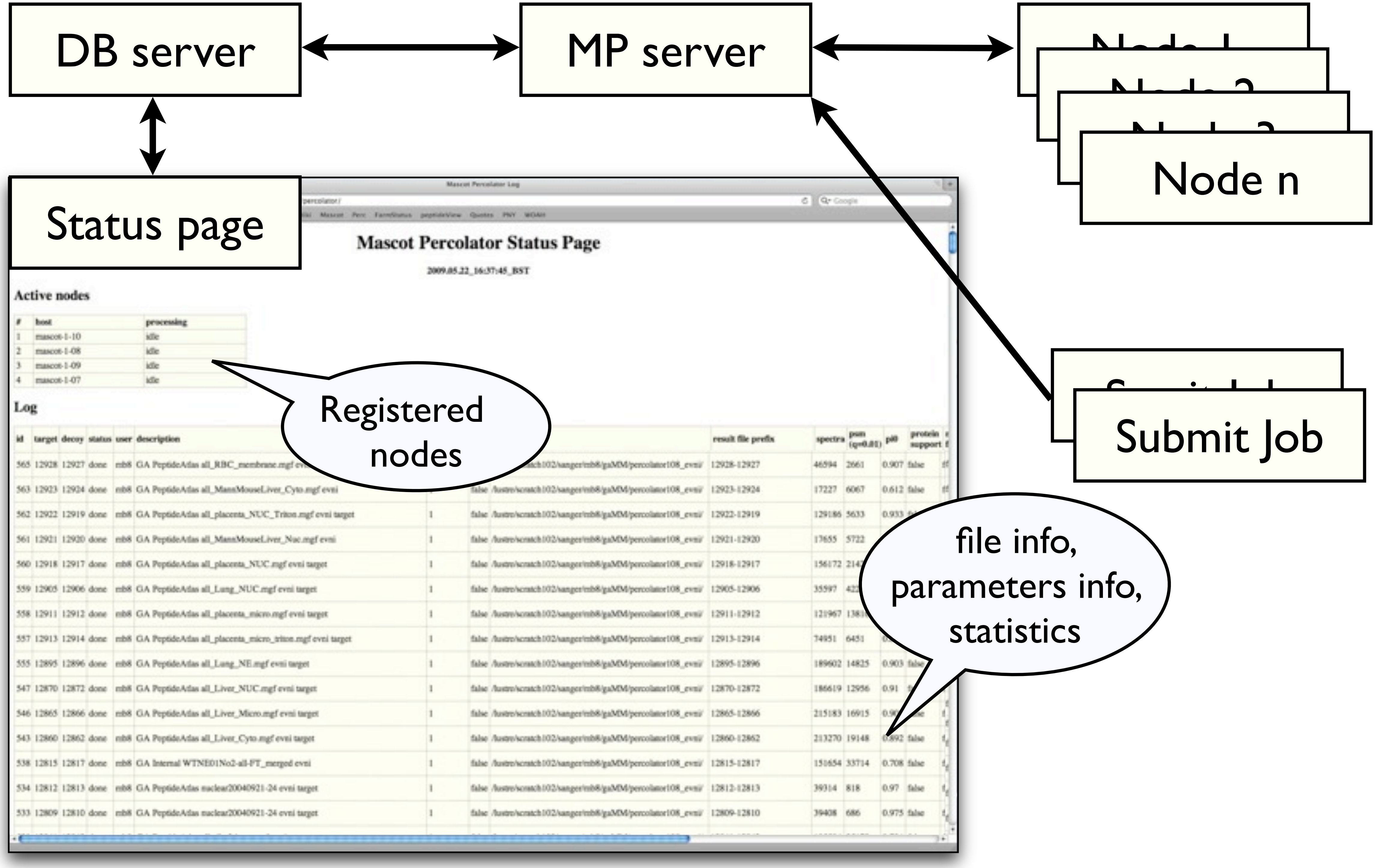
p = PEP = 0.05; MIT 13
p = PEP = 0.01; MIT 20

FDR

Score:
-10log₁₀(PEP)

Expect = PEP

Distributing Mascot Percolator Jobs



Mascot Percolator Status Page

2009.05.22_16:37:45_BST

Active nodes

#	host	processing
1	mascot-1-10	idle
2	mascot-1-08	idle
3	mascot-1-09	idle
4	mascot-1-07	idle

Log

id	target	decoy	status	user	description	[illegible]	[illegible]	[illegible]	result file prefix	spectra	psm (q=0.01)	pi0	protein support
565	12928	12927	done	mb8	GA PeptideAtlas all_RBC_membrane.mgf ev	[illegible]	[illegible]	atch102/sanger/mb8/gaMM/percolator108_evni/	12928-12927	46594	2661	0.907	false
563	12923	12924	done	mb8	GA PeptideAtlas all_MannMouseLiver_Cyto.mgf evni	[illegible]	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12923-12924	17227	6067	0.612	false
562	12922	12919	done	mb8	GA PeptideAtlas all_placenta_NUC_Triton.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12922-12919	129186	5633	0.933	[illegible]
561	12921	12920	done	mb8	GA PeptideAtlas all_MannMouseLiver_Nuc.mgf evni	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12921-12920	17655	5722	[illegible]	[illegible]
560	12918	12917	done	mb8	GA PeptideAtlas all_placenta_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12918-12917	156172	214	[illegible]	[illegible]
559	12905	12906	done	mb8	GA PeptideAtlas all_Lung_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12905-12906	35597	42	[illegible]	[illegible]
558	12911	12912	done	mb8	GA PeptideAtlas all_placenta_micro.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12911-12912	121967	1381	[illegible]	[illegible]
557	12913	12914	done	mb8	GA PeptideAtlas all_placenta_micro_triton.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12913-12914	74951	6451	0.	[illegible]
555	12895	12896	done	mb8	GA PeptideAtlas all_Lung_NE.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12895-12896	189602	14825	0.903	false
547	12870	12872	done	mb8	GA PeptideAtlas all_Liver_NUC.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12870-12872	186619	12956	0.91	f
546	12865	12866	done	mb8	GA PeptideAtlas all_Liver_Micro.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12865-12866	215183	16915	0.9	[illegible]
543	12860	12862	done	mb8	GA PeptideAtlas all_Liver_Cyto.mgf evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12860-12862	213270	19148	0.892	false
538	12815	12817	done	mb8	GA Internal WTNE01No2-all-FT_merged evni	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12815-12817	151654	33714	0.708	false
534	12812	12813	done	mb8	GA PeptideAtlas nuclear20040921-24 evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12812-12813	39314	818	0.97	false
533	12809	12810	done	mb8	GA PeptideAtlas nuclear20040921-24 evni target	1	false	/lustre/scratch102/sanger/mb8/gaMM/percolator108_evni/	12809-12810	39408	686	0.975	false

Acknowledgements

- Lukas Käll (Stockholm University)   WOE 3:10 pm
- Erik Deutsch (Institute of Systems Biology)
- Jyoti Choudhary & Tim Hubbard (Sanger)
- Members of team 17 and 119, in particular Lu Yu (Sanger)
- Matrix Science
- Wellcome Trust for funding

Lukas Käll - WOE 3:10 pm

Markus Brosch - MPB 063 - mb8@sanger.ac.uk

Mascot Percolator Website:

http://www.sanger.ac.uk/Software/analysis/MascotPercolator/

http://tinyurl.com/mascotpercolator/