BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(8)

Comparison of methods for identifying differentially expressed genes across multiple conditions from microarray data

Yuande Tan[1] & Yin Liu[2]*

[1]School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, United States of America; [2]Department of Neurobiology and Anatomy, University of Texas Medical School at Houston, Houston, Texas, United States of America; Yin Liu- Email: yin.liu@uth.tmc.edu; *Corresponding author

Selected publications from Asia Pacific Bioinformatics Network (APBioNet) 10th International Conference on Bioinformatics (InCoB 2011), Malaysia, November 30 to December 02, 2011

Received November 23, 2011; Accepted December 07, 2011; Published December 21, 2011

Abstract:
Identification of genes differentially expressed across multiple conditions has become an important statistical problem in analyzing large-scale microarray data. Many statistical methods have been developed to address the challenging problem. Therefore, an extensive comparison among these statistical methods is extremely important for experimental scientists to choose a valid method for their data analysis. In this study, we conducted simulation studies to compare six statistical methods: the Bonferroni (B-) procedure, the Benjamini and Hochberg (BH-) procedure, the Local false discovery rate (Localfdr) method, the Optimal Discovery Procedure (ODP), the Ranking Analysis of F-statistics (RAF), and the Significant Analysis of Microarray data (SAM) in identifying differentially expressed genes. We demonstrated that the strength of treatment effect, the sample size, proportion of differentially expressed genes and variance of gene expression will significantly affect the performance of different methods. The simulated results show that ODP exhibits an extremely high power in indentifying differentially expressed genes, but significantly underestimates the False Discovery Rate (FDR) in all different data scenarios. The SAM has poor performance when the sample size is small, but is among the best-performing methods when the sample size is large. The B-procedure is stringent and thus has a low power in all data scenarios. Localfdr and RAF show comparable statistical behaviors with the BH-procedure with favorable power and conservativeness of FDR estimation. RAF performs the best when proportion of differentially expressed genes is small and treatment effect is weak, but Localfdr is better than RAF when proportion of differentially expressed genes is large.

Background:

Identifying genes differentially expressed across multiple conditions is one of the major goals in many microarray experiments. Because microarray data usually consist of ten thousand or more of genes, they are beyond the scope of conventional statistical methods for single tests **[1].** To address the challenging statistical problem rising in the large-scale data, a variety of multiple-testing procedures have been adopted to microarray data analysis. Some of these procedures, such as the Bonferroni procedure, control the family-wise-error-rate (FWER). The other multiple-testing procedures, such as the Benjamini and Hochberg (BH) procedure, control the false discovery rate (FDR) **[2].** Another challenging aspect of microarray data analysis is to choose appropriate test statistics for different types of responses and covariates obtained from the datasets. The commonly used statistics including the t-statistic and the F-statistic were originally designed for performing a single test but are not appropriate for large-scale data analysis. This motivated the development of many new statistics that borrow information across multiple genes for identifying differentially expressed genes, including a modified t-statistic used in the Significance Analysis of Microarrays

(SAM) approach **[3],** the regularized t-test **[4],** and the shrunken F-test **[5].** More recently, Storey et al. **[6]** developed a new approach based on the Optimal Discovery Procedure (ODP), which aims to maximize the expected number of true positive genes for each fixed level of expected false positives.

Therefore, at current stage, it seems even more important to effectively compare existing microarray data analysis methods than to develop new ones, simply because experimental scientists are faced with a seemingly endless choice of methods for their data analyses **[7].** There have been some studies done in this area, while the simulated or real microarray data were utilized to compare a list of methods **[8],** but we note the comparison could indicate neither the power of gene identification, nor the FDR estimation accuracy of the methods. Ge et al. **[9]** and Dudoit et al. **[10]** compared a set of multiple hypothesis testing methods using theoretical analysis and simulated data. A recent study applied the same FDR methodology to the gene-ranking methods including SAM [3], Shrunken F-test **[5],** Localfdr **[11],** ODP **[12]** and empirical Bayesian method **[13],** and demonstrated that the ODP method identified significantly more genes than the other methods. However, one natural concern is whether a statistical method achieves such a high power at the tradeoff of identifying too many false positives. Unfortunately, this concern was not addressed in the study. Regarding to the datasets used for method comparison, Pearson **[7]** and Astrand et al **[14]** used a golden spike dataset **[15],** but it has been criticized for containing artifact factors **[16].** Another most recent study introduced the controlled fold changes into the real data **[17]** and successfully demonstrated the effects of fold changes and the sample sizes on the performance of different methods. However, the number of data scenarios investigated in this study was limited. To objectively and comprehensively evaluate different methods, we need to extend our scope to multiple data scenarios consisting of different levels of treatment effect, proportion of differentially expressed genes, sample size, and noise. To this end, simulation study seems to be the most appropriate way to achieve this goal because the set of truly differentially expressed genes across different conditions is known and different data scenarios to be studied can be controlled.

In this study, our comparison focuses on the existing methods for identifying genes differentially expressed among multiple conditions. We compare the results from three leading methods including the ODP procedure **[2],** the SAM approach **[3]** and the Localfdr method **[4, 15]** along with the Ranking Analysis of F-statistics (RAF) method we developed previously **[18].** The original Localfdr method has been extended by McLachlan et al **[19]** which used a Z-statistics and has been applied on multiple-class microarray data. Therefore, we chose this extended version of Localfdr method in our comparison. For the ODP, RAF and SAM methods, we kept their original implementation by combining the original test statistics associated with these methods and the original FDR methods used by them. This makes the comparison results mostly useful for the biologists since many of them would prefer using the original implementation than modifying either the test statistics or the FDR method for microarray data analysis. We also included two multiple-testing procedures, the Bonferroni (B-) procedure and BH-procedure in our comparison by coupling them with the traditional F-statistics due to the fact that they are most typical and widely-used ones among multiple-testing procedures.

Methodology:

Microarray Dataset Simulation

We obtained two real microarray datasets of 3,770 genes that were expressed among four groups with each having 6 biological replicates **[18, 20].** We estimated about 10-15% genes are differentially expressed across multiple conditions of stroke susceptibility in the datasets. We first used one group-mean and error variance for each gene to simulate a dataset of pure noise. Then, treatment effect τ = AU was randomly assigned to a proportion of genes, where A is set as the maximum treatment effect level and U is a uniform random variable. Therefore, treatment effects in differentially expressed (DE) genes are uniformly distributed in $0 \le \tau \le A$. We generating our simulation scenarios by setting two proportions of DE genes (10% and 20%), two treatment effect levels (τ = 100U and 200U), three replication levels (4, 6, and 12 replicate samples), and two levels of expression noise variances (large and small) according to the real datasets **[18, 20].** Combining these different settings, we considered totally six simulation scenarios. Scenario 1: 10% DE genes, 100U, 6 samples, large variance; Scenario 2: 20% DE genes, 100U, 6 samples, large variance; Scenario 3: 20% DE genes, 200U, 6 samples, large variance; Scenario 4: 20% DE genes, 100U, 12 samples, large variance; Scenario 5: 20% DE genes, 200U, 4 samples, large variance; Scenario 6: 20% DE genes, 100U, 6 samples, small variance. For each scenario, we generate 30 datasets using a normal pseudorandom generator. With these simulation scenarios, we examined the responses of these statistical methods to treatment effects and the proportion of DE genes, the impacts of sample sizes on the performance of these statistical methods, and the robustness of these statistical methods to gene expression noises.

Metrics for Methods Comparison

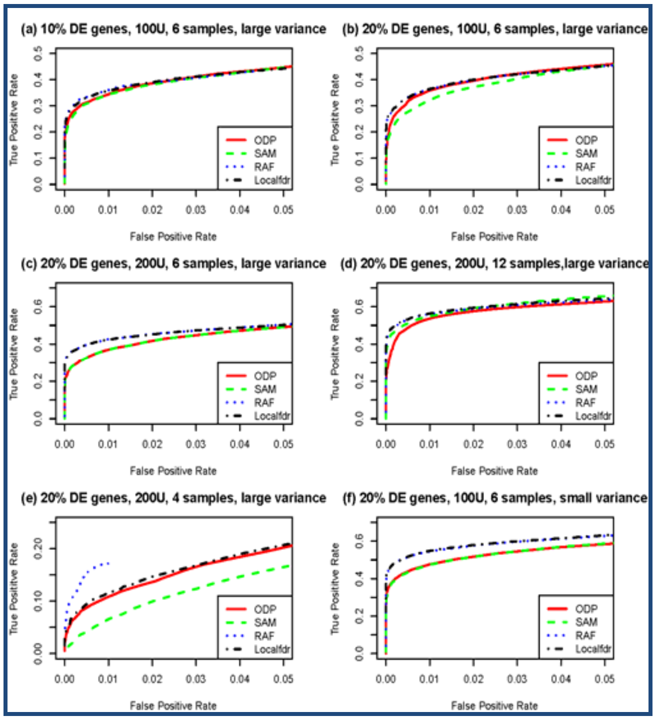
We applied six statistical methods (B-procedure, BH procedure, Localfdr, RAF, ODP and SAM) to simulated datasets. We first summarized the comparison results in Receiver Operating Characteristics (ROC) curves and computed the area under the curve (AUC) up to 0.01, 0.05 and 0.1 false positive rates. To examine the FDR estimation accuracy of these methods, we set the cutoff α=0.05 for Bonferroni procedure and BH procedure. For all other methods, FDR values at the level of 0.04 < FDR ≤ 0.05 were used as the cutoffs. We collected the number of identified positives (N_p), the estimated (N_{EFP}) and the true (N_{TFP}) number of false positives, and the differences ($d = N_{EFP} - N_{TFP}$) between N_{EFP} and N_{TFP} across 30 simulation datasets under each scenario. Then we calculated means and standard deviations of N_p, N_{EFP} and N_{TFP} for each method. We also measured the conservativeness of FDR estimation of a method by the conservative degree C ($d \ge 0$) **[18],** defined as the proportion of simulations with $d \ge 0$ at a given FDR cutoff as given in equation 1 **(see supplementary material).**

Discussion:

Comparison of sensitivities and specificities of different methods

The comparison results were summarized in Receiver Operating Characteristics (ROC) curves **(Figure 1).** We also computed the area under the curve (AUC) up to different false positive rates (**Table 2 see supplementary material**). We find the rank of these methods with respect of their performance are consistent across different scenarios, with the Localfdr and RAF performing overall better than the ODP and SAM in all six scenarios. In the case of large sample size (scenario 4), SAM performs the best among all four methods at the FPR level of 0.1, but it performs the worst compared to other methods when the sample size is small (scenario 5), indicating its performance is sensitive to the sample size (**Table 2 see supplementary material**). In all other scenarios, the Localfdr and RAF take turns to have the highest AUC, depending on the FPR cutoff values up to which the AUC is calculated.

Responses of different methods to the changes in proportion of differentially expressed (DE) genes


(a) 10% DE genes, 100U, 6 samples, large variance

(b) 20% DE genes, 100U, 6 samples, large variance

(c) 20% DE genes, 200U, 6 samples, large variance

(d) 20% DE genes, 200U, 12 samples,large variance

(e) 20% DE genes, 200U, 4 samples, large variance

(f) 20% DE genes, 100U, 6 samples, small variance

Figure 1: ROC curves of the four methods for six simulated datasets.

The comparison between scenario 1 and scenario 2 demonstrates the response of different methods to the changes in proportion of DE genes **(Table 1 and 2, see supplementary material).** The ODP approach displays highest power with largest value of N_p. At the first glance, we were concerned the results are not consistent with that shown in ROC curves, since the AUC of ODP was not shown significantly better than any of the methods we compared. To explore the seemingly discrepancy, we examined the FDR estimation of different methods (see methodology section). We found that ODP underestimated false positives on average, while other methods tend to overestimate the number of false positives **(Table 1 see supplementary material).** Therefore, the high power of ODP at the estimated FDR cutoff of 0.05 could be actually achieved at the true FDR value much higher than 0.05. The B-procedure did not obtain any false positives but had the lowest power among these six methods. In both scenarios, the BH-procedure is completely conservative in FDR estimation, but has about two times of findings more than the B-procedure. RAF identifies more genes with a higher degree of conservativeness than SAM in both scenarios. With the same level of conservativeness as RAF, the Localfdr method identifies fewer genes when there are 10% DE genes, but demonstrates higher power than RAF when there are 20% DE genes, which is consistent with the results shown in **Figure 1** and **Table 2, (see supplementary material)**.

Sensitivities of different methods to treatment effects

We investigated the responses of methods to different levels of treatment effects (scenarios 2 and 3 in **Table 1 see supplementary material**). The B-procedure and the BH-procedure always have conservativeness of 100%, indicating that their FDR estimation are absolutely conservative and their conservativeness are insensitive to treatment effects. The ODP shows high power and low degree of conservativeness compared to the other methods in both scenarios. SAM demonstrates lower power and lower degree of conservativeness than Localfdr and RAF under both treatment effects. So both ODP and SAM methods are insensitive to treatment effects. When τ =100 u with 20% DE genes (scenario 2), Localfdr has a higher power than RAF, but they both have similar power and conservativeness with strong treatment effect (scenario 3). Therefore, both Localfdr and RAF methods may be sensitive to treatment effects.

Impact of sample size on the performance of the statistical methods

The performance of methods was compared in datasets with different sample sizes when other conditions were fixed (**scenarios 4 and 5 in Table 1 see supplementary material).** ODP still shows its highest power with extremely poor conservativeness compared to the other methods. In the samples of 4 replicates, SAM has the poorest power among all the methods. The BH-procedure, Localfdr and RAF have similar powers but RAF possesses of higher degrees of conservativeness than the other two methods and has the best performance when the sample size is small.

Robustness of different methods to noise in microarray datasets

We also assessed the robustness of methods to different levels of expression variances of genes. The first dataset has a large expression variance (~ 10^5) and the second one has a small variance (~10^4) while other conditions are fixed. Based on the comparison **(Scenarios 2 and 6 in Table 1 and 2, see supplementary material**), we found in datasets with small expression variances, all methods obviously improve their powers while SAM and B-procedure have most significant changes. ODP shows the highest power but poor conservativeness in both scenarios, while the B-procedure has smallest number of findings with conservativeness of 100%. The BH-procedure, Localfdr and RAF have similar power and high degree of conservativeness in both scenarios.

Conclusion:

In this study, we have evaluated and compared six statistical methods: the B-procedure, the BH-procedure, the Localfdr, the ODP, the RAF, and the SAM method. Our study shows that the B-procedure is over conservative but has an extremely low power in any scenario; on the other hand, the ODP method displays an extremely high power but low degree of conservativeness in all cases. Therefore the B-procedure would be selected only if we prefer very conservative finding, while ODP would be selected if the power is a sole criterion for the DE genes identification. SAM is sensitive to the quality of microarray data. It shows better performance in the data with small noise variances, but works poor in the data with large noise variances or with sample sizes smaller than 6. Localfdr and RAF are two stable methods with high power and high degree of conservativeness in most situations we tested. RAF is robust to find positive genes of interest in the scenarios of sample sizes ≤ 6, weak treatment effects, and/or low proportion of genes differentially expressed. Localfdr outperforms RAF when proportion of differentially expressed genes is larger and treatment effect is stronger. The BH-procedure performs very similar with Localfdr and RAF in most cases except it has lower power when the sample size is small. We expect the results of this simulation study will provide a critical guideline for the biologists to make the choices of methods for microarray data analysis under different experimental scenarios.

Acknowledgement:
This work is supported in part by NIH grant R01 LM010022 and the seed grant from the University of Texas health Science Center at Houston. The authors thank Dr. Myriam Fornage for kindly providing the microarray datasets as the basis for the simulation study.

References:

[1] Efron BJ *Am Stat Assoc*. 2004 **99**: 96
[2] Benjamini Y *et al. JRStatist Soc B*. 1995 **57**: 289
[3] Tusher VG *et al. Proc Natl Acad Sci U S A*. 2001 98: 5116 [PMID: 11309499]
[4] Baldi P & Long AD. *Bioinformatics*. 2001 17: 509 [PMID: 11395427]
[5] Cui X *et al. Biostatistics*. 2005 6: 59 [PMID: 15618528]
[6] Storey JD *et al. Biostatistics*. 2007 8: 414 [PMID: 16928955]
[7] Pearson RD. *BMC Bioinformatics*. 2008 **9**: 164 [PMID: 18366762]
[8] Broberg P. *Genome Biol*. 2003 **4**: R41 [PMID: 12801415].
[9] Ge Y *et al. TEST*. 2003 **12**: 1
[10] Dudoit S *et al. Statistical Science*. 2003 18: 71
[11] Efron B *et al. Journal of the American Statistical Association*. 2001 **96**: 1151
[12] Storey JD. *JRStatist Soc B*. 2007 69: 347
[13] Lonnstedt I & Speed T. *Statistica Sinica*. 2002 **12**: 31
[14] Astrand M *et al. BMC Bioinformatics*. 2008 **9**: 156. [PMID: 18366694]
[15] Choe SE *et al. Genome Biol*. 2005 **6**: R16. [PMID: 15693945]
[16] Dabney AR & Storey JD. *Genome Biol*. 2006 **7**: 401. [PMID: 16563185]
[17] Dozmorov MG *et al. PLoS One*. 2010 9: e12657 [PMID: 20844739]
[18] Tan YD *et al. BMC Bioinformatics*. 2008 **9**: 142. [PMID: 18325100]
[19] McLachlan GJ *et al. Bioinformatics*. 2006 **22**: 1608[PMID: 16632494]
[20] Fornage M *et al. Physiol Genomics*. 2003 **15**: 75[PMID: 12902546]

Edited by TW Tan
Citation: Tan & Liu . Bioinformation 7(8): 400-404 (2011)

Supplementary material:

$$C(d \geq 0) = \sum_{k=1}^{N_{FDR}} \mathrm{I}(d_k \geq 0) / N_{FDR}$$

→ (1)

Where N_{FDR} is the total number of FDR values in the interval $0.04 < \mathrm{FDR} \leq 0.05$ across 30 simulated datasets. For each FDR value, we obtained the d value accordingly as $d = N_{EFP} - N_{TFP}$. I is an indicator of the k^{th} d value d_k, where I= 1 if $d_k \geq 0$, and I = 0 otherwise.

Table 1: Comparison among the statistical methods in identifying differentially expressed (DE) genes when estimated FDR < 0.05 under different scenarios. The numbers in the parenthesis indicate the standard deviation of Np, NEFP and NTFP (see Methodology section).

	Scenario 1: 10% DE, 100U, 6 samples, large variance						Scenario 2: 20% DE, 100U, 6 samples, large variance					
	N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$			N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$		
				Mean	Var	C(d>0)				Mean	Var	C_{ab}(d>0)
B-procedure	55.3(7.6)	0(0)	0(0)	0	0	100.0	111.9(7.3)	0(0)	0(0)	0	0	100.0
BH-procedure	94.6(15.2)	4.7(0.8)	0.7(0.9)	1.0	0.7	100.0	220.1(18.5)	11.0(2.6)	4.3(2.6)	6.4	5.8	100.0
Localfdr	92.8(13.5)	4.3(0.7)	0.8(1.1)	3.5	1.8	100.0	222.7(17.8)	10.9(0.8)	5.5(2.4)	5.4	5.6	100.0
ODP	128.0(14.9)	5.8(0.9)	10.1(3.6)	-4.6	9.2	10.4	276.7(46.0)	12.4(2.2)	18.5(5.3)	-6.1	15.6	7.3
RAF	107.5(15.9)	4.9(0.9)	2.4(1.6)	2.5	1,9	100.0	213.3(18.1)	9.8(1.2)	3.9(2.2)	5.9	4.6	100.0
SAM	99.3(14.4)	4.6(0.8)	3.5(2.0)	1.8	2.0	52.9	204.9(14.0)	9.1(0.8)	9.1(2.1)	1.7	0.8	47.4
	Scenario 3: 20% DE, 200U, 6 samples, large variance						Scenario 4: 20% DE, 200U, 12 samples, large variance					
	N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$			N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$		
				Mean	Var	C(d>0)				Mean	Var	C(d>0)
B-procedure	175.0(14.4)	0(0)	0(0)	0	0	100.0	184.2(7.1)	0(0)	0(0)	0	0	100.0
BH-procedure	286.2(20.4)	14.3(1.0)	6.9(3.4)	7.3	8.6	100.0	315.7(21.3)	15.2(1.3)	7.1(3.1)	8.5	7.2	100.0
Localfdr	289.9(16.3)	13.8(1.1)	7.8(3.3)	6.0	8.7	100.0	332.7(16.4)	17.3(1.0)	7.7(2.1)	9.8	4.9	100.0
ODP	304.7(20.5)	15.8(1.5)	28.5(7.8)	-11.5	40.9	8.4	343.2(18.3)	17.2(1.6)	30.1(8.9)	-12.3	45.3	7.6
RAF	286.9(19.7)	13.9(1.0)	8.1(2.3)	5.8	4.1	100.0	301.2(17.2)	14.6(1.1)	7.9(2.4)	8.2	5.1	100.0
SAM	244.5(16.6)	11.7(1.4)	11.7(3.3)	-1.9	2.5	33.5	329.8(16.4)	15.2(1.8)	13.9(4.1)	2.7	4.6	64.3
	Scenario 5: 20% DE, 200U, 4 samples, large variance						Scenario 6: 20% DE, 100U, 6 samples, small variance					
	N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$			N_p	N_{EFP}	N_{TFP}	$d = N_{EFP} - N_{TFP}$		
				Mean	Var	C(d>0)				Mean	Var	C(d>0)
B-procedure	12.9(3.7)	9(0)	0(0)	0	0	100.0	233.8(13.5)	0(0)	0(0)	0	0	100.0
BH-procedure	33.6(5.9)	1.7(0.3)	1.1(0.8)	1.6	0.1	0.8	380.3(28.2)	19.0(1.4)	7.4(3.0)	11.6	5.3	100.0
Localfdr	31.7(6.5)	1.5(0.5)	1.1(0.9)	0.7	0.4	66.7	380.1(17.6)	18.1(1.1)	8.0(2.4)	10.1	5.5	100.0
ODP	75.4(9.6)	3.4(0.5)	8.6(2.3)	-5.2	4.7	0.0	446.8(25.3)	20.1(1.9)	31.8(9.3)	-11.7	66.3	6.5
RAF	31.4(9.4)	1.4(0.5)	0.6(0.8)	0.8	0.3	78.6	382.3(19.5)	18.6(1.0)	8.8(3.6)	9.8	9.9	100.0
SAM	6.3(2.7)	0.0(0.0)	0.7(1.0)	-0.7	1.0	60.0	410.3(23.9)	18.6(2.0)	16.9(5.7)	4.0	8.2	61.5

Table 2: ROC scores computed as the areas under ROC curves. FPR, False Positive Rate. Scenario 1: 10% DE genes, 100U, 6 samples, large variance; Scenario 2: 20% DE genes, 100U, 6 samples, large variance; Scenario 3: 20% DE genes, 200U, 6 samples, large variance; Scenario 4: 20% DE genes, 100U, 12 samples, large variance; Scenario 5: 20% DE genes, 200U, 4 samples, large variance; Scenario 6: 20% DE genes, 100U, 6 samples, small variance. Note: In scenario 4, the AUC up to 0.05 and 0.1 FPR for the RAF method cannot be calculated, since the RAF cannot reach to those high FPR levels under this scenario.

FPR	Methods	Areas under ROC curves					
		Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6
0.01	ODP	3.0	3.0	3.2	4.6	0.8	4.3
	SAM	2.9	2.7	3.2	5.0	0.4	4.6
	RAF	3.2	3.2	3.9	5.2	1.4	5.1
	Localfdr	3.2	3.2	3.9	5.2	0.9	5.1
0.05	ODP	19.3	19.7	20.9	28.3	7.2	25.8
	SAM	19.0	18.6	20.9	29.5	5.2	26.8
	RAF	19.5	19.7	22.7	29.2	-	28.8
	Localfdr	19.4	19.8	22.8	29.6	7.5	28.6
0.1	ODP	43.3	44.2	47.4	60.9	19.2	56.8
	SAM	42.9	43.3	47.4	63.9	15.6	60.0
	RAF	43.3	44.4	48.8	62.5	-	61.4
	Localfdr	42.8	44.5	49.2	63.1	19.3	61.7