Methods for DMR analysis

- B. Data defined regions
  - *ProbeLasso* R package
    - https://bioconductor.org/packages/release/bioc/html/ChAMP.html
  - *DMRcate* R package
    - http://bioconductor.org/packages/release/bioc/html/DMRcate.html
  - bumphunting implemented in *minfi* R package
    - https://www.bioconductor.org/help/course-materials/2015/BioC2015/methylation450k.html
- Procedures
  - (1) computes p-values for each CpG
  - (2) identifies regions in the genome enriched with consecutive small p-values

Simulation study

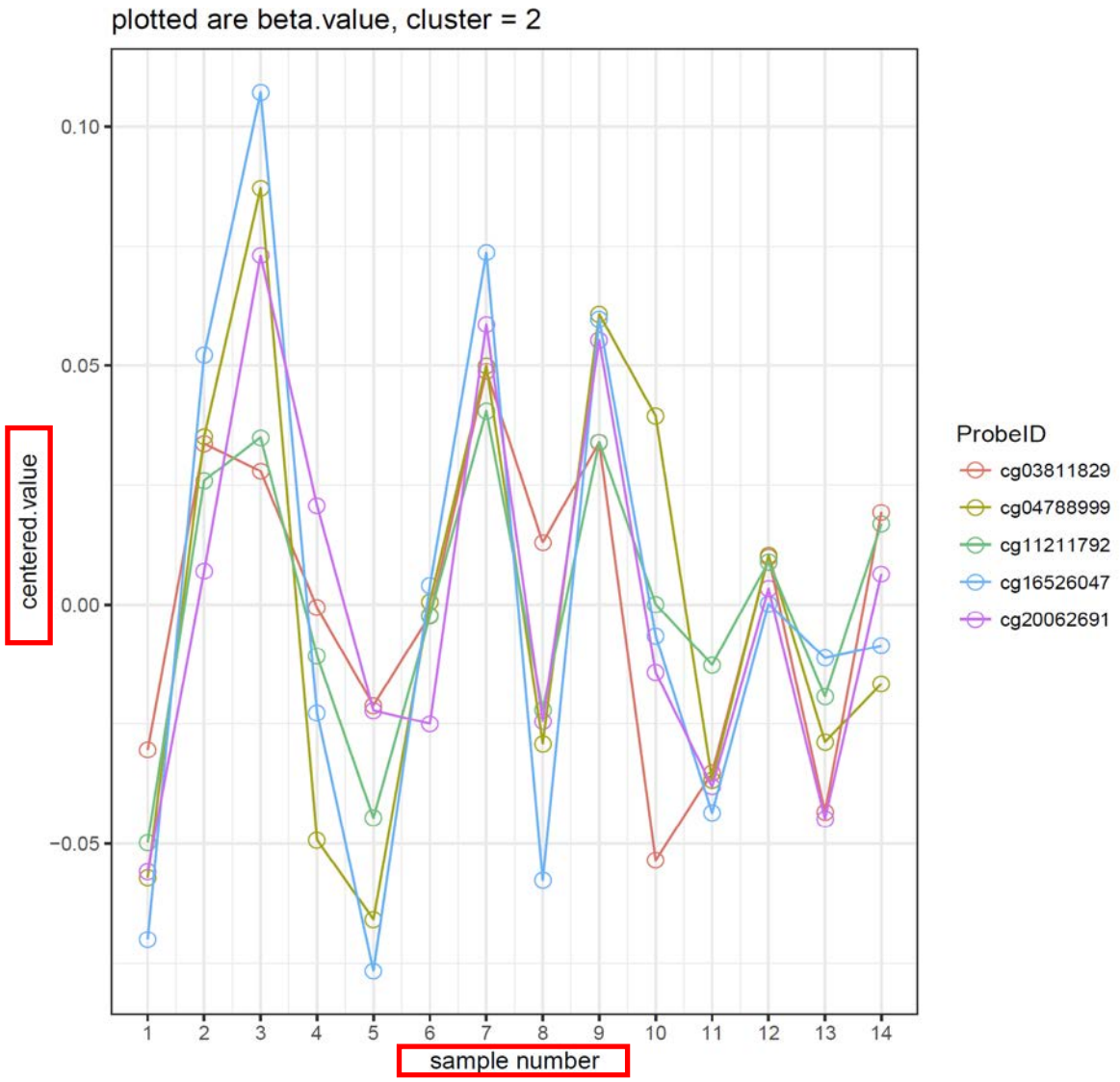
1. 14 samples of normal samples with similar ages (GSE41169)
2. A-clustering software identified 3063 co-methylated clusters

- the clustering is performed by cycling through the sites, ordered by location, and merging together neighboring clusters (e.g. those within 200bp) if the distance measure (e.g. 1- spearman correlation) between them is smaller than a predefined threshold (e.g. 0.5)

A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure FREE

Tamar Sofer ✉, Elizabeth D. Schifano, Jane A. Hoppin, Lifang Hou, Andrea A. Baccarelli
Author Notes

Bioinformatics, Volume 29, Issue 22, 15 November 2013, Pages 2884–2891, https://doi.org/10.1093/bioinformatics/btt498
Published: 29 August 2013 **Article history** ▾


plotted are beta.value, cluster = 2
0.10
0.05
0.00
−0.05
centered.value
1
2
3
4
5
6
7
8
9
10
11
12
13
14
sample number
ProbeID
cg03811829
cg04788999
cg11211792
cg16526047
cg20062691

Simulation Study

3. Choose 500 random clusters
4. For each cluster, randomly divide samples into 2 groups
5. Compare group means, increase beta values in the group with higher mean by $\mu = \{0, 0.025, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4)$
6. Repeat 5 times

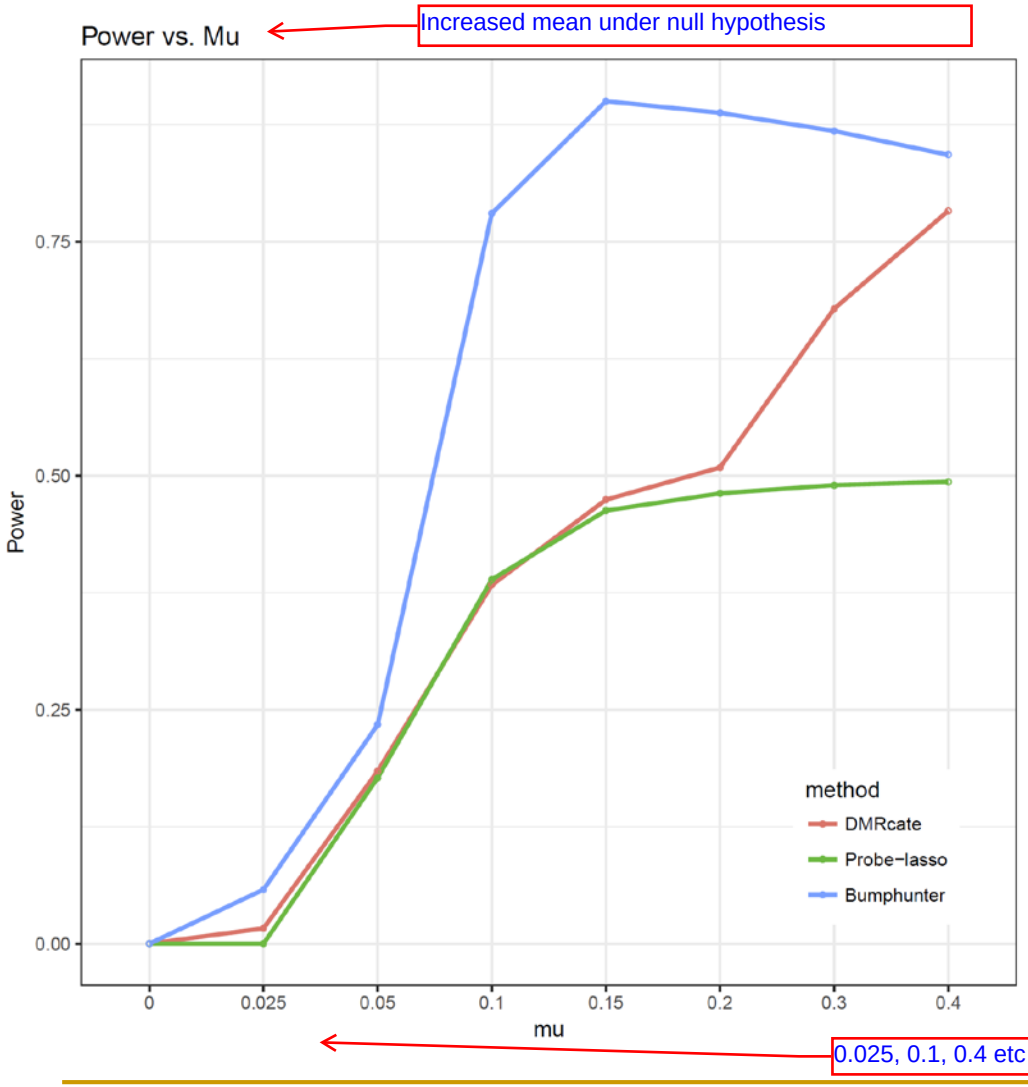
A total of 40 simulation datasets

= 8 values for μ $\times$ 5 repetitions

	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

(a) Confusion Matrix

Comparison of 3 type B methods - Power



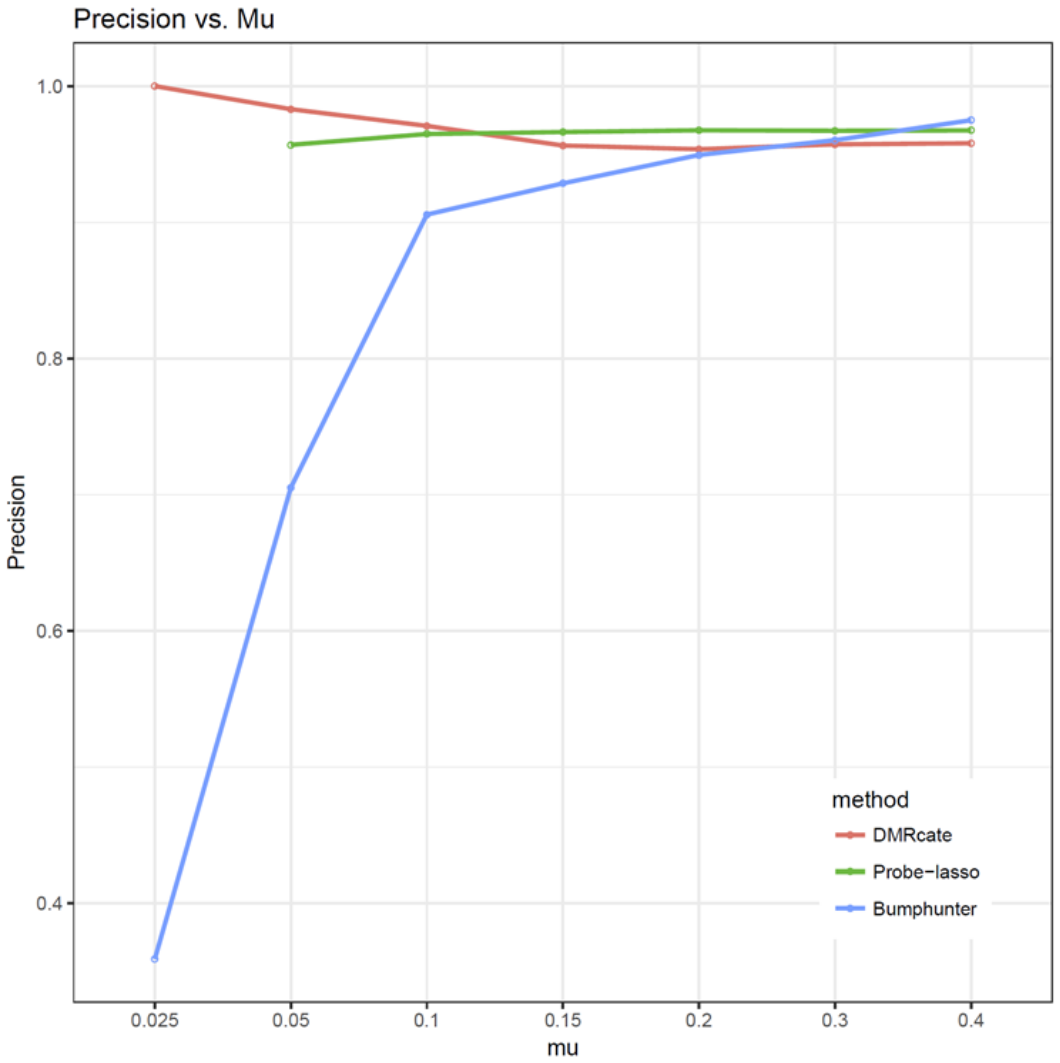
However, bumphunter is a permutation based method, very memory intensive.

Took about 8 min for 14 samples using parallel computing with 18 cores on a windows machine with 64G memory.

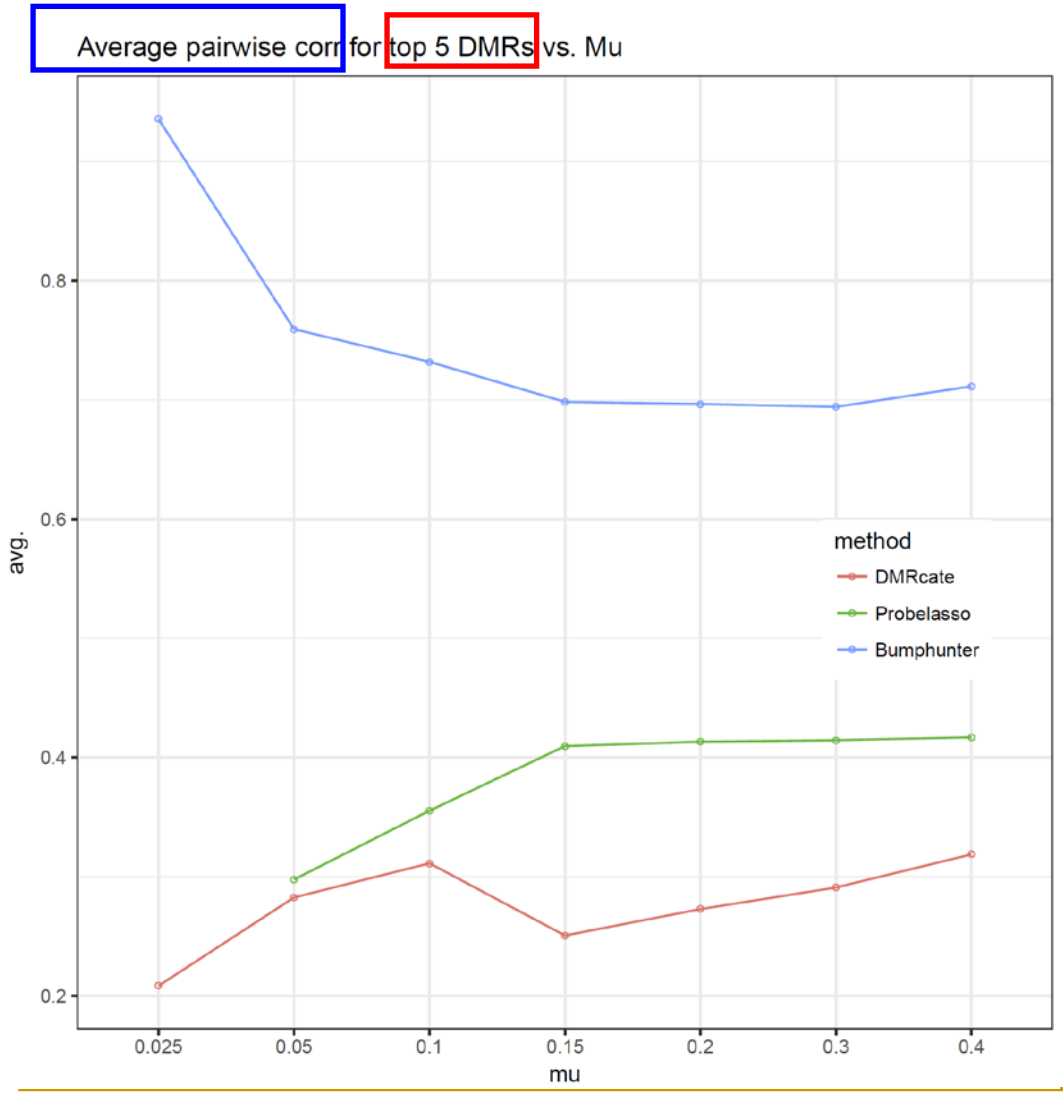
The other two methods took about 0.5 min without parallel computing.

Power = Probability (Predicted Positive | Actual Positive)

Comparison of 3 type B methods - Precision



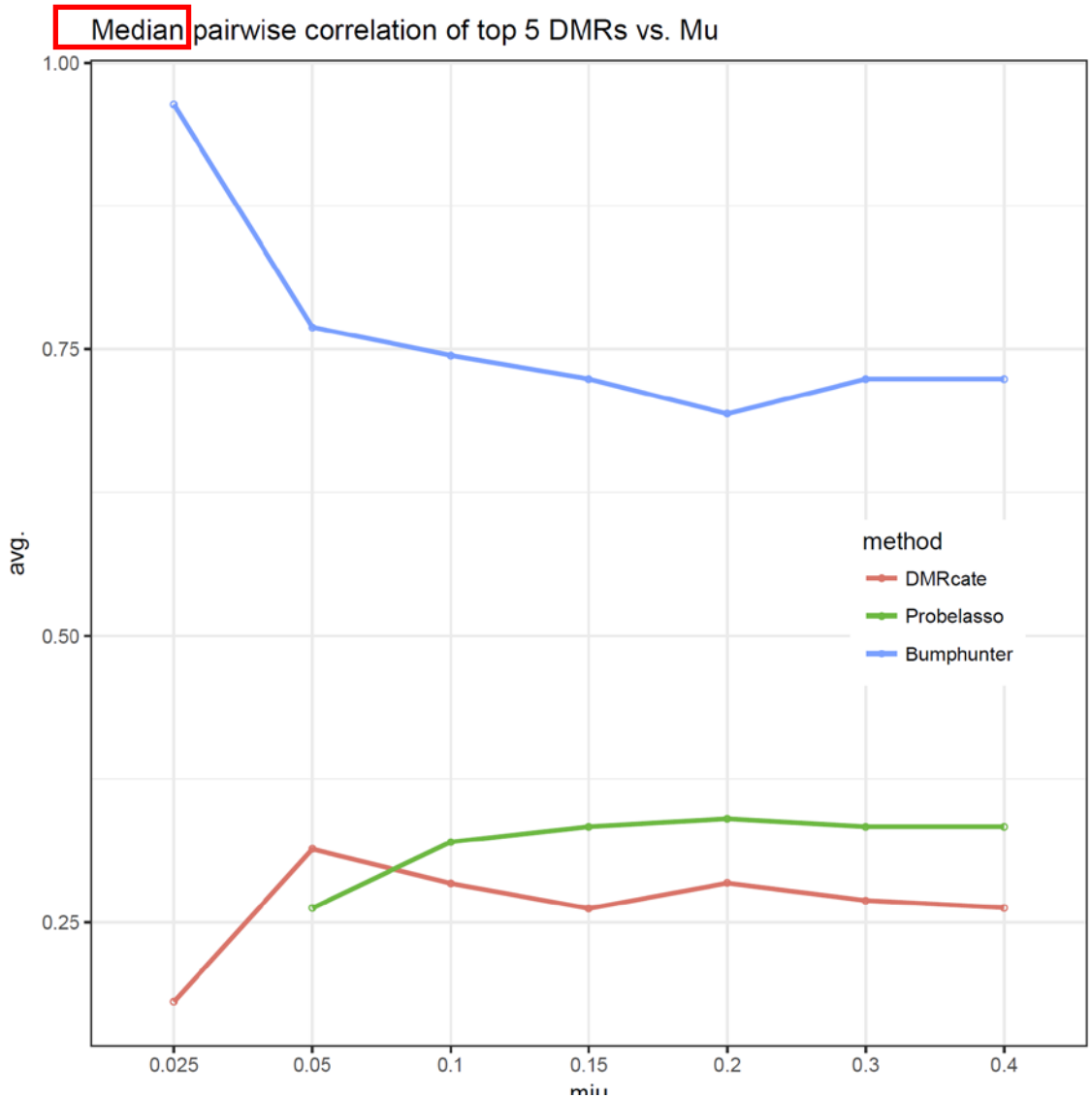
Precision = Probability (Actual Positive | Predicted Positive)

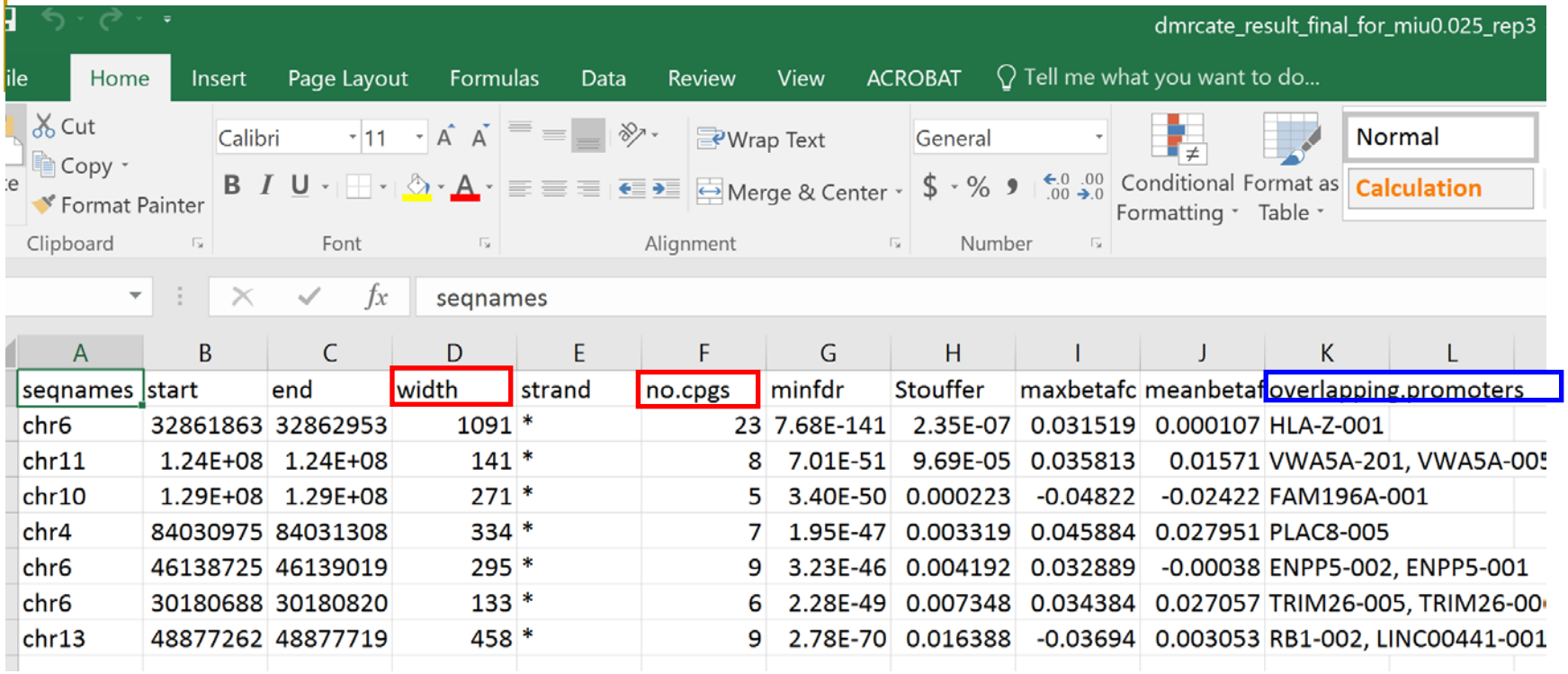


Take 5 most significant DMRs found by a method

extract pairwise correlations between cpgs within each DMR

Take average


Median pairwise correlation of top 5 DMRs vs. Mu
1.00
0.75
0.50
0.25
avg.
0.025
0.05
0.1
0.15
0.2
0.3
0.4
miu
method
DMRcate
Probelasso
Bumphunter



seqnames	start	end	width	strand	no.cpgs	minfdr	Stouffer	maxbetafc	meanbetaf	overlapping.promoters
chr6	32861863	32862953	1091	*	23	7.68E-141	2.35E-07	0.031519	0.000107	HLA-Z-001
chr11	1.24E+08	1.24E+08	141	*	8	7.01E-51	9.69E-05	0.035813	0.01571	VWA5A-201, VWA5A-005
chr10	1.29E+08	1.29E+08	271	*	5	3.40E-50	0.000223	-0.04822	-0.02422	FAM196A-001
chr4	84030975	84031308	334	*	7	1.95E-47	0.003319	0.045884	0.027951	PLAC8-005
chr6	46138725	46139019	295	*	9	3.23E-46	0.004192	0.032889	-0.00038	ENPP5-002, ENPP5-001
chr6	30180688	30180820	133	*	6	2.28E-49	0.007348	0.034384	0.027057	TRIM26-005, TRIM26-00
chr13	48877262	48877719	458	*	9	2.78E-70	0.016388	-0.03694	0.003053	RB1-002, LINC00441-001

DMRcate result – note that this only includes information on a subset of genes in the genome

Comments on type B methods

- All methods did well in terms of precision when effect size (μ) is moderate (i.e. > 0.1)
- Bumphunter had highest power, but was also most memory intensive
- DMRs detected by bumphunter also had the highest level of co-methylation
- Recommend bumphunter for datasets with moderate to large effect sizes (μ > 0.1)

Comments on type A & B approaches

- In contrast to gene expression, methylation regions are often poorly defined, so approaches in *B* (data defined regions) might have more power

- On the other hand, approaches in *A* (user defined regions) might be better suited for mega or integrative analysis