Supplementary material for:

annotatr: Associating genomic regions with genomic annotations

Raymond G. Cavalcante[1,*] and Maureen A. Sartor[1,2]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109; [2]Department of Biostatistics, University of Michigan

Supplementary Methods:

1. Construction of Genomic Annotations

We extracted CpG island (CGI), knownGene, and gap tables for hg19, hg38, mm9, and mm10 from the UCSC Table Browser (Karolchik *et al.*, 2004). We defined CpG shores as 2kb upstream and downstream of the CpG island boundaries and excluding CpG islands. CpG shelves are defined as an addition 2kb upstream and downstream of the furthest upstream and downstream boundaries of the CpG shores. Again, this excludes regions already annotated as CpG islands and CpG shores. See Supplementary Figure 1A for a detailed description of the UCSC CpG annotations. UCSC knownGenes (Hsu *et al.*, 2006) annotations include 1-5kb upstream of a TSS, promoter (<1Kb upstream of a TSS), 5'UTR, exons, introns, and 3'UTR. Intergenic annotations are taken to be the complement of the aforementioned knownGene annotations union the gaps tracks. We allow all UCSC knownGene annotations to overlap. See Supplementary Figure 1B for a detailed description of the UCSC knownGene annotations.

2. Benchmarking with microbenchmark

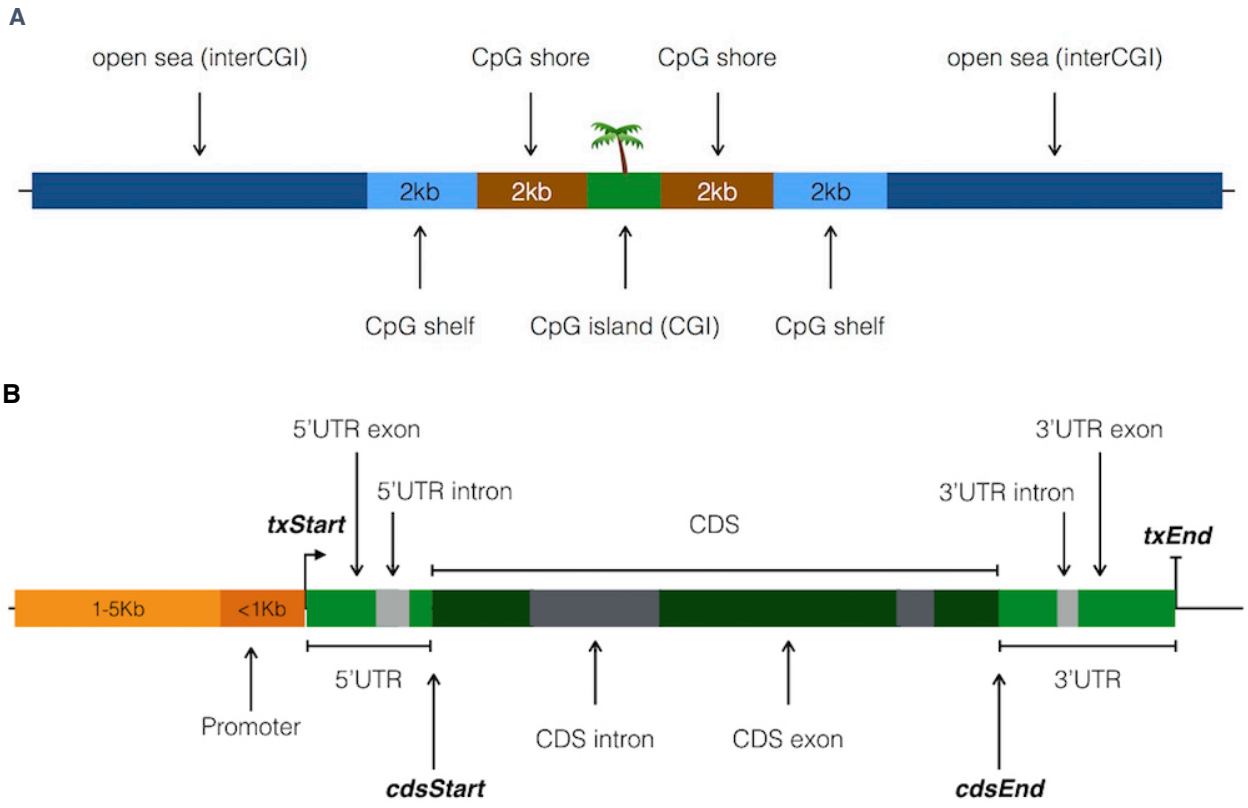
The microbenchmark R package (Mersmann, 2015) was used on four data sets to compare 10 runs of annotatr v2.0.0 and ChIPpeakAnno v3.4.1 (Zhu *et al.*, 2010). Benchmarks were run on a MacBook Pro with a 2.7Ghz Intel Core i7, 8GB RAM, and a solid-state hard drive. The four data sets, ranging in size from 27,000 to 25,000,000 lines, are:

1. A ~27,000 line ChIP-seq peak file from ENCODE for Pol2 in the Gm12878 cell line (ENCODE Consortium, 2012).
2. A ~365,000 line file of differential methylation tests resulting from methylSig (Park *et al.*, 2014) on GEO dataset GSE52945 (Figueroa *et al.*, 2010).
3. A ~4,000,000 line CpG methylation report from Bismark (Krueger & Andrews, 2011) on a whole genome bisulfite sequencing run (unpublished data).
4. A ~25,000,000 line file representing classification of genomic ranges into methylation types (unpublished data).

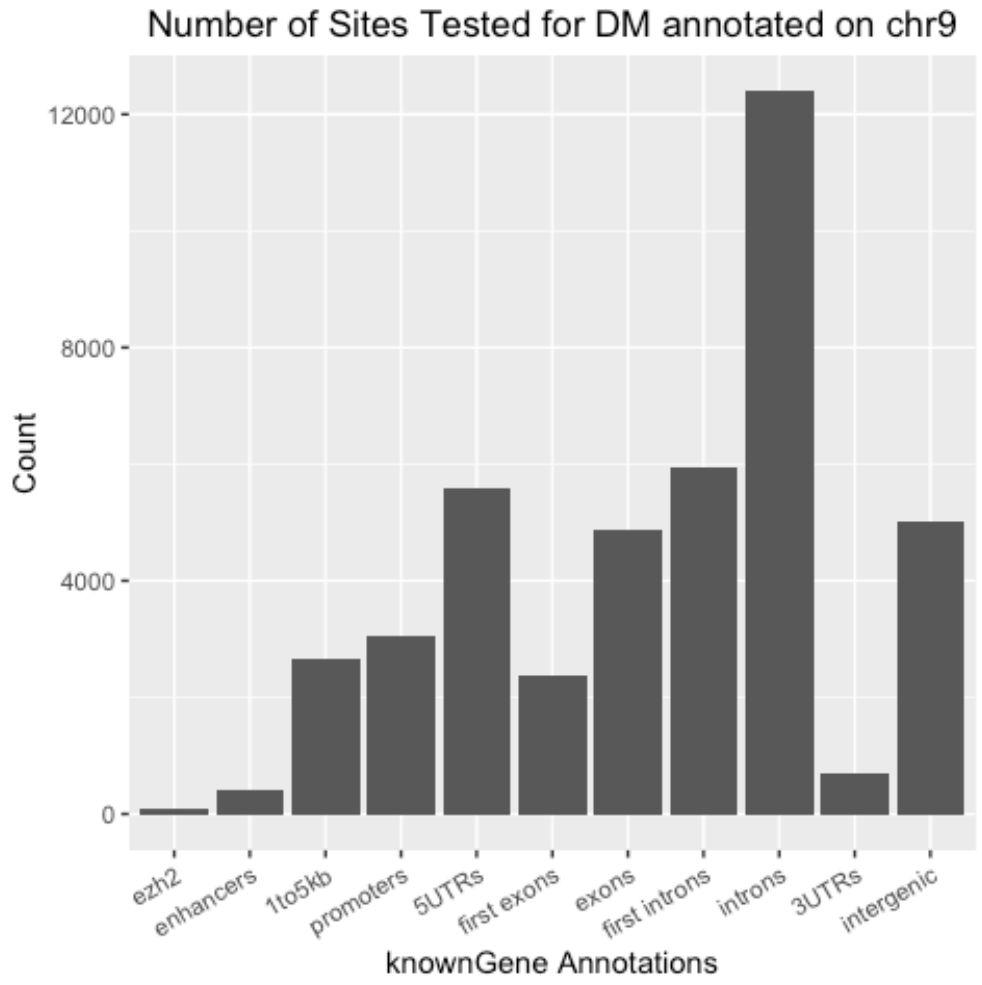
Supplementary Table 1. Benchmarking (in seconds, over 10 runs and 4 datasets) of ChIPpeakAnno versus annotatr using the microbenchmark R package.

File Size (lines)	Software	Min (s)	Mean (s)	Max (s)	Mean Speedup
27k	ChIPpeakAnno	2.72	3.16	4.01	1.62x
	annotatr	1.77	1.95	2.29	
365k	ChIPpeakAnno	26.91	28.74	30.93	4.93x
	annotatr	5.01	5.83	6.76	
4m	ChIPpeakAnno	329.08	367.98	388.39	7.81x
	annotatr	42.31	47.13	57.44	
25m	ChIPpeakAnno	2107.82	2526.71	3065.98	11.37x
	annotatr	189.89	222.21	312.02	

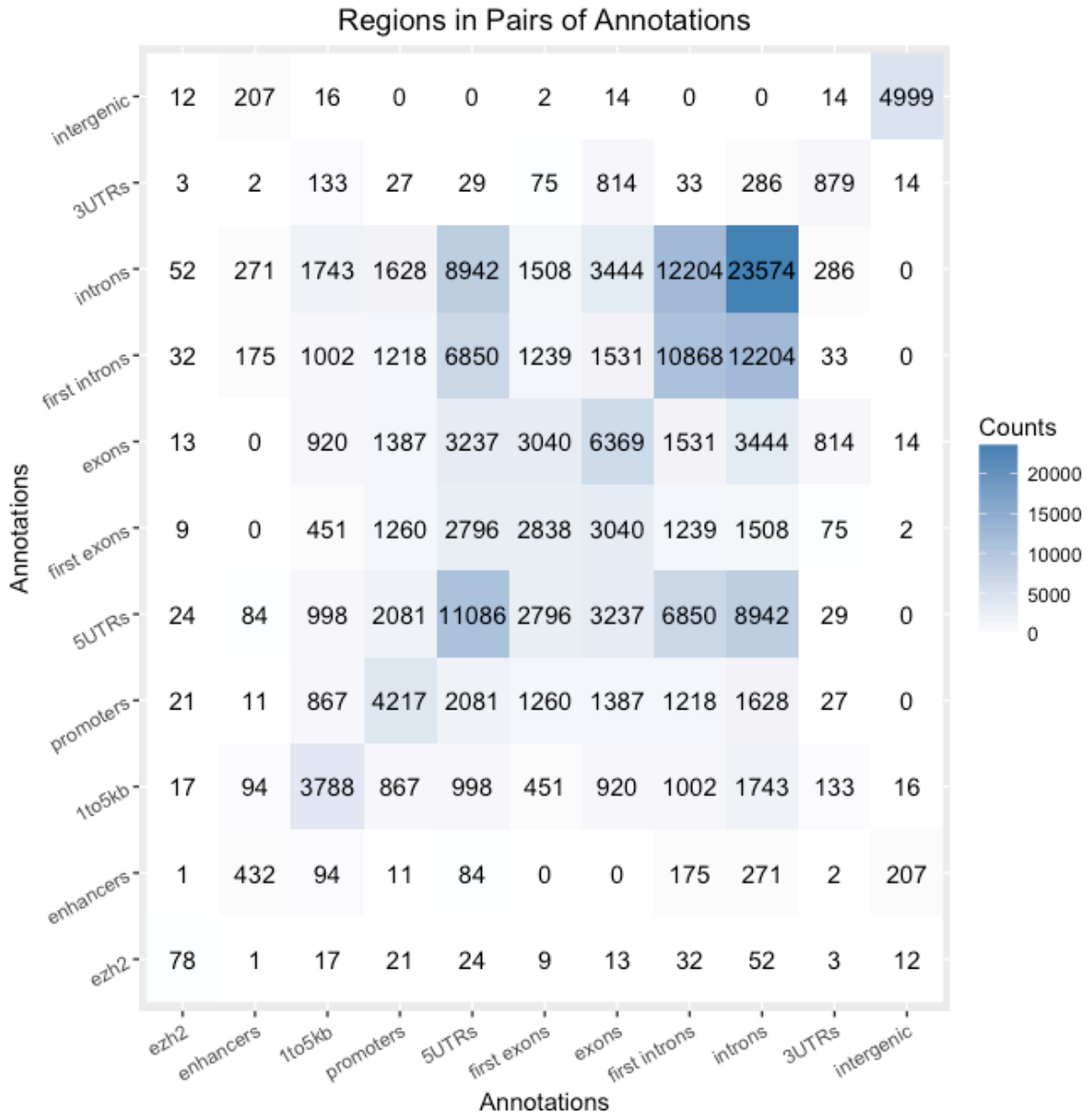
Supplementary Figure 1: (A) Schematic of the UCSC CpG annotations used in annotatr. The CpG islands are contained in the UCSC CpG island track for each genome. CpG shores are considered the 2kb extension upstream and downstream of the CpG island boundaries, less any CpG islands. The CpG shelves are a further 2kb extension upstream and downstream of the furthest upstream and downstream boundaries of the CpG shores, less any CpG island and shore annotations. The complement of the CpG islands, shores, and shelves make up the "open sea" or interCGI annotation. (B) A detailed schematic of the UCSC knownGene annotations used in annotatr. Bold and italicized text (e.g. txStart, cdsStart, etc.) indicates columns from the table downloaded from the UCSC Genome Browser. *Basic genes*: 1-5Kb, promoter, 5'UTR, exons, introns, and 3'UTR compose the basic gene annotation. Exons and introns are denoted by green and gray in the gene body. *Detailed genes*: 1-5Kb, promoter, 5'UTR exon, 5'UTR intron, CDS exon, CDS intron, 3'UTR exon, and 3'UTR intron compose the detailed gene annotation. Annotations may overlap one another from the same or from different transcripts.



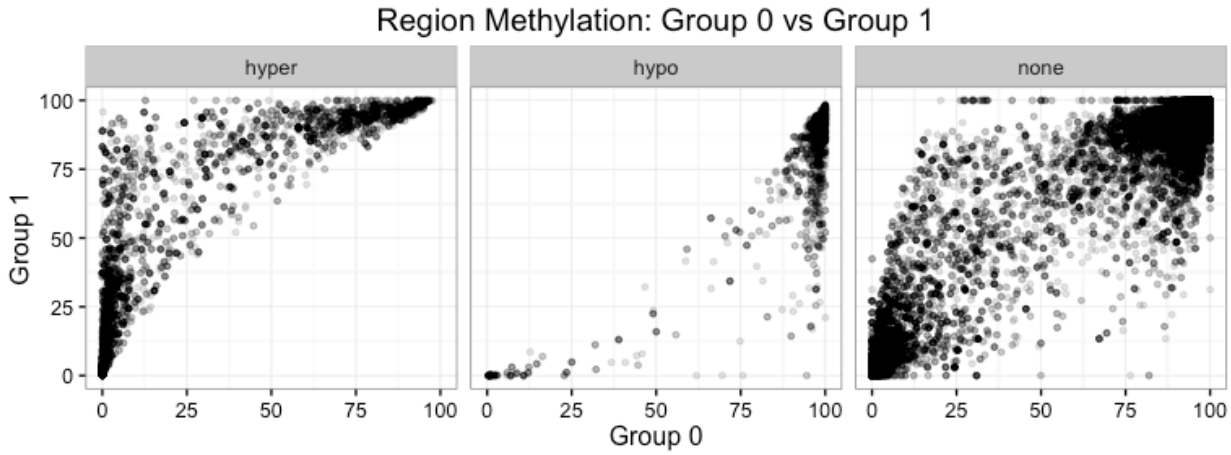
Supplementary Figure 2: The counts of regions per annotation type.



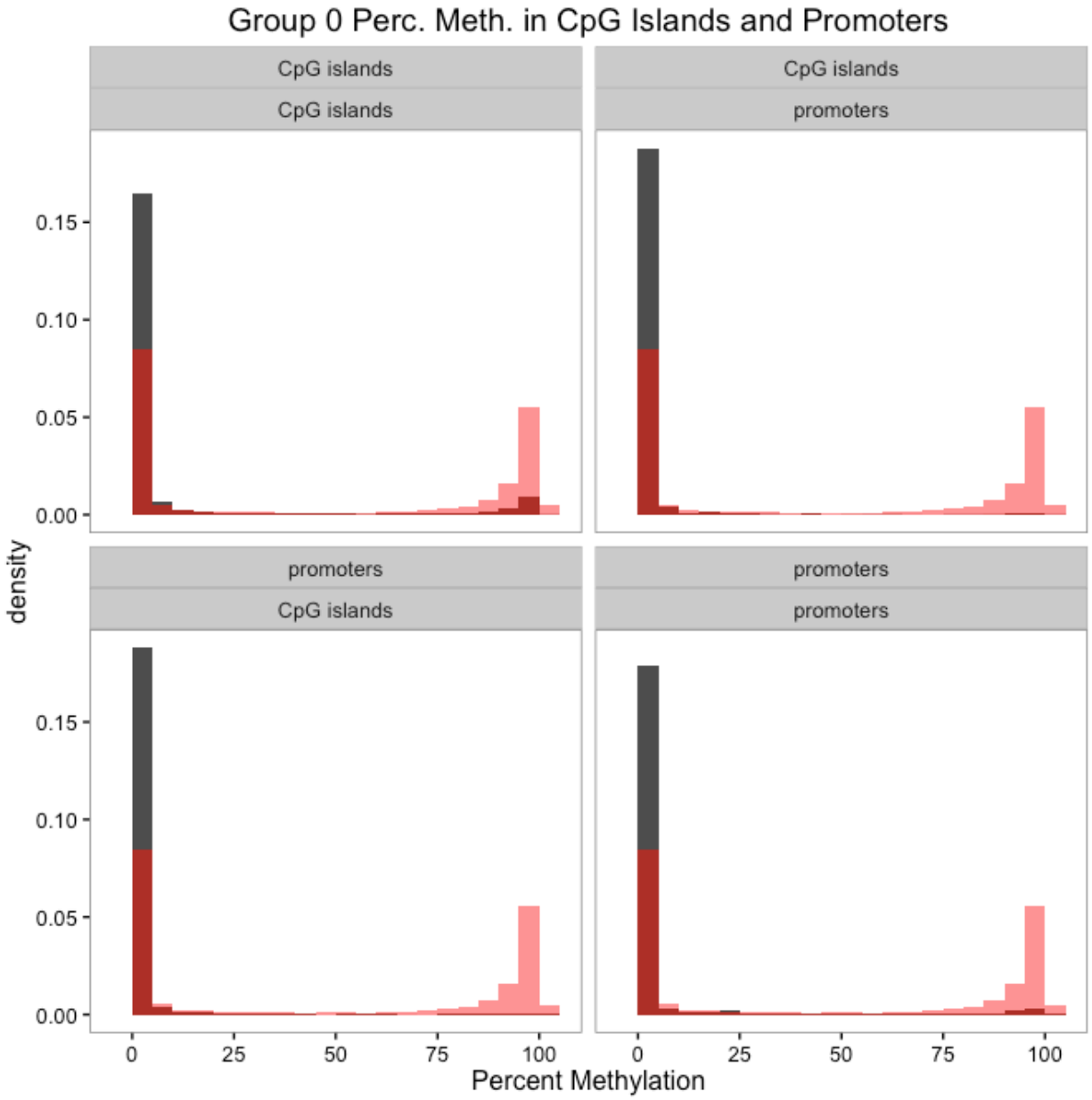
Supplementary Figure 3: The counts of regions occurring in pairs of annotations.



Supplementary Figure 4: The distribution of methylation rates for group 1 and group 0 across the categorical variable denoting differential methylation status.



Supplementary Figure 5: The distribution of methylation rates for regions that occur in just CpG islands, just promoters, and both CpG islands and promoters.



Supplementary References

ENCODE Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*, 57–74.

Figueroa, M. E., Abdel-Wahab, O., Lu, C., Ward, P. S., Patel, J., Shih, A., & Melnick, A. (2010). Leukemic IDH1 and IDH2 Mutations Result in a Hypermethylation Phenotype, Disrupt TET2 Function, and Impair Hematopoietic Differentiation. *Cancer Cell*, *18*, 553–567.

Hsu, F., Kent, W. J., Clawson, H., Kuhn, R. M., Diekhans, M., & Haussler, D. (2006). The UCSC Known Genes. *Bioinformatics*, *22*(9), 1036–1046.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., & Kent, W. J. (2004). The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, *32*(90001), 493D–496.

Krueger, F., & Andrews, S. R. (2011). Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*, *27*(11), 1571–1572.

Mersmann, O. (2015). microbenchmark: Accurate Timing Functions. R package version 1.4-2.1. http://CRAN.R-project.org/package=microbenchmark

Park, Y., Figueroa, M. E., Rozek, L. S., & Sartor, M. A. (2014). MethylSig: a whole genome DNA methylation analysis pipeline. *Bioinformatics*, *30*(17), 2414–2422.

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, *11*, 237.