

Supplementary figures and tables

motifTestR: A Bioconductor package for the analysis of transcription factor binding motifs in sequence data

Stevie Pederson

Correspondence: stevie.pederson@adelaide.edu.au

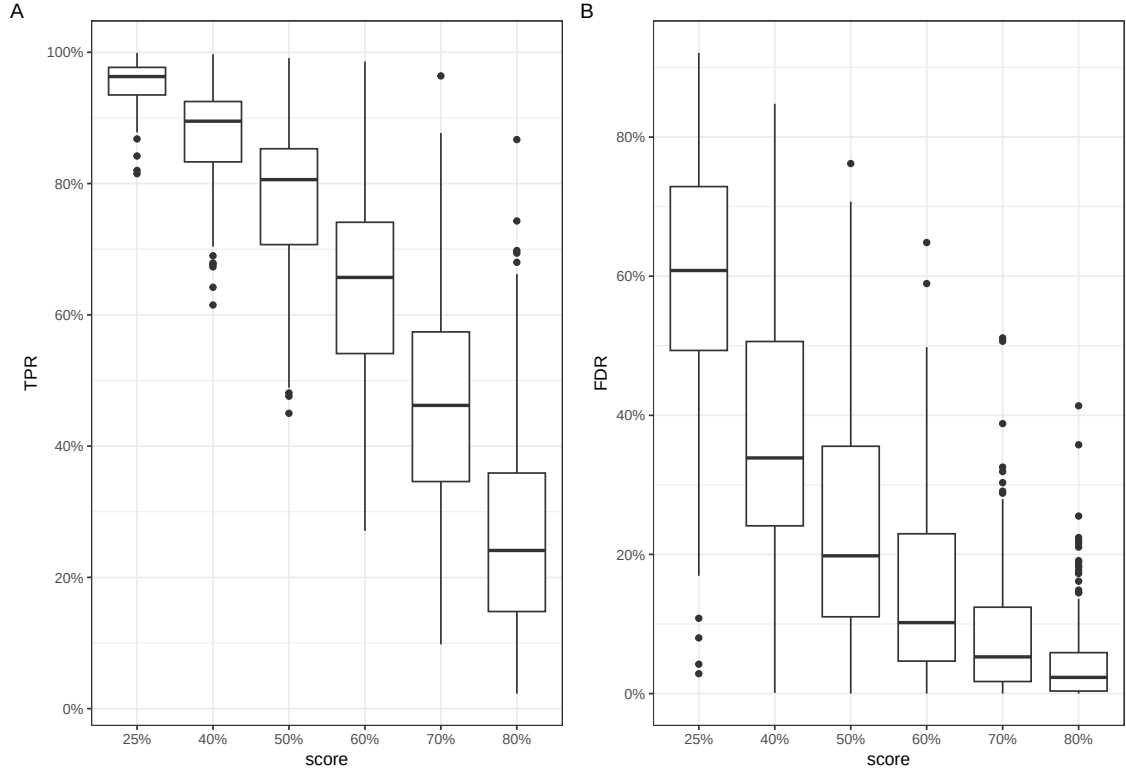


Figure S2. True Positive Rates (A) and False Discovery Rates (B) at different score thresholds. Using the algorithm from `getPwmMatches`, the TPR and FDR were calculated for each of the 181 motifs at score thresholds of 25%, 40%, 50%, 60%, 70% and 80%, with boxplots showing the distribution of TPR and FDR across the motifs at each threshold. At 50% the median TPR was 80% with a median FDR of 20%, providing a good balance between TPR and FDR across the motifs tested. These values will be applicable to motif scanning using the package `motifTestR` and `centrimo` [2], but may differ under AME [3].

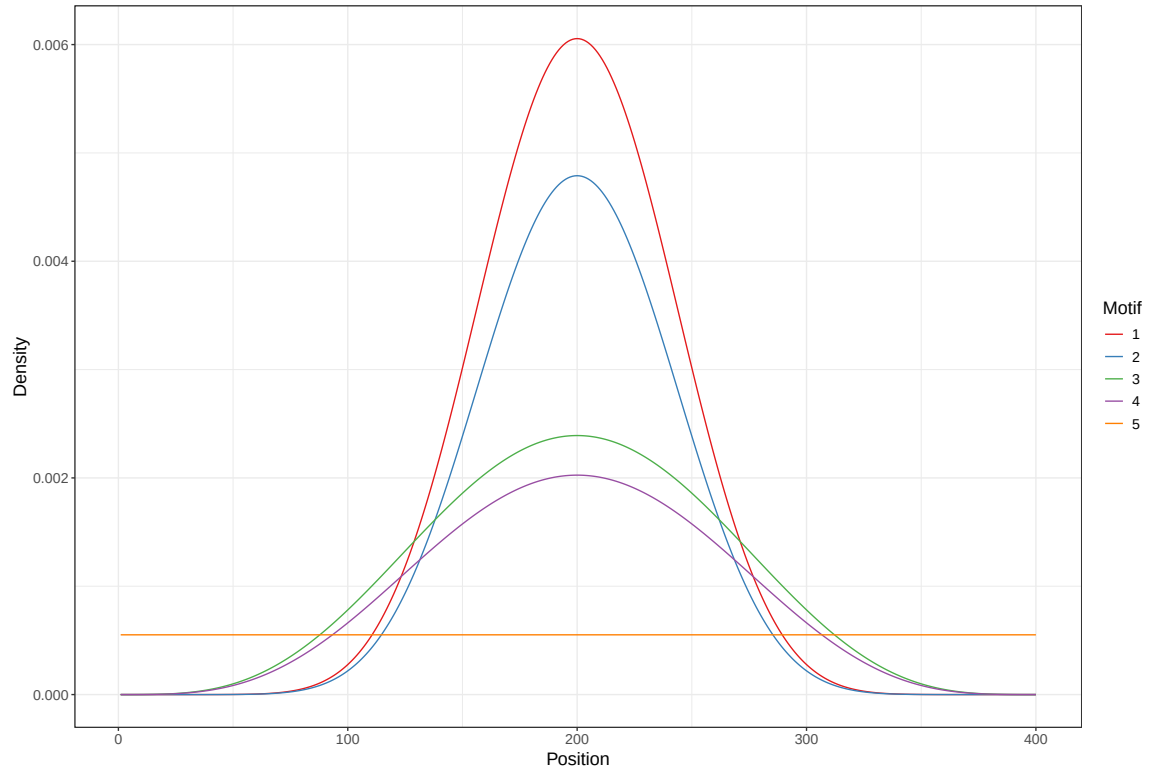


Figure S3. Motif distributions within simulated sequences Five motifs were simulated within 1000 sequences at varying frequencies and positional distributions, with motifs overlaid at positions determined by a Beta-Binomial distribution, and at frequencies determined by either a Poisson or Negative Binomial distribution, as described more fully in Supplementary Table [S1](#).

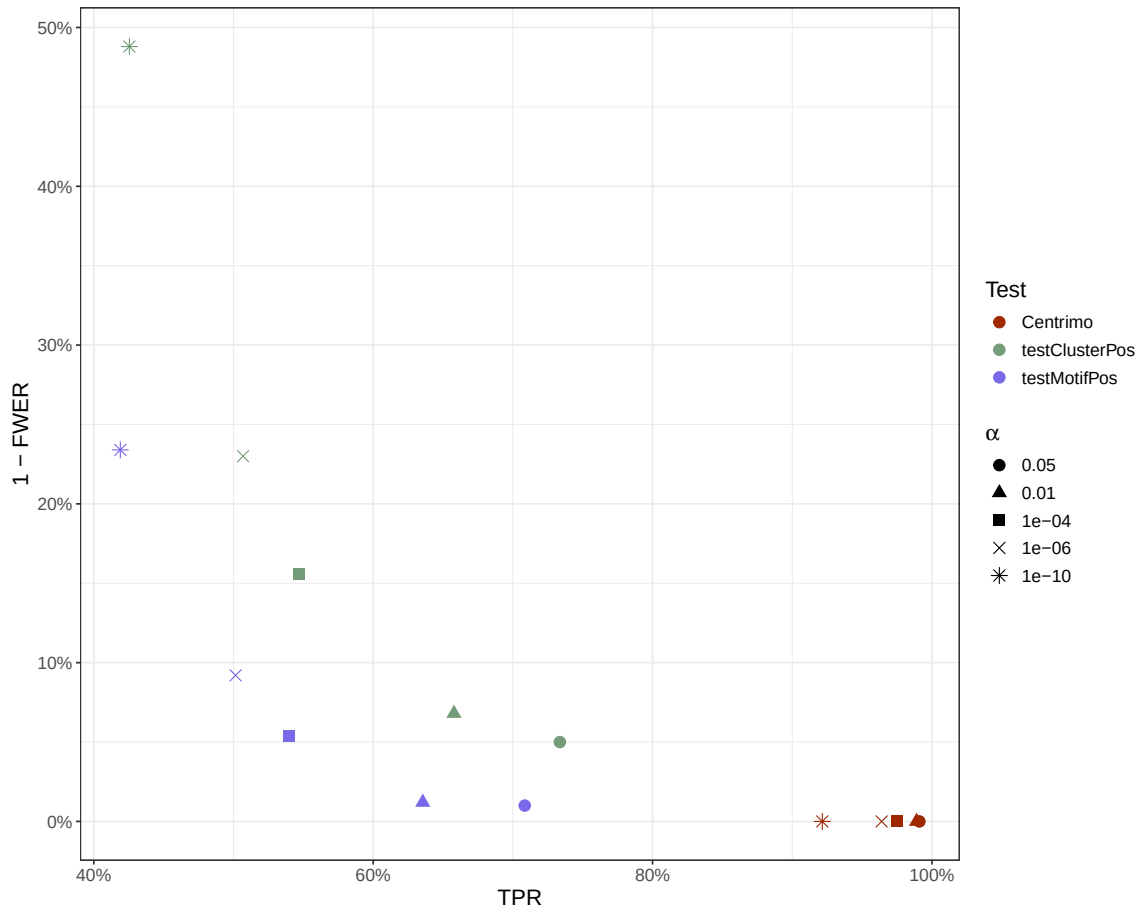


Figure S4. Power and Family-Wise Error Rate (FWER) control for testing positional bias. Using different thresholds for significance, the power and FWER were calculated for each of the three methods for testing positional bias. Centrimo remained consistently powerful at all thresholds but never returned a result without false positives. Conversely, the two methods implemented in `motifTestR` were less powerful, but did return results with partial FWER control, whilst still being far below the level of desired FWER control

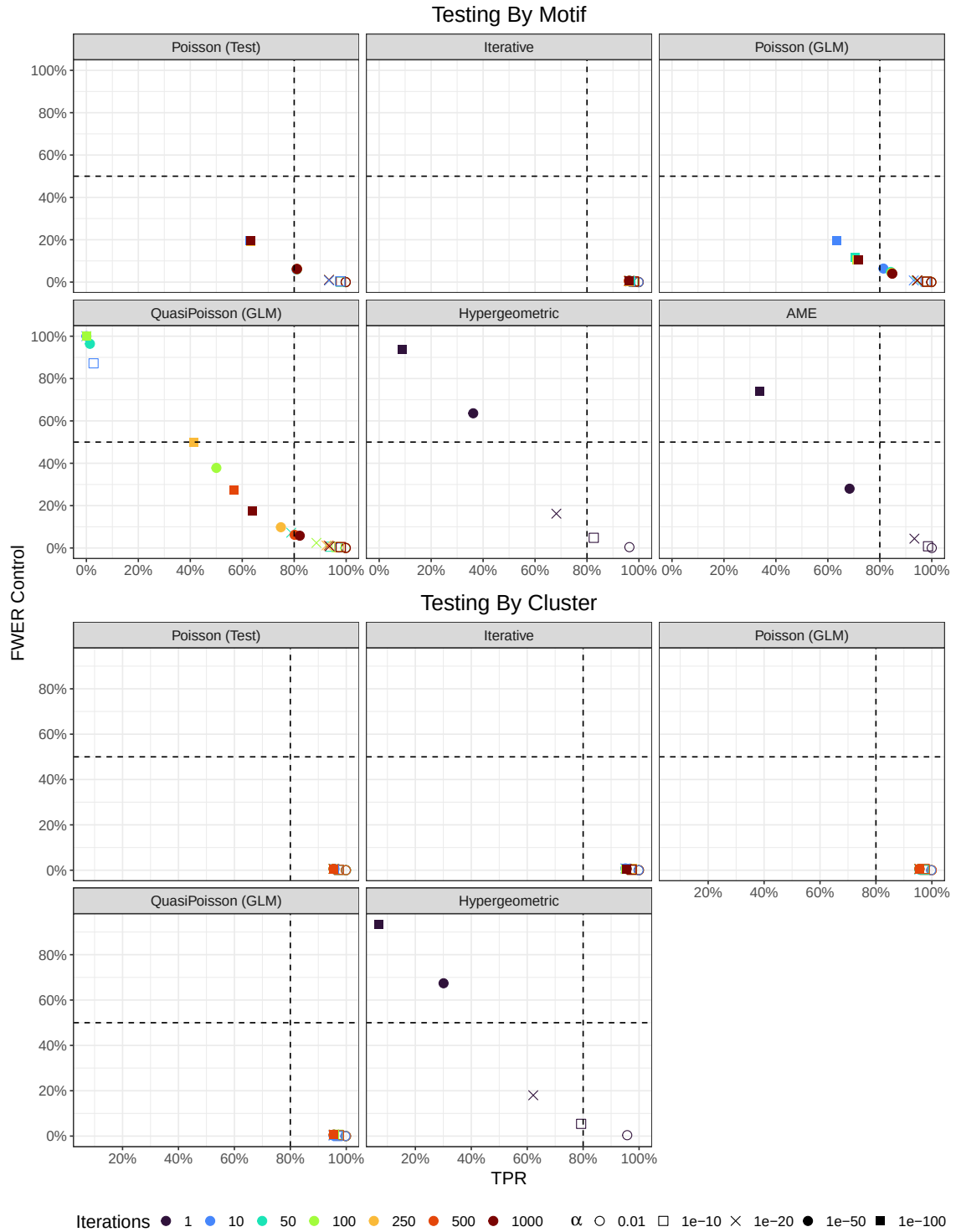


Figure S5. Power and Family-Wise Error Rate (FWER) control when testing for enrichment. Using different thresholds for significance, the power and FWER were calculated for each of the methods for testing motif enrichment, using Bonferroni-adjusted p-values. The FWER is not controlled adequately by any method, with testing by motif appearing to be preferable for controlling the FWER, and with hypergeometric tests appearing to show the best tradeoff between the two metrics.

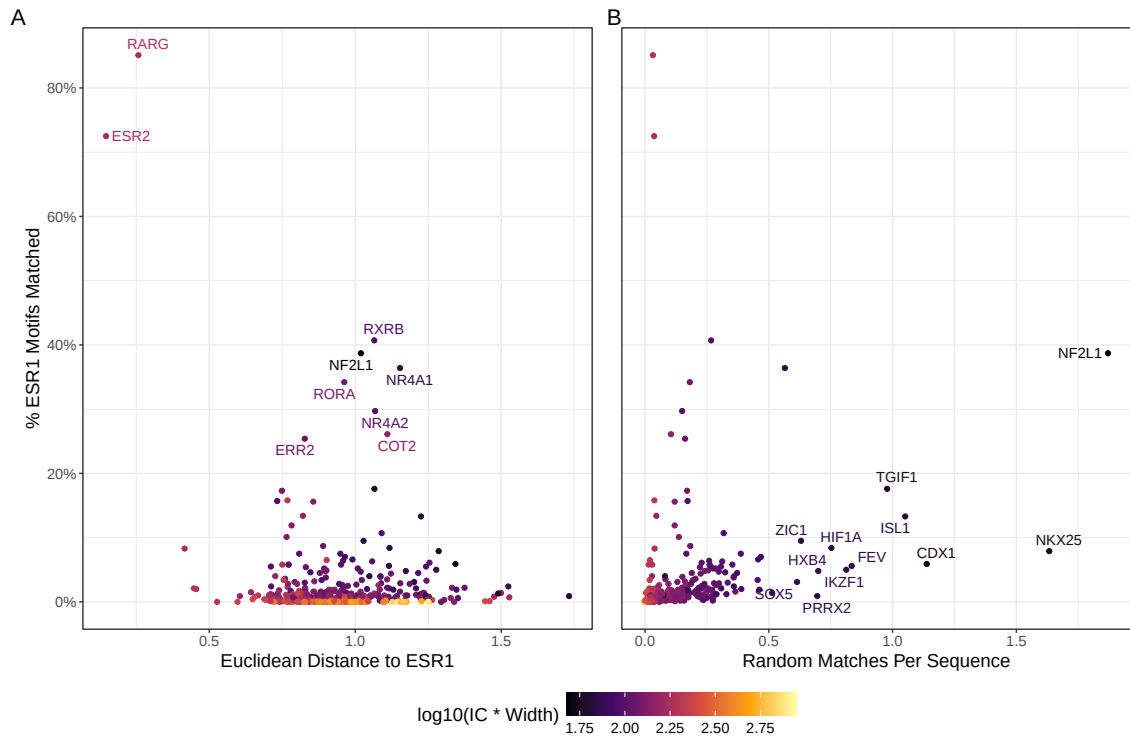


Figure S6. Spurious motif detections for sequences simulated with an ESR1 motif
 (A) Motifs with small Euclidean distances to the ESR1 motifs were often erroneously detected, with ESR2 and RARG being the most striking. (B) An additional set of motifs were commonly detected within the sequences, without being detected at the site of a simulated ESR1 motif, and were considered to matches to random sequences. These motifs tended to be short with low information content, as summarised by using the product of these two values to colour points.

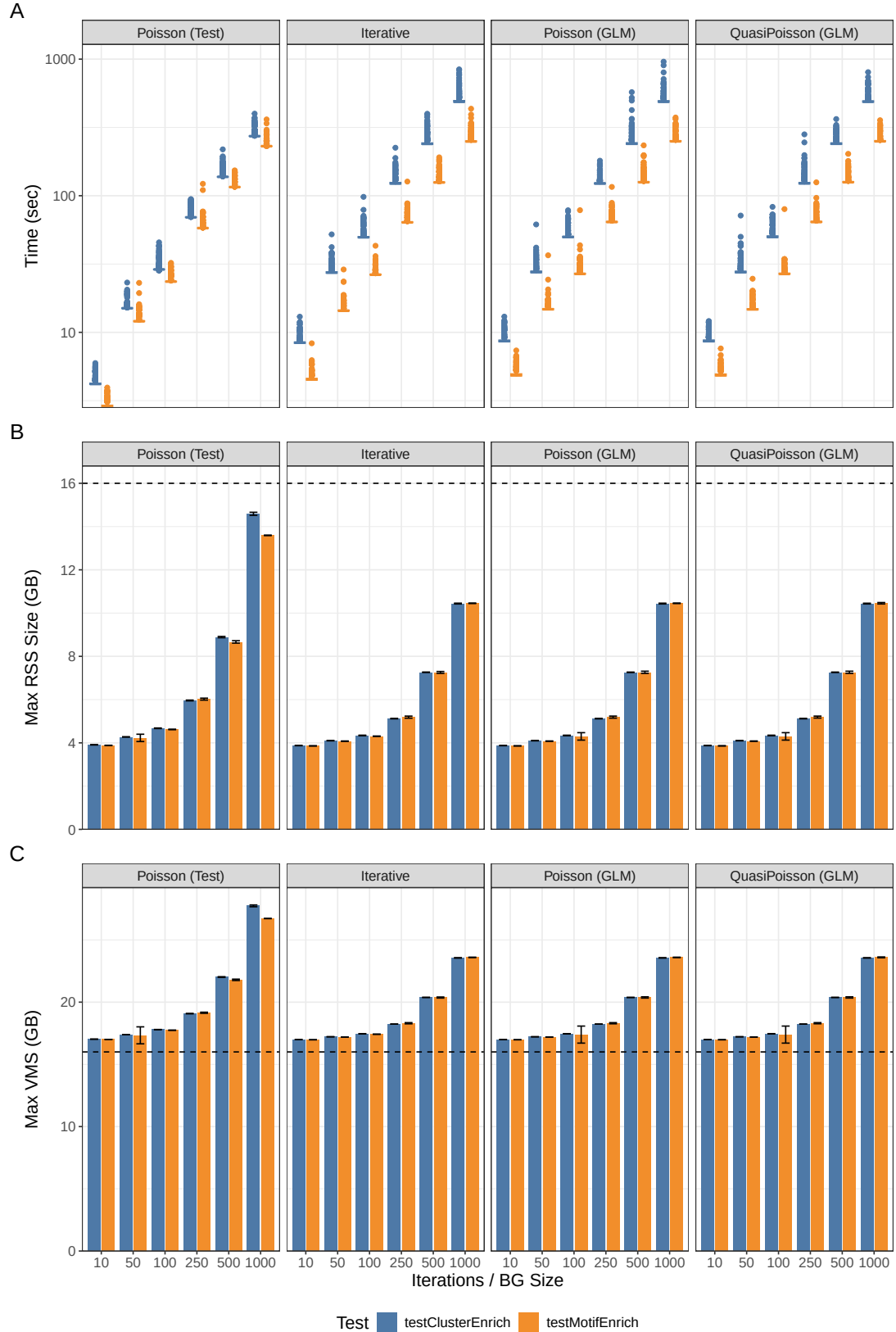


Figure S7. Computational benchmarks for enrichment analysis at both the motif and cluster level. All benchmarks were recorded within an R session running on the Adelaide University HPC **phoenix** cluster, analysing 1000×400nt sequences with a motif database of 181 motifs. (A) The time taken for an analysis using 4 threads in seconds. (B) The max RSS usage for an analysis using 4 threads in gigabytes. (C) Max VMS usage for the same process, including SWAP memory, in gigabytes.

Table S1. Summary of simulated motif sets. Five sets of 1000 sequences were simulated containing 5 motifs in each set, with motifs overlaid at positions determined by a Beta-Binomial distribution, with shape parameters as described. Motif frequencies were determined by either a Poisson or Negative Binomial distribution, with rate parameters as given, and overdispersion as parameterised by θ from `rnegbin` in the package *MASS* [4]

Motif	Enrichment	Peak	Distribution	% With Motif	Shape1	Shape2
Motif 1	Strong	Tight	Poisson(1)	63.2%	12.5	12.5
Motif 2	Strong	Tight	NegBin(1, 1)	50.0%	12.5	12.5
Motif 3	Moderate	Broad	Poisson(0.5)	39.3%	5	5
Motif 4	Moderate	Broad	NegBin(0.5, 1)	33.3%	5	5
Motif 5	Weak	None	Poisson(0.25)	22.1%	1	1

References

- [1] Kulakovskiy IV, Vorontsov IE, Yevshin IS, Sharipov RN, Fedorova AD, Rumynskiy EI, et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* 2018 Jan;46(D1):D252–D259.
- [2] Bailey TL, Machanick P. Inferring direct DNA binding from ChIP-seq. *Nucleic Acids Res.* 2012 Sep;40(17):e128.
- [3] McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics.* 2010 Apr;11(1):165.
- [4] Venables WN, Ripley BD. *Modern Applied Statistics with S.* 4th ed. New York: Springer; 2002. ISBN 0-387-95457-0. Available from: <https://www.stats.ox.ac.uk/pub/MASS4/>.