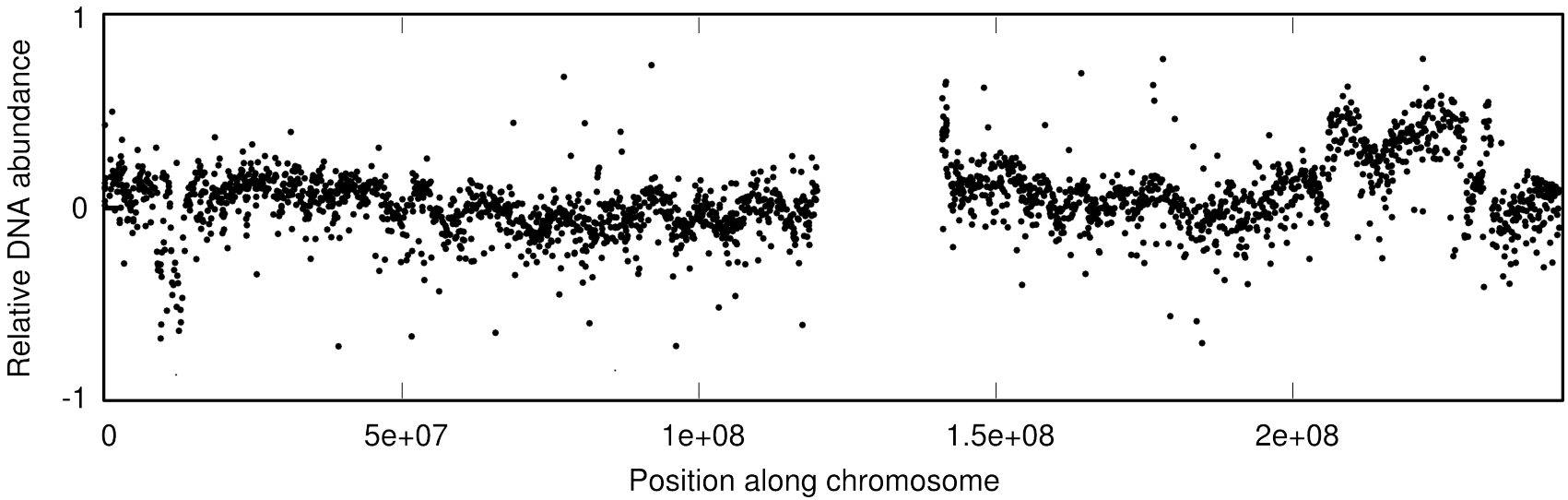
HaMMLET - Fast Bayesian Inference of Copy Number Variants using Hidden Markov Models with Wavelet Compression

John Wiedenhoeft, Eric Brugel, Alexander Schliep

Department of Computer Science
Rutgers University

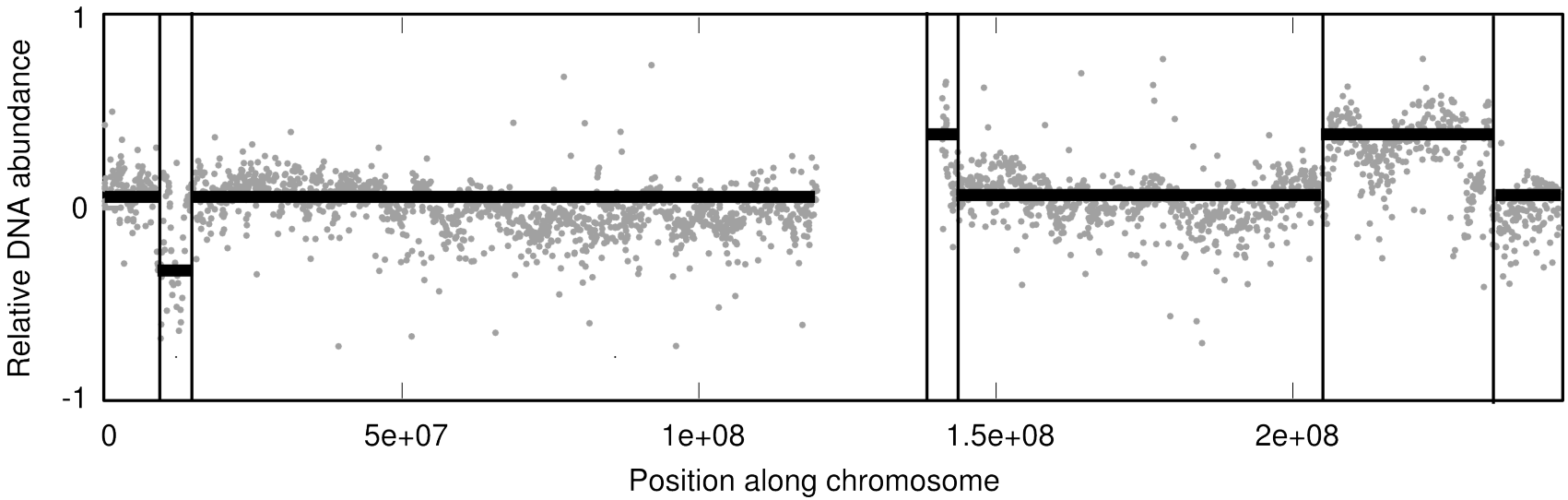
RECOMB 2016

Genomic Copy Number Variants (CNV)



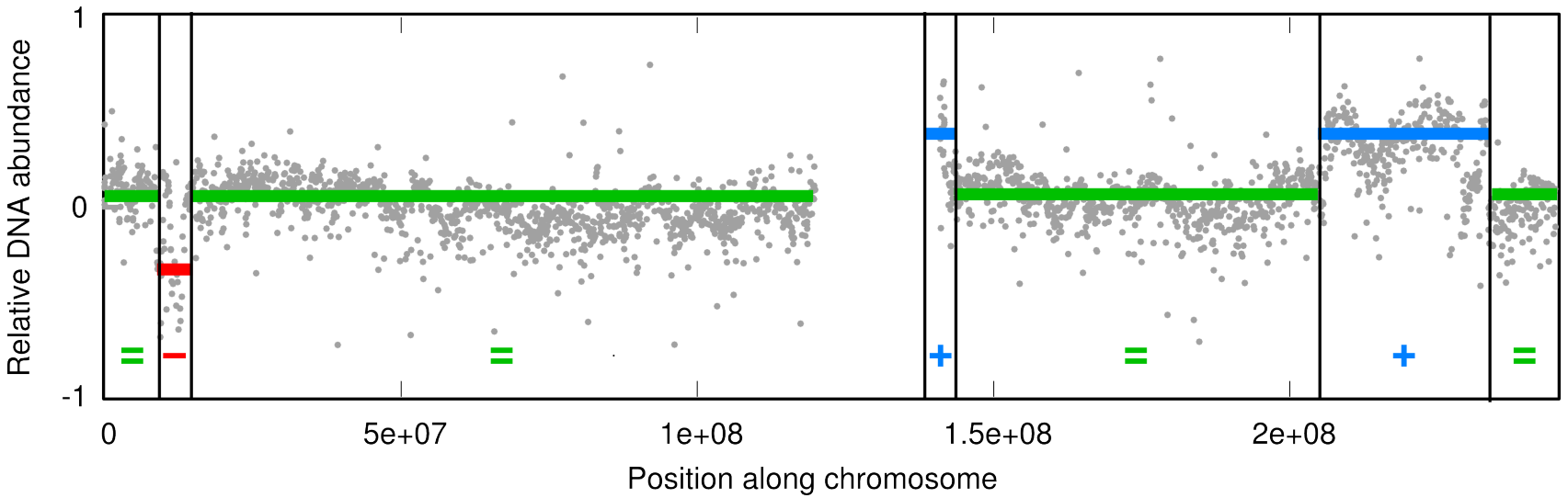
- Non-diploid genomic segments (duplications, deletions)
- Data from aCGH, SNP arrays, exome sequencing, WGS,...
- Measures relative or absolute DNA abundance
- Piecewise constant data with measurement noise

Genomic Copy Number Variants (CNV)



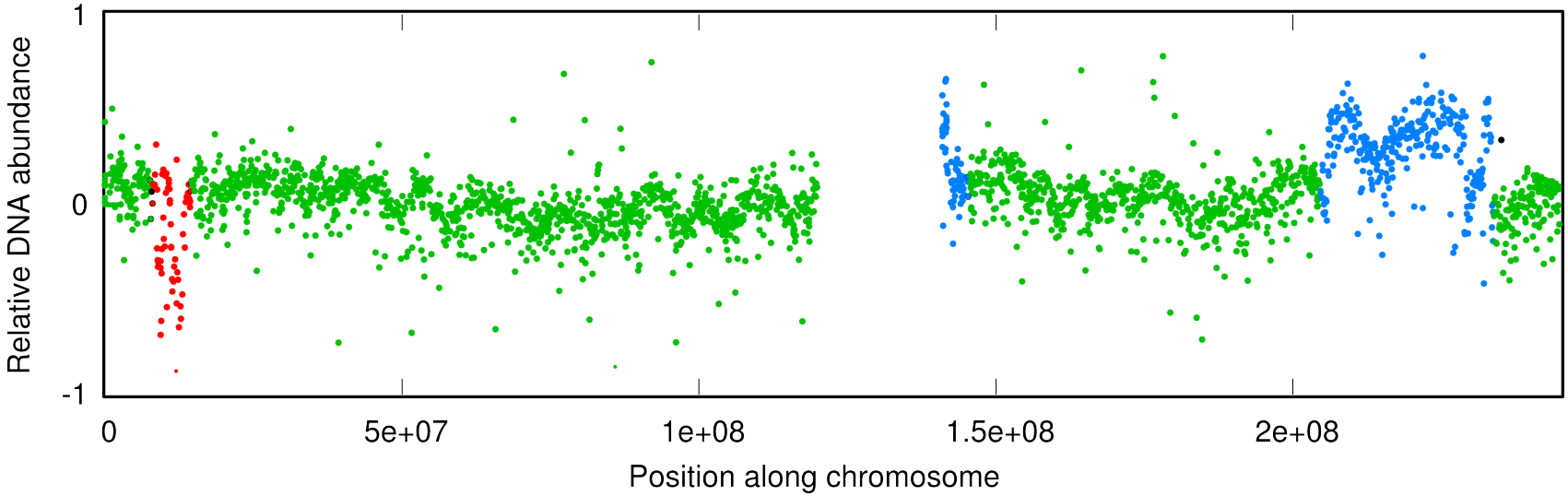
- Non-diploid genomic segments (duplications, deletions)
- Data from aCGH, SNP arrays, exome sequencing, WGS,...
- Measures relative or absolute DNA abundance
- Piecewise constant data with measurement noise

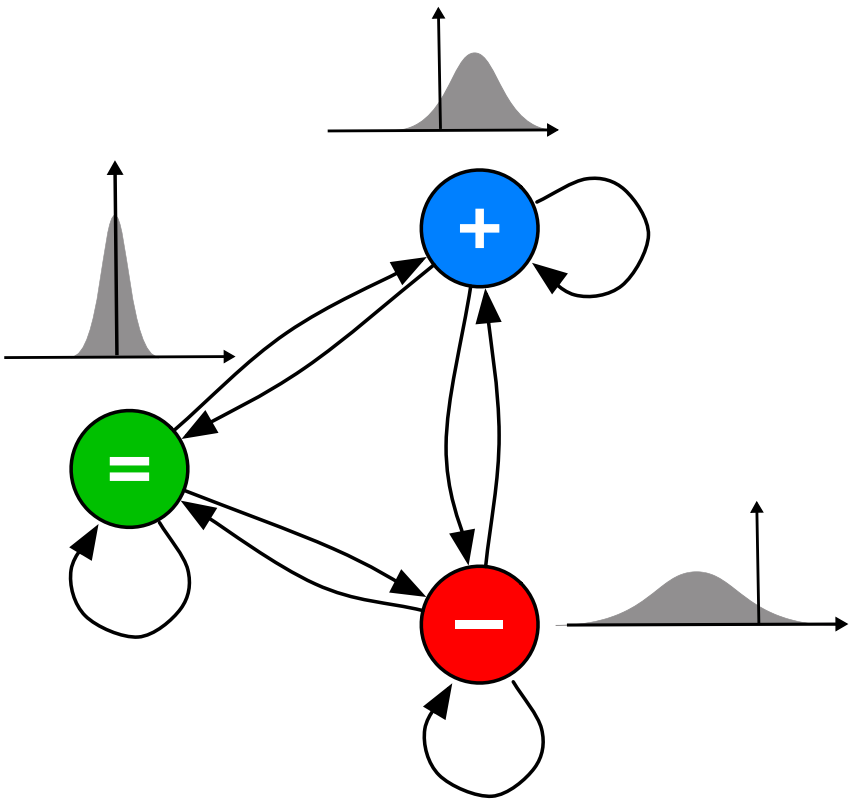
Genomic Copy Number Variants (CNV)



- Non-diploid genomic segments (duplications, deletions)
- Data from aCGH, SNP arrays, exome sequencing, WGS,...
- Measures relative or absolute DNA abundance
- Piecewise constant data with measurement noise

Hidden Markov Model





- Latent CNV state variable for each data point ("state sequence")
- Different noise parameters for each state
- Model parameters are unknown

Frequentist inference of the state sequence

1. Maximum likelihood estimate of the parameters using Baum-Welch[1] (EM)
2. Viterbi decoding[2] to obtain the most likely state sequence

- Likelihood function not convex (multiple reinitializations)
- ML parameters tend to overfit majority class
- Segmentation assumed to come from ML parameters
- Segmentation relies on single parameter estimate
- Viterbi yields single solution

[1] Rabiner. "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE* 77 (1989), pp. 257–286. DOI: 10.1109/5.18626.

[2] Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2 (Apr. 1967), pp. 260–269. DOI: 10.1109/TIT.1967.1054010.

Frequentist inference of the state sequence

1. Maximum likelihood estimate of the parameters using Baum-Welch[1] (EM)
2. Viterbi decoding[2] to obtain the most likely state sequence

- Likelihood function not convex (multiple reinitializations)
- ML parameters tend to overfit majority class
- Segmentation assumed to come from ML parameters
- Segmentation relies on single parameter estimate
- Viterbi yields single solution

---

[1] Rabiner. "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE* 77 (1989), pp. 257–286. DOI: 10.1109/5.18626.

[2] Viterbi. "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm". In: *IEEE Transactions on Information Theory* 13.2 (Apr. 1967), pp. 260–269. DOI: 10.1109/TIT.1967.1054010.

Bayesian inference of the state sequence

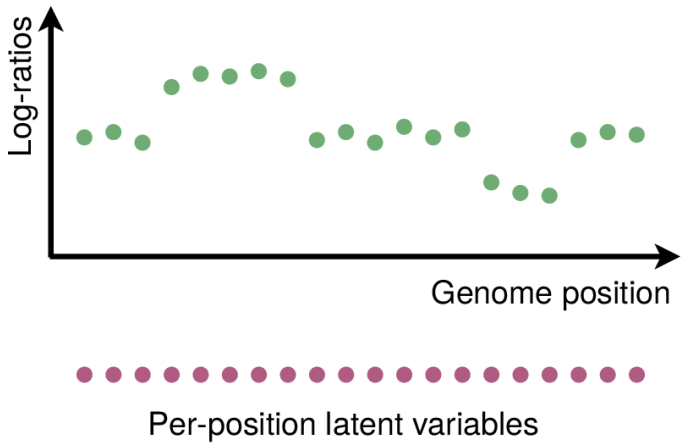
$$\mathbb{P}(\text{states} \,|\, \text{data}) = \int_{\text{parameters}} \mathbb{P}(\text{states, parameters} \,|\, \text{data})$$

- Integration over entire parameter space
- Full marginal CNV state distribution for each position
- Quantifies uncertainties and alternative interpretations
- Explicit inductive bias due to hyperparameters
- No feasible analytic expression $\Rightarrow$ Markov Chain Monte Carlo

Forward-Backward Gibbs sampling

- For each iteration
  1. Sample parameters from HMM conditioned on state sequence
  2. Sample state sequence from HMM conditioned on parameters
- Tally state counts for each position to approximate marginals

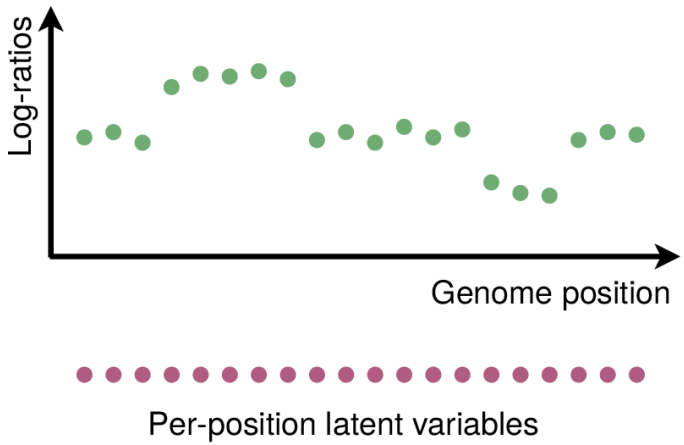
Compressed Forward-Backward Sampling[3]

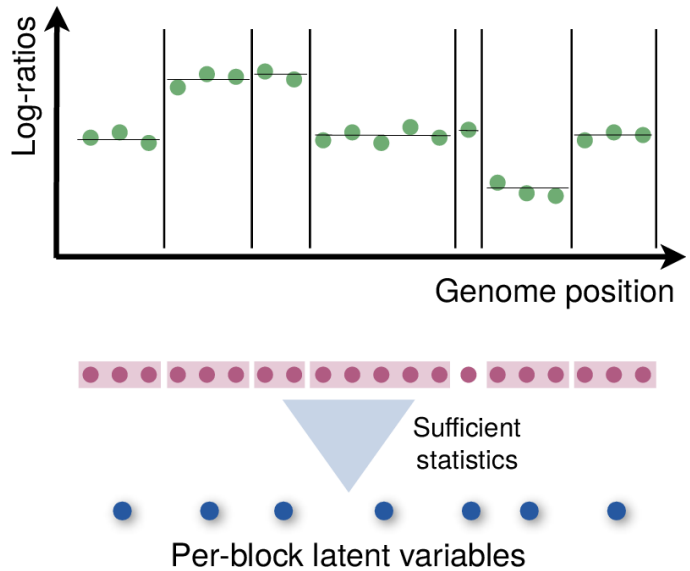


$$\Theta(\#\text{states}^2 \times \#\text{values})$$

[3] Mahmud and Schliep. "Fast MCMC sampling for Hidden Markov Models to determine copy number variations". In: *BMC Bioinformatics* 12 (Jan. 2011), p. 428. DOI: 10.1186/1471-2105-12-428.

Compressed Forward-Backward Sampling[3]



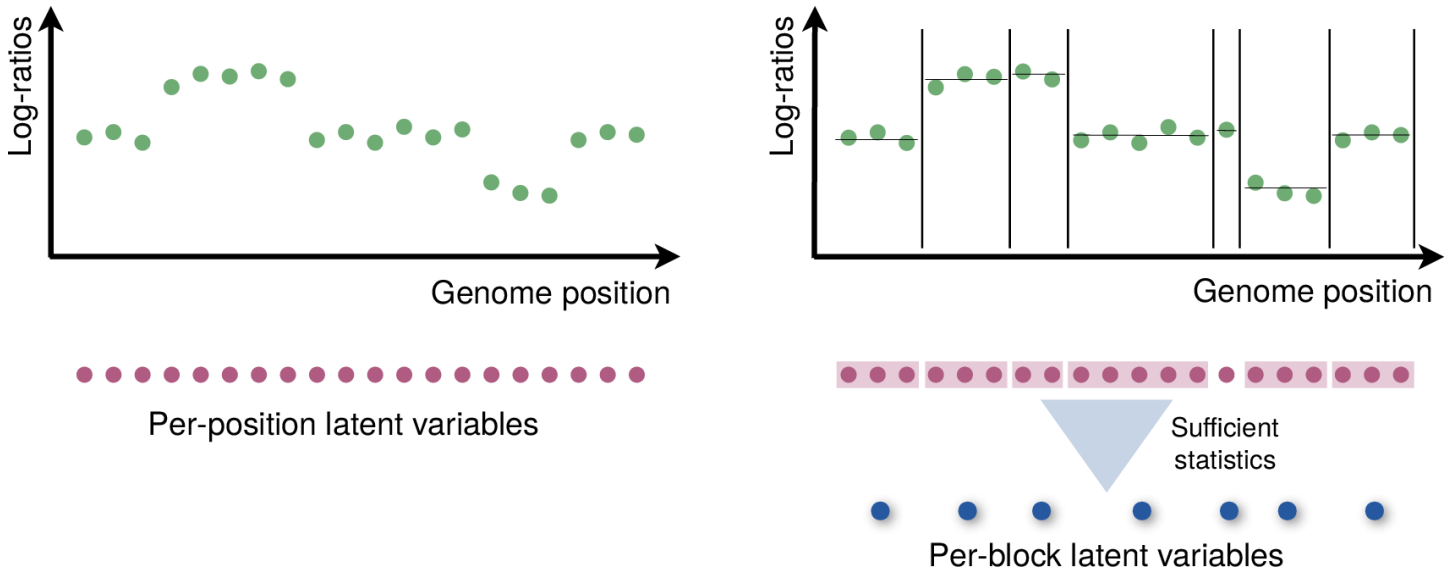


$\Theta(\#\text{states}^2 \times \#\text{values})$

$\Theta(\#\text{states}^2 \times \#\text{blocks})$

[3] Mahmud and Schliep. "Fast MCMC sampling for Hidden Markov Models to determine copy number variations". In: *BMC Bioinformatics* 12 (Jan. 2011), p. 428. DOI: 10.1186/1471-2105-12-428.

Compressed Forward-Backward Sampling[3]

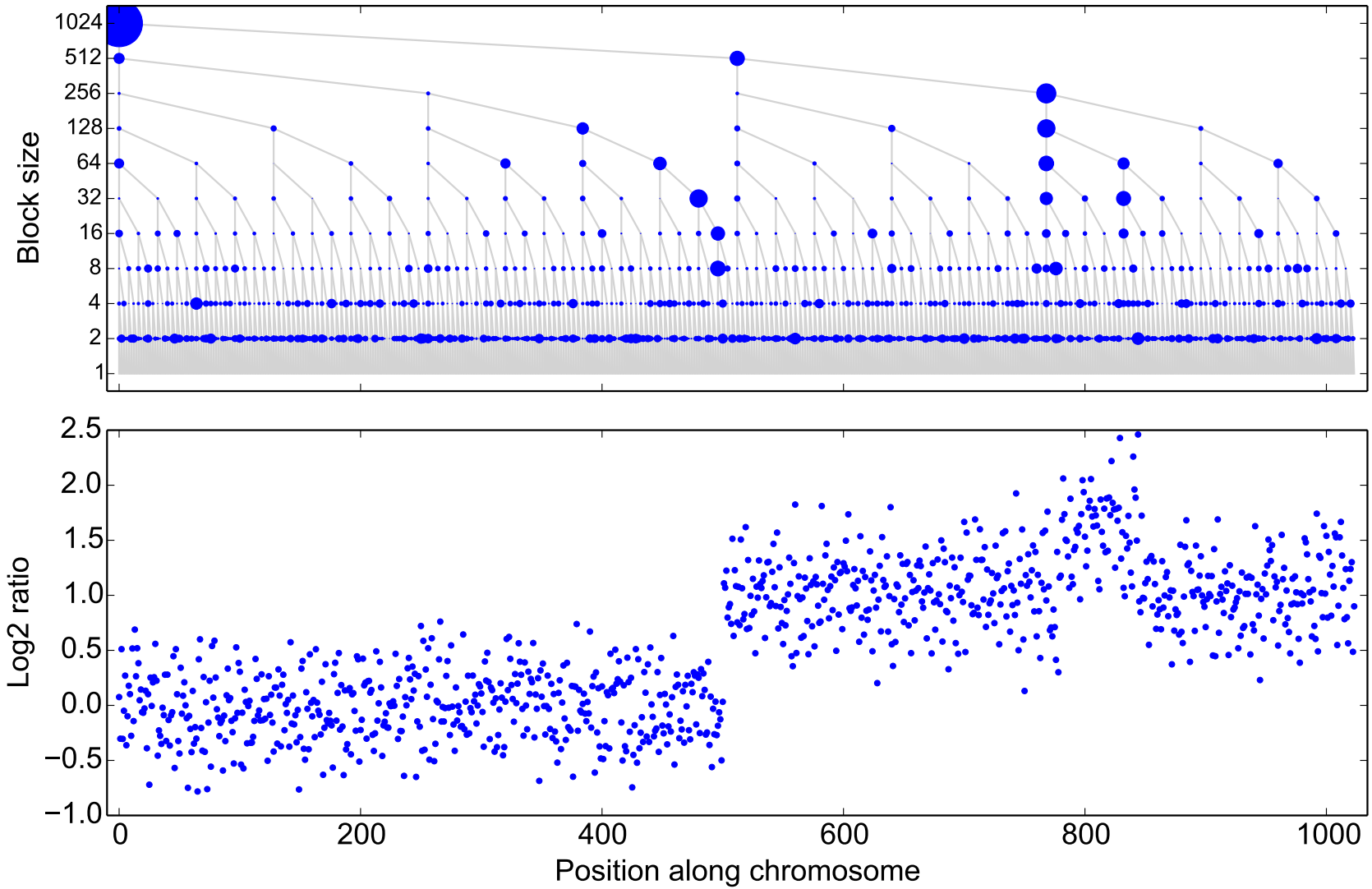


Conflicting objectives

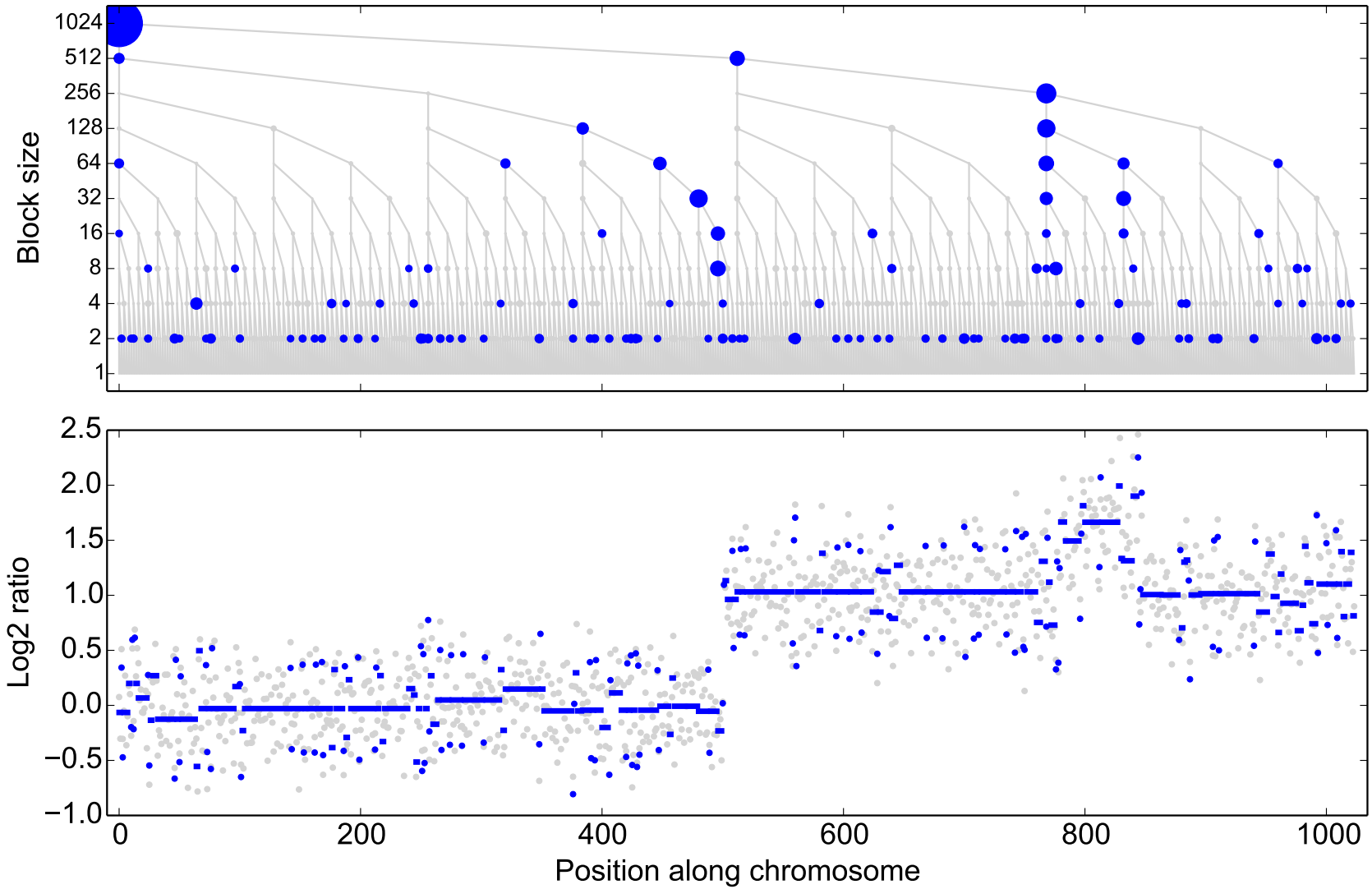
- Treat noise as potential break point ⇒ poor compression
- Treat variation due to break point as noise ⇒ wrong segmentation

[3] Mahmud and Schliep. "Fast MCMC sampling for Hidden Markov Models to determine copy number variations". In: *BMC Bioinformatics* 12 (Jan. 2011), p. 428. DOI: 10.1186/1471-2105-12-428.

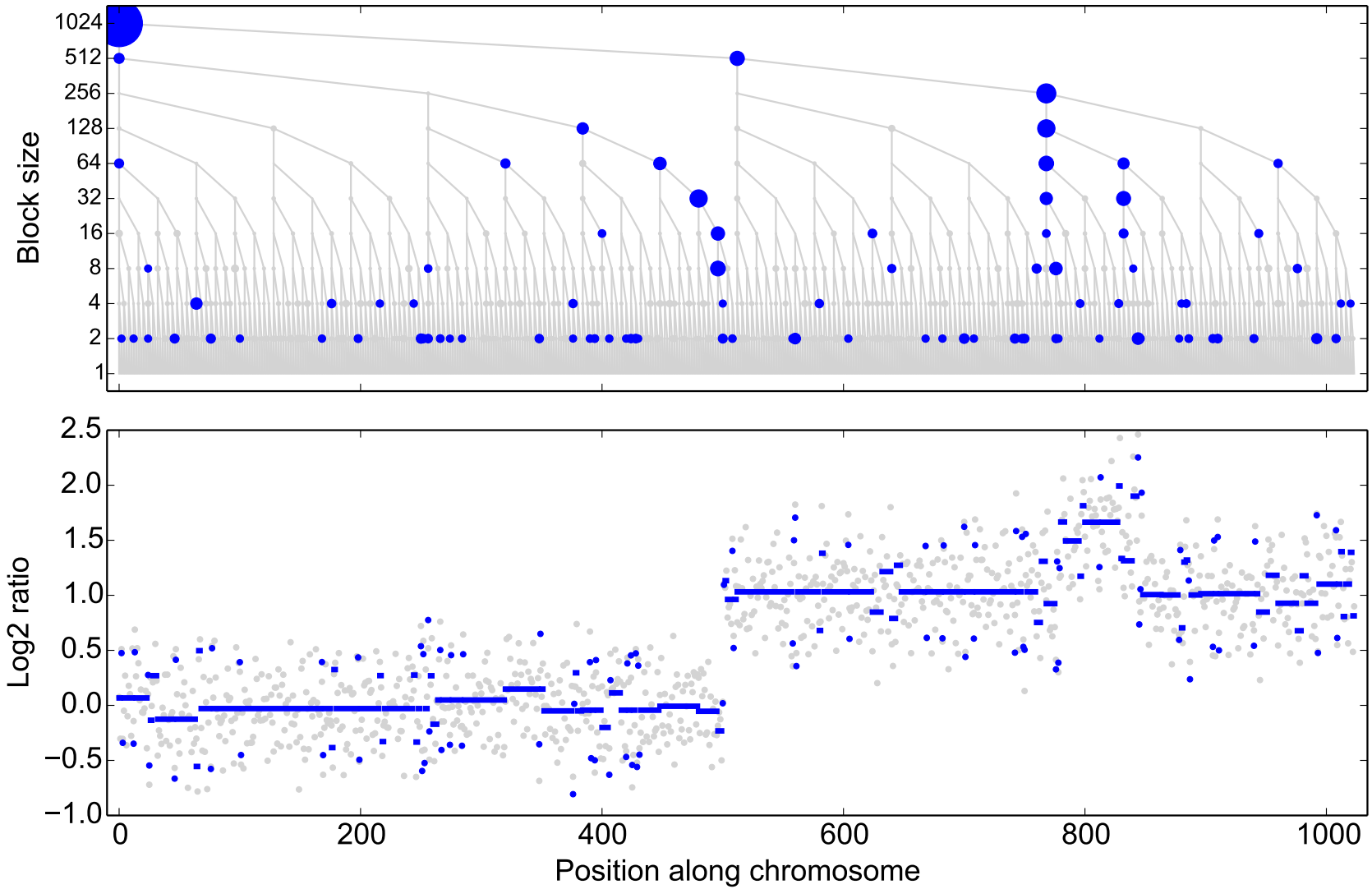
Haar wavelet regression for given noise variance



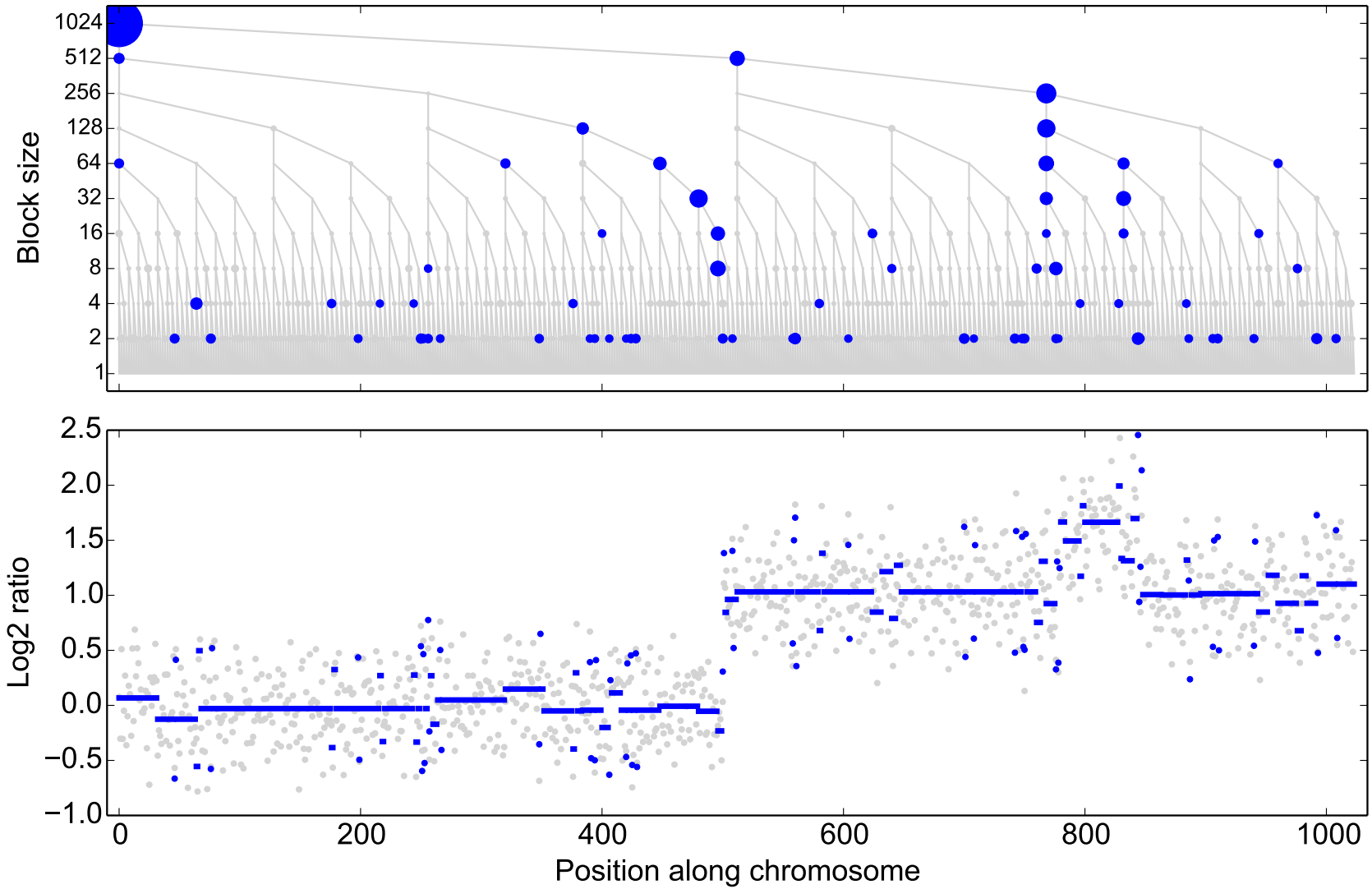
Haar wavelet regression for given noise variance



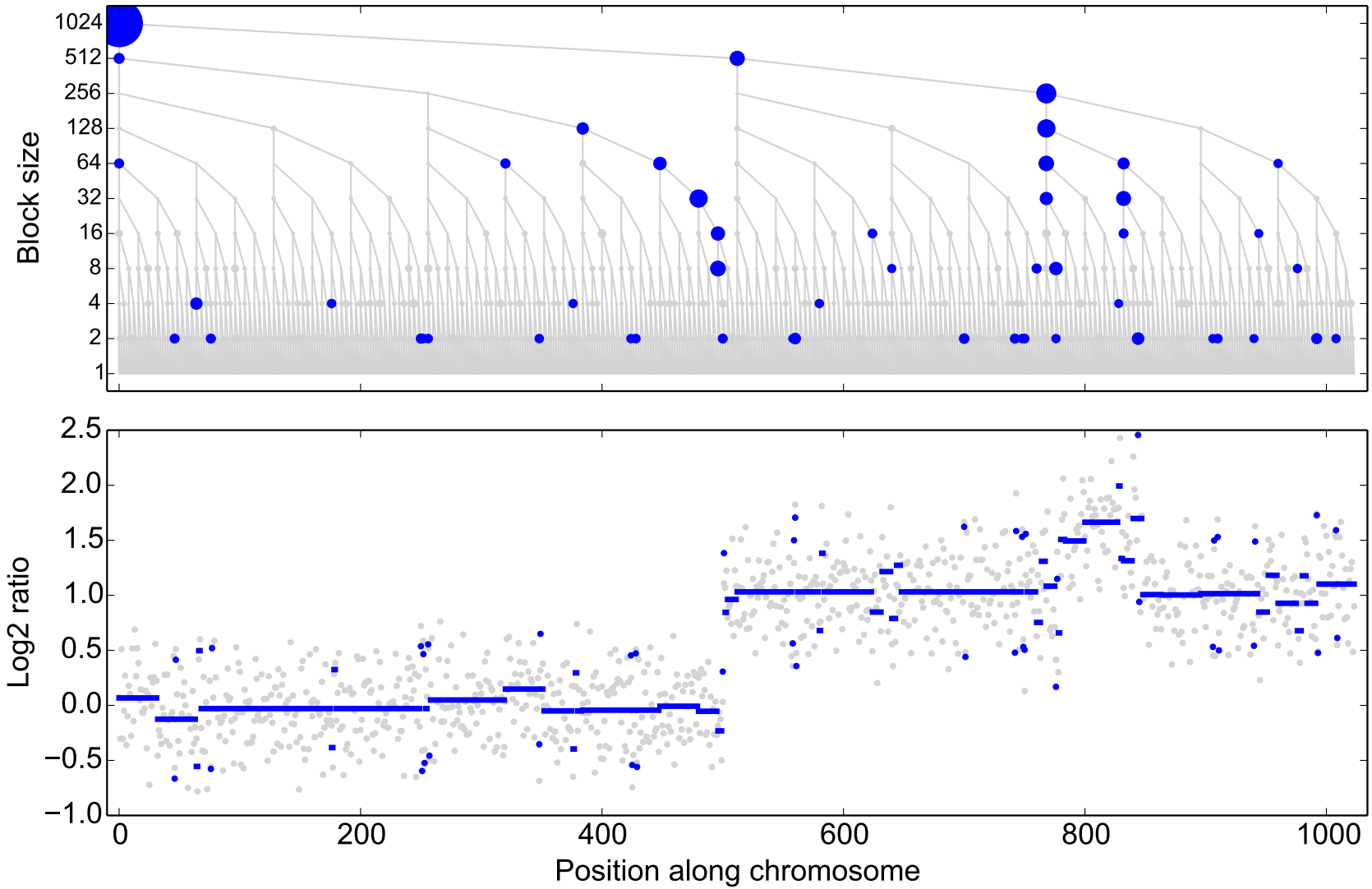
Haar wavelet regression for given noise variance



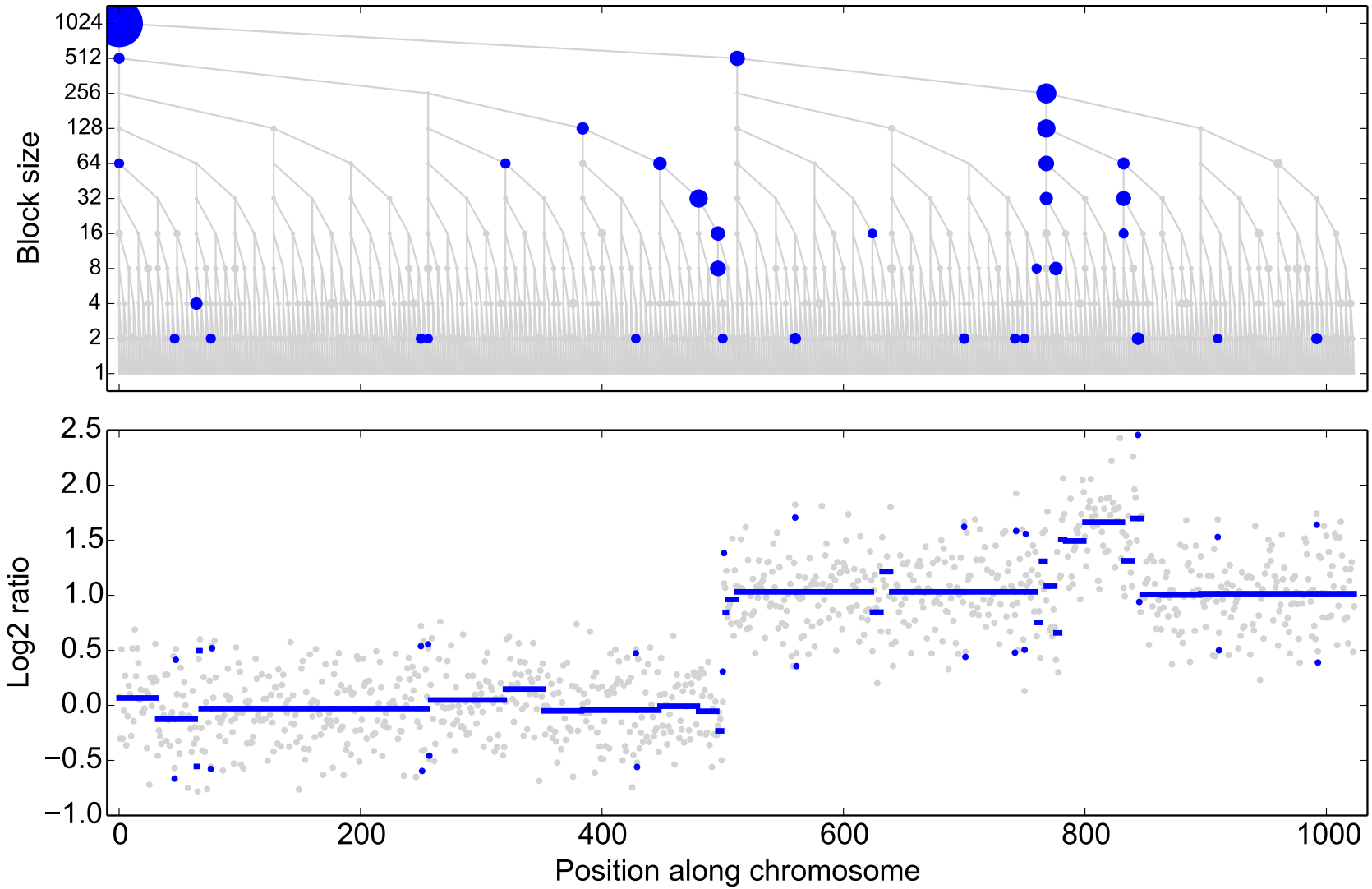
Haar wavelet regression for given noise variance



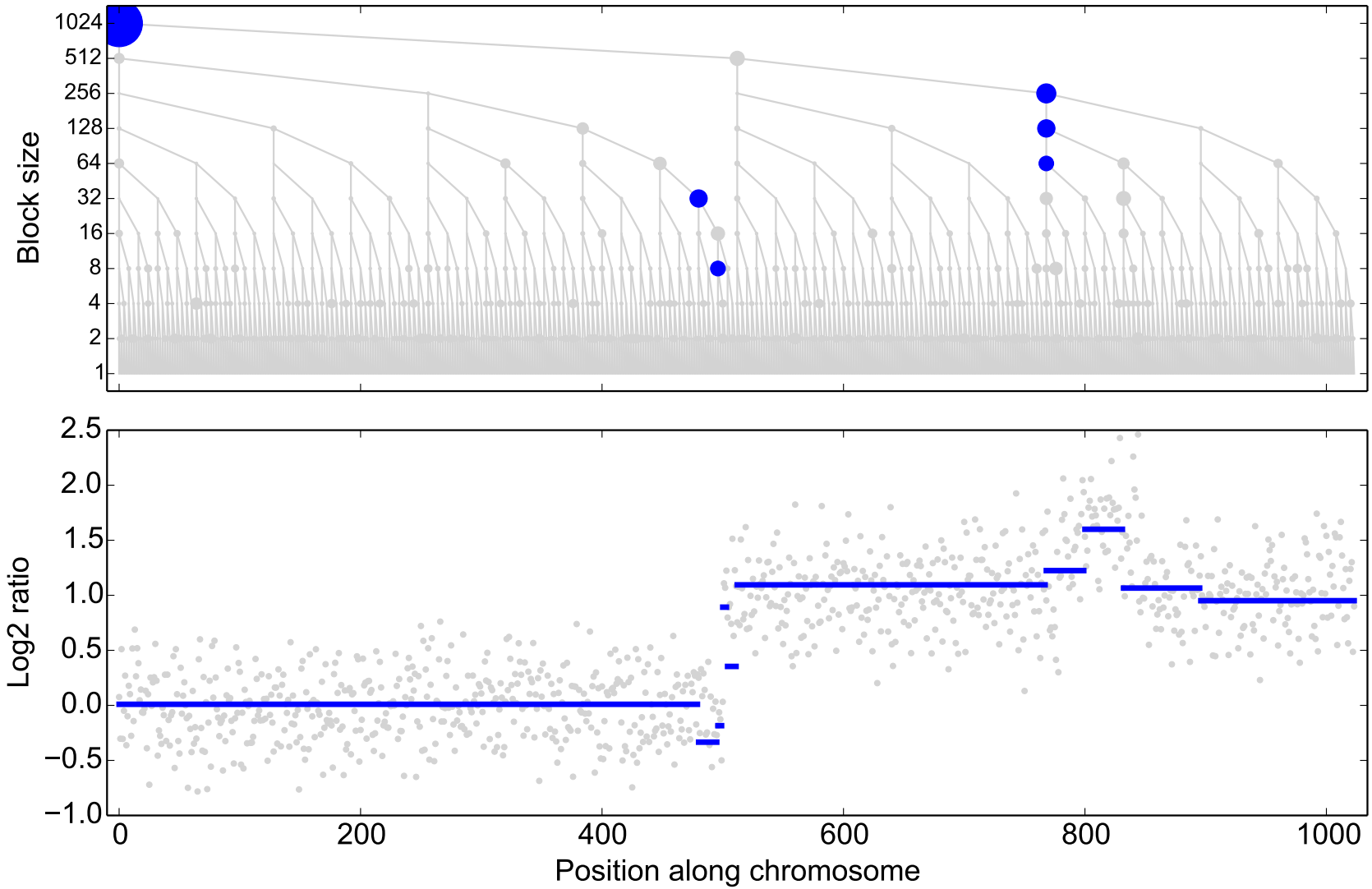
Haar wavelet regression for given noise variance



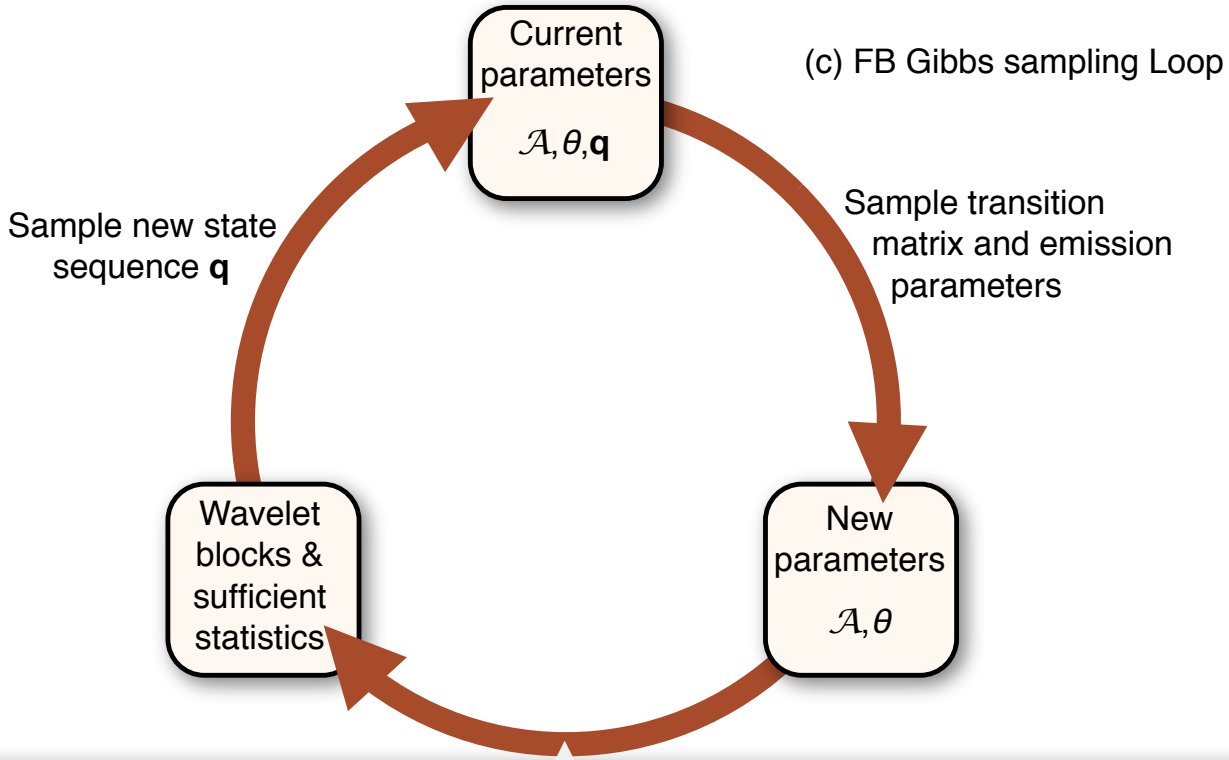
Haar wavelet regression for given noise variance

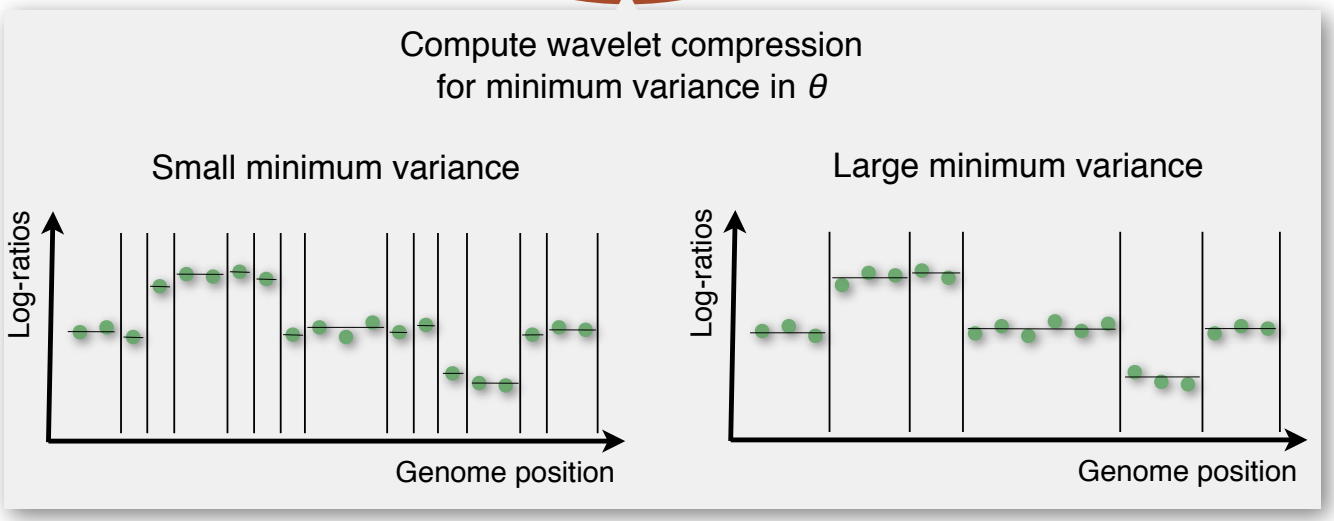


Haar wavelet regression for given noise variance

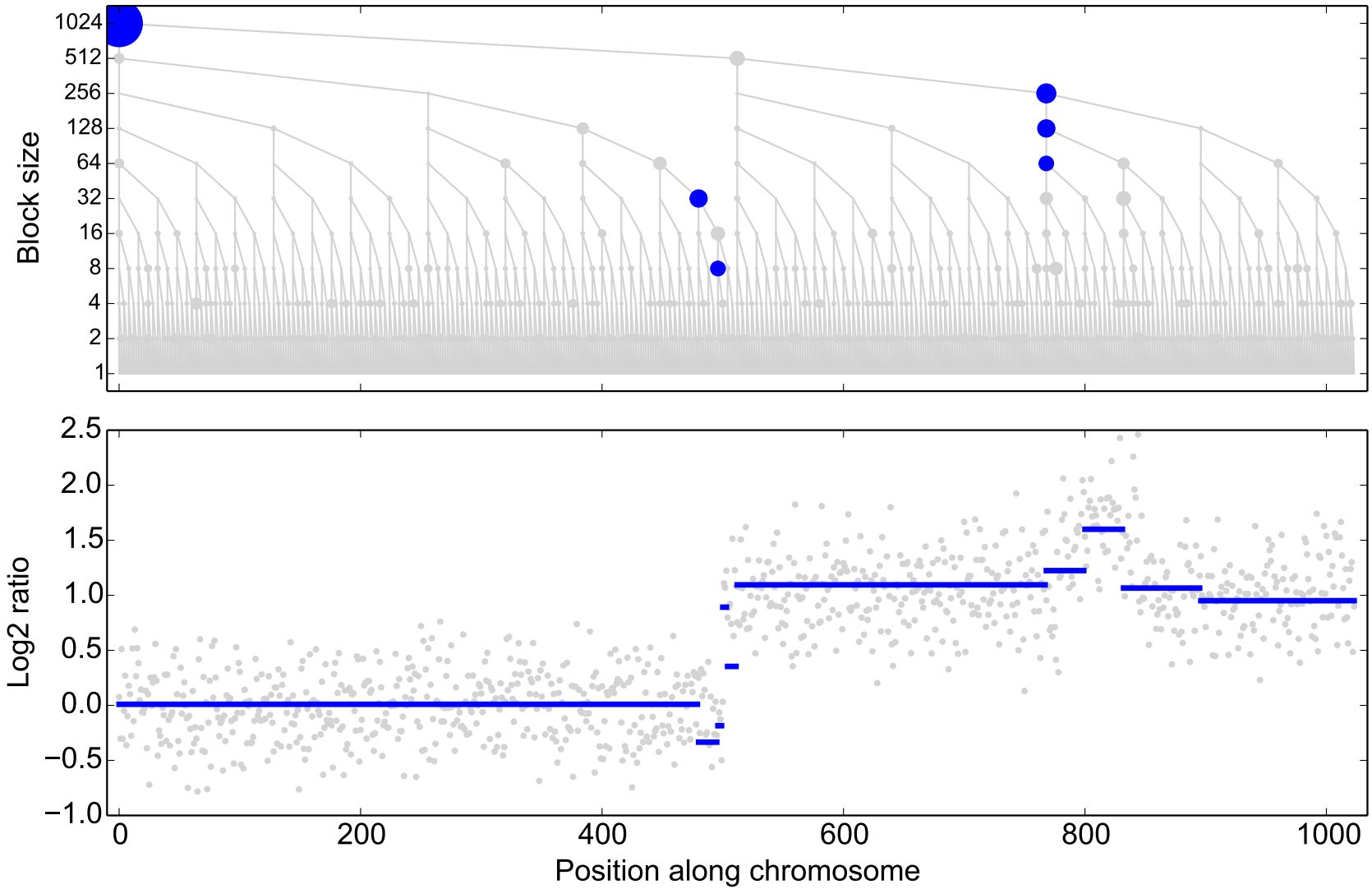


HaMMLET – Dynamically compressed FBG

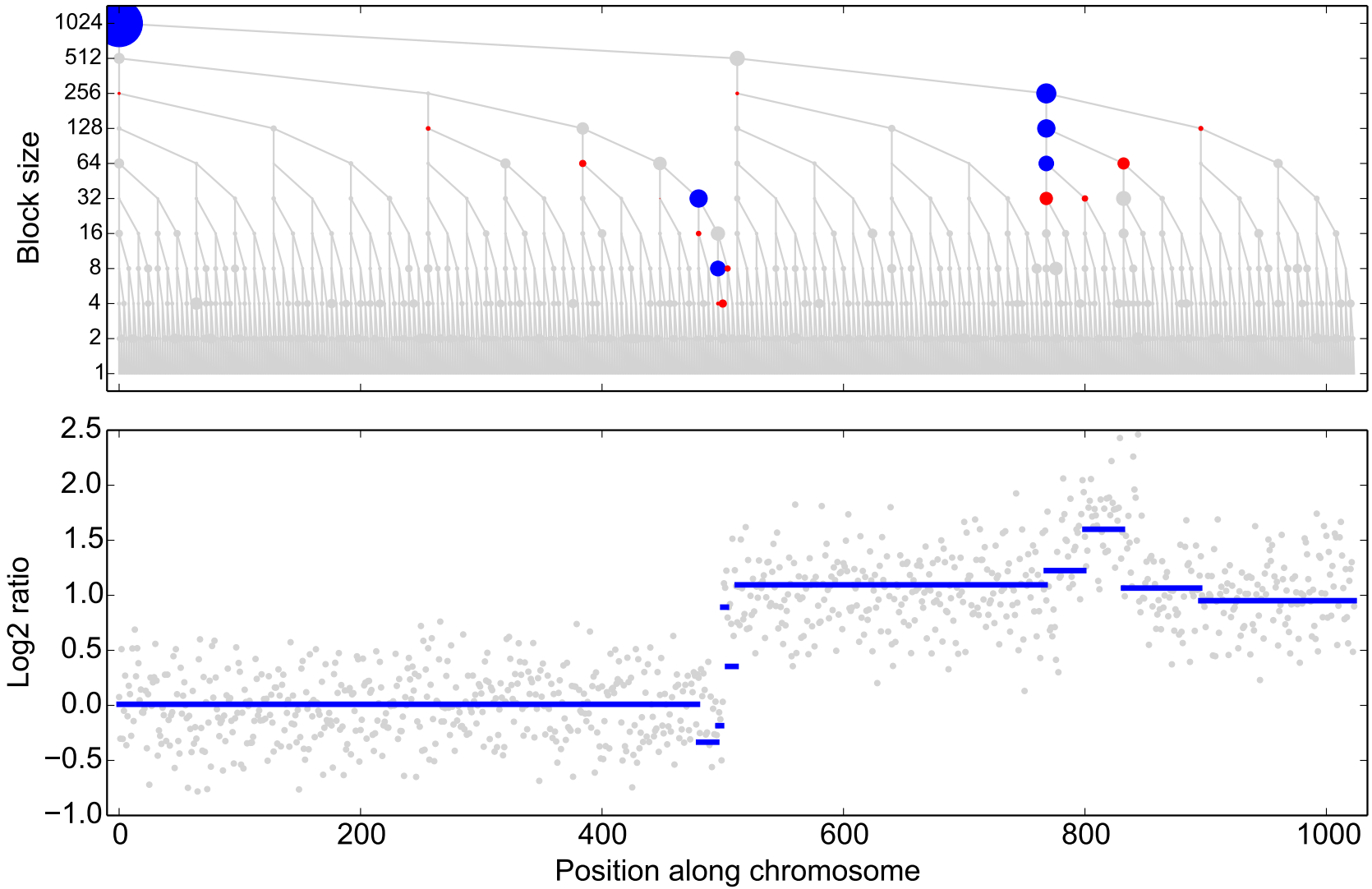




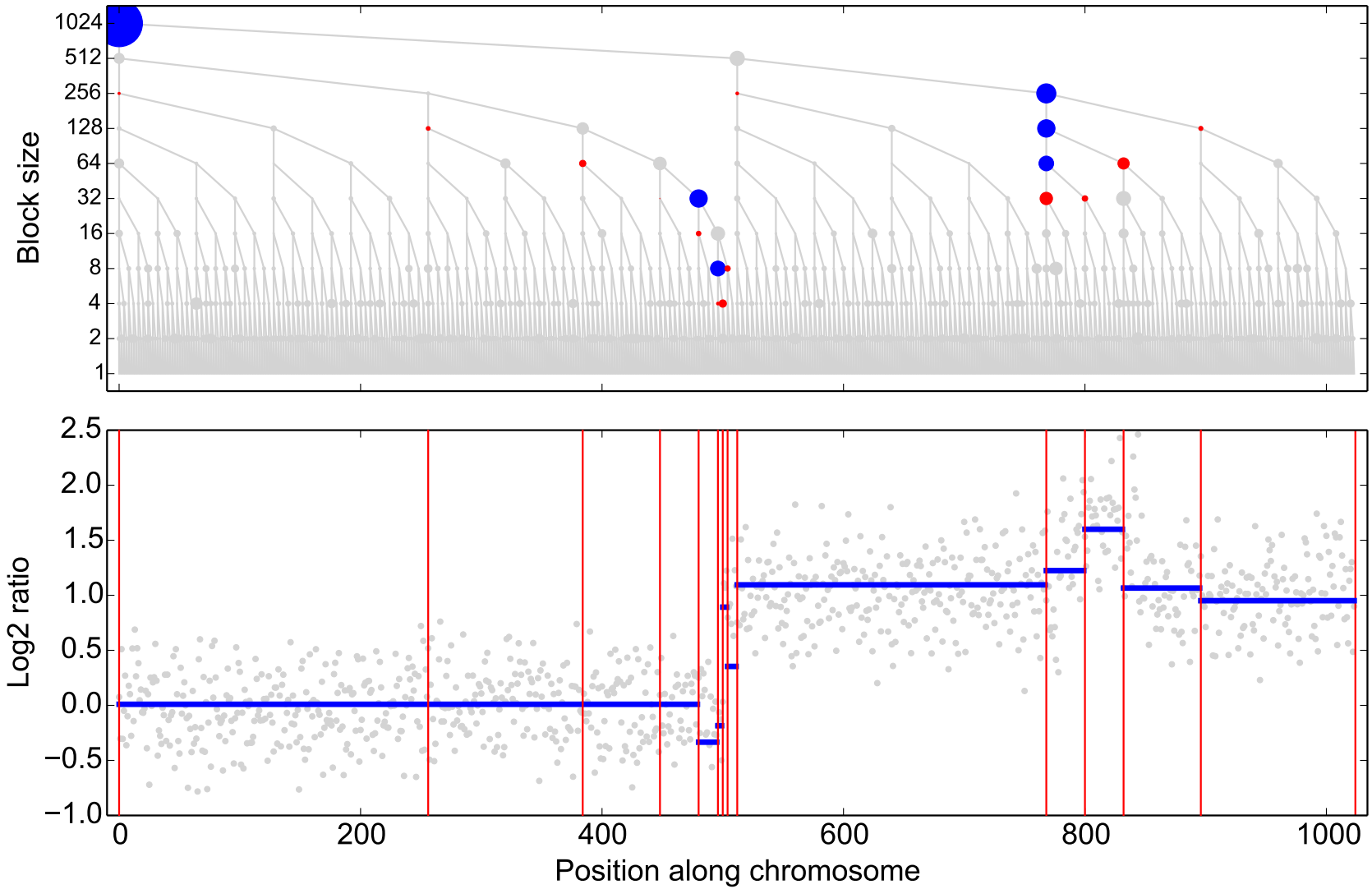
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



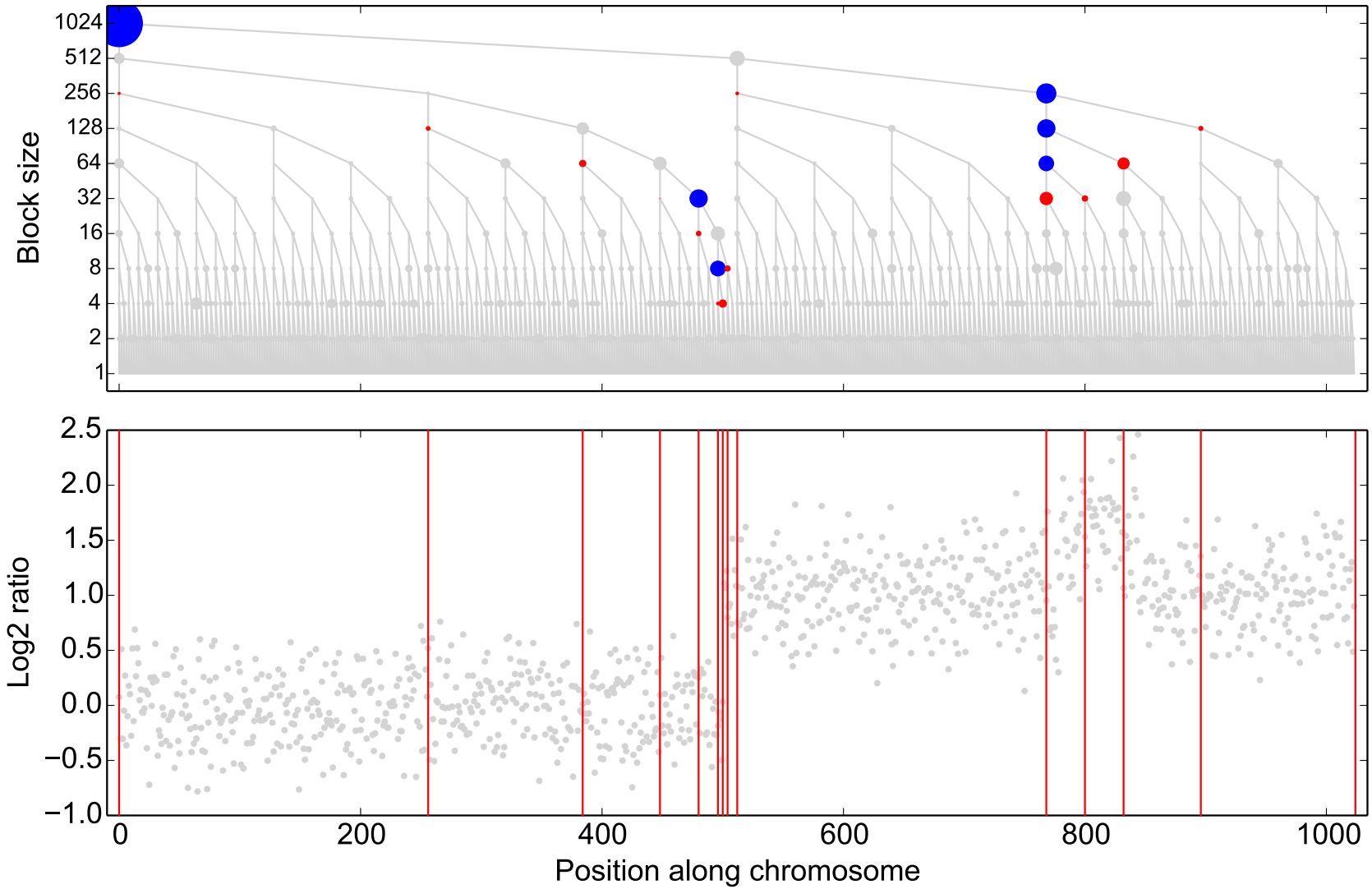
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



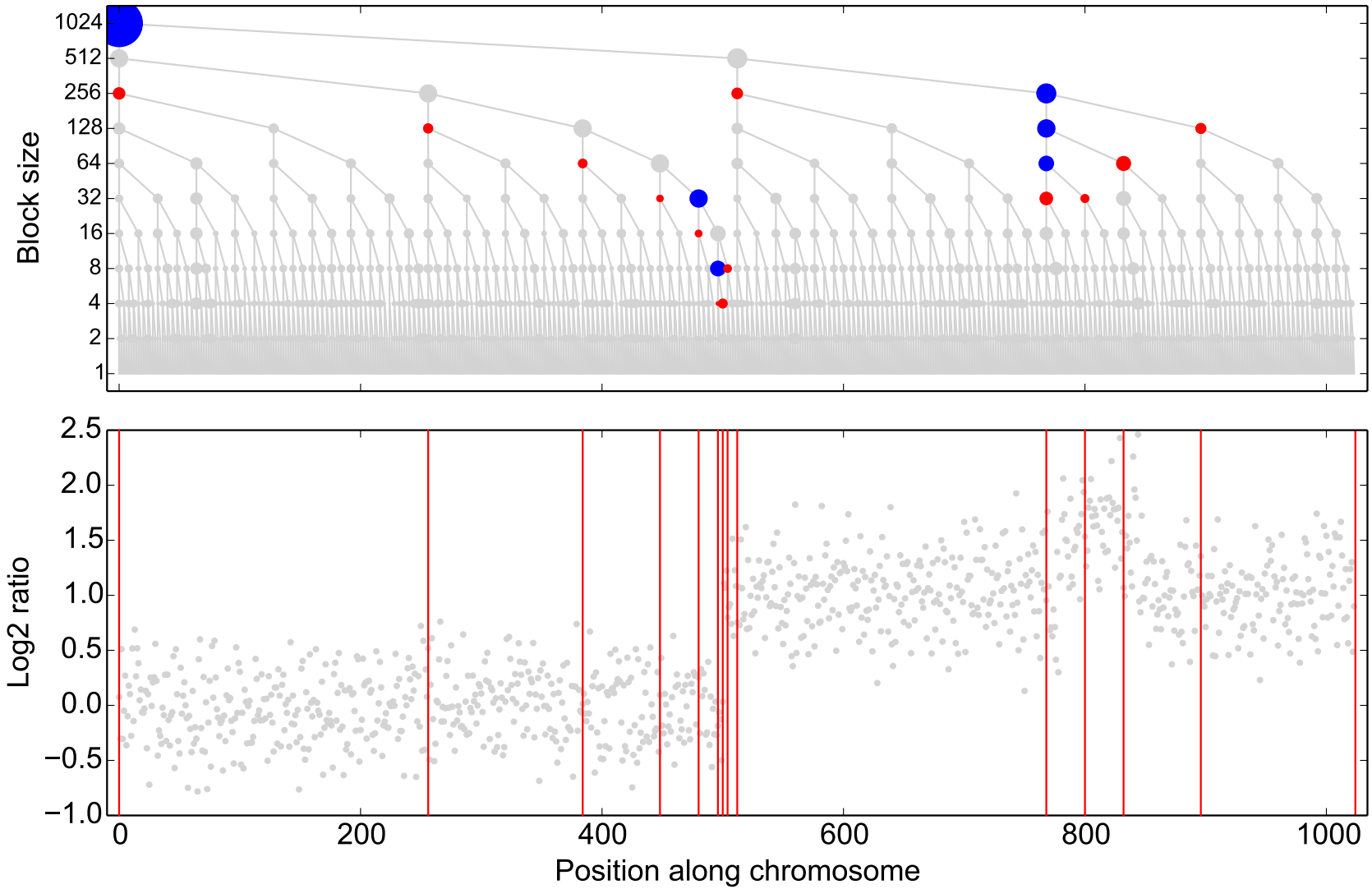
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



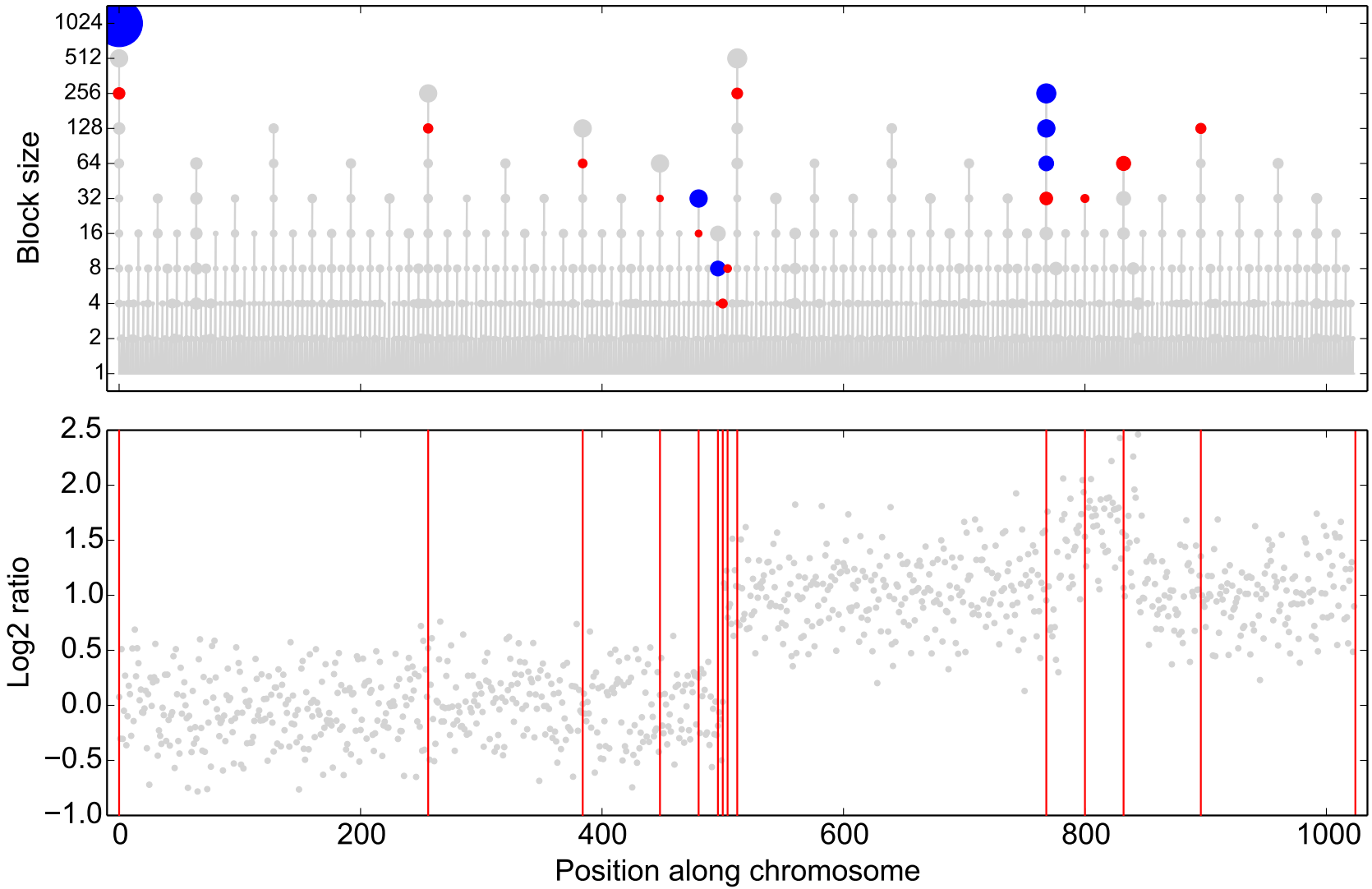
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



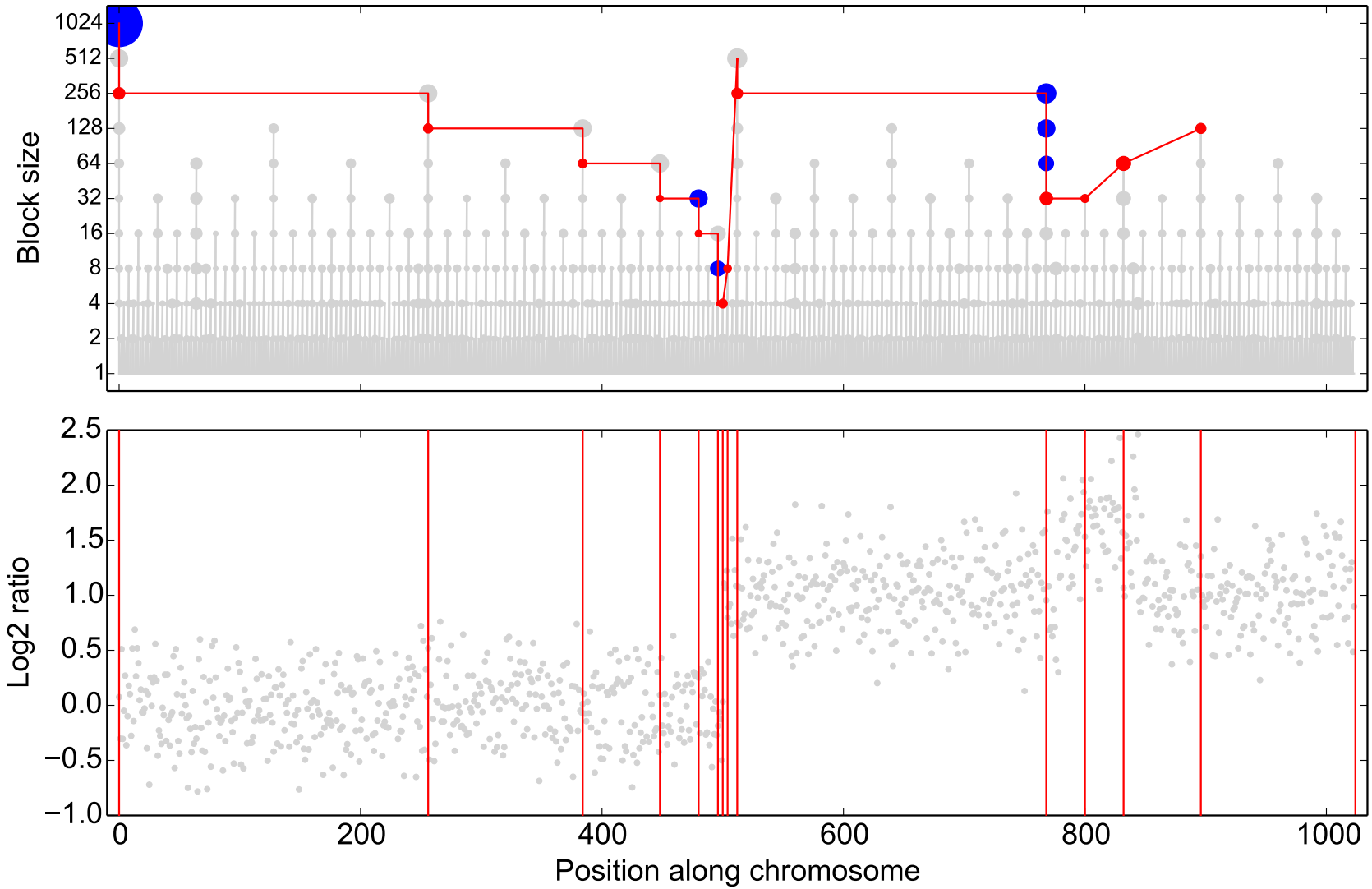
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



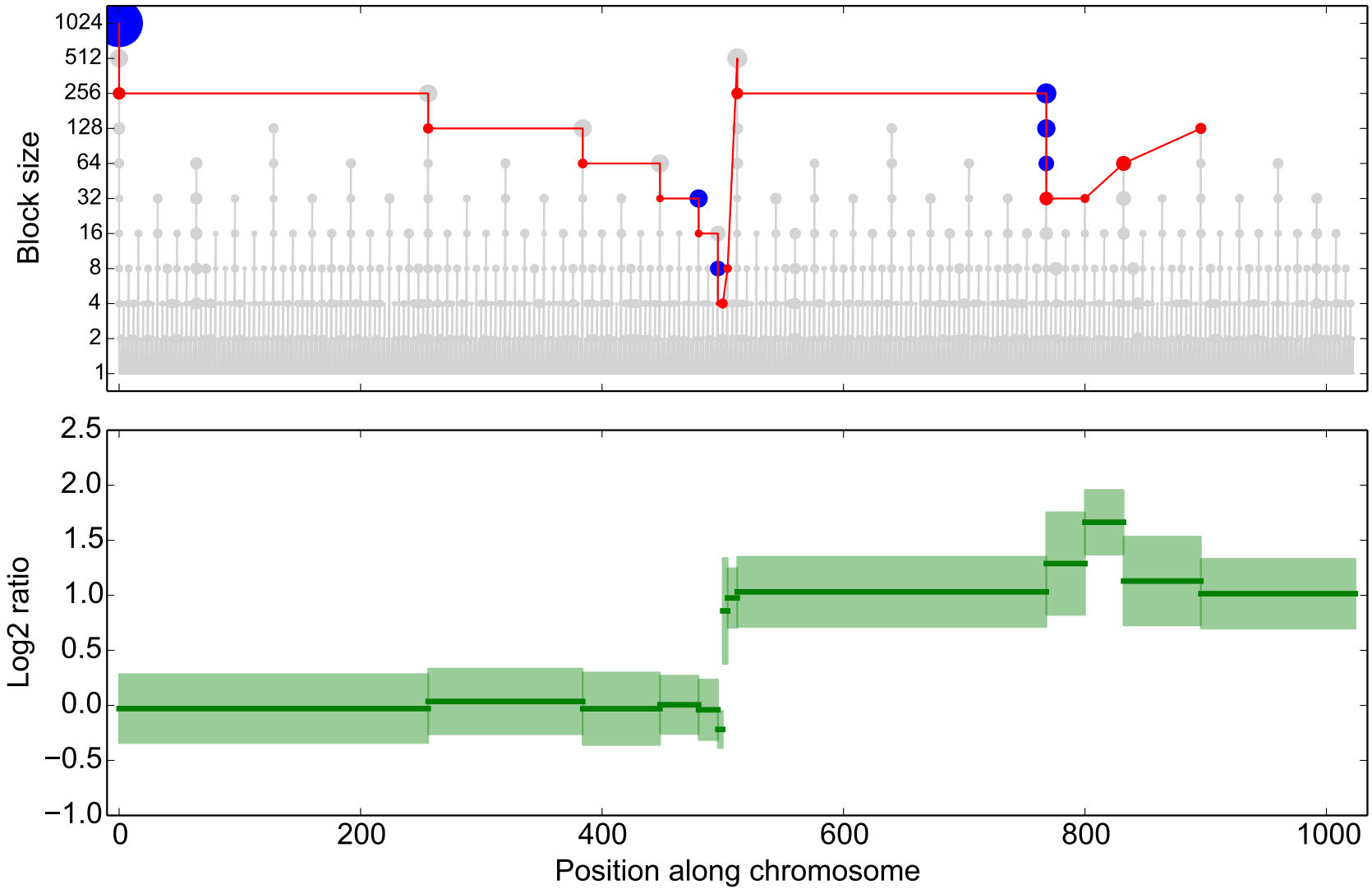
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



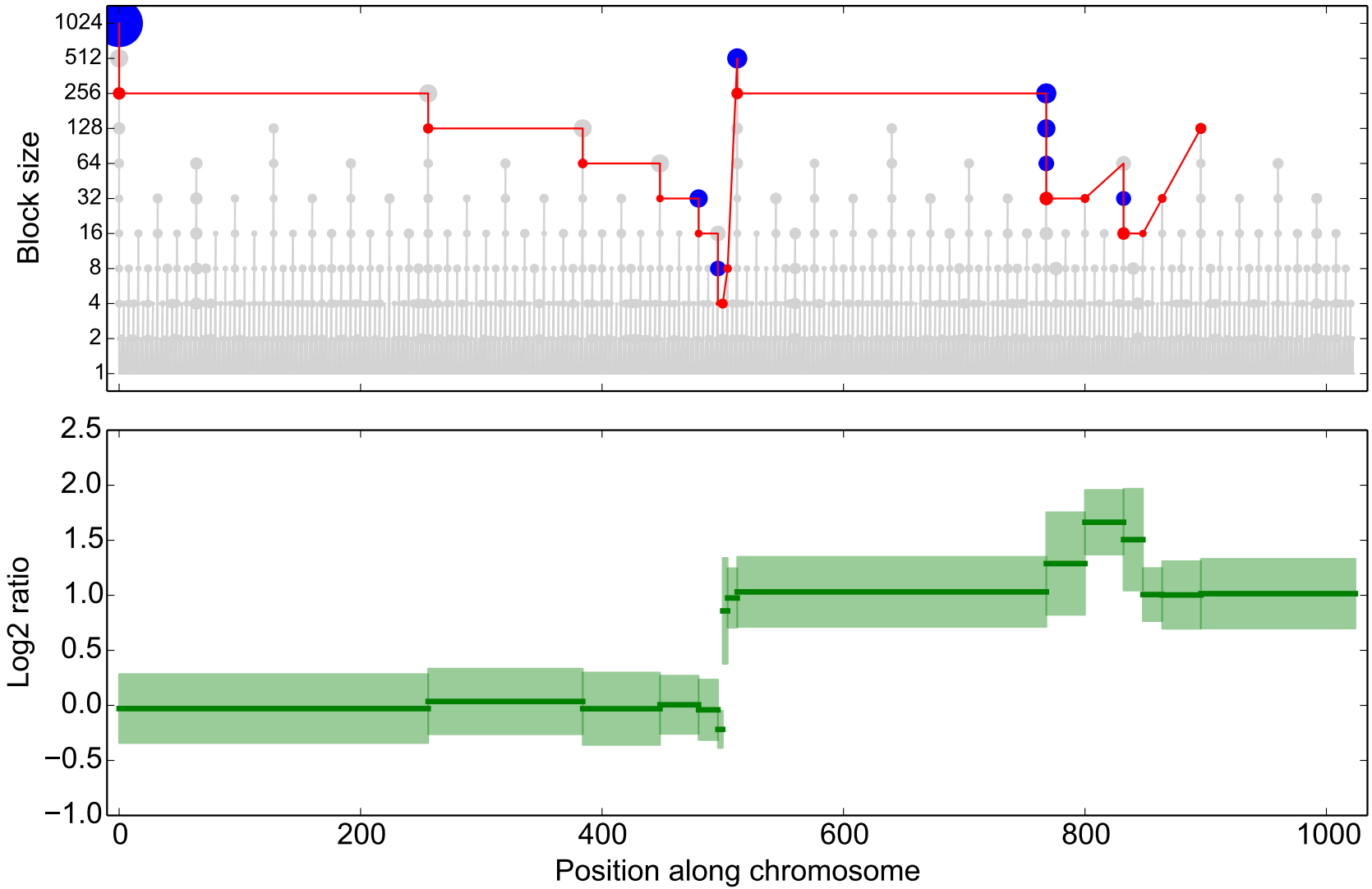
Create blocks of sufficient statistics in $\Theta(\#\text{blocks})$



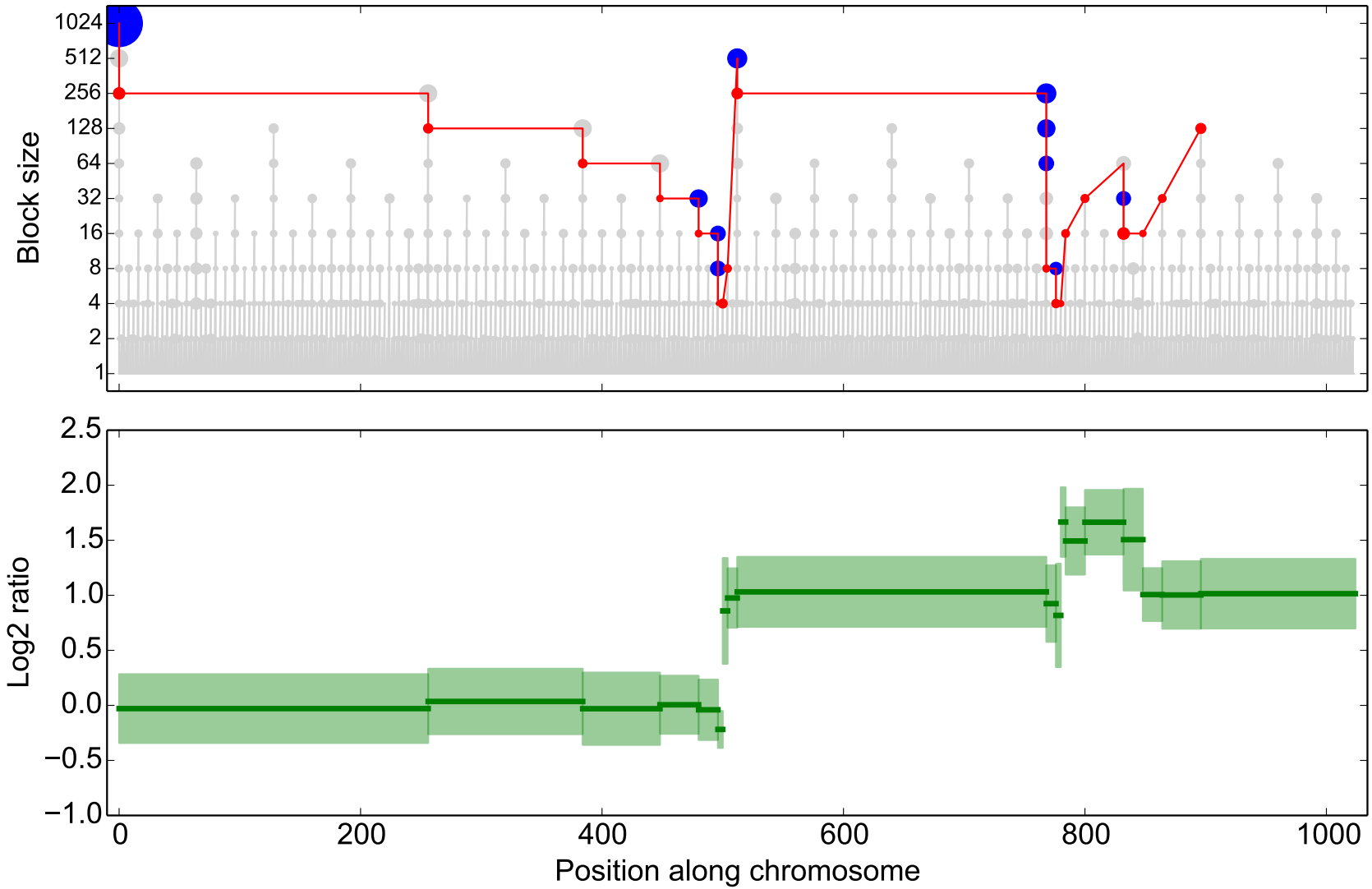
Sufficient statistics in wavelet tree



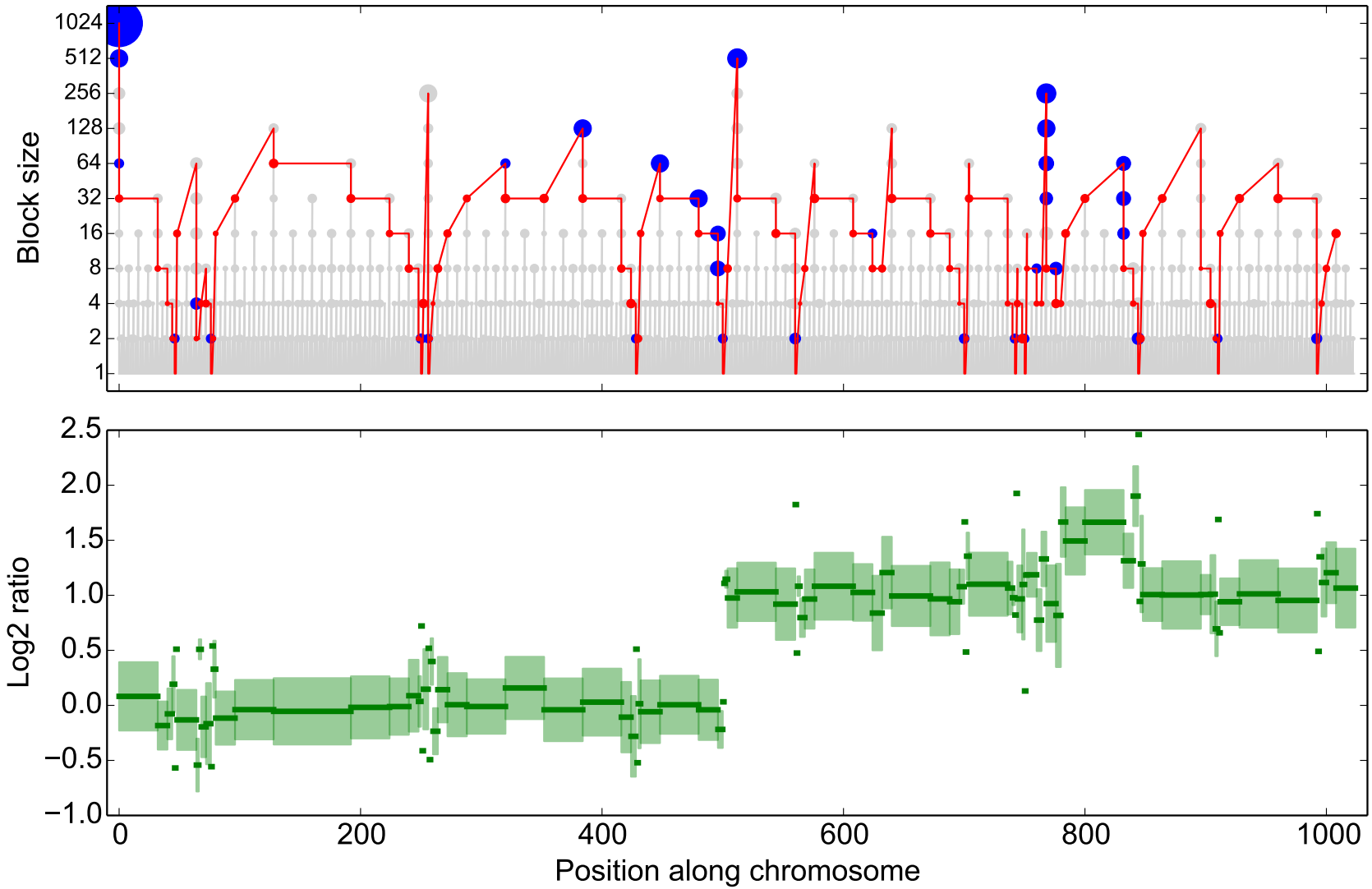
Sufficient statistics in wavelet tree



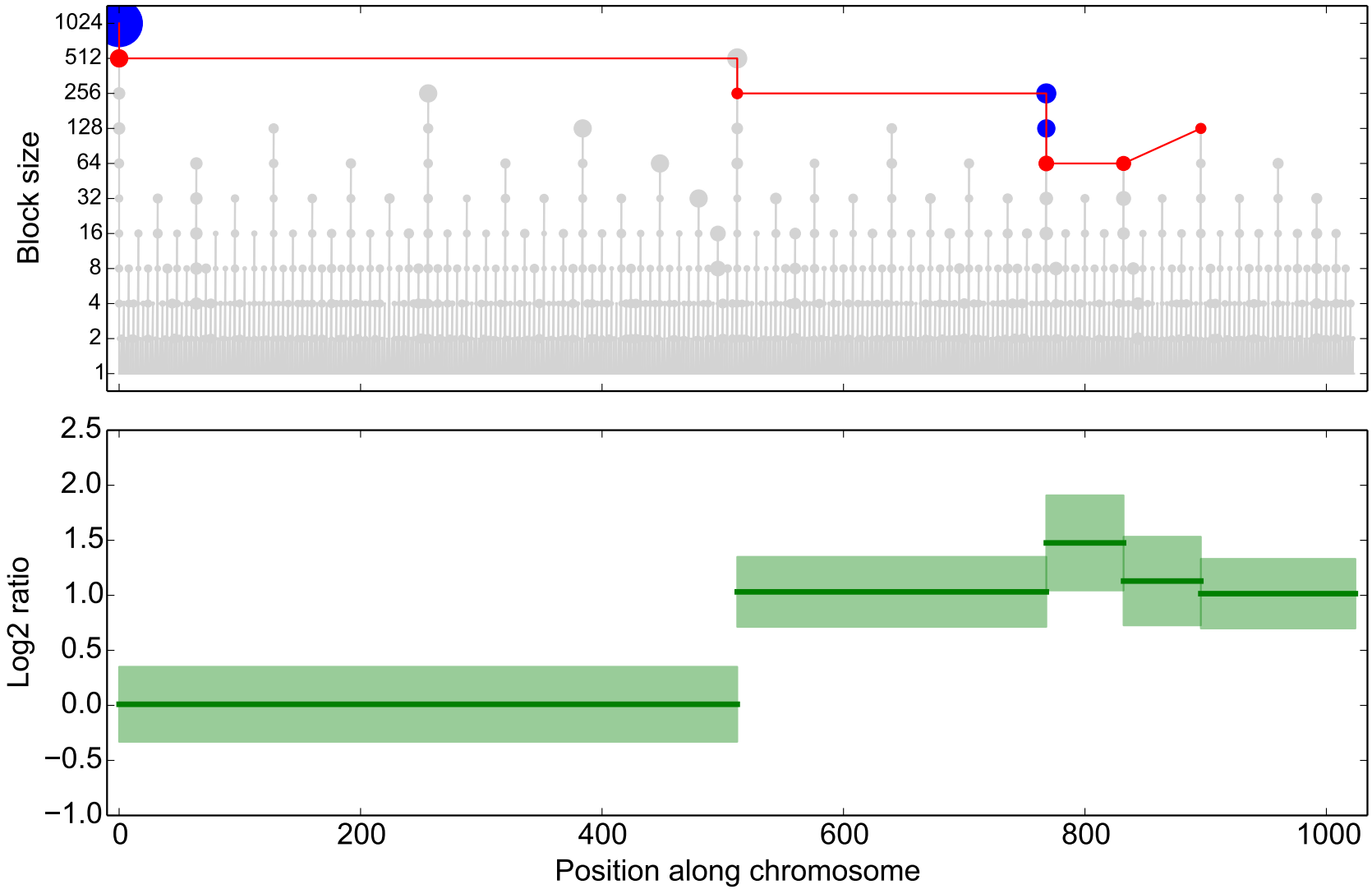
Sufficient statistics in wavelet tree



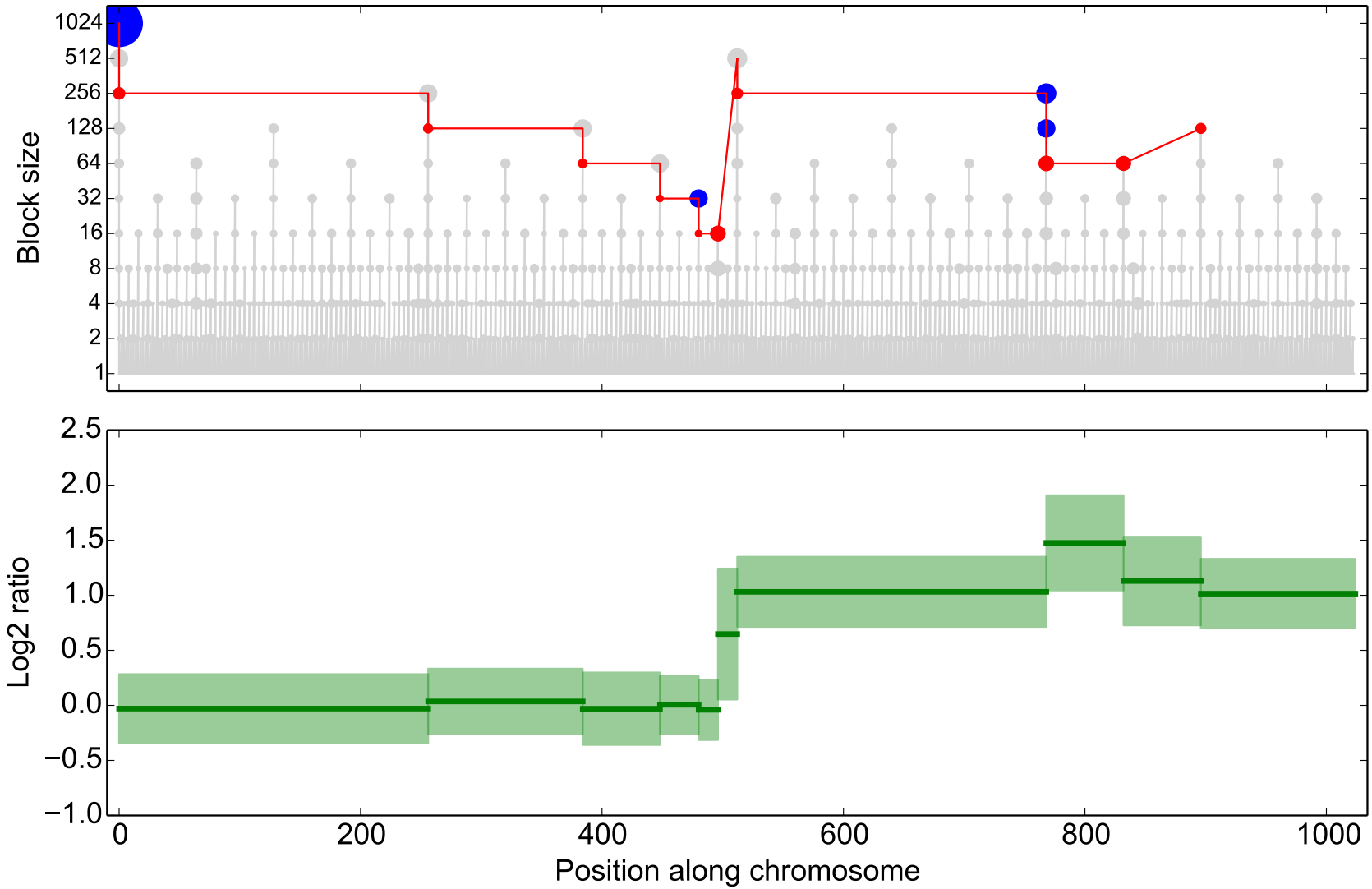
Sufficient statistics in wavelet tree



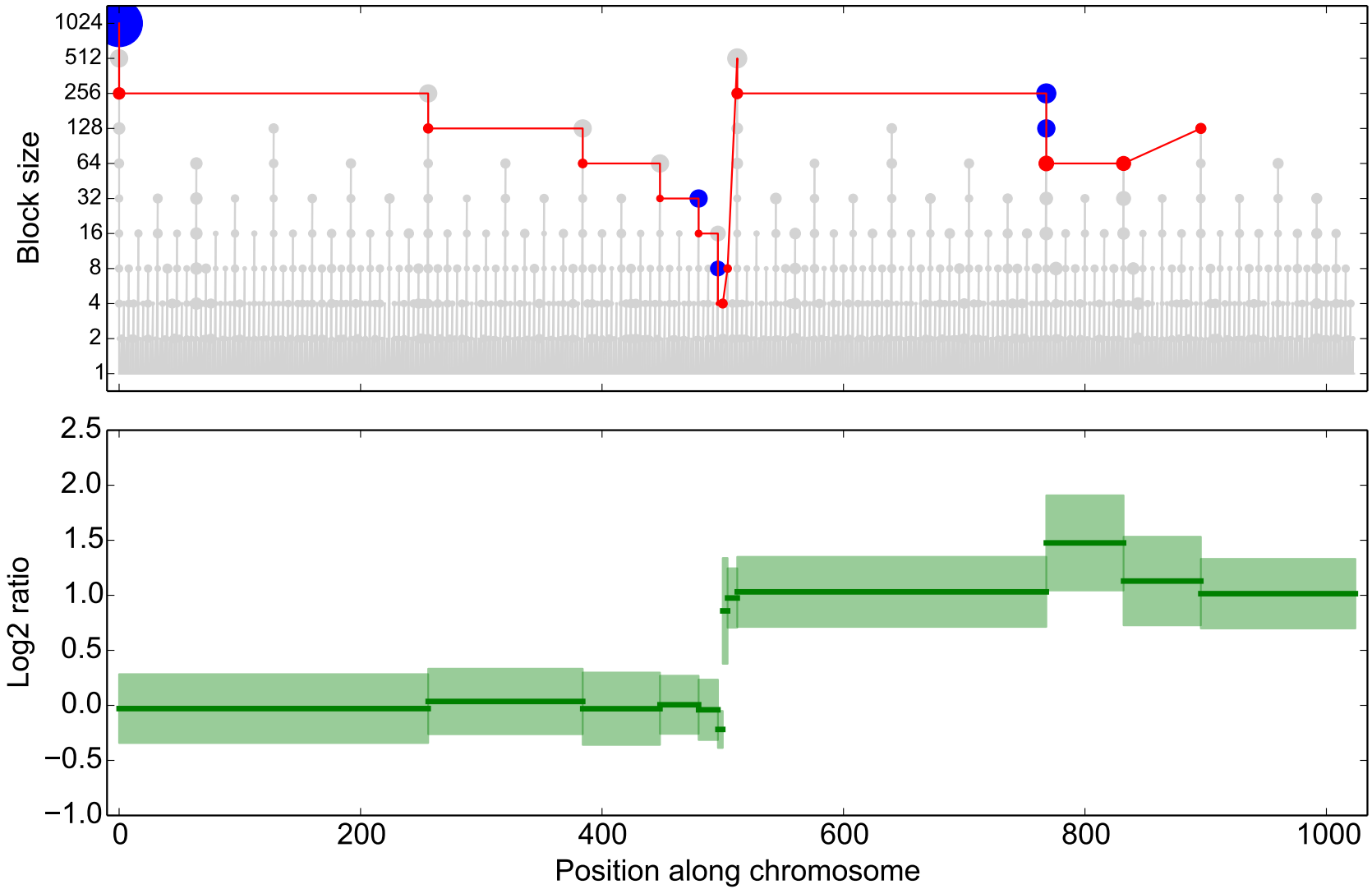
Sufficient statistics in wavelet tree



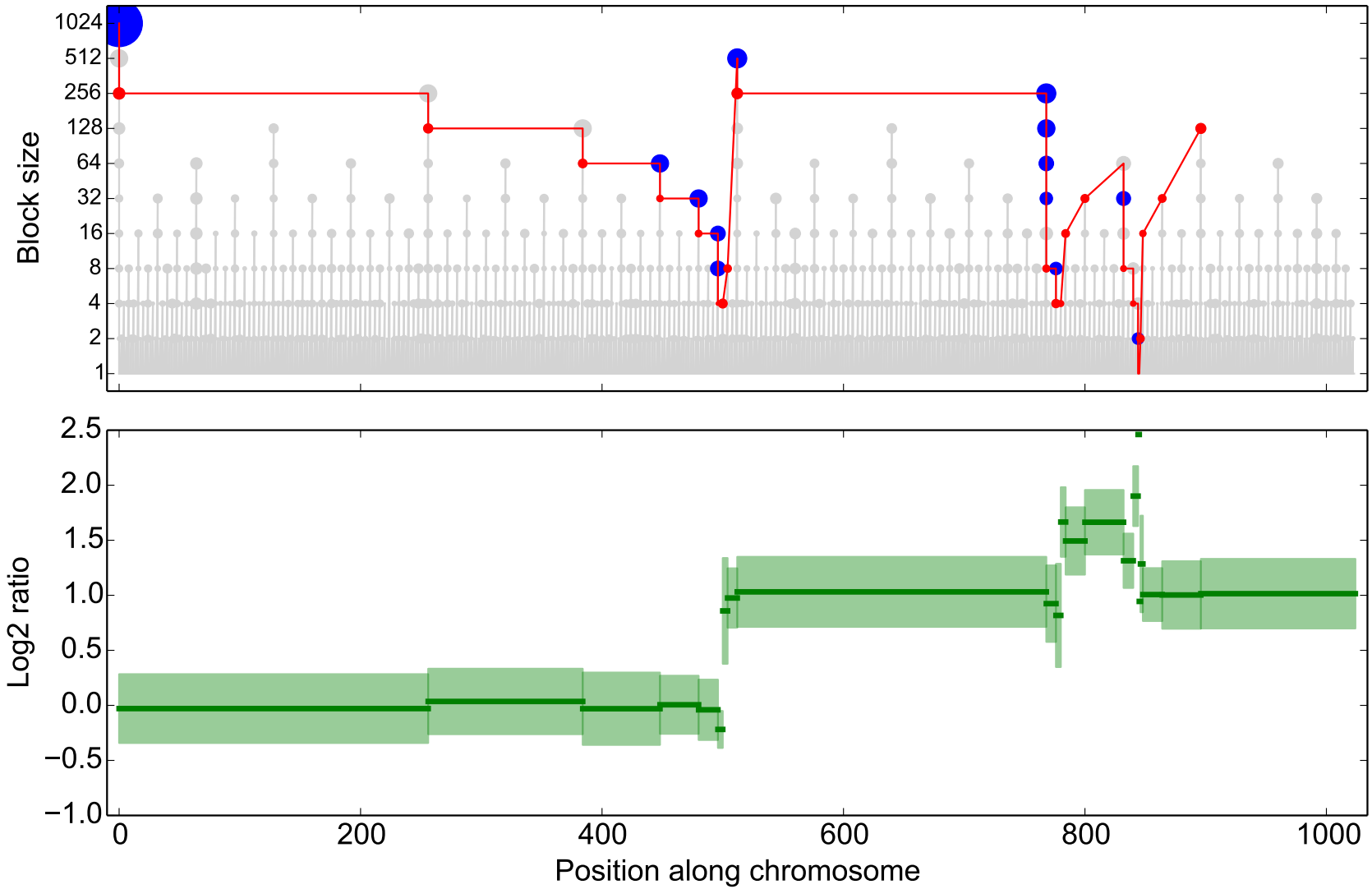
Sufficient statistics in wavelet tree



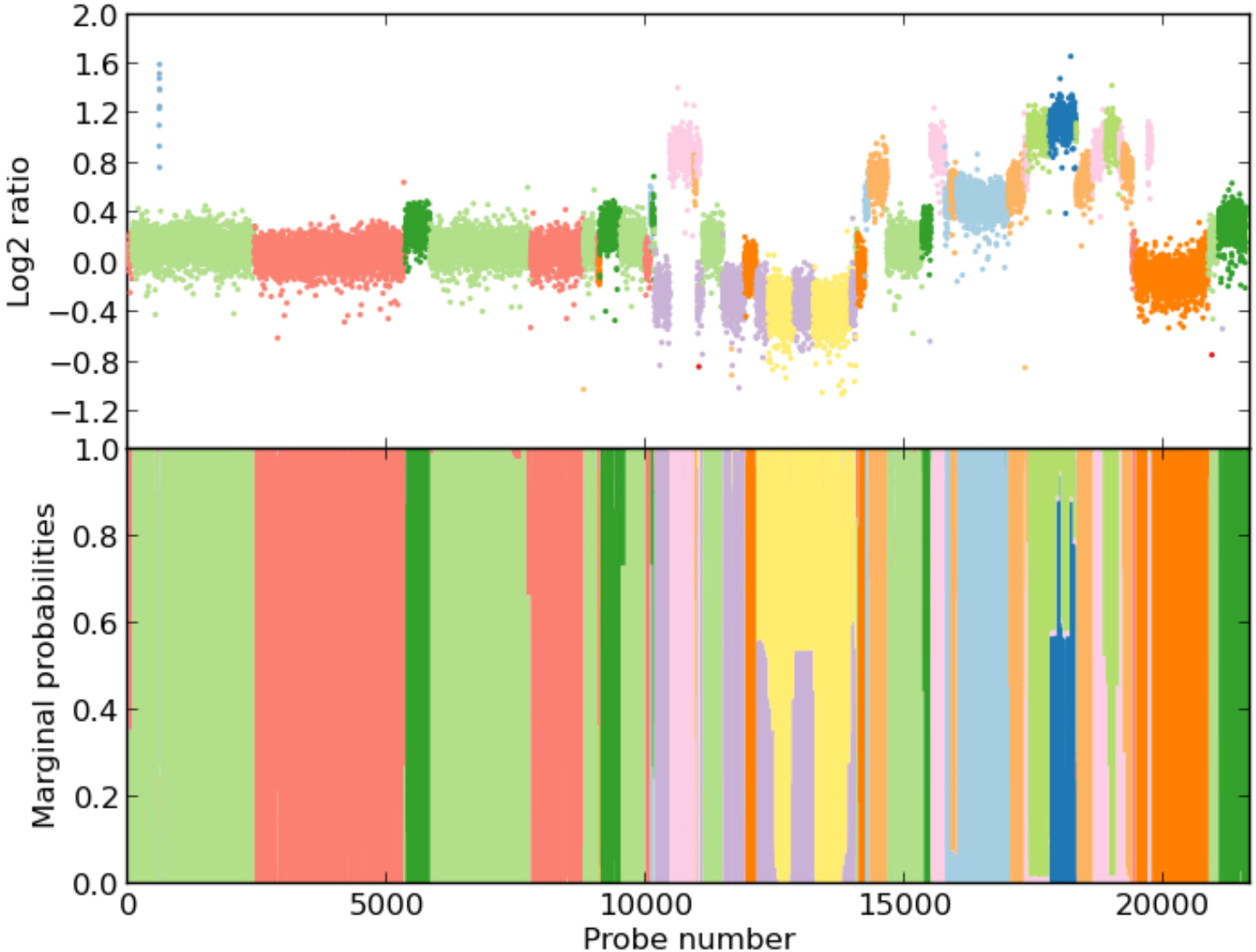
Sufficient statistics in wavelet tree



Sufficient statistics in wavelet tree



aCGH of invasive ductal carcinoma cell line (BT-474)[4]



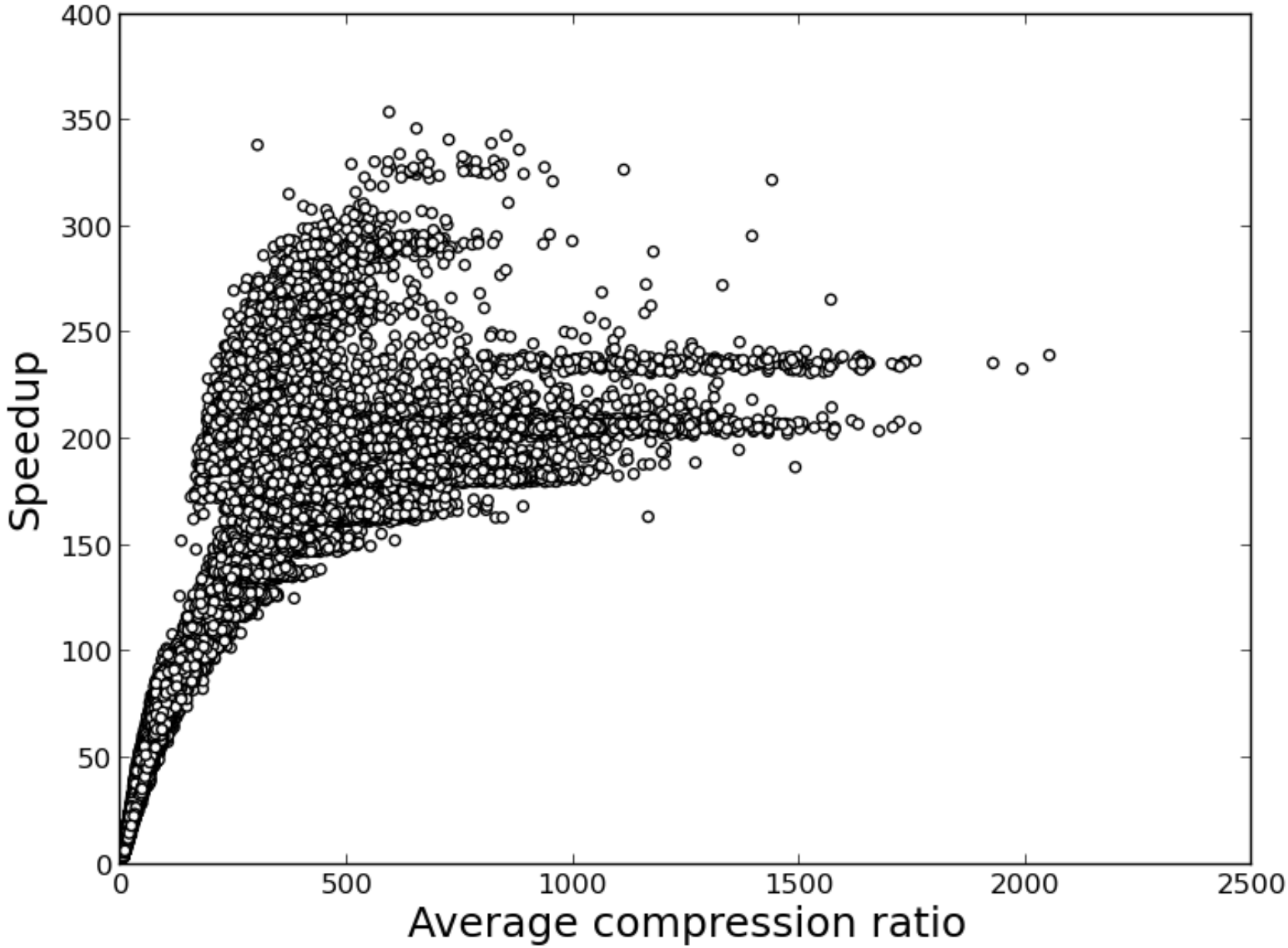
[4] Edgren et al. "Identification of fusion genes in breast cancer by paired-end RNA-sequencing." In: *Genome Biology* 12.1 (Jan. 2011), R6. DOI: 10.1186/gb-2011-12-1-r6.

Evaluation on simulated data

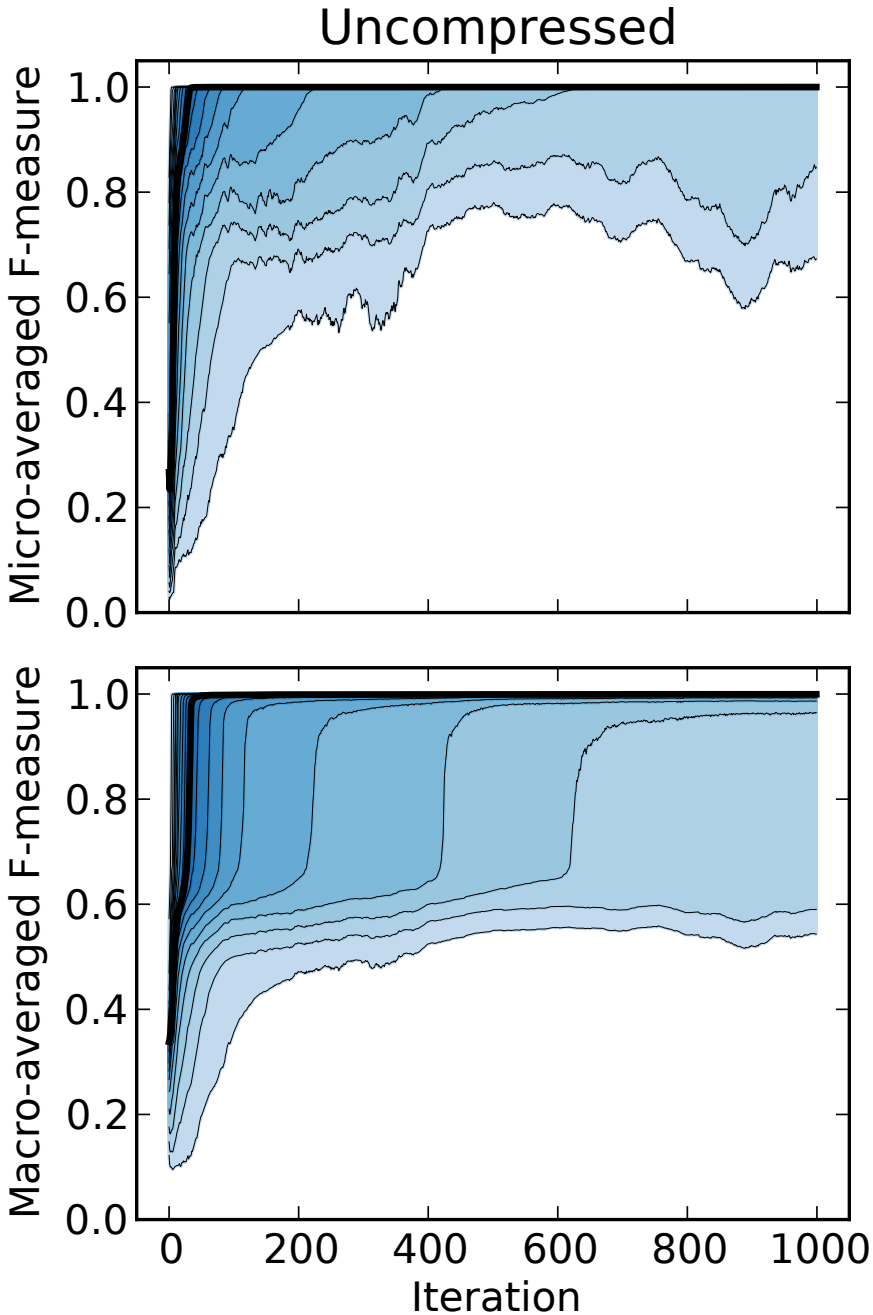
- Simulation
  - 129,600 different data sets (4.2 billion data points)
  - 3 states
  - Univariate data $T = 2^{11} = 32,768$ with Gaussian noise
  - Different combinations of noise parameters
  - $\{1, \ldots, 6\}$ gains of total length $\in \{100, 250, 500, 750, 1000\}$
  - $\{1, \ldots, 6\}$ losses of total length $\in \{100, 250, 500, 750, 1000\}$
- Inference
  - Various Dirichlet priors for transitions and state distribution
  - Normal-Inverse Gamma noise priors automatically derived from wavelet transform
- Evaluation:
  - Compressed vs. uncompressed HMM
  - Speedup through compression
  - Convergence in terms of multi-class F-measures[5]

[5] Özgür, Özgür, and Güngör. "Text Categorization with Class-Based and Corpus-Based Keyword Selection". In: *Proceedings of the 20th International Conference on Computer and Information Sciences* 3733 (2005), pp. 606–615. DOI: 10.1007/11569596.

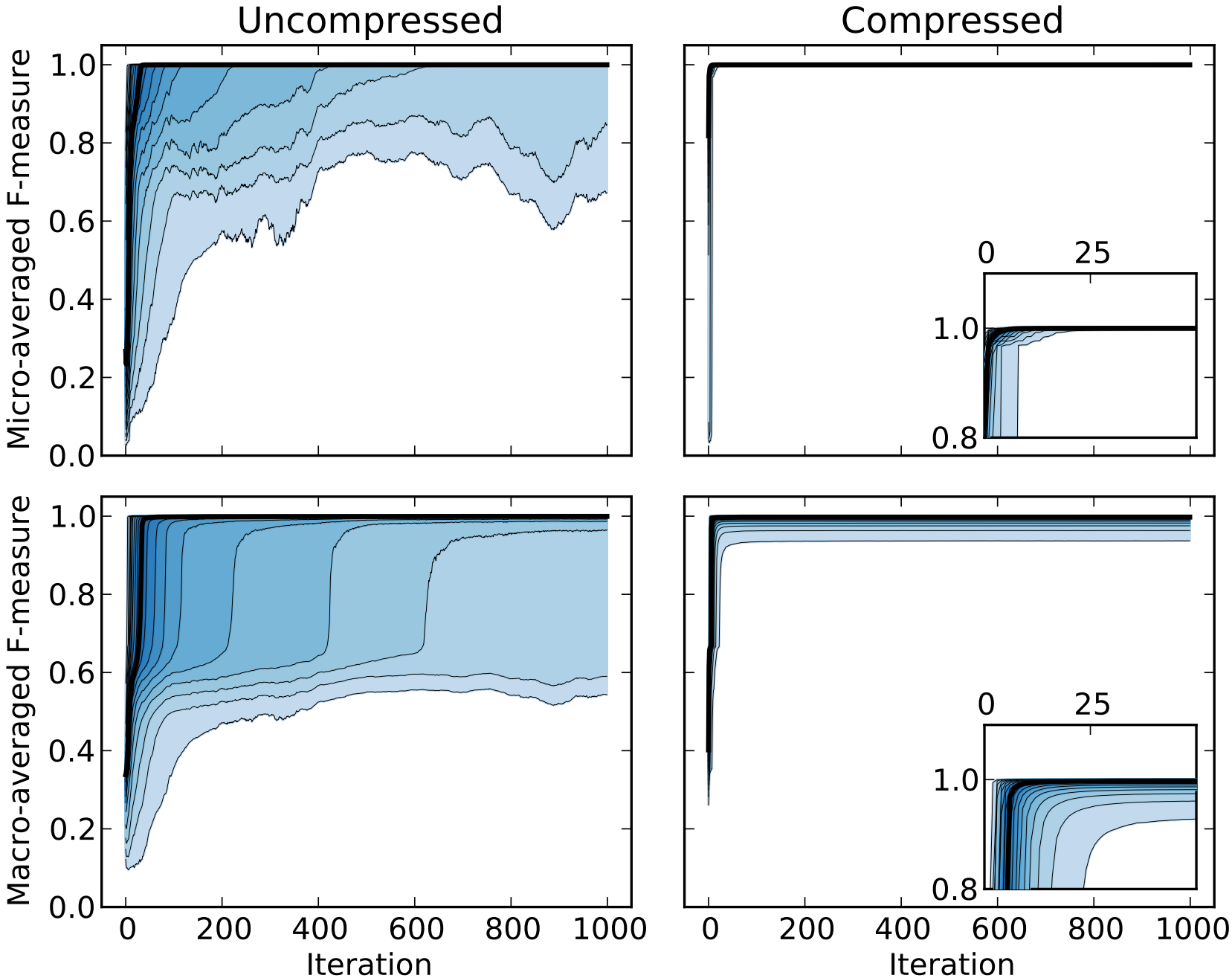
Speedup per iteration



Convergence



Convergence



Summary

- Adaptive, dynamic wavelet-based compression for Forward-Backward Gibbs sampler
- Improved speed and convergence of Bayesian inference in HMM
- Preprocessing linear in data size
- Block creation and query of sufficient statistics linear number of blocks (compression)

Thank you!

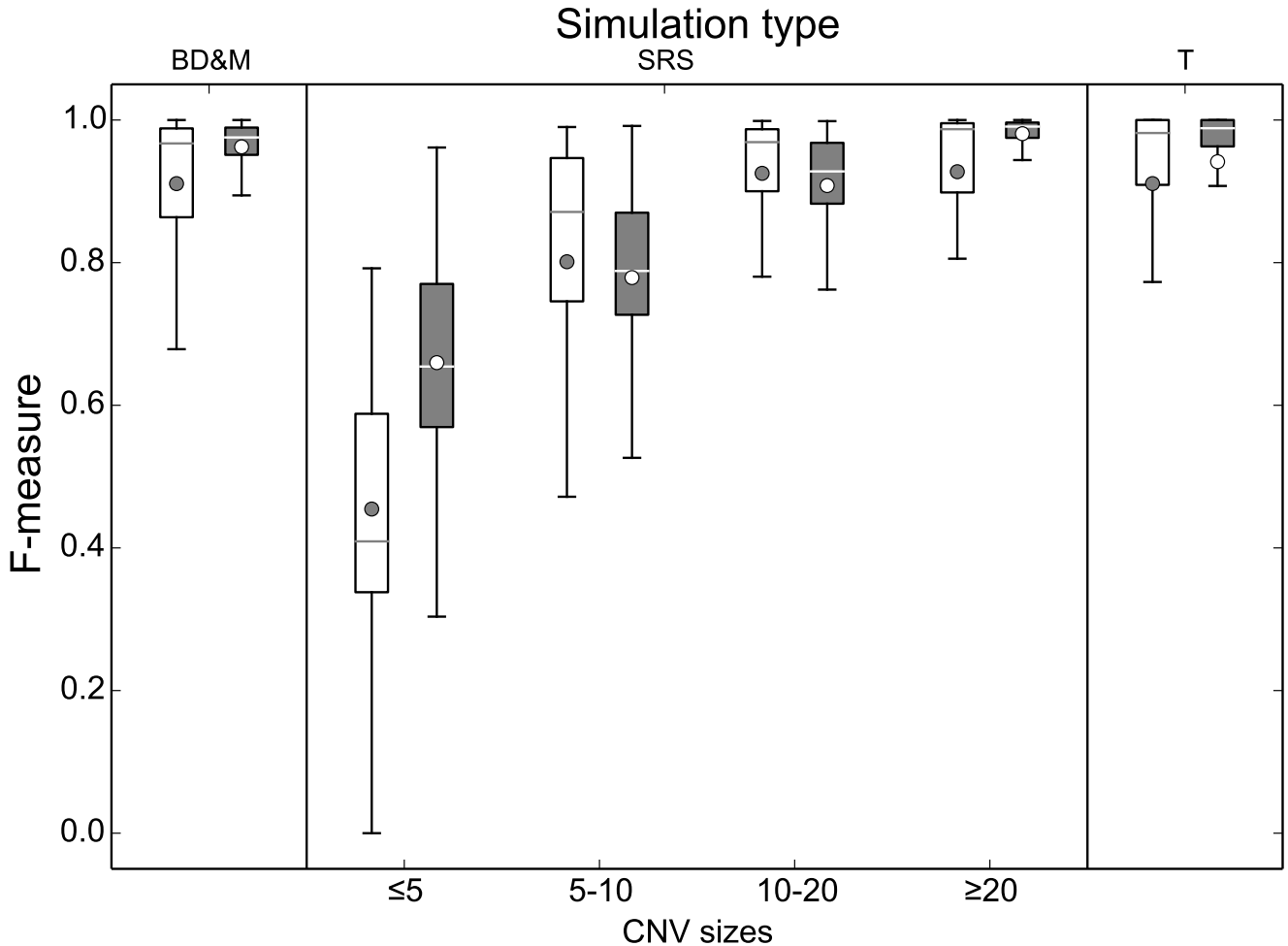
Wiedenhoeft, Brugel, and Schliep. "Fast Bayesian Inference of Copy Number Variants using Hidden Markov Models with Wavelet Compression". In: PLOS Computational Biology (2016, accepted for publication). Preprint: http://biorxiv.org/content/early/2015/07/31/023705

Software: http://schlieplab.org/Software/HaMMLET

Funding

- National Institute of Health: "Meaningful Data Compression and Reduction of High-Throughput Sequencing Data", award 1 U01 CA198952-01
- National Science Foundation: "Research Experience for Undergraduates", award 1263082
- RECOMB 2016 Travel Support

Comparison with CBS[6]



[6] Venkatraman and Olshen. "A faster circular binary segmentation algorithm for the analysis of array CGH data." In: *Bioinformatics* 23.6 (Mar. 2007), pp. 657–63. DOI: 10.1093/bioinformatics/btl646, Willenbrock and Fridlyand. "A comparison study: applying segmentation to array CGH data for downstream analyses." In: *Bioinformatics* 21.22 (Nov. 2005), pp. 4084–91. DOI: 10.1093/bioinformatics/bti677.