Supplementary Text – 1

Details of Implementation of the Algorithm for Finding "Variant" Motifs, with Special Reference to Case-Control Data

Consider data matrices $((a_{ij}))_{N\times L}$ and $((b_{ij}))_{N\times L}$, where a_{ij} denotes a nucleotide (A,T,G or C) at the j^{th} polymorphic site ($j = 1, 2, \ldots, L$) for the i^{th} individual ($i = 1, 2, \ldots, N$) among the cases and b_{ij} denotes the same for the i^{th} individual ($i = 1, 2, \ldots, N$) among normal controls. These data matrices are generated from aligned DNA sequences of a specific homologous genomic segment of $2N$ individuals (N cases and N controls), from which all monomorphic sites have been removed. For simplicity we have considered the sample size (N) to be same for cases and controls, but these can be different in practice. For genotype data, we can numerically recode genotypes, e.g., AA as 0, AG as 1 and GG as 2 (assuming that there are two variant nucleotides A and G at the locus under consideration) or we can estimate haplotypes and carry out analyses where a_{ij} denotes the nucleotide (A,T,G or C) at the j^{th} polymorphic site ($j = 1, 2, \ldots, L$) for the i^{th} haplotype ($i = 1, 2, \ldots, N$) among the cases or among controls. We also note that if disjoint segments of DNA are to be simultaneously examined for motif finding, then appropriate segments may be separately aligned and the aligned segments concatenated in the data matrix.

As before, let $V = \{1, 2, \ldots, L\}$ denote the set of all L polymorphic sites in the data. Let Π_p denote the set of all possible combinations of p sites in V. In general, $\Pi_p = \{V_p^k\}$, $k = 1, 2, \ldots, \binom{L}{p}$ and $V_p^k = \{x_1^k, x_2^k, \ldots, x_p^k : x_i^k \in V\}$. For a fixed k, we define the *modal sequence among cases* on V_p^k as that particular combination of nucleotides at the sites $\{x_1^k, x_2^k, \ldots, x_p^k\}$ included in V_p^k, $k = 1, 2, \ldots, \binom{L}{p}$, which has the highest frequency (f_{case}) among the cases. In the data matrix of Supplementary Table 1, the modal sequence, for example, on $V_2^1 = \{1, 2\}$ is AG with frequency 2, on $V_2^2 = \{1, 3\}$ is AT with frequency 3, etc.

Since our goal is to simultaneously look for a sequence that occurs with a high frequency among controls and is variant to that occurring at a large frequency among the cases, we evaluate the following:

Given a specific string, S_p, of length p, we first find the *modal sequence among cases* (ξ_{case}) and its frequency f_{case}. Let us denote that sequence by ξ_{case}. Then we enumerate from the control data all possible sequences $\xi_{l,p}$ $(l = 1, 2, \ldots)$ of nucleotides, at the sites included in S_p. For each such sequence $\xi_{l,p}$, we calculate its frequency $f_{l,p}$. We then calculate the number of mismatches, $m_{l,p}$, of each of these sequences $\xi_{l,p}$ $(l = 1, 2, \ldots)$ with the reference sequence. Since we are interested in identifying a motif that is different from the reference sequence (in this case ξ_{case}), we need to take these $m_{l,p}$ values into account. We provide a greater weightage to a sequence that has a larger number of mismatches with the reference sequence. In this problem, therefore, we have used a modified objective function of the form $G(S_p) = g(f_1, f_2, m)$, where f_1, f_2 and m denote, respectively, the frequencies of the string of nucleotides at the sites in S_p among cases and controls, respectively, and m denotes the number of mismatches between these two nucleotide sequences among cases and controls. The specifics of the use of such an objective function are explained below with the help of an example.

In the data matrix of Supplementary Table 1, the modal sequence among cases (ξ_{case}), on $V_2^2 = \{1, 3\}$ is AT with frequency 3. For this set of sites $V_2^2 = \{1, 3\}$ there are 2 distinct sequences, AT and GC, in the control data set with frequencies 3 and 1, respectively. The number of mismatches of these two sequences with ξ_{case}, on $V_2^2 = \{1, 3\}$ are, respectively, 0 and 2. Hence the value of the objective function for this choice of two candidate sites is: maximum of ((0.75+1)(0.75+1)(0) and (0.75+1)(0.25+1)(2)). Therefore, although the sequence AT among controls occur at a high frequency, it is not chosen as a candidate sequence because of its smaller number of mismatches with ξ_{case}. AT among cases and GC among the controls are the preferred candidates for variant motifs if these two sites are chosen. We then use the Metropolis-Hastings algorithm to choose the set of sites which globally maximises the objective function.

Supplementary Table 1 An Example of a Case-Control Data Matrix

	Sequence/	Variant Site No.						
	Individual No.	1	2	3	4	5	6	7
Cases	1	A	A	T	T	G	C	C
	2	A	G	T	C	G	C	T
	3	A	G	T	T	A	C	T
	4	G	G	C	C	A	T	T
Controls	1	A	G	T	T	G	C	C
	2	A	T	T	C	G	C	T
	3	A	T	T	T	A	C	T
	4	G	G	C	T	A	T	T

Supplementary Text – 2

Details on the Method of Generating Synthetic Data Set 1

Our synthetic Data Set 1, comprises a data-matrix of size $N \times L$, corresponding to data on N individuals at L binary polymorphic sites. At each site, for each individual, we assigned a binary digit (0 or 1) with probability 0.5. (However, as noted in the Introduction, the assumption of each polymorphic site being binary is not crucial to this algorithm.) A motif of length p was planted in a fraction u of the N individuals. To do this, we selected p sites randomly from the L sites; that is, we chose p columns of the $N \times L$ data-matrix randomly. Then, $[N \times u]$ rows were randomly chosen (where $[x]$ denotes the largest integer contained in x), and the elements of each of the p chosen columns corresponding to each of these chosen rows were replaced by 1. For a given set of values of (N, L, and p), 1000 independent synthetic data matrices were thus generated.

Supplementary Table 2. Performance of the algorithm on Synthetic Data Set 1 with *N*=200 for different values of the variables *L* (number of segregating sites) and *u* (proportion of the planted motif among *N*), and for different values of the control parameter *c*. (The mean and s.d. of the number of sweeps to convergence and the % of simulation runs in which the planted motif was correctly identified are based on 1000 independent simulation runs for each combination of values of the variables and the parameter. The minimum and maximum values of the significance level as the motif length was increased from 9 to 10 and from 10 to 11 were also calculated over 1000 simulation runs.)

L	***u***	***c*=50**					***c*=100**					***c*=200**				
		No. of Sweeps		**% correct**	**Sig. Level* (min,max)**		**No. of Sweeps**		**% correct**	**Sig. Level* (min,max)**		**No. of Sweeps**		**% correct**	**Sig. Level* (min,max)**	
		Mean	**s.d.**		**p: 9 → 10**	**p: 10 → 11**	**Mean**	**s.d.**		**p: 9 → 10**	**p: 10 → 11**	**Mean**	**s.d.**		**p: 9 → 10**	**p: 10 → 11**
50	0.3	130.13	57.34	100	.024, .027	529, 532	72.52	48.71	100	.024, .027	529, 532	66.26	49.29	100	.024, .027	529, 532
	0.4	95.74	43.44	100	.15, .17	527, 530	72.42	35.69	100	.15, .17	527, 530	63.92	41.62	100	.15, .17	527, 530
	0.5	81.41	34.88	100	.20, .22	526, 528	70.68	39.15	100	.20, .22	526, 528	63.08	39.06	100	.20, .22	526, 528
	0.6	77.10	47.42	100	.41, .45	526, 527	70.88	42.77	100	.41, .45	526, 527	62.73	43.35	100	.41, .45	526, 527
	0.7	71.85	35.55	100	1.1, 1.5	527, 527	62.72	39.23	100	1.1, 1.5	527, 527	62.08	43.78	100	1.1, 1.5	527, 527
100	0.3	728.88	352.03	100	.024, .027	529, 532	216.63	99.06	100	.024, .027	529, 532	127.69	100.62	53	.024, .027	529, 532
	0.4	363.97	149.04	100	.16, .17	527, 530	200.23	123.80	100	.16, .17	527, 530	143.26	87.27	83	.16, .17	527, 530
	0.5	288.40	124.22	100	.20, .22	526, 528	157.20	96.93	100	.20, .22	526, 528	143.27	91.69	95	.20, .22	526, 528
	0.6	246.02	120.07	100	.41, .45	526, 527	161.04	93.10	100	.41, .45	526, 527	141.54	88.79	98	.41, .45	526, 527
	0.7	213.92	104.83	100	1.1, 1.5	527, 527	161.30	90.30	100	1.1, 1.5	527, 527	139.05	90.39	100	1.1, 1.5	527, 527
150	0.3	190.86	97.07	100	.024, .027	529, 532	435.49	202.93	100	.024, .027	529, 532	1736.32	900.54	32	.024, .027	529, 532
	0.4	149.52	67.88	100	.15, .17	527, 530	331.85	173.77	100	.15, .17	527, 530	954.16	497.42	40	.15, .17	527, 530
	0.5	189.40	117.54	100	.20, .22	526, 528	287.73	133.55	100	.20, .22	526, 528	620.00	250.47	55	.20, .22	526, 528
	0.6	228.16	130.53	100	.41, .45	526, 527	259.56	163.36	100	.41, .45	526, 527	461.06	190.07	74	.41, .45	526, 527
	0.7	217.61	138.81	100	1.1, 1.5	527, 527	239.62	130.98	100	1.1, 1.5	527, 527	436.14	227.85	88	1.1, 1.5	527, 527
200	0.3	369.50	200.10	100	.024, .027	529, 532	720.90	294.71	100	.024, .027	529, 532	2932.18	1218.12	4	.024, .027	529, 532
	0.4	301.17	140.12	100	.15, .17	527, 530	610.35	264.93	100	.15, .17	527, 530	2028.86	986.43	14	.15, .17	527, 530
	0.5	206.58	135.76	100	.20, .22	526, 528	416.3	231.80	100	.20, .22	526, 528	1252.38	474.05	22	.20, .22	526, 528
	0.6	224.83	126.04	100	.41, .45	526, 527	404.4	194.50	100	.41, .45	526, 527	817.95	203.28	46	.41, .45	526, 527
	0.7	283.26	183.31	100	1.1, 1.5	527, 527	415.52	242.10	100	1.1, 1.5	527, 527	721.78	263.08	62	1.1, 1.5	527, 527

* All values are multiplied by 10 $^{-3}$

Supplementary Text – 3

Validation and Performance of the Proposed Method of Assessment of Statistical Significance of Motif Discovery

Before proceeding further, we provide some general results pertaining to the proposed method of assessing the statistical significance of a motif discovered by our algorithm. To estimate the statistical significance of a motif discovered by our algorithm in relation to a random data set of "similar" structure, we created a data set with N=20, L=10 and planted a motif of length p=5 (=1,1,1,1,1) at 5 randomly-chosen sites from among the 10 sites with a frequency u (among N). Three values of u were used; these were 0.3, 0.5 and 0.7. The remaining cells in the $N \times L$ data matrix were filled with 1 or 0, each with probability 0.5. We shall refer to this data set as the "real" data set. We then used our algorithm to discover a motif in this "real" data set. Next, we created 10,000 $N \times L$ replicate data matrices in which the 1's and the 0's in cells in column i were randomly permuted so that the total number of 1's and 0's occurring in the column remain same as that in the "real" data. In each of the 10,000 replicate random data matrices, we searched for the "best" motif of length 5 *by complete enumeration*. The large-deviation probability, that is, the probability that the best or the discovered motif occurs in a random data set with a frequency that is greater than or equal to that of the motif discovered by our algorithm in the "real" data set, is $\leq$ 0.0001, for all values of u (Supplementary Table 3). These findings further indicate that our algorithm performs well.

Since for a large data set, it is not possible to search for the "best" motif by complete enumeration, we additionally sought to evaluate our algorithm by the above statistical-significance criteria using an approximate method. In this approximate method, the only change that was made is that instead of searching for the "best" motif by complete enumeration in a random data set, we applied our algorithm to discover the "best" motif. The "real" data sets were generated in the same way as the Synthetic Data Set 1, with N=200, L=50 and 200, and p=10. Three values of u were used – 0.3, 0.5 and

0.7. Motif search was performed using c=100 in both the “real” and the random data sets. In all cases, the estimated probability of existence of a motif in a random data set with a higher frequency than the motif discovered in the real data set was $< 10^{-7}$.

Supplementary Table 3. Probabilities that the best motif discovered in a random data set has a frequency (*f**) greater than the frequency (*f*) of the motif discovered in the real data set of size 20×10, for different values of $u^{1.2}$

u	Prob (*f** $\geq$ *f*)
0.3	< 0.00001
0.5	0.0001
0.7	<0.00001

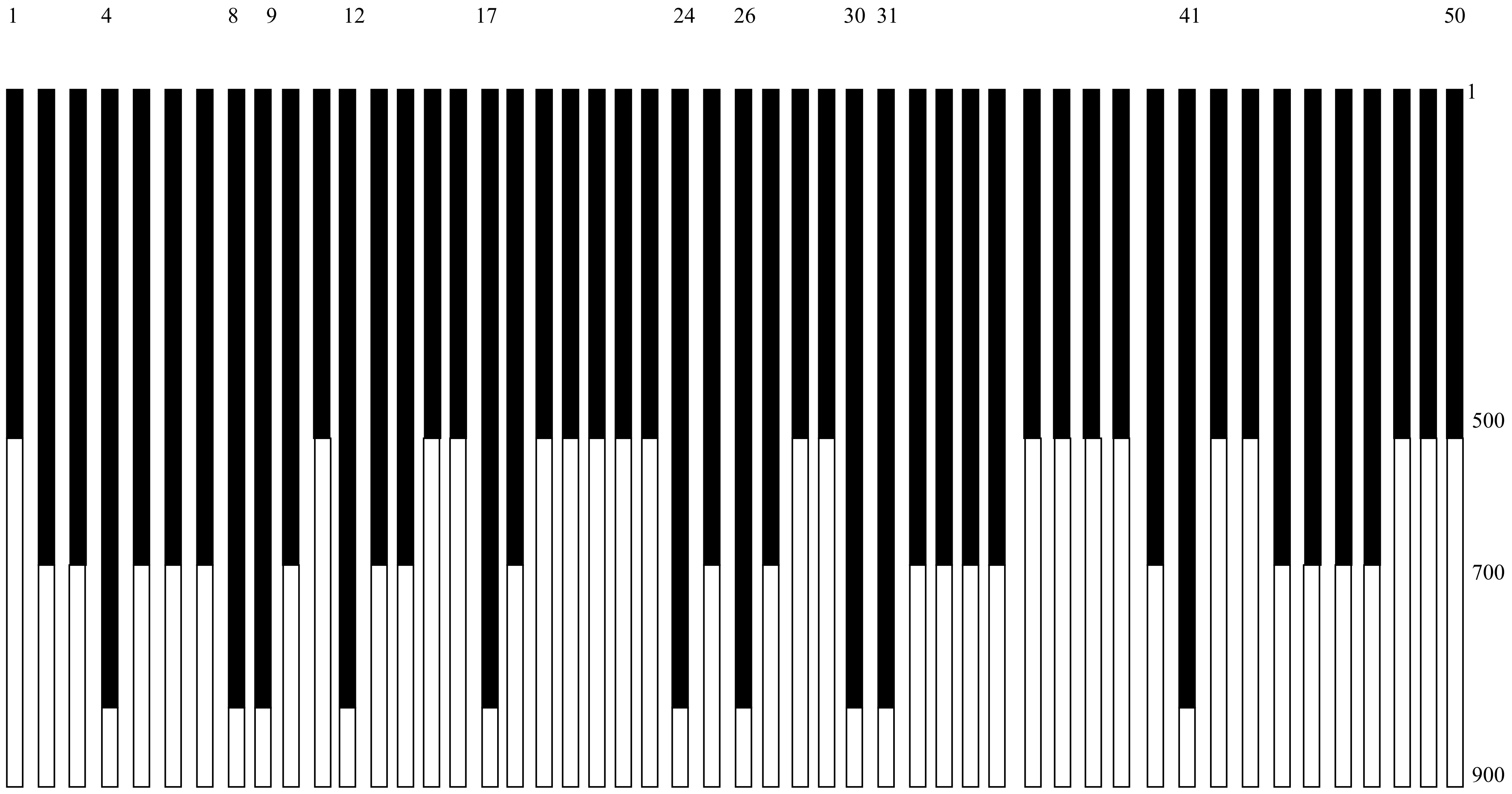
[1] Results are based on 1000 real data sets for each value of *u* and 10000 random data sets for each real data set.

[2] In each case, the motif discovered in the real data set coincided with the planted motif, and its proporation was also close to *u*.

Supplementary Text – 4

Performance of the Algorithm in the Presence of a Large Number of Local Optima

We generated synthetic data matrices of structures similar to that given in Supplementary Figure 1. The motivation for analyzing these synthetic data sets was to assess the performance of our algorithm when the search space comprises a large number of local optima. In this data matrix of size 1000 × 50, 10 columns were randomly chosen. Each chosen column, was filled with 900 1s and 100 0s. In other words, the first 900 elements of each chosen column were 1, and the remaining 100 were 0. Of the remaining 40 columns of the data matrix, 20 were randomly chosen, and each chosen column was filled with 700 1s and 300 0s. Each of the remaining 20 columns, was filled with 500 1s and 500 0s. For the example data matrix presented in Supplementary Figure 1, the set of columns, S, filled with 900 1s and 100 0s is: $S = \{4, 8, 9, 12, 17, 24, 26, 30, 31, 41\}$. Thus, this set of sites comprises the motif $(1, 1, \ldots, 1)$ of length $p = 10$, with $G(S) = 900$. However, this motif is clearly almost impossible to find, because there is exactly one such among $\binom{50}{10}$ possibilities. Among these $\binom{50}{10}$ points in the search space, there are $\left[\binom{50}{10} - \binom{30}{10}\right]$ 10-site combinations at which the sequence will be $((1))_{1\times10}$ with a frequency of 500, and $\left[\binom{30}{10} - 1\right]$ 10-site combinations at which the sequence will be $((1))_{1\times10}$ with a frequency of 700. Thus, the set of 10 out of 50 sites versus the frequency distribution of individuals has a very discrete structure: there is only one element in this set with 900 individuals, a very large number of elements with 700 individuals and a similarly large number of elements with 500 individuals. Clearly, therefore, to find the element with 900 individuals is nearly impossible. In 1000 runs of our algorithm with p=10 and different initial values of c, we were never able to discover the correct motif. Invariably, the convergence was to a string with frequency either 500 or 700. However, when $(m_{500}, m_{700}, m_{900})$, where m_i denotes the number of columns in each of which there are i 1s and $(1000 - i)$ 0s $(i = 500, 700, 900; m_{500} + m_{700} + m_{900} = 50)$, was changed from (20,20,10) to other sets of values, the proportions of runs in which the correct motif of length 10 was discovered increased. In other words, when the structure of the search space was slightly changed so that there were multiple - not just one - elements in the space with 900 individuals, the algorithm converged correctly and identified the motif. Simulation experiments were performed with various values of the control parameter c, the results of which are presented in Supplementary Table 4. Best results were obtained with c=200.



Supplementary Figure 1. Model structure of the synthetic data sets with multiple local optima. [Dark boxes are filled with 1; white boxes are filled with 0. Ten columns have 900 1s and 100 0s. Expected motif is (1,1,1,1,1,1,1,1,1,1) at sites (4, 8, 9, 12, 17, 24, 26, 30, 31, 41).]

Supplementary Table 4. Results of 1000 Simulation Runs for Different Structures of Synthetic Data Set 2 as Specified by $(m_{500}, m_{700}, m_{900})$ for Different Values of the Control Parameter c[1]

$(m_{500}, m_{700}, m_{900})$	c	% runs in which the motif was correctly identified	Mean±s.d. of the number of sweeps to convergence
(17,17,16)	10	36.8	1101.44±540.73
	30	50.8	986.43±574.89
	50	51.0	976.62±538.11
	100	52.0	926.69±589.32
	200	55.0	896.77±571.40
(16,16,18)	10	84.5	823.33±504.37
	30	92.0	627.87±489.17
	50	93.9	590.74±489.17
	100	92.5	638.92±503.21
	200	93.6	621.96±489.63
(15,15,20)	10	99.4	458.32±354.85
	30	99.8	270.20±237.87
	50	99.9	295.04±268.50
	100	100	293.07±268.81
	200	100	288.20±267.71

[1] The significance level of the motif, when correctly identified, was $\approx 10^{-15}$ when assessed by the "drop" procedure, and was = 0 when assessed by the bootstrap procedure (indicating that no motif "better" than that identified by the algorithm existed in the data).

Supplementary Text – 5

Performance of the Algorithm when the Motif Proportion or Sample Size is Small

We generated multiple synthetic Data Sets 1 with u=0.1; that is, only 10% of individuals carry a known motif of size p=10. Further, we genenerated multiple synthetic Data Sets 1 with N=50; that is, data sets with small sample sizes. Although, both these scenarios are somewhat unrealistic, we carried out these simulation experiments to examine the limits to which our algorithm can be pushed. The results are given in Supplementary Table 5(a) and (b). We have used relatively small values of c, which is what be prescribe should be used when the motif frequency or the sample size is small. We find that even in these extreme cases, our algorithm performs well.

Supplementary Table 5. Performance of the algorithm on Synthetic Data Set 1 with (a) *N*=200 for different values of the variables *L* (number of segregating sites) and *u* (proportion of the planted motif among *N*) = 0.1; and, (b) *N*=50, *L*=200 and various values of u. (The mean and s.d. of the number of sweeps to convergence and the % of simulation runs in which the planted motif was correctly identified are based on 1000 independent simulation runs for each combination of values of the variables and the parameter.)

(a)

L	*c*=50			*c*=100		
	No. of Sweeps			**No. of Sweeps**		
	Mean	**s.d.**	**% correct**	**Mean**	**s.d.**	**% correct**
50	1203.67	575.49	100	208.36	94.30	100
100	3660.30	1157.10	100	990.04	494.44	87

→

(b)

u	*c*=50			*c*=100		
	No. of Sweeps			**No. of Sweeps**		
	Mean	**s.d.**	**% correct**	**Mean**	**s.d.**	**% correct**
0.3	2067.20	1847.40	74	1983.21	1541.60	85
0.5	551.36	275.80	100	657.20	301.60	100
0.7	398.60	223.80	100	382.10	212.50	100

Supplementary Text – 6

Method of Creating Synthetic Data Set 2

Two separate data matrices, each of size $N \times L$, corresponding to the cases and controls, were created. Elements of each column of each matrix were randomly filled with 1 or 0; the proportion of 1s occurring in any column was taken to be 0.5. Then p columns (polymorphic sites) were chosen at random. In the first data matrix corresponding to the cases, a set of $[N \times u_1]$ rows were randomly chosen, where $0 < u_1 < 1$. In each of these rows, the elements corresponding to the p chosen columns were replaced with 1. Thus, we planted, in the case data matrix, a motif (1,1,. . . ,1) of length p in a proportion of u_1 individuals. Under the common-disease, common-variant model (Collins et al. 1998), each of the p sites (SNPs) carries a small relative risk, RR, to the disease, that collectively results in a large haplotype (motif) relative risk. If v_i and w_i denote, respectively, the number of 1s (that is, the specific nucleotide that confers a higher risk) at the i^{th} site $(1 \leq i \leq p)$, among cases and controls, then RR $= v_i/w_i$. (We have assumed that the site-specific relative risk is the same for each of the p sites.) Hence, in the data matrix corresponding to the controls, for the i^{th} of the p sites (that is, the i^{th} column), we placed 1s in $[N \times w_i]$, where $w_i = v_i/RR$, randomly chosen rows, and filled the remaining elements in that (i^{th}) column with 0s.

Supplementary Text – 7

Method of Creating Synthetic Data Set 3

For constructing the data set, we first created a data matrix of size 1000 × 50, and assigned a value of 0 or 1 to each cell with probability 0.5. Then, we randomly selected 10 sites (that is, 10 columns of the data matrix) from the set (Π) of 50 sites, and changed the 0s to 1s at each site (column) so that at each of these sites the proportion of 1s among the 1000 individuals was ≥ 0.8. This resulted in the data matrix, D_1, of the ancestral population which expectedly has a motif of length 10 comprising the set of the 10 randomly selected sites, which we shall denote as Π_1. We then created two daughter populations of this ancestral populations. The data matrix, of size 1000 × 50, corresponding to the first daughter population was initially created by sampling 1000 rows (each with 50 columns), with replacement, from the data matrix of the ancestral population. We then selected a set (Π_2) of 5 sites randomly from $\Pi \backslash \Pi_1$, and at these selected sites we randomly replaced 0s by 1s in the initial data matrix of the first daughter population such that the proportions of 1s among the 1000 individuals at each of these 5 sites was ≥ 0.8. This yielded the final data matrix, D_2, corresponding to the first daughter population in which the motif expectedly comprises sites belonging to $\Pi_1 \cup \Pi_2$ of length 15. For the second daughter population, the initial data matrix was similarly created. A set (Π_3) of 5 sites were chosen from $\Pi \backslash (\Pi_1 \cup \Pi_2)$ and the final data matrix, D_3, was similarly created. In the second daughter population, the expected motif of length 15 has sites belonging to $\Pi_1 \cup \Pi_3$.

Supplementary Table 6. Mean ± s.d. of the Number of Sweeps to Convergence in 1000 Independent Simulation Runs for Synthetic Data Set 3

Population (Data Matrix)	Mean ± s.d. of the number of sweeps to convergence
Population 1 (D_1)	45.76 ± 32.15
Population 2 (D_2)	33.25 ± 12.96
Population 3 (D_3)	32.86 ± 13.07