SUPPORTING INFORMATION for

H. T. Vy and Y. Kim, A composite likelihood method for detecting incomplete selective sweep from population genomic data, submitted to Genetics

File S1

Scripts to generate simulated data sets:

1. For testing power of CL, iHS, nSL:

- $R = 4Nr = 2000$

  - Neutral model:
    /msms 20 10000 -N 100000 -s 3000 -r 2000 100000

  - Selective sweep model:
    /msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 1000 -SaA 500 -SF 0 0.5 -Sp 0.5 –Smark
    /msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 2000 -SaA 1000 -SF 0 0.5 -Sp 0.5 –Smark
    /msms 20 10000 -N 100000 -s 2999 -r 2000 100000 -SAA 4000 -SaA 2000 -SF 0 0.5 -Sp 0.5 –Smark

- $R = 4 = 4000$

  - Neutral model:
    /msms 20 10000 -N 100000 -s 3000 -r 4000 100000

  - Selective sweep model:
    /msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 1000 -SaA 500 -SF 0 0.5 -Sp 0.5 –Smark
    /msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 2000 -SaA 1000 -SF 0 0.5 -Sp 0.5 –Smark
    /msms 20 10000 -N 100000 -s 2999 -r 4000 100000 -SAA 4000 -SaA 2000 -SF 0 0.5 -Sp 0.5 –Smark

2. For generating neutral data matching the sample size, mean recombination rate, and the mean density of polymorphic sites to those of Drosophila genome data (to calculate T_1 when apply composite likelihood test to Drosophila genomes):

/ms 22 20 -t 35000 -r 60000 5000000

3. To simulate data under different demographic assumptions:

- Population bottleneck:

  - With different severities:
    /ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.2 -eN 0.1 1.0

/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.1 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.05 -eN 0.1 1.0

- With different recombination rates:

/ms 20 1000 -s 3000 -r 4000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 6000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 8000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 10000 100000 -eN 0.05 0.05 -eN 0.1 1.0
/ms 20 1000 -s 3000 -r 12000 100000 -eN 0.05 0.05 -eN 0.1 1.0

- Exponential population growth:

  - With different growth rates:

/ms 20 1000 -s 3000 -r 4000 100000 -G 500 -eG 0.0032 0.0
/ms 20 1000 -s 3000 -r 4000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 4000 100000 -G 10 -eG 0.016 0.0

  - With different recombination rates:

/ms 20 1000 -s 3000 -r 4000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 6000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 8000 100000 -G 100 -eG 0.016 0.0
/ms 20 1000 -s 3000 -r 10000 100000 -G 100 -eG 0.016 0.0

- Population subdivision:

  - With different migration rates:

/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 1.0
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 10

  - With different recombination rates:

/ms 20 1000 -s 3000 -r 1000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 2000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 4000 100000 -I 2 20 0 0.1
/ms 20 1000 -s 3000 -r 6000 100000 -I 2 20 0 0.1

Table S1: List of putative loci under incomplete selective sweeps in *D. melanogaster* Rwanda population inferred from *iHS* test.

Chromosome	Cluster start - end	Minimum standardized *iHS*	Site of minimum *iHS*	Derived allele frequency
2L	1547008 - 1557247	-3.75	1547204	15/22
	4824296 - 4860577	-4.03	4824431	10/22
	5810884 - 5825009	-3.34	5815486	8/22
	6020532 - 6055868	-3.34	6055868	9/22
	9509499 - 9831699	-5.90	9582539	10/20
	11022970 - 11036917	-3.82	11036917	9/21
	11866168 - 11892750	-3.52	11892750	9/21
	12804885 - 12840609	-3.76	12840609	10/21
	16019930 - 16020004	-3.17	16019930	9/21
	17230328 - 17297935	-3.84	17297624	12/21
	17602339 - 17603635	-3.32	17603635	13/21
2R	5649989 - 5718051	-3.33	5708056	11/20
	7114975 - 7137571	-3.87	7127281	11/22
	7828412 - 8658752	-5.19	7911386	12/22
	10135366 - 10665005	-3.86	10665005	9/22
	11001899 - 11016632	-3.28	11006196	9/22
	12723745 - 12768255	-4.38	12734718	10/21
	13832525 - 13835190	-3.49	13832748	11/22
3L	3103656 - 3145740	-4.84	3142414	9/22
	4472504 -4490155	-4.35	4490062	8/22
	5960208 - 5974572	-4.04	5966477	8/22
	6072249 - 6129005	-5.13	6109760	10/21
	6537946 - 6595815	-4.26	6548571	12/21
	8126905 - 8182746	-4.18	8134344	10/22
	14430470 - 14434825	-4.12	14431036	11/18
	16070044 - 16095749	-4.48	16095749	12/22
	19210723 - 19236655	-4.14	19210732	8/22
3R	8567616 - 8567660	-3.70	8567660	9/19

	9033505 - 9157259	-3.63	9125222	8/19
	10333644 - 10389581	-3.46	10386821	10/22
	12063858 - 12066090	-3.12	12063858	14/22
	13905933 - 13911447	-3.84	13910288	12/22
	15427616 - 15503321	-3.42	15442743	13/22
	16254275 - 16259857	-3.40	16254275	11/22
	18973474 - 18973529	-3.49	18973529	13/22
X	737336 - 1229075	-4.50	1081402	8/22
	2814828 - 2836819	-5.68	2832651	9/22
	17230521 - 17234450	-4.88	17230521	8/21
	18636049 - 18639500	-5.23	18636156	8/20
	19077625 - 19104409	-5.80	19093150	8/20

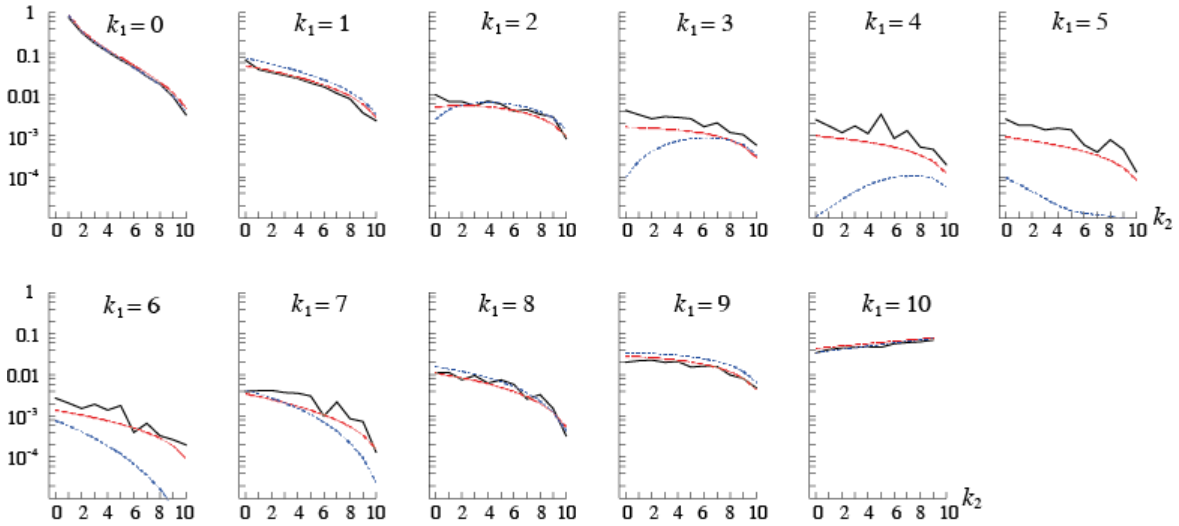
Table S2: List of putative loci under incomplete selective sweeps in *D. melanogaster* Rwanda population inferred from nS_L test.

Chromosome	Cluster start – end	Minimum standardized nS_L	Site of minimum nS_L	Derived allele frequency
2L	1147263-1152901	-3.77	1147263	8/22
	1545621-1548645	-3.68	1545621	14/22
	5812475-5816415	-3.15	5815486	8/22
	6596769-6603746	-3.99	6603091	9/22
	9433919-9610983	-3.33	9433919	17/22
	12372465-12377872	-3.46	12375980	11/21
	12718277-12844111	-3.10	12834789	8/21
	16798951-16894905	-3.35	16894905	12/21
	17234502-17245049	-3.45	17235226	9/21
	17602170-17619087	-3.72	17603916	16/21
	19727261-19768977	-3.44	19730058	8/21
2R	3786646-3886834	-2.42	3886479	8/21
	5254576-5283840	-2.54	5266445	8/21
	5548485-5552974	-2.42	5548485	9/21
	7119351-7133870	-2.47	7126724	12/22
	7816110-8669818	-3.29	8133130	17/22
	12721950-12740626	-2.83	12727369	8/22
	13826522-13834107	-2.46	13826676	8/22
	18089346-18541764	-2.80	18089697	8/22
3L	2997574-2998073	-3.82	2998073	8/22
	3136553-3149179	-4.49	3144835	11/22
	6087588-6129005	-4.51	6129005	16/21
	6538594-6567487	-4.18	6547881	8/21
	11654477-11679598	-3.59	11654544	9/21
	11826677-11839249	-3.88	11833069	12/18

	13427088-13432591	-3.44	13430725	11/22
	16076838-16137484	-4.20	16094714	12/22
	19210376-19236655	-3.80	19210732	8/22
3R	10230101-10904118	-3.14	10333644	14/22
	12062855-12066090	-3.37	12062855	12/22
	13330699-13334656	-3.14	13330699	15/22
	13906691-13911447	-3.41	13908193	10/22
	15432404-15445674	-3.46	15443091	12/22
	16561264-16575140	-3.22	16569519	13/21
	19838638-19884976	-3.17	19839035	12/22
	20872678-20878483	-3.49	20876949	12/22
X	2814198-2838349	-5.45	2833040	9/22
	10210768-10225115	-4.21	10219139	8/22
	14151360-14164063	-4.50	14157513	16/21
	14969055-14996063	-4.16	14975977	13/21
	16948533-17158766	-4.25	17124249	8/19

Figure S1

A. $r/s = 0.01$



B. $r/s = 0.04$

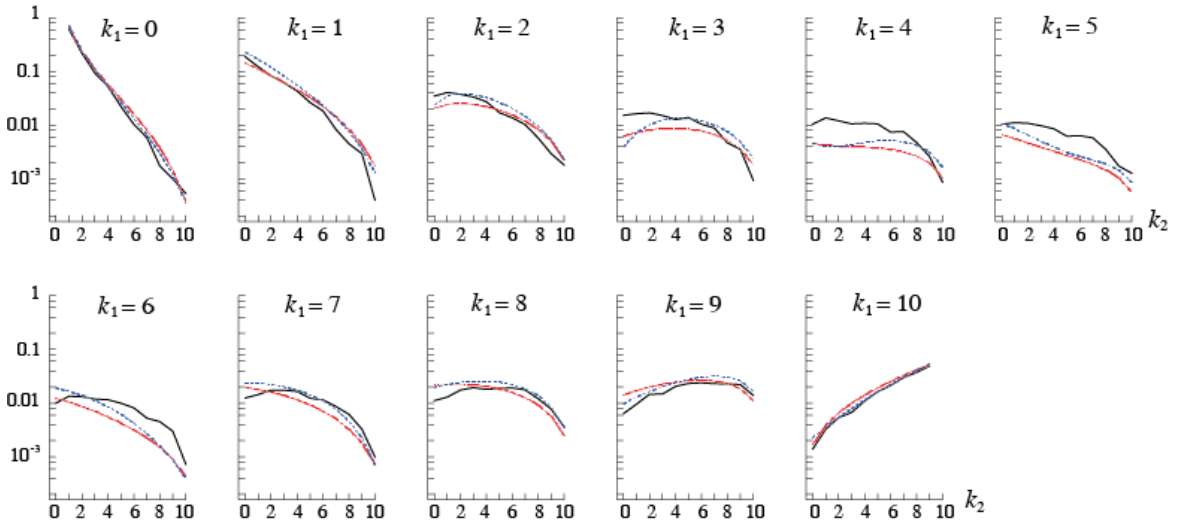


Figure S1 legend: Joint sampling probability under incomplete selective sweep for $n_1 = n_2 = 10$ and $r/s = 0.01$ or 0.04. ϕ_{S1} (blue) and ϕ_{S2} (red) are compared against simulation result (black).

Figure S2

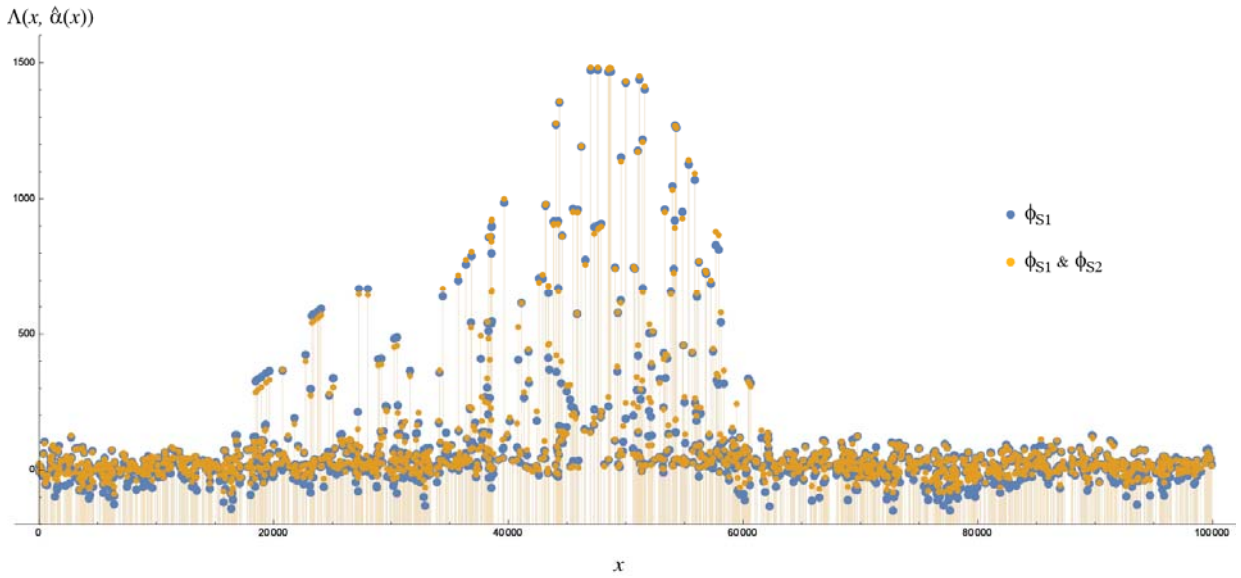


Figure S2 legend: Composite likelihood ratio calculated for a simulated data set of 20 DNA sequences of 100kb long (R = 4,000). Advantageous mutation with α = 4,000 is located in the middle (50kb). Blue dots are CLR calculated using ϕ_{S1}, approximation suggested by Nielsen et al. (2005), and yellow dots are CLR calculated using ϕ_{S1} for $r/s < 0.03$ but ϕ_{S2}, approximation based on Etheridge et al. (2006), for $r/s \geq 0.03$.

Figure S3

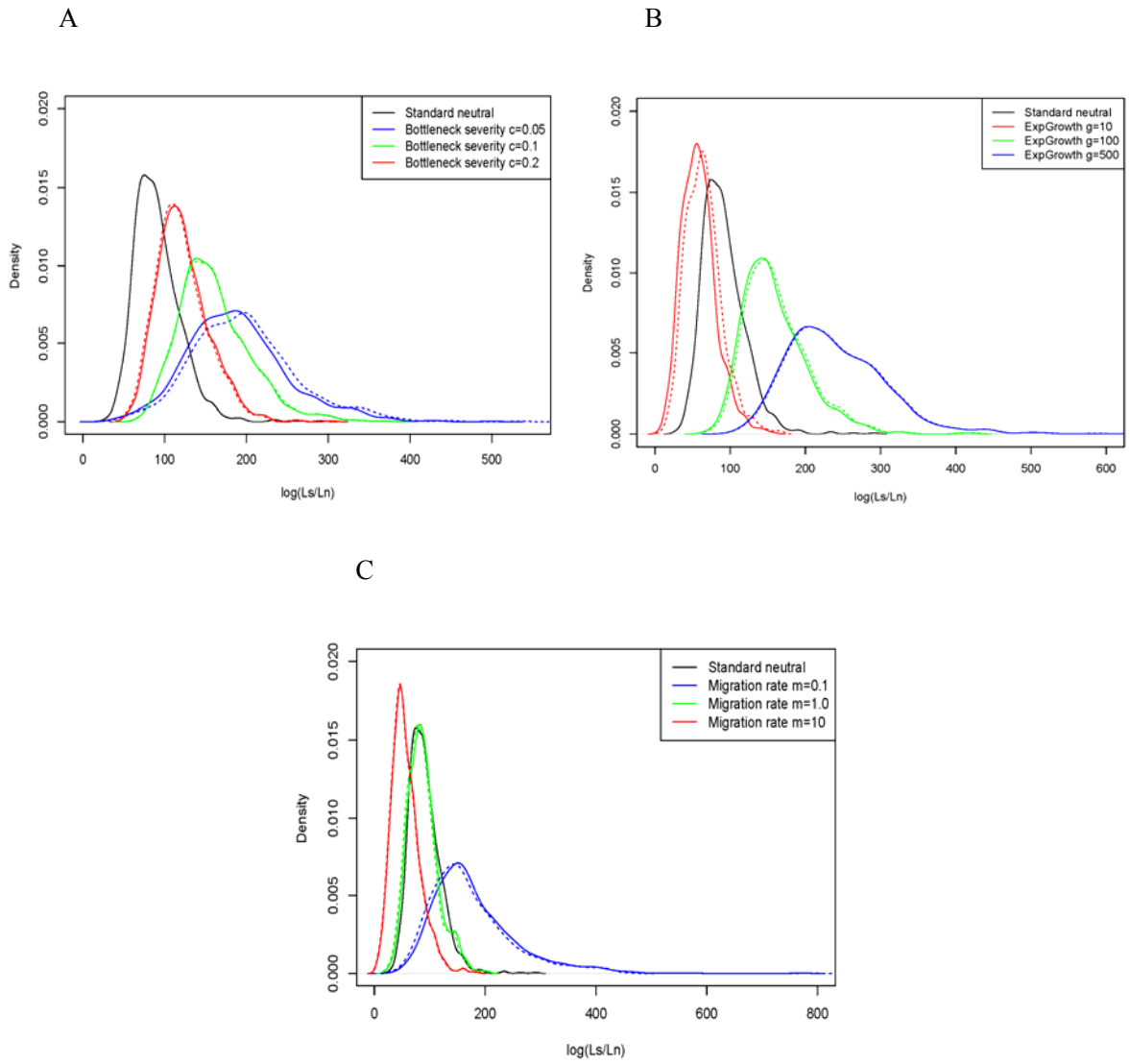


Figure S3 legend: Distributions of maximum CLR, $T_0 = \max_{x \in S[10]} \log(L_{IS} / L_N)$ where the maximum was obtained over the set of polymorphic sites with $n_1 = 10$ ($S[10]$ for each replicate), for samples generated under different demographic models: A, population bottleneck model with different bottleneck severities $c = 0.05$, 0.1, and 0.2; B, exponential population growth with different growth rates $g = 10$, 100 and 500; C, population subdivision model with different migration rates $m = 0.1$, 1, and 10 between 2 subpopulations. Recombination rate $4Nr_n = 0.04$ ($R = 4{,}000$) was used to generate all data sets. Distribution of T_0 for standard neutral model is plotted in each figure (black lines) for comparison. Distributions of T_0 calculated from empirical frequency spectrum (option B) are shown by dashed curves.

Figure S4

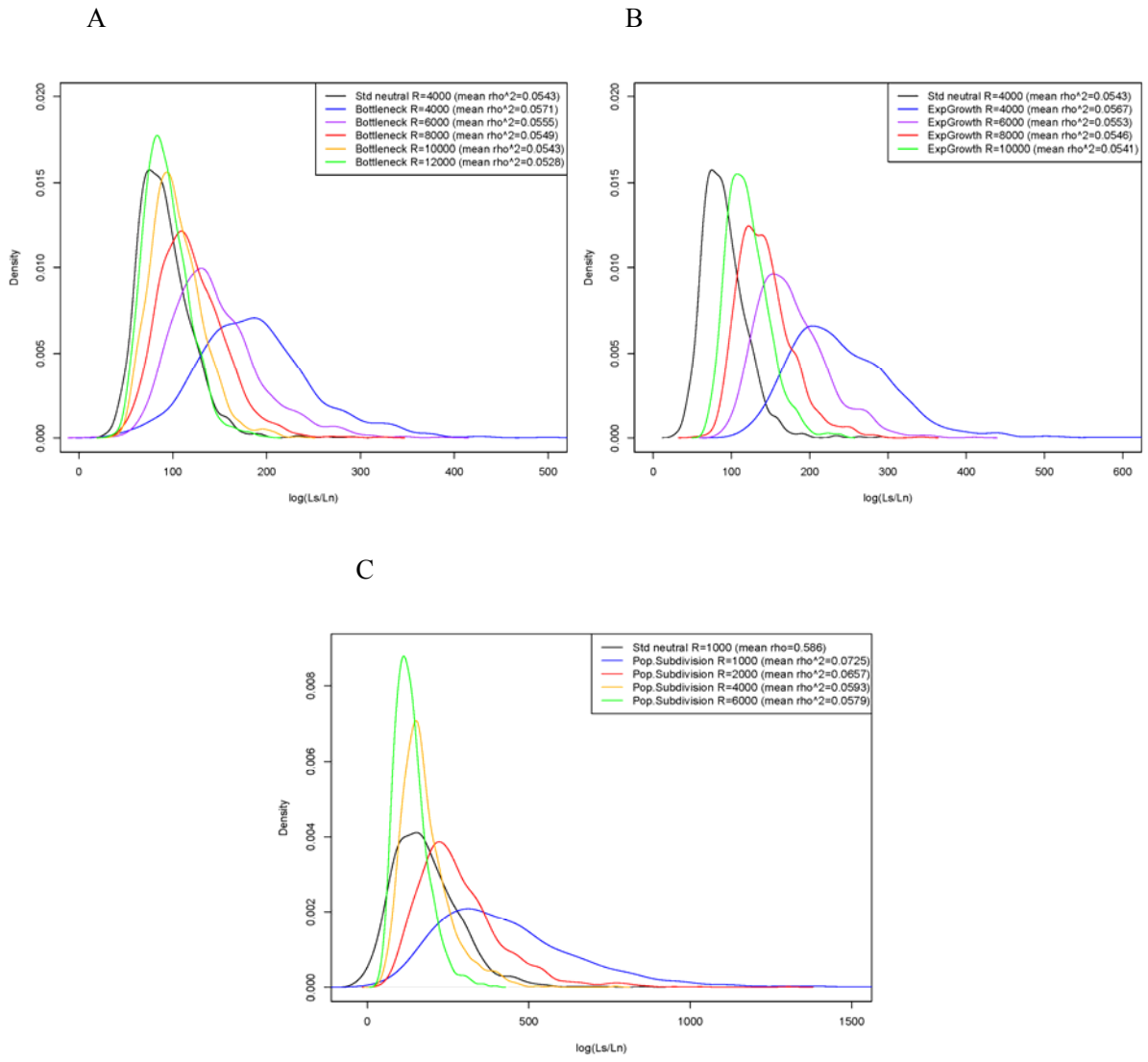


Figure S4 legend: Changes in the distributions of T_0 with varying recombination rates in different demographic models: A, population bottleneck model with bottleneck severity $c = 0.05$; B, exponential population growth with growth rate $g = 500$; C, population subdivision model with migration rate between two subpopulations $m = 0.1$. Mean correlation coefficient of LD among polymorphic sites (average ρ^2) for each model is shown in parenthesis.

Figure S5

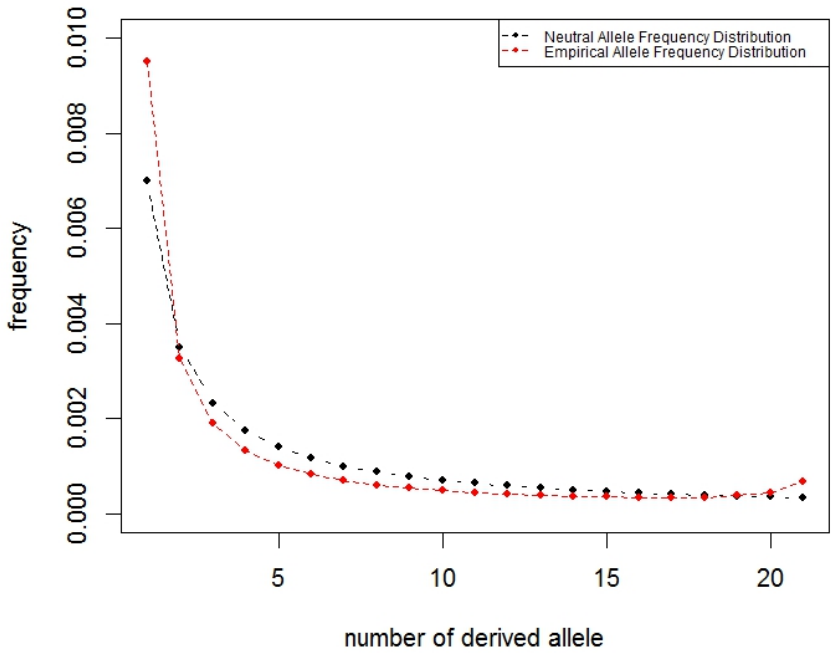


Figure S5 legend: Genome-wide empirical distribution of derived-allele frequency in the Rwanda D. *melanogaster* sample (22 sequences) in comparison with the standard neutral distribution for a sample of same size.

Figure S6

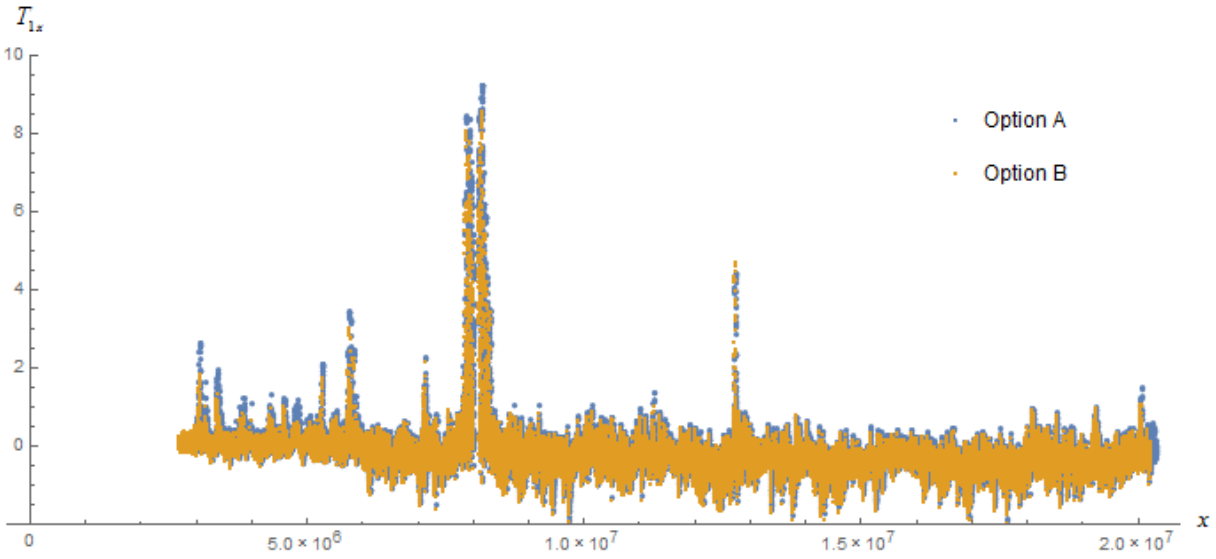
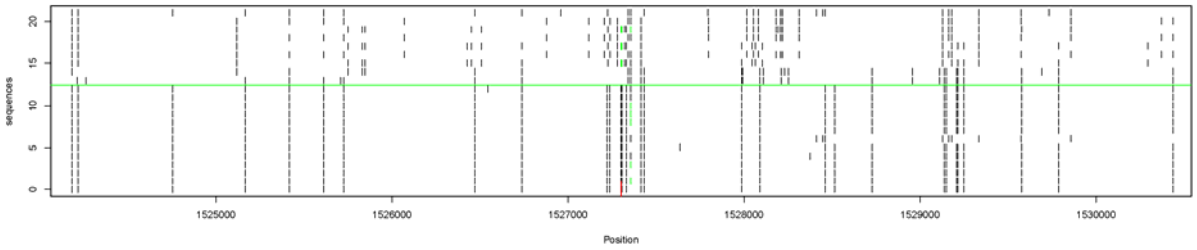


Figure S6 legend: Composite Likelihood Ratio (T_{1x}) calculated for chromosome 2R. T_{1x} was calculated based on sampling probabilities assuming neutral equilibrium (option A) or empirical frequency spectrum (option B) at the start of a selective sweep.

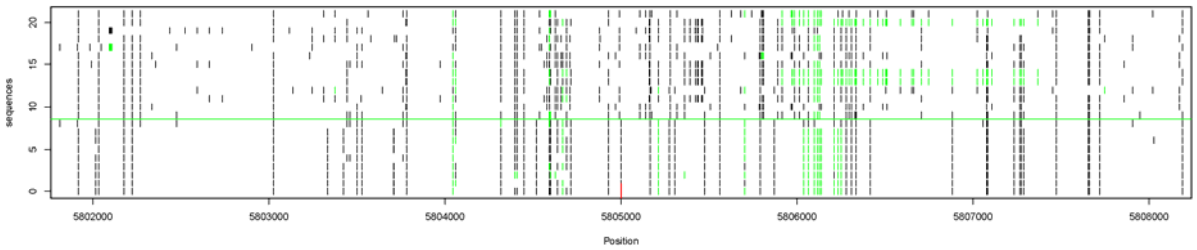
Figure S7

2L Patterns:

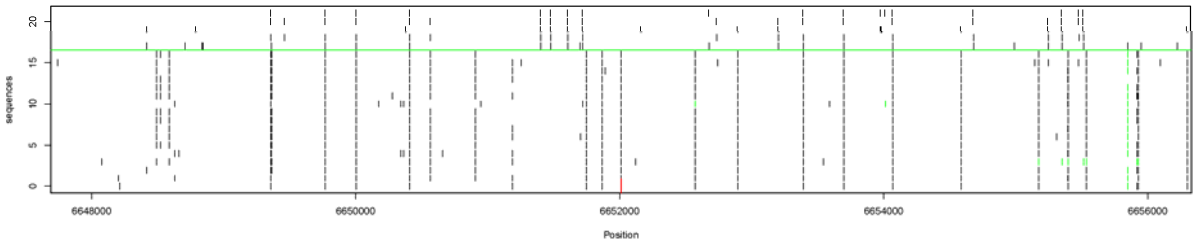
Putative site: 1527302*, closest gene: halo (1517533 – 1518148)



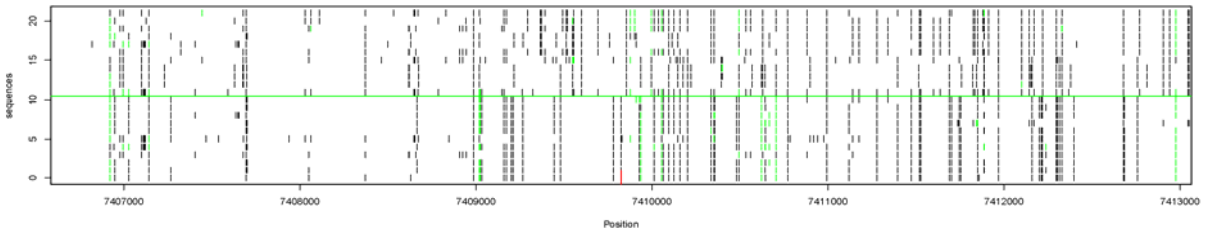
Putative site: 5805001*[#$], closest gene: CG11034 (5805395 – 5809063)



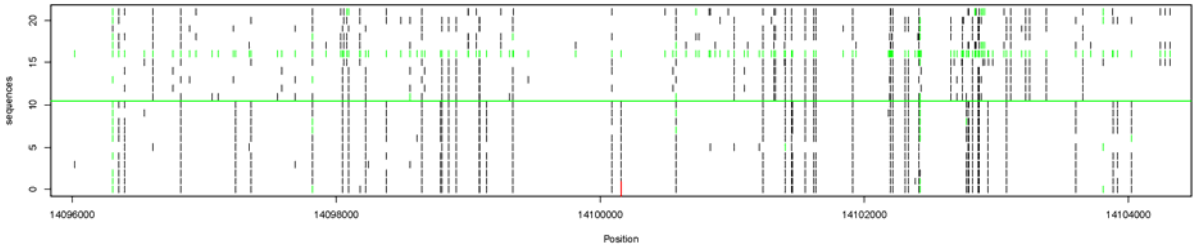
Putative site: 6652011[*$], closest gene: Tango1 (6649388 – 6654574)



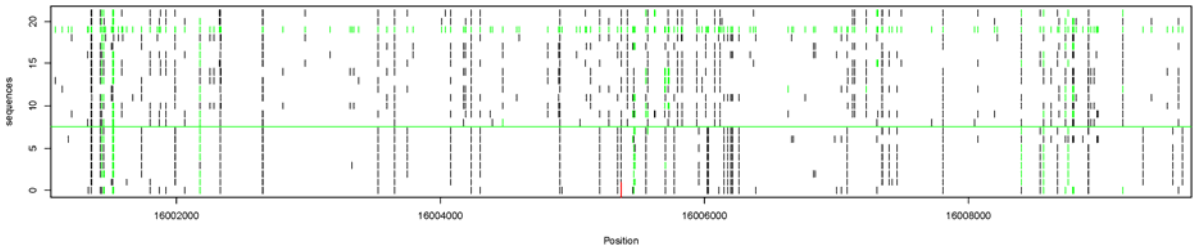
Putative site: 7409825, closest gene: CG5181 (7408533 – 7409809)



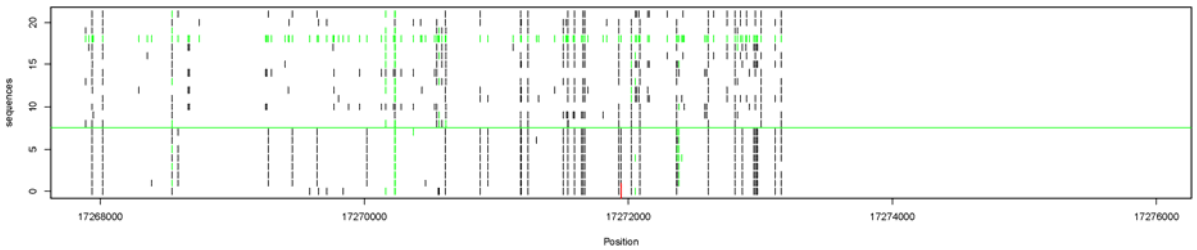
Putative site: 14100158*, closest gene: nAChRalpha5 (14040170 – 14094401)



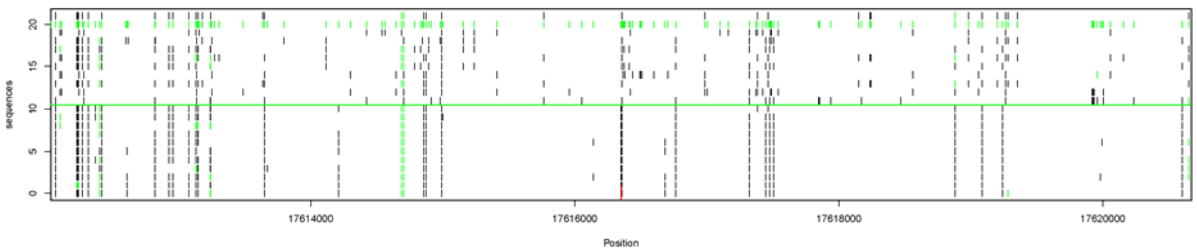
Putative site: 16005369[#], closest gene: Beat-Ic (16000291 – 16041703)



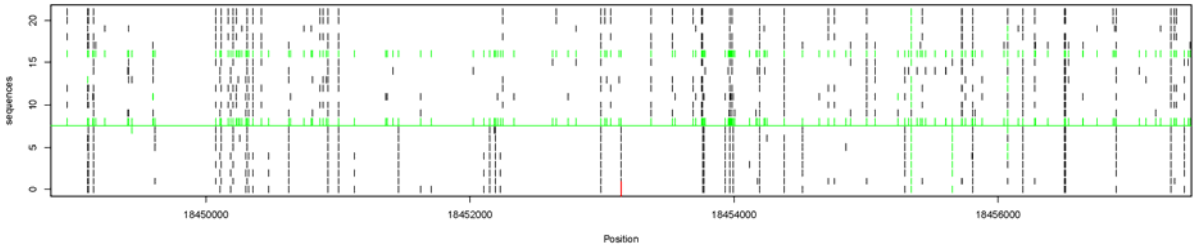
Putative site: 17271945[#$], closest gene: CG6380 (17291075 – 17292202)



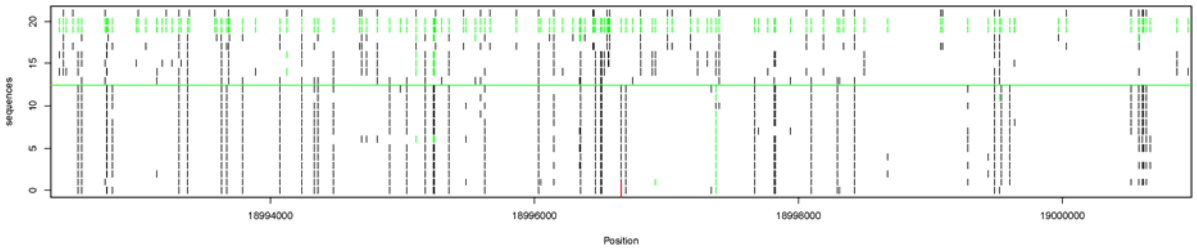
Putative site: 17616351[#$], closest gene: Sytalpha (17592260 – 17604387)



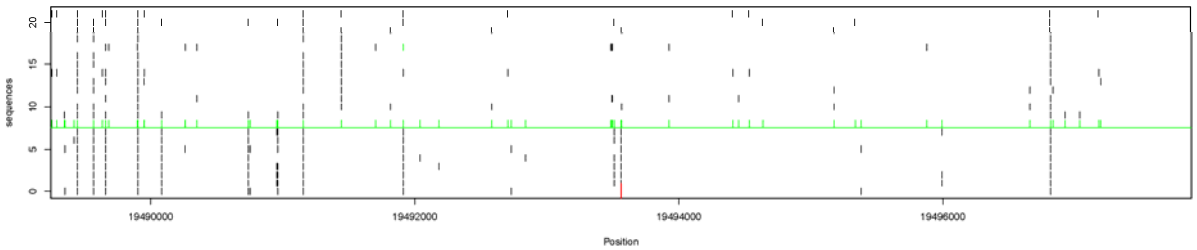
Putative site: 18453145, closest gene: bsf (18449517 – 18454587)



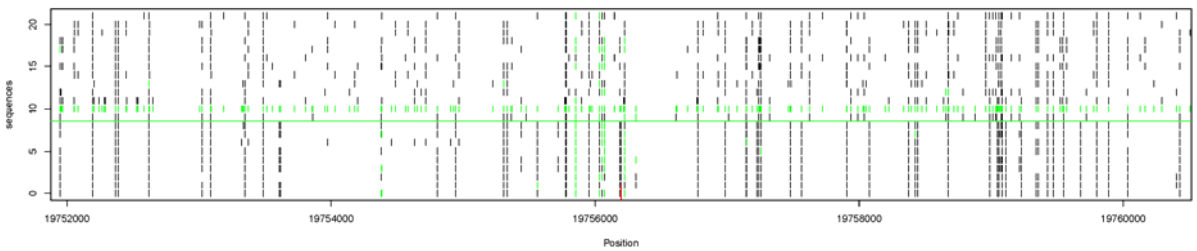
Putative site: 18996657, closest gene: CG10650 (18993360 – 18995934)



Putative site: 19493563, closest gene: swm (19493251 – 19497978)

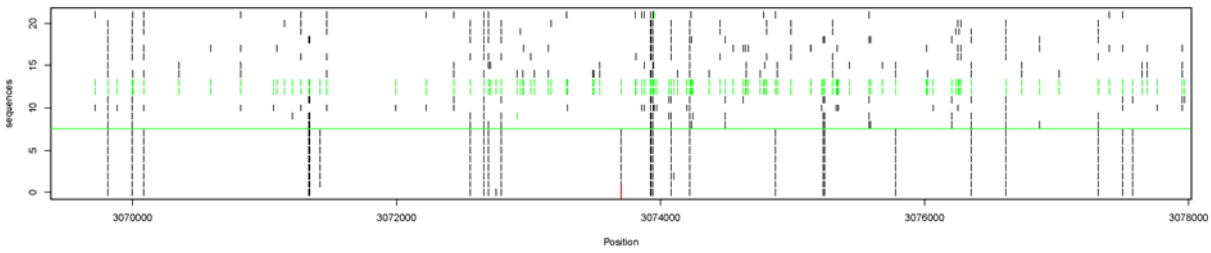


Putative site: 19756197[#$], closest gene: CG10631 (19742817 – 19756904)

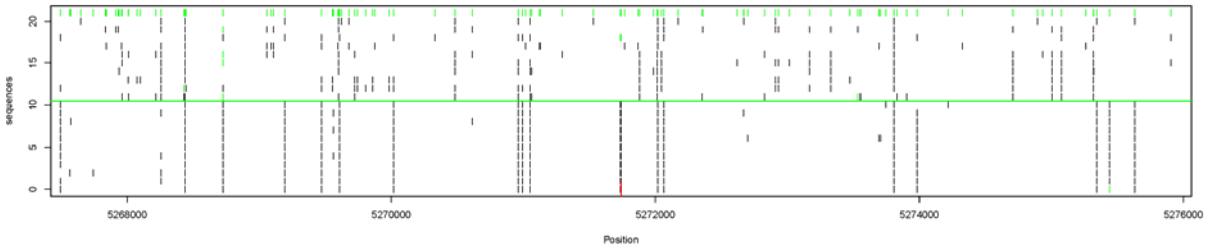


2R Patterns:

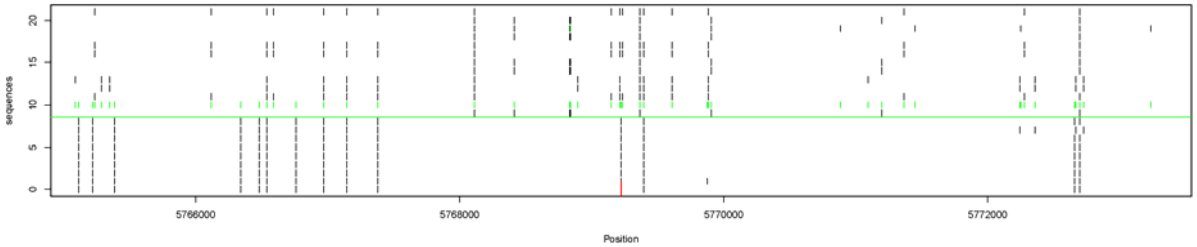
Putative site: 3073701, closest gene: diddum (3387652 – 3396130)



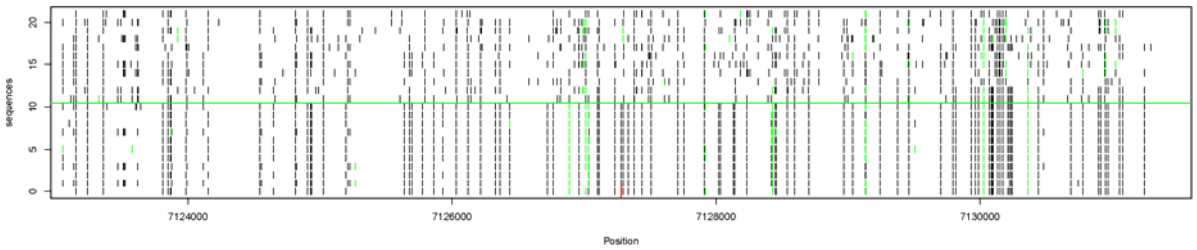
Putative site: 5271741[$], closest gene: CG13954 (5196801 – 5276972)



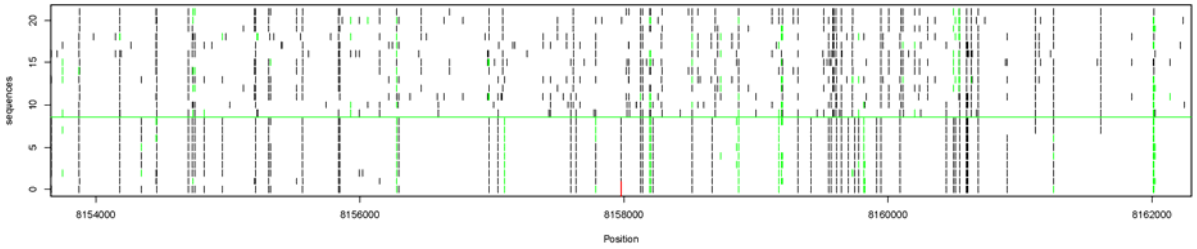
Putative site: 5769223, closest gene: Sec24AB (5763737 – 5769862)



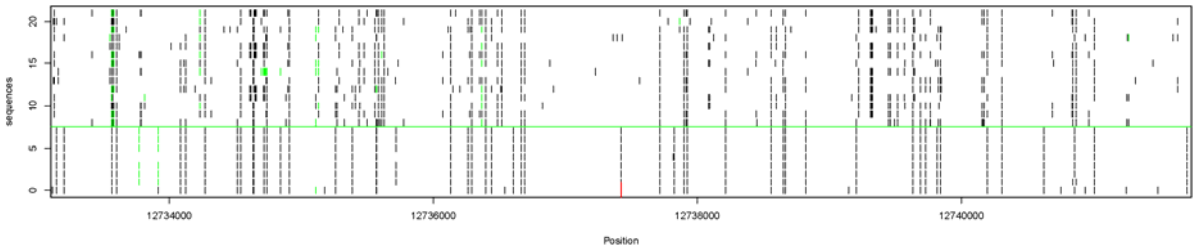
Putative site: 7127281[#$], closest gene: CG13215 (7126999 – 7127619)



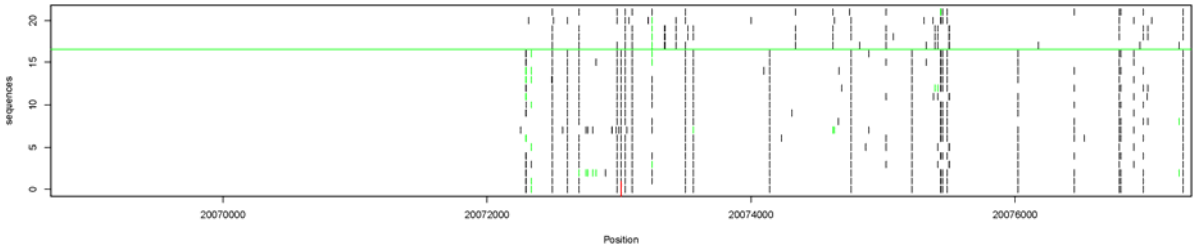
Putative site: 8157979[#$], closest gene: otk (7888978 – 7907351)



Putative site: 12737423*[#$], closest gene: IntS8 (12737942 – 12741609)

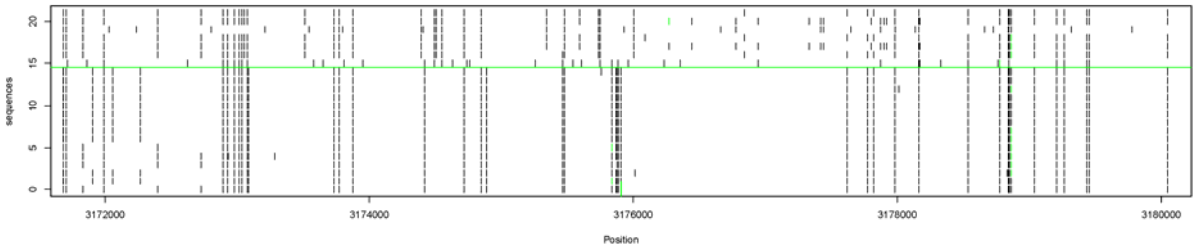


Putative site: 20073016*, closest gene: Nop60B (20062400 – 20073866)

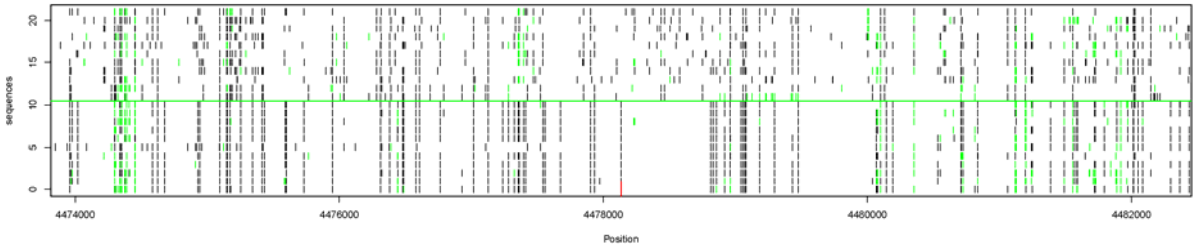


3L Patterns:

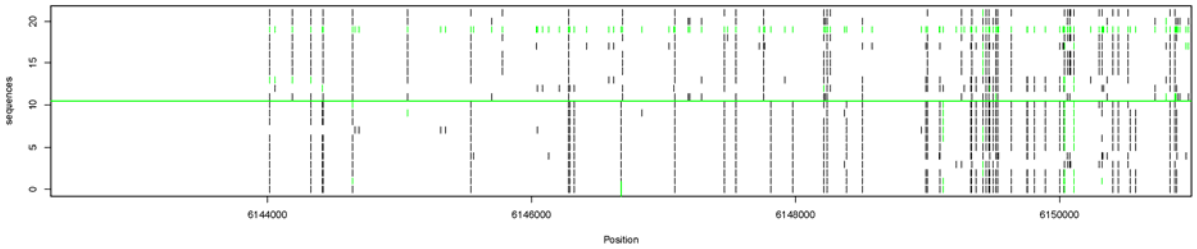
Putative site: 3175908*, closest gene: Girdin (3178930 – 3185287)



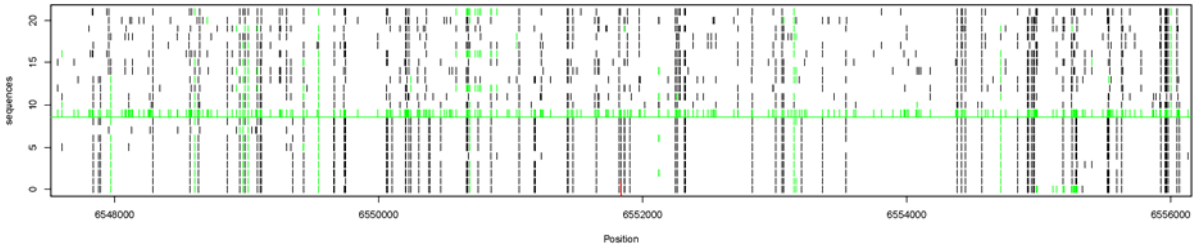
Putative site: 4478135*[#], closest gene: CG7465 (4480283 – 4481487)



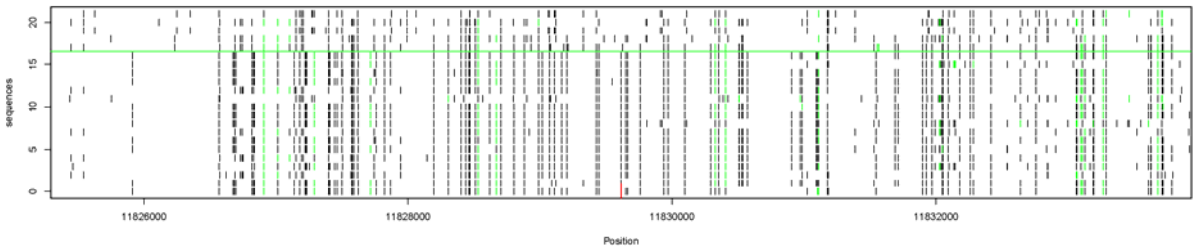
Putative site: 6146679*[#$], closest gene: Lcp65Ag2 (6126090 – 6126693)



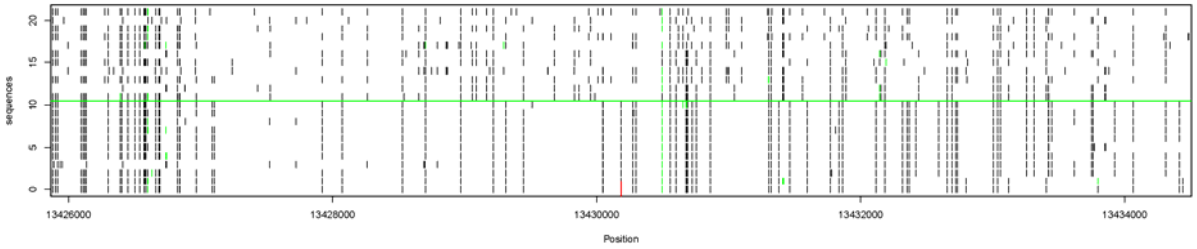
Putative site: 6551837*[#$], closest gene: CG18769 (6543838 – 6587040)



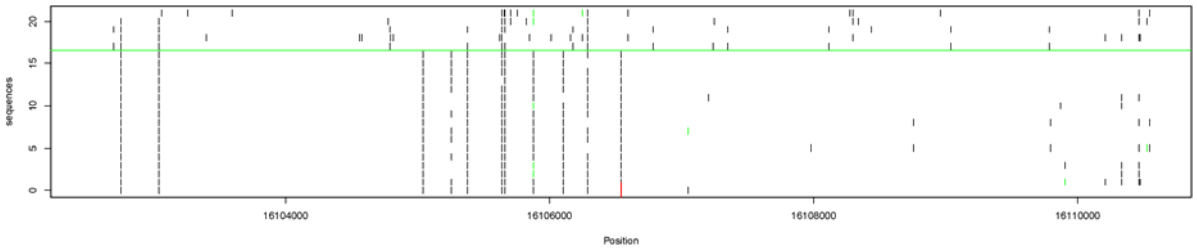
Putative site: 11829615*[$], closest gene: CG43064 (11828293 – 11829821)



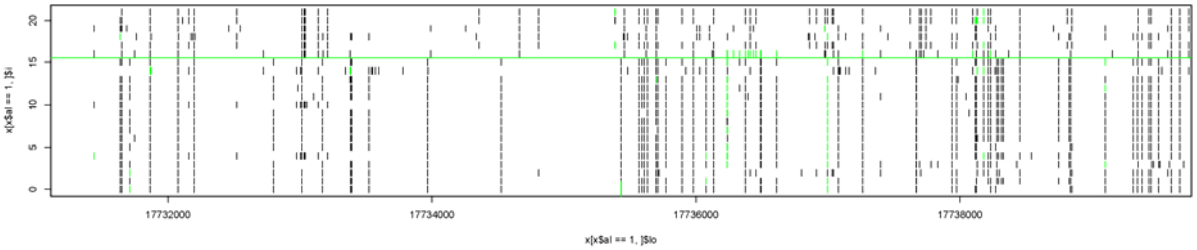
Putative site: 13430186*$, closest gene: CG10713 (13421939 – 13428329)



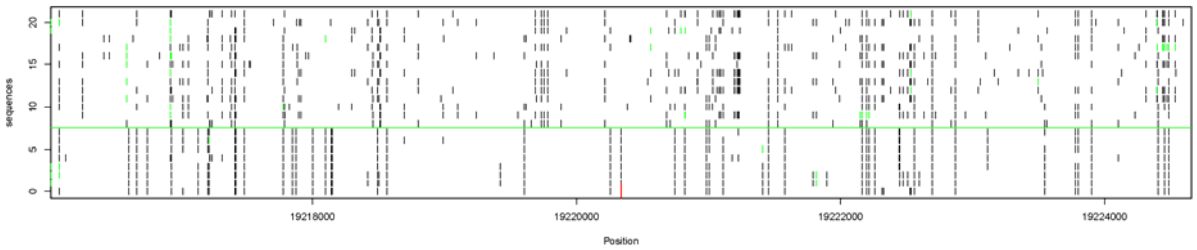
Putative site: 16106542#$, closest gene: Taf4 (16106312 – 16114751)



Putative site: 17735433, closest gene: CG7460 (17733640 – 17735640)

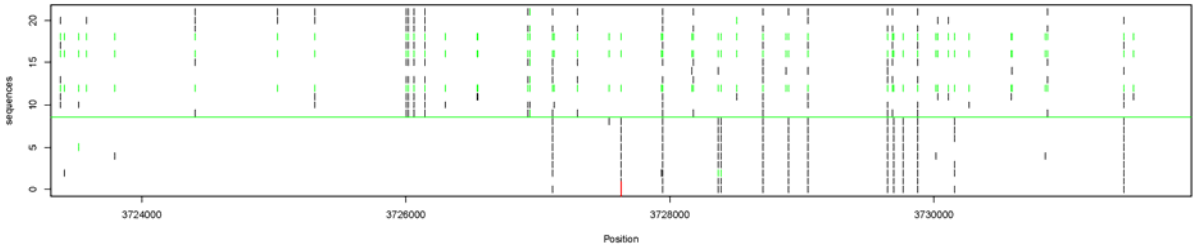


Putative site: 19220338#$, closest gene: fz2 (19134075 – 19228473)

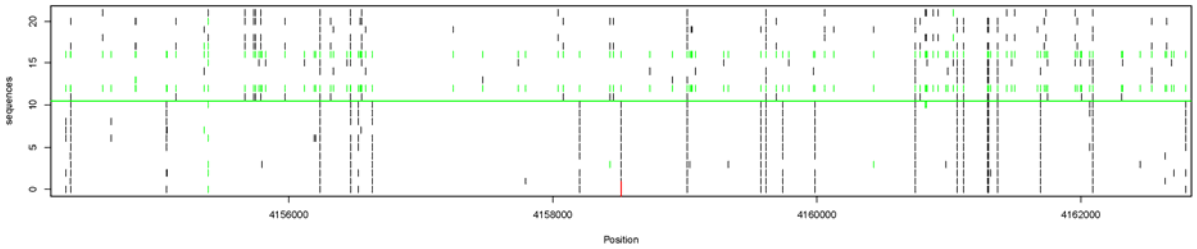


3R Patterns:

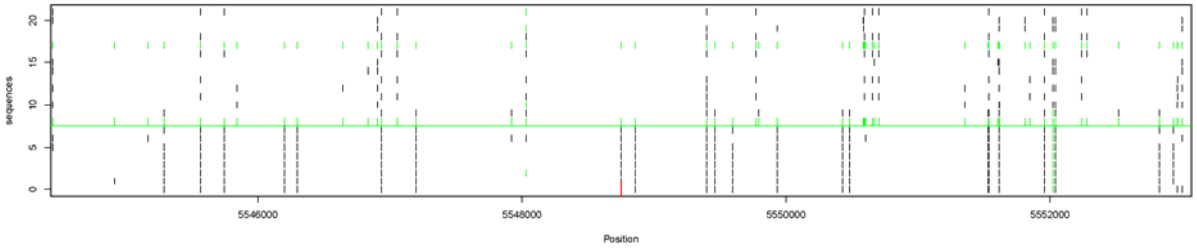
Putative site: 3727631, closest gene: mRpS9 (3714999 – 3728389)



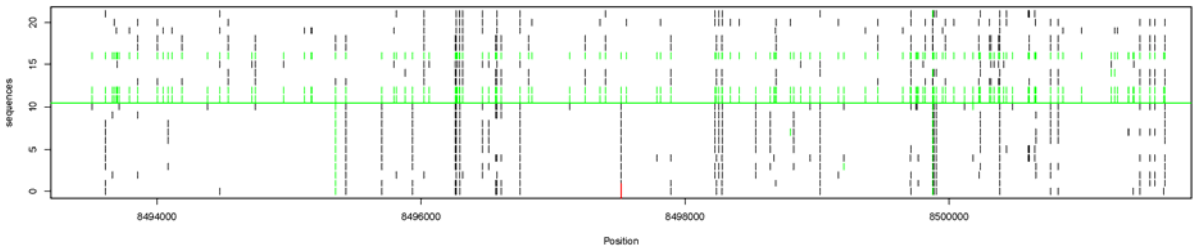
Putative gene: 4158518, closest gene: CG9601 (4167383 – 4169238)



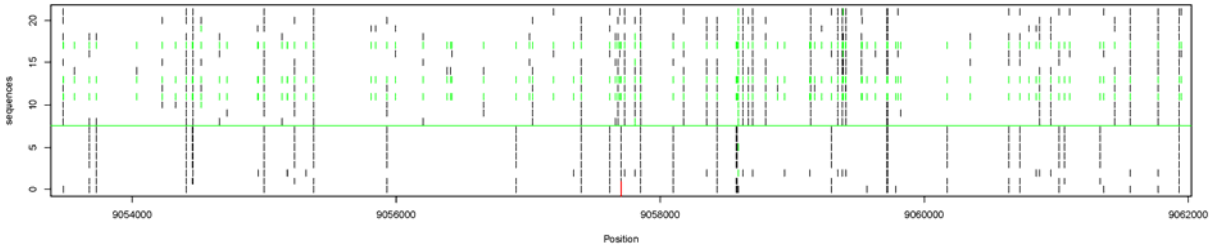
Putative site: 5548751, closest gene: CG8478 (5589372 – 5591857)



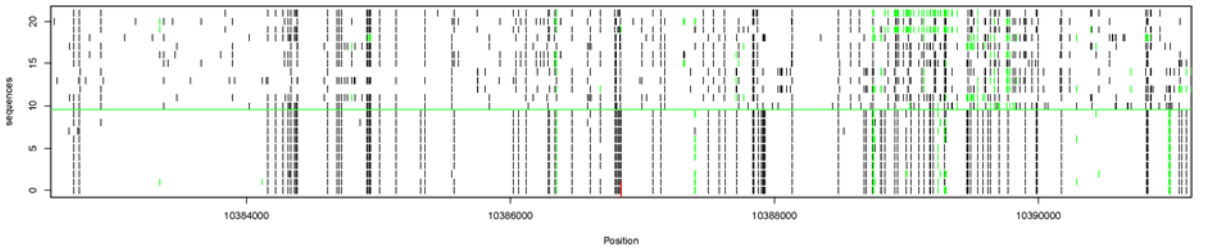
Putative site: 8497516, closest gene: CG14395 (8488553 – 8499681)



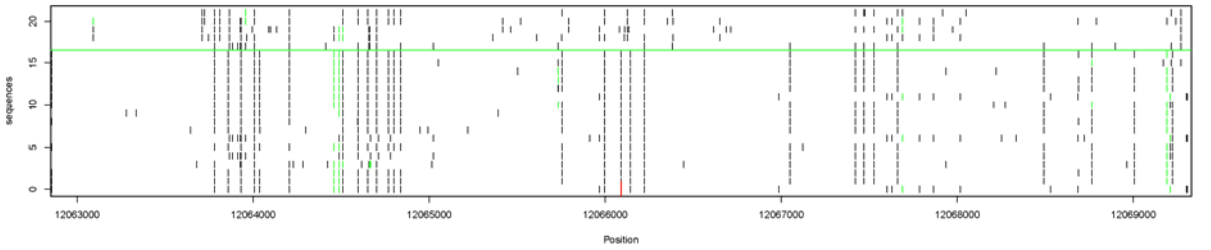
Putative site: 9057704[#], closest gene: Ace (9048673 – 9085239)



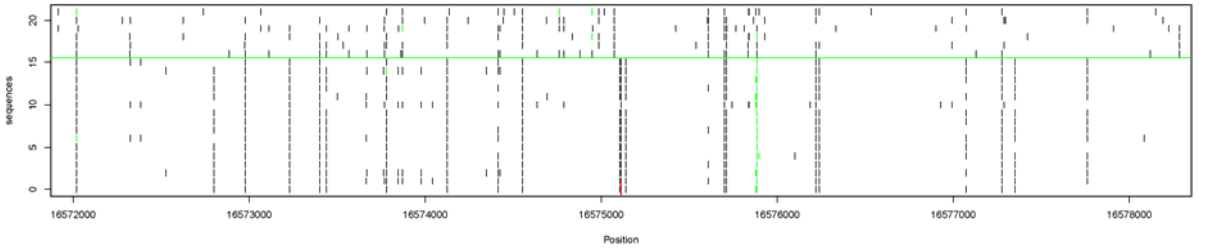
Putative site: 10386839[#$], closest gene: Pde6 (10339623 – 10384026)



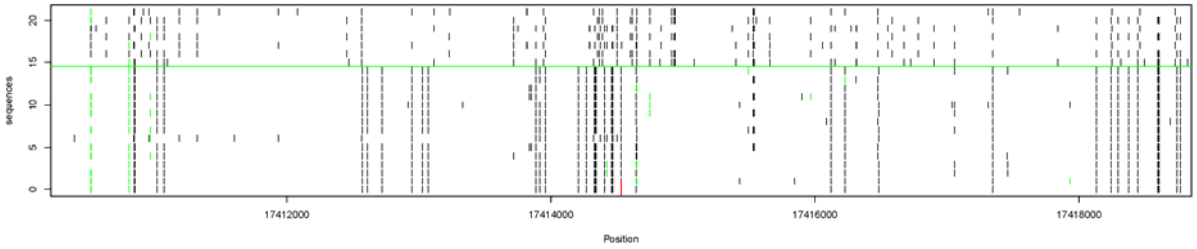
Putative site: 12066090[#$], closest gene: tara (12051373 – 12086051)



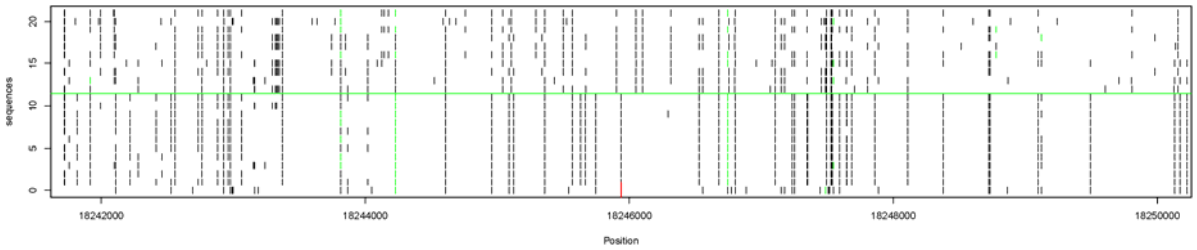
Putative site: 16575113*[$], closest gene: CG42322 (16565830 – 16582361)



Putative site: 17414532*, closest gene: InR (17395970 – 17445043)

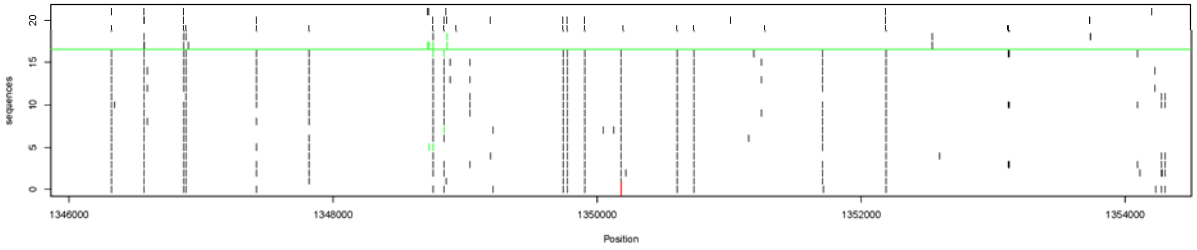


Putative site: 18245938*, closest gene: IqfR (18237023 – 18244773)

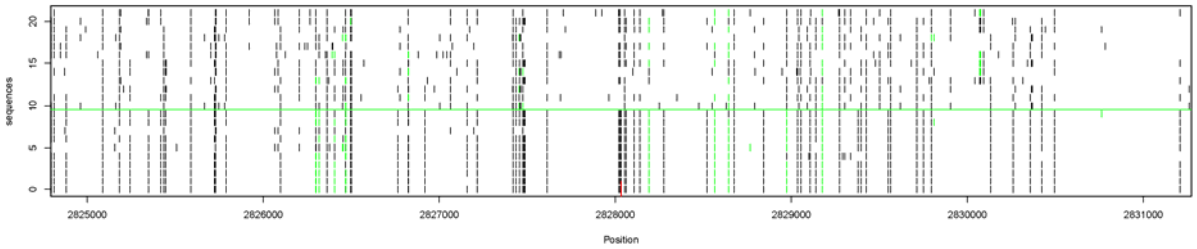


X Patterns:

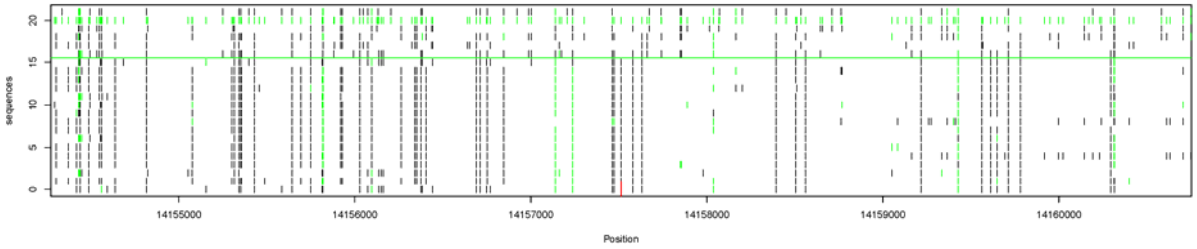
Putative site: 1350182[#], closest gene: MED18 (1759942 – 1760920)



Putative site: 2828033[#$], closest gene: kirre (2634417 – 3028565)



Putative site: 14157513*[$], closest gene: CG1461 (14155256 – 14159412)



Putative site: 15620351*, closest gene: CG8184 (15606661 – 15625968)

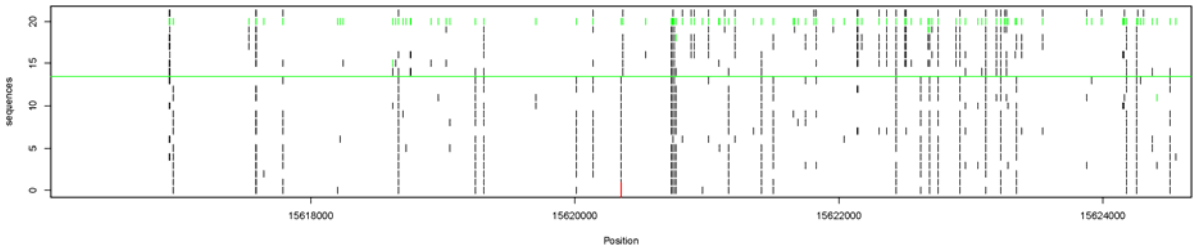
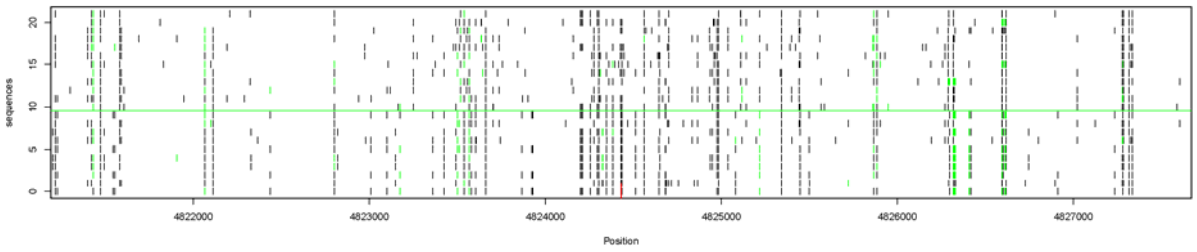


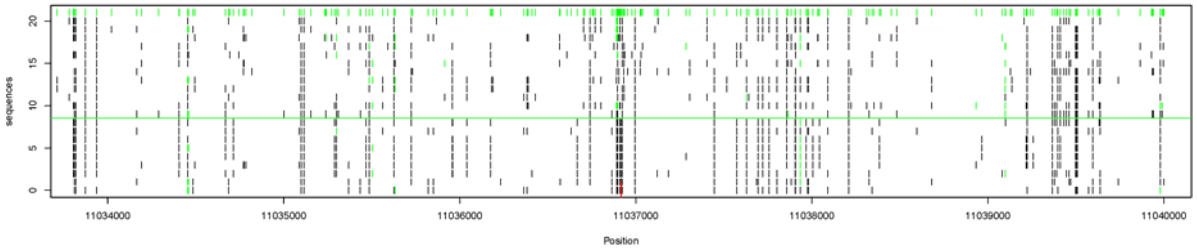
Figure S7 legend: Polymorphism pattern surrounding the site of strongest signal detected by CLR test within each cluster listed in Table 2. In each figure, 22 chromosomes are aligned (number from 0 to 21 vertically) and the putative site under selection (site with strongest signal) is located in the middle (red tick on horizontal axis). Chromosomes are arranged below or above a green line according to allele type (derived or ancestral, respectively) at the putative site. Derived alleles and missing base calls at polymorphic sites are represented by black and green bars, respectively. Whether each region overlaps with a candidate region of complete selective sweep, with a cluster detected by *iHS* test, and by nS_L test are indicated by *, [#], and [$], respectively.

Figure S8

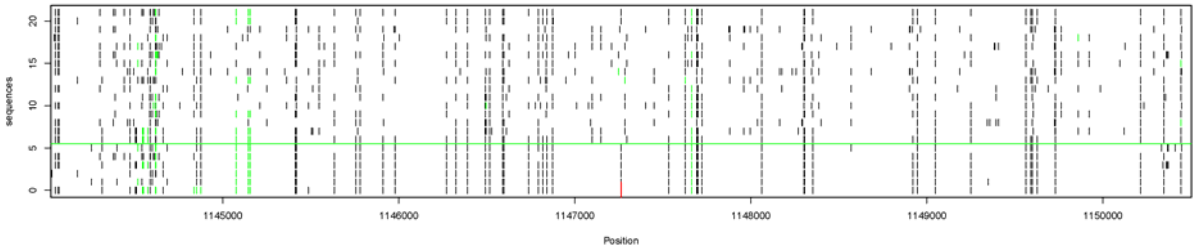
Chromosome: 2L, putative site: 4824431, detected by *iHS* (-4.03)



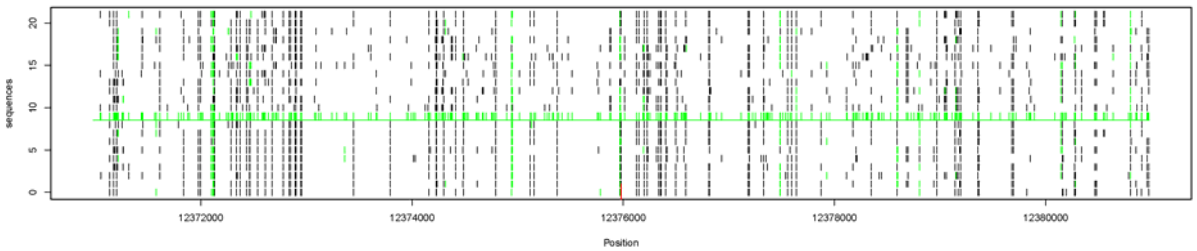
Chromosome: 2L, putative site: 11036917, detected by *iHS* (-3.82)



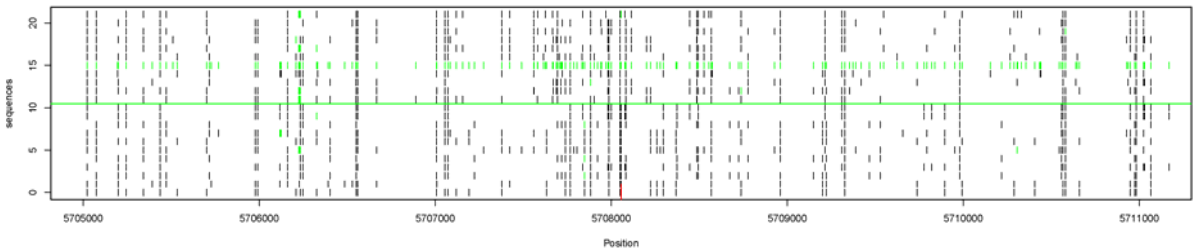
Chromosome: 2L, putative site: 1147263, detected by nS_L (-3.77)



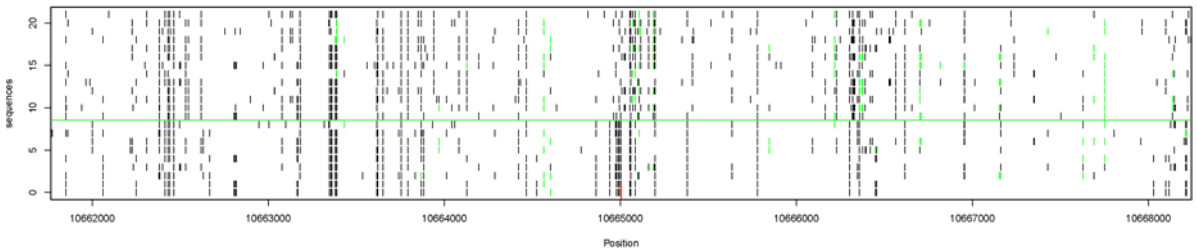
Chromosome: 2L, putative site: 12375980, detected by nS_L (-3.46)



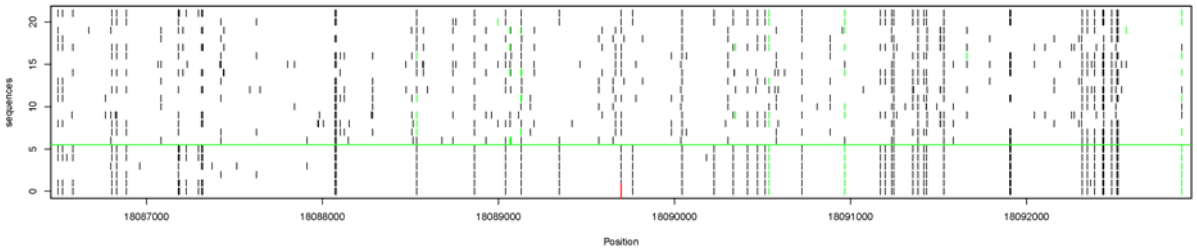
Chromosome: 2R, putative site: 5708056, detected by *iHS* (-3.33)



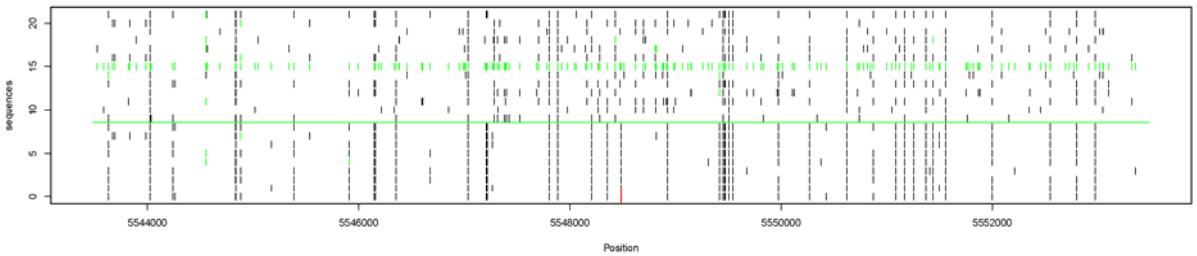
Chromosome: 2R, putative site: 10665005, detected by *iHS* (-3.86)



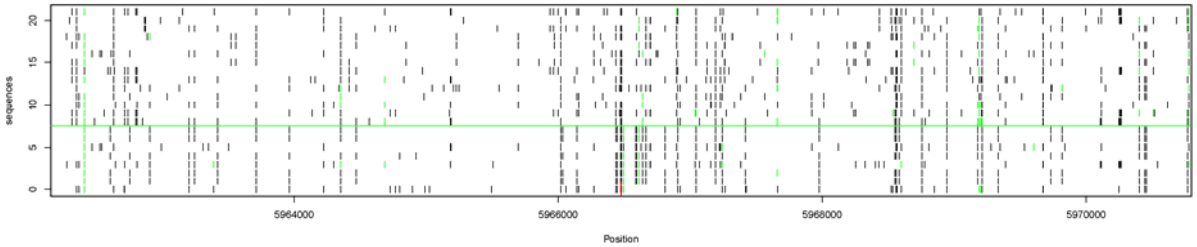
Chromosome: 2R, putative site: 18089697, detected by nS_L (-2.80)



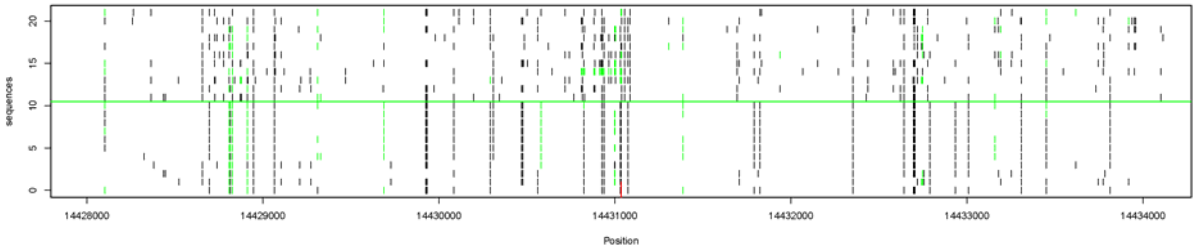
Chromosome: 2R, putative site: 5548485, detected by nS_L (-2.42)



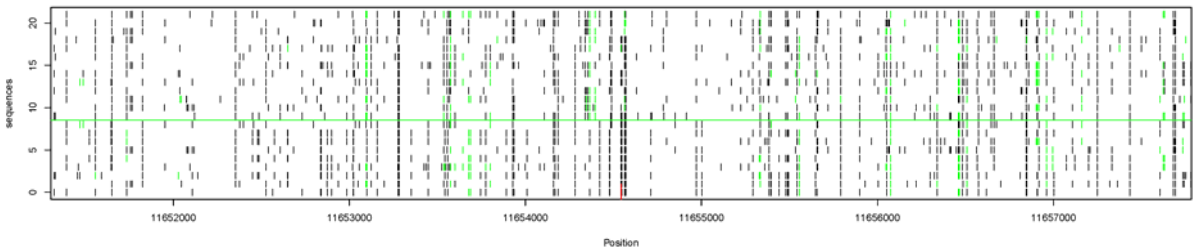
Chromosome: 3L, putative site: 5966477, detected by *iHS* (-4.18)



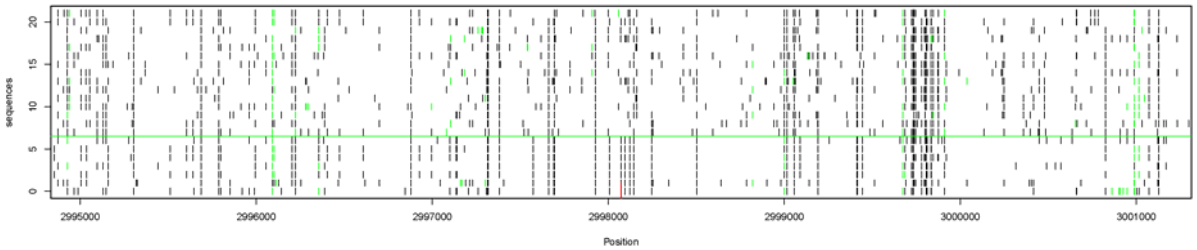
Chromosome: 3L, putative site: 14431036, detected by *iHS* (-4.12)



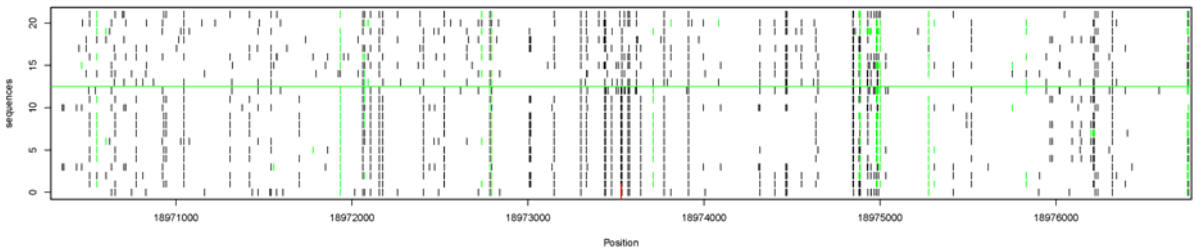
Chromosome: 3L, putative site: 11654544, detected by nS_L (-3.59)



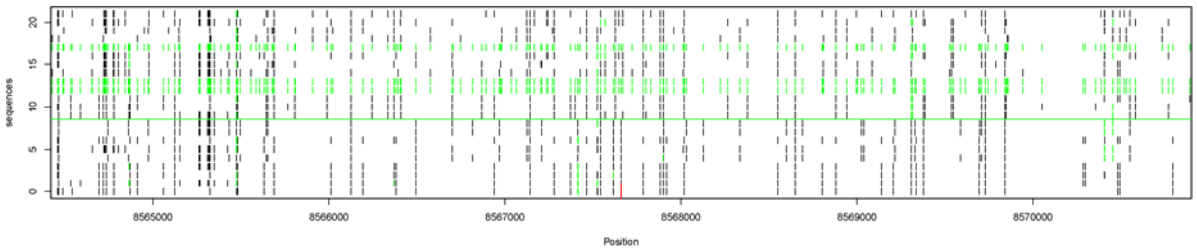
Chromosome: 3L, putative site: 2998073, detected by nS_L (-3.82)



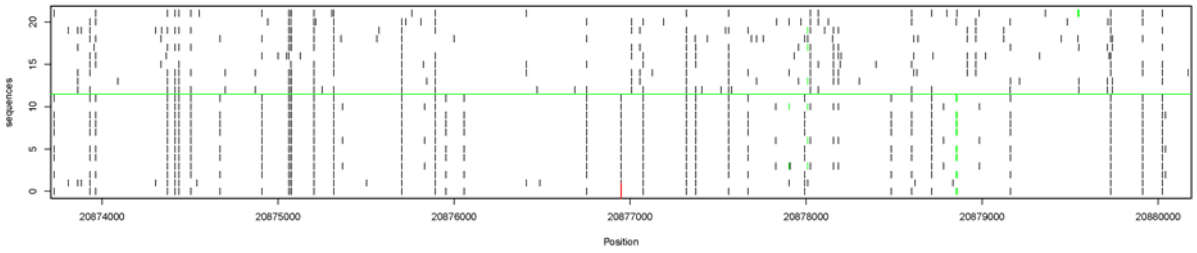
Chromosome: 3R, putative site: 18973529, detected by *iHS* (-3.49)



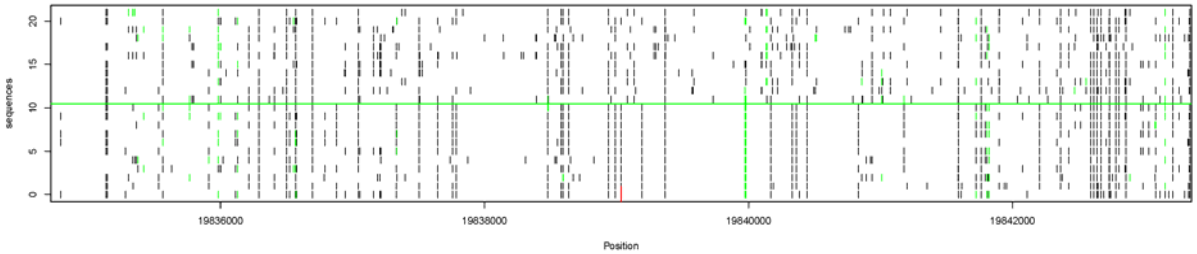
Chromosome: 3R, putative site: 8567660, detected by *iHS* (-3.70)



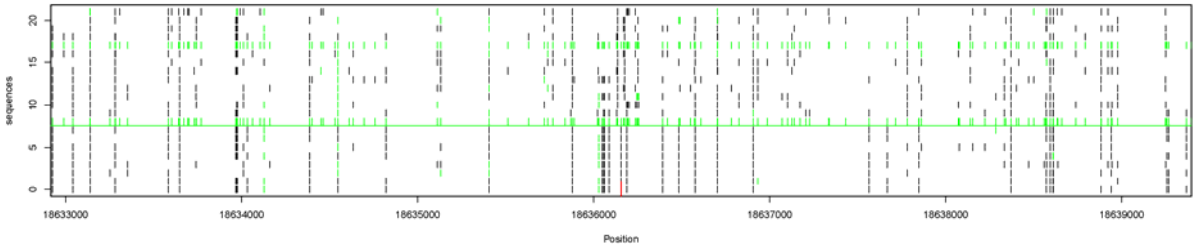
Chromosome: 3R, putative site: 20876949, detected by nS_L (-3.49)



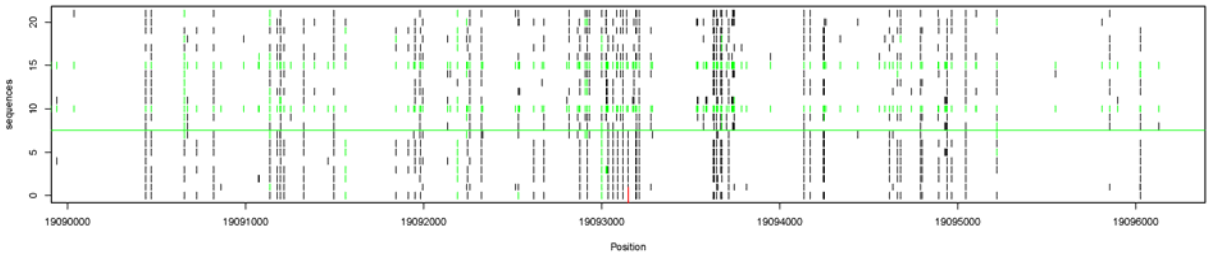
Chromosome: 3R, putative site: 19839035, detected by nS_L (-3.17)



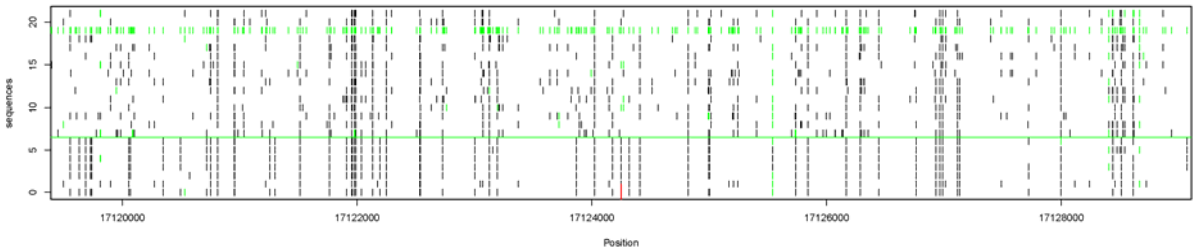
Chromosome: X, putative site: 18636156, detected by *iHS* (-5.23)



Chromosome: X, putative site: 19093150, detected by *iHS* (-5.80)



Chromosome: X, putative site: 17124249, detected by nS_L (-4.25)



Chromosome: X, putative site: 14975977, detected by nS_L (-4.16)

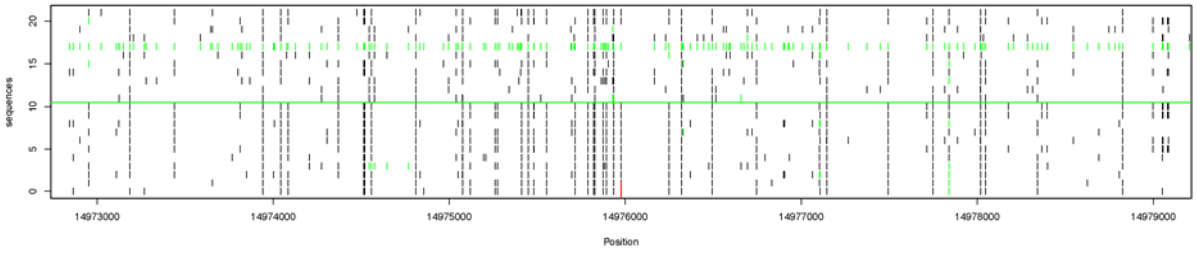


Figure S8 legend: Polymorphism patterns of genome areas surrounding the putative site under selection detected exclusively by *iHS* or *nSL* method. For each chromosome arm, top two candidate loci with strongest signals by each method, however not significant by the other tests, are shown.

Figure S9

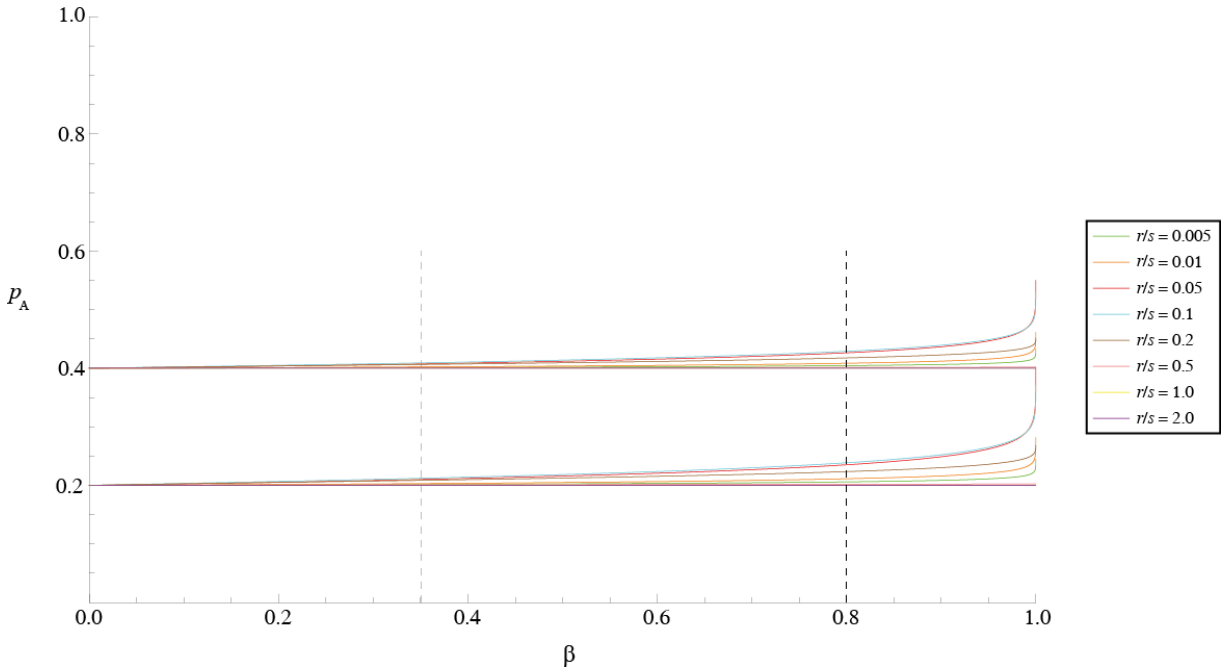


Figure S9 Legend: Deterministic changes in p_A, the frequency of a linked neutral derived allele in the subpopulation of chromosomes carrying the ancestral allele of the *S* locus during the course of a selective sweep. β is the frequency of the beneficial mutation in the population. The frequencies were obtained from equation (12b) of Stephan et al. (1992) for two different values of *p* (frequency of neutral derived allele at the beginning of sweep): 0.2 and 0.4, with different values of *r/s* (recombination rate/selection coefficient). Dashed lines (at β = 0.35 and β = 0.8) mark the interval of beneficial allele frequency at the *S* locus for which composite likelihood test is performed.