Tutorial of the STRUCTURE software

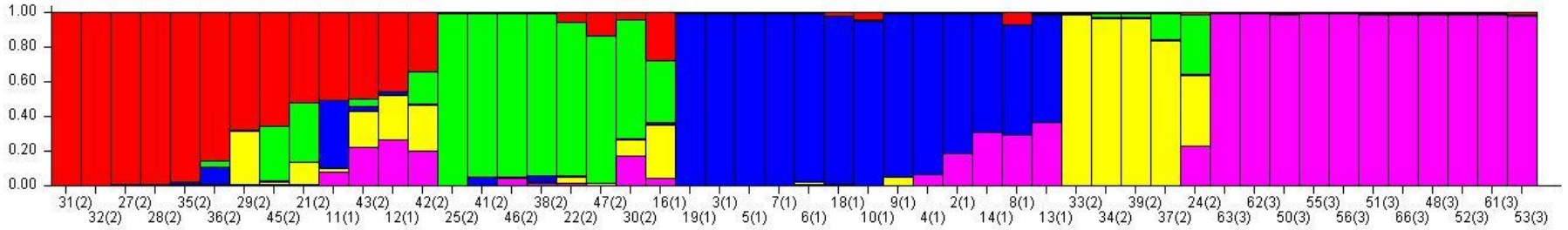
Dr. Sung-Chur Sim

Tomato Genetics and Breeding program

The Ohio State Univ., OARDC

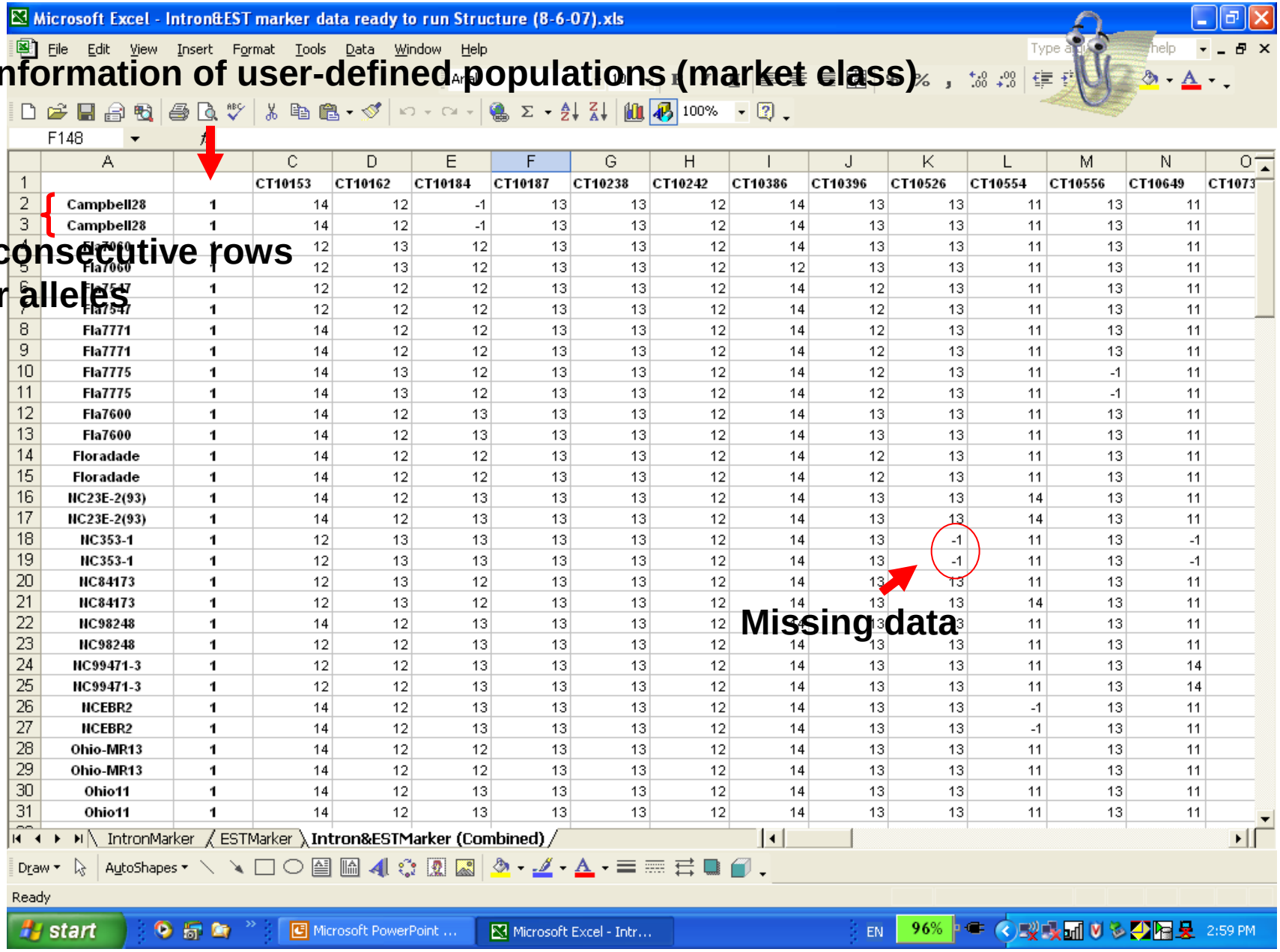
STRUCTURE software

- A model-based clustering method (Pritchard et al. 2000)
  - Free software (http://pritch.bsd.uchicago.edu/software/structure2_1.html)
  - Bayesian approach (MCMC: Markov Chain Monte Carlo)
  - Detects the underlying genetic population among a set of individuals genotyped at multiple markers
  - Computes the proportion of the genome of an individual originating from each inferred population (quantitative clustering method)



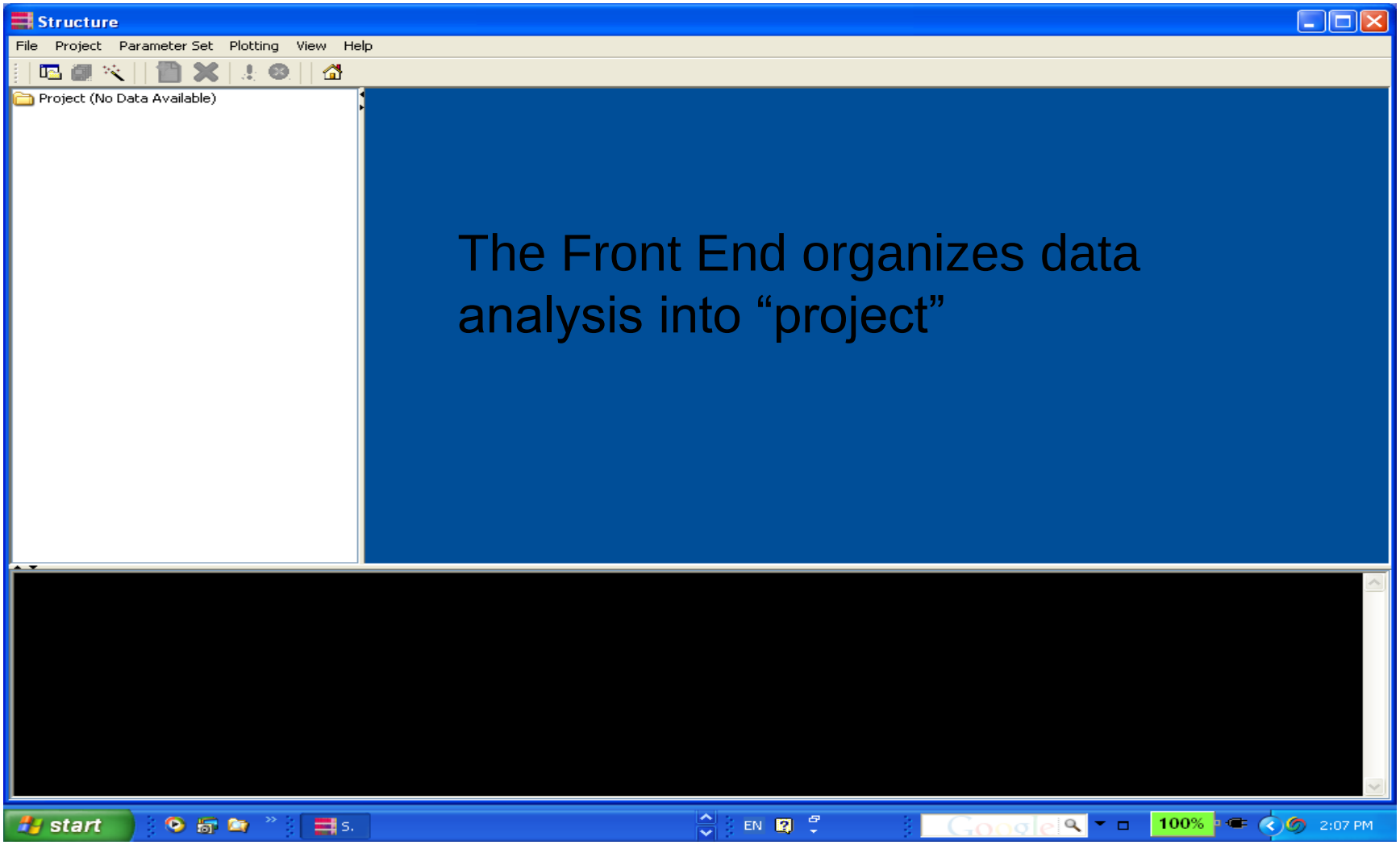
Input data

- A matrix where the data for individuals are in rows, the loci are in column
  - ***n* consecutive rows** have the data for each individual of *n*-ploid species
  - **Integer** should be used for coding genotype
  - Missing data should be indicated by **a number** which doesn't occur elsewhere in the data (e.g. -1)
  - The data file should be a **text file (.txt)** not an excel file (.xls) for running STRUCTURE

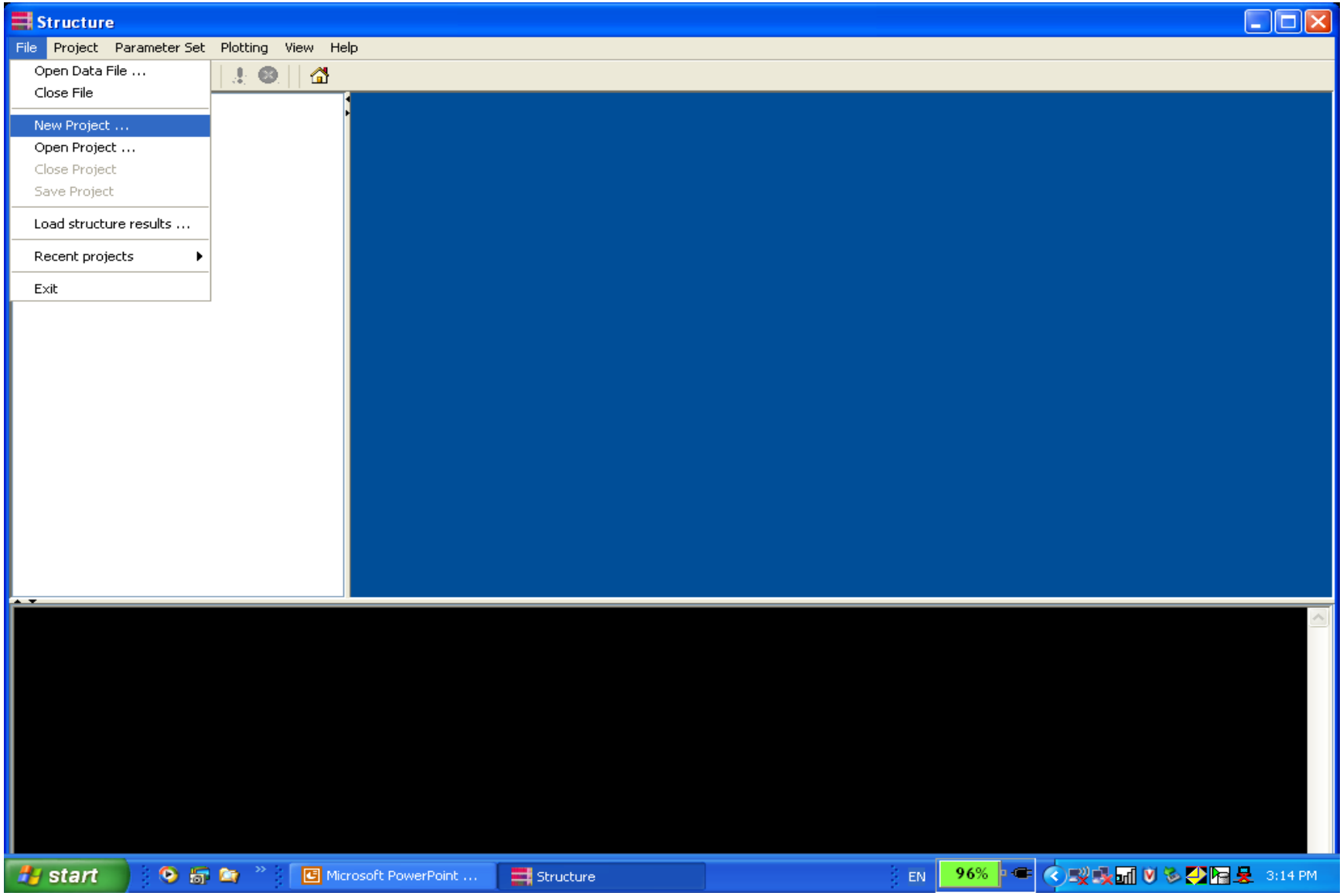


	A		C	D	E	F	G	H	I	J	K	L	M	N	O
1			CT10153	CT10162	CT10184	CT10187	CT10238	CT10242	CT10386	CT10396	CT10526	CT10554	CT10556	CT10649	CT1073
2	Campbell28	1	14	12	-1	13	13	12	14	13	13	11	13	11	
3	Campbell28	1	14	12	-1	13	13	12	14	13	13	11	13	11	
4	Fla7060	1	12	13	12	13	13	12	14	13	13	11	13	11	
5	Fla7060	1	12	13	12	13	13	12	12	13	13	11	13	11	
6	Fla7547	1	12	12	12	13	13	12	14	12	13	11	13	11	
7	Fla7547	1	12	12	12	13	13	12	14	12	13	11	13	11	
8	Fla7771	1	14	12	12	13	13	12	14	12	13	11	13	11	
9	Fla7771	1	14	12	12	13	13	12	14	12	13	11	13	11	
10	Fla7775	1	14	13	12	13	13	12	14	12	13	11	-1	11	
11	Fla7775	1	14	13	12	13	13	12	14	12	13	11	-1	11	
12	Fla7600	1	14	12	13	13	13	12	14	13	13	11	13	11	
13	Fla7600	1	14	12	13	13	13	12	14	13	13	11	13	11	
14	Floradade	1	14	12	12	13	13	12	14	12	13	11	13	11	
15	Floradade	1	14	12	12	13	13	12	14	12	13	11	13	11	
16	NC23E-2(93)	1	14	12	13	13	13	12	14	13	13	14	13	11	
17	NC23E-2(93)	1	14	12	13	13	13	12	14	13	13	14	13	11	
18	NC353-1	1	12	13	13	13	13	12	14	13	-1	11	13	-1	
19	NC353-1	1	12	13	13	13	13	12	14	13	-1	11	13	-1	
20	NC84173	1	12	13	12	13	13	12	14	13	13	11	13	11	
21	NC84173	1	12	13	12	13	13	12	14	13	13	14	13	11	
22	NC98248	1	14	12	13	13	13	12	14	13	13	11	13	11	
23	NC98248	1	12	12	13	13	13	12	14	13	13	11	13	11	
24	NC99471-3	1	12	12	13	13	13	12	14	13	13	11	13	14	
25	NC99471-3	1	12	12	13	13	13	12	14	13	13	11	13	14	
26	NCEBR2	1	14	12	13	13	13	12	14	13	13	-1	13	11	
27	NCEBR2	1	14	12	13	13	13	12	14	13	13	-1	13	11	
28	Ohio-MR13	1	14	12	12	13	13	12	14	13	13	11	13	11	
29	Ohio-MR13	1	14	12	12	13	13	12	14	13	13	11	13	11	
30	Ohio11	1	14	12	13	13	13	12	14	13	13	11	13	11	
31	Ohio11	1	14	12	13	13	13	12	14	13	13	11	13	11	

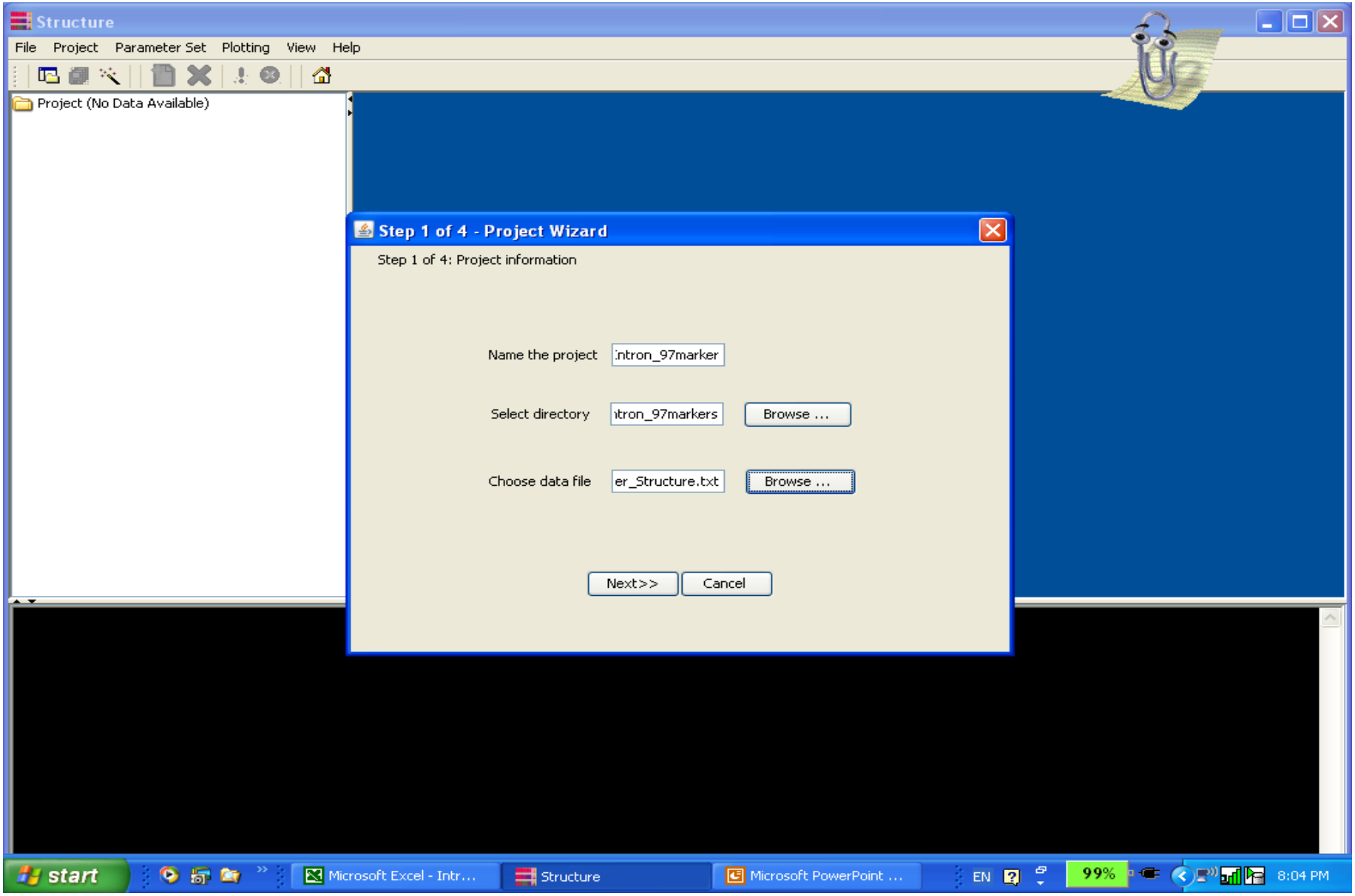
Running STRUCTURE from a graphical interface, Front End



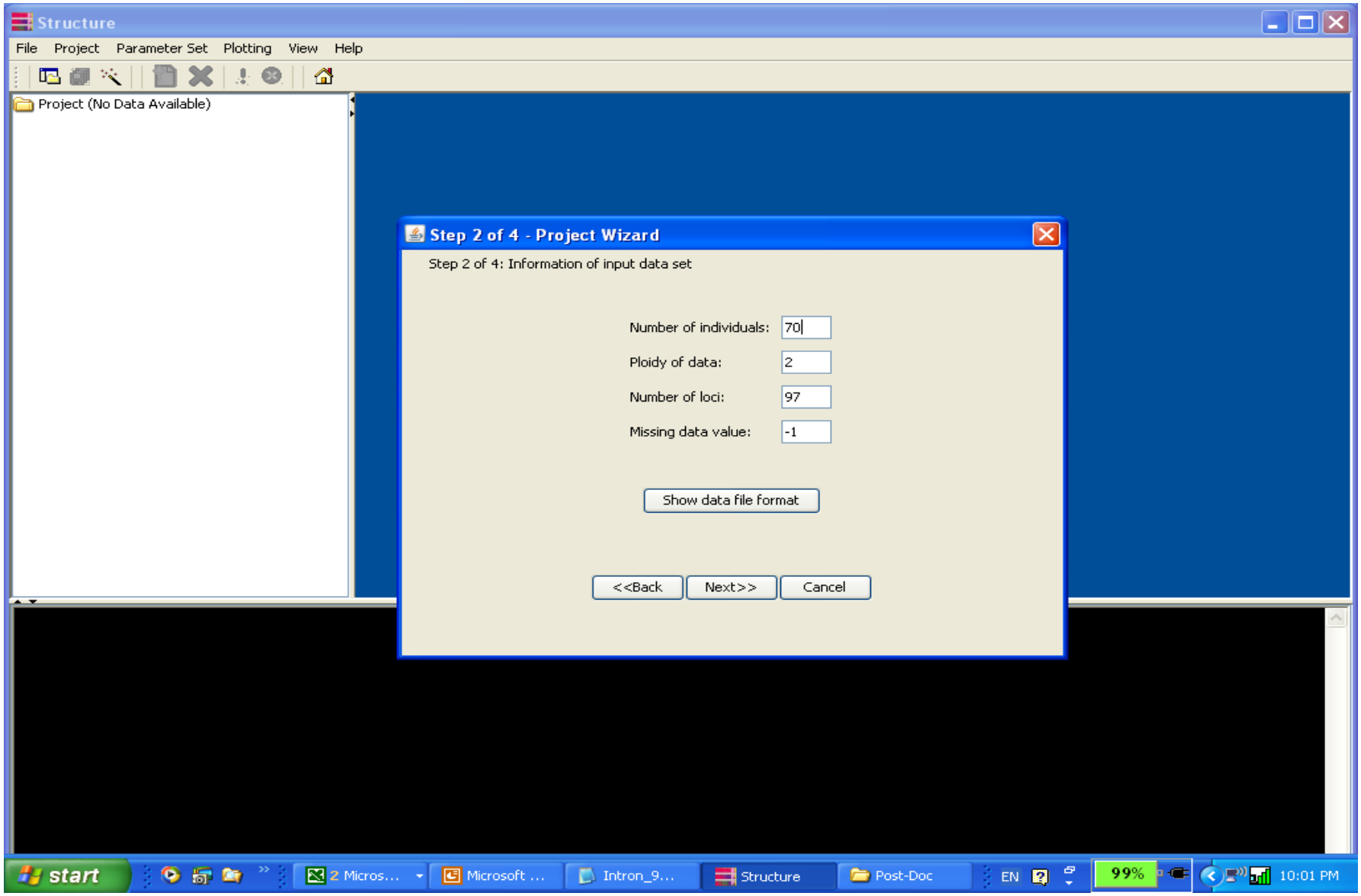
Importing input data into a project



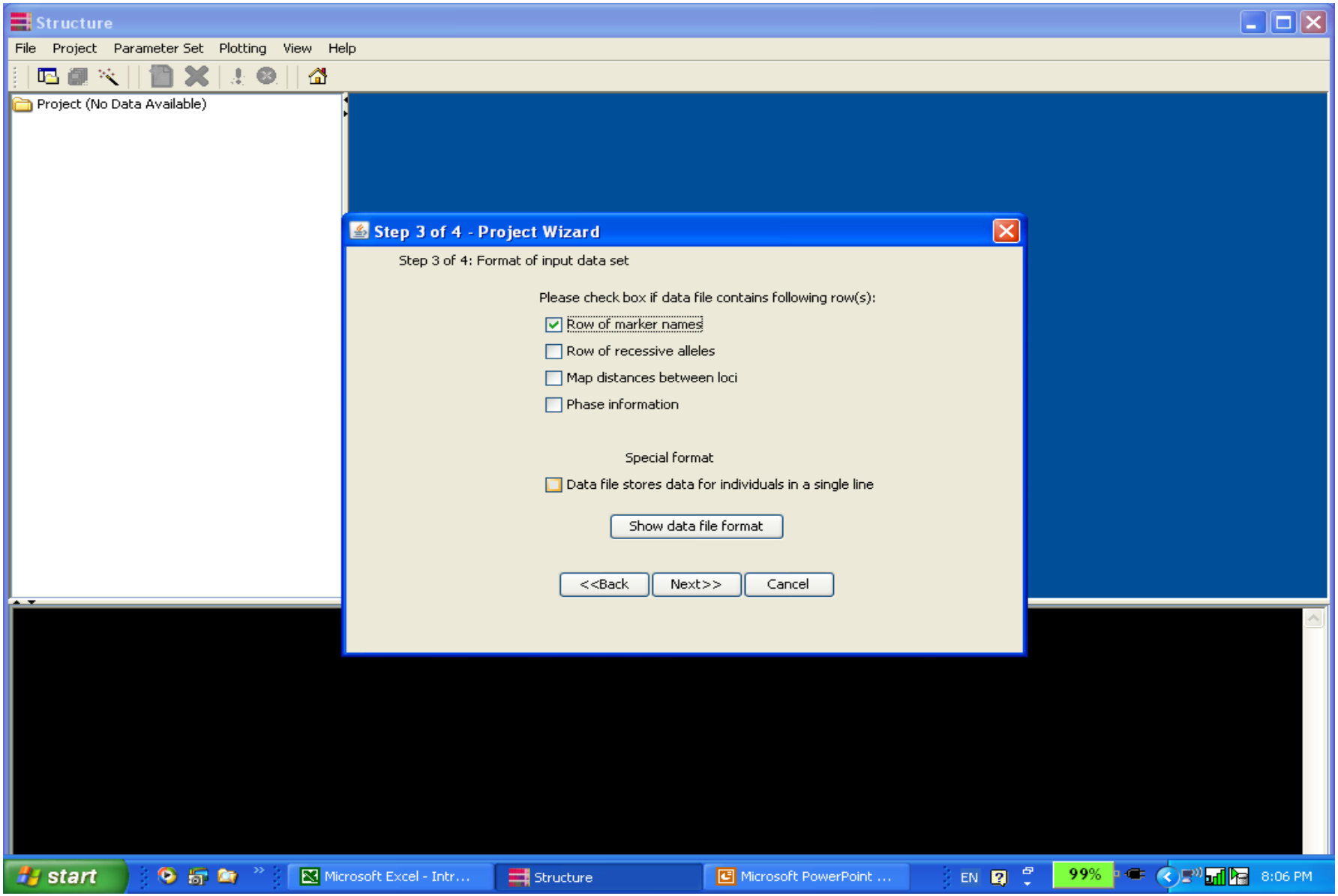
Importing input data into a project (cont.)



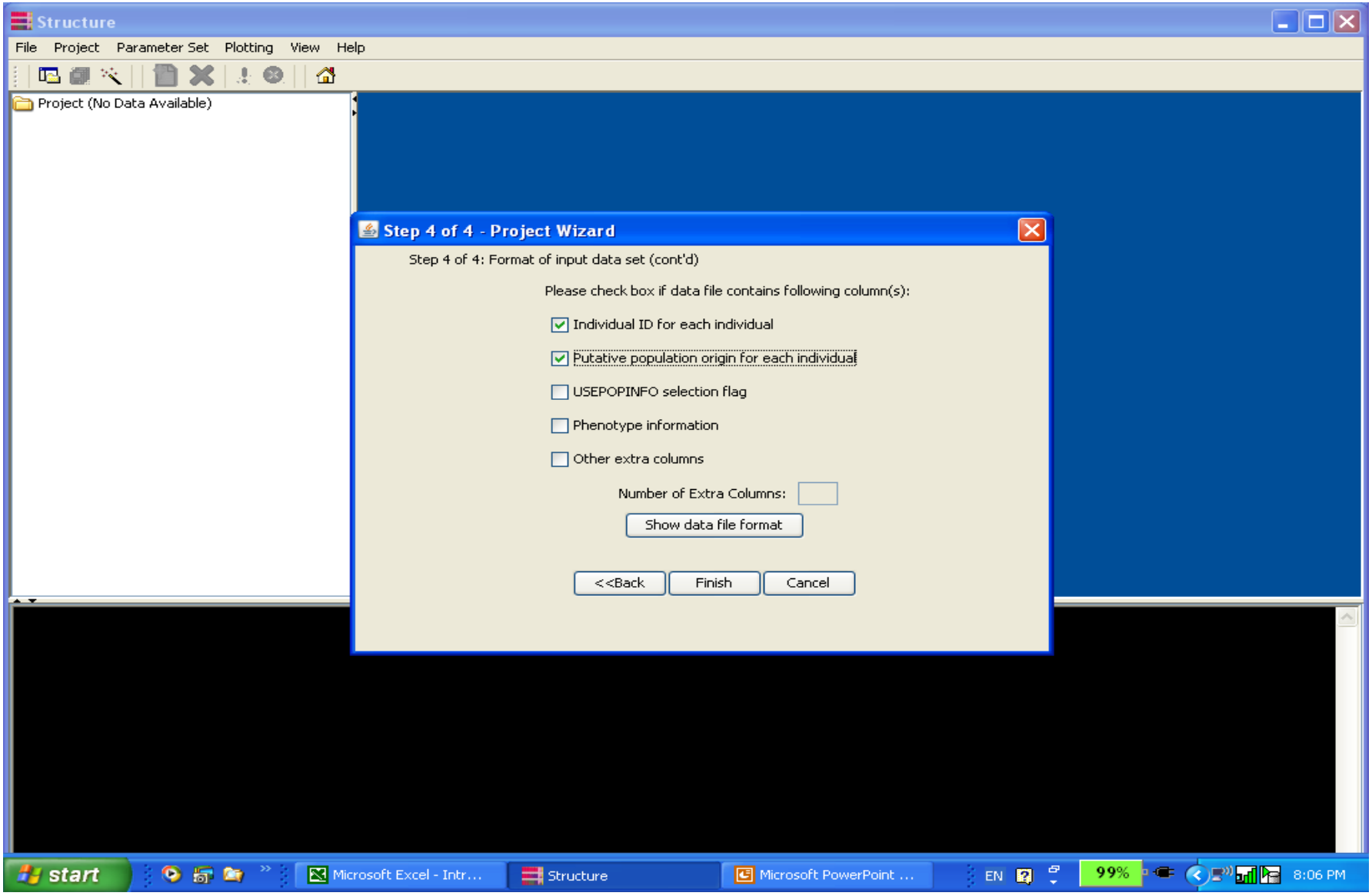
Importing input data into a project (cont.)



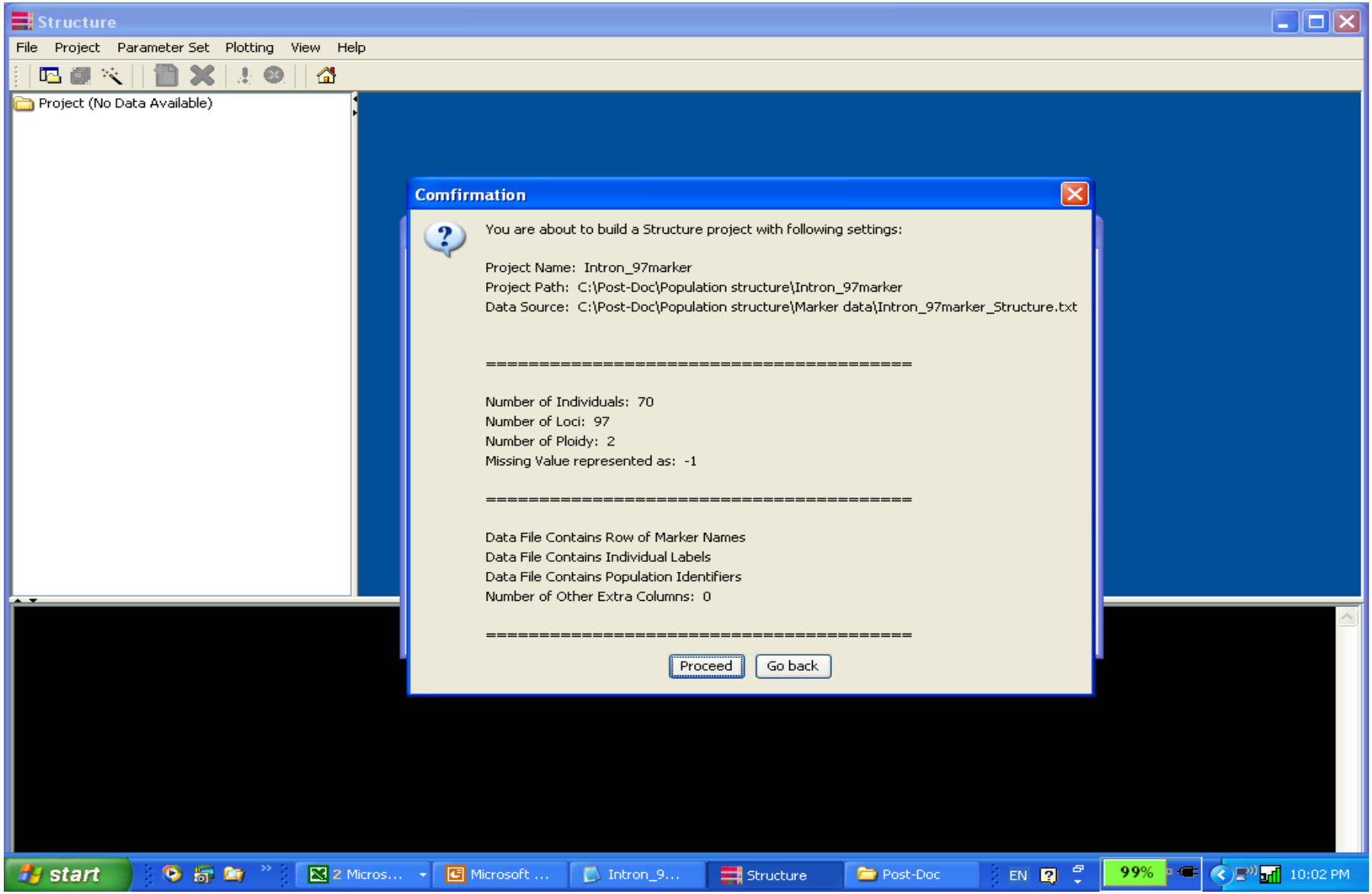
Importing input data into a project (cont.)



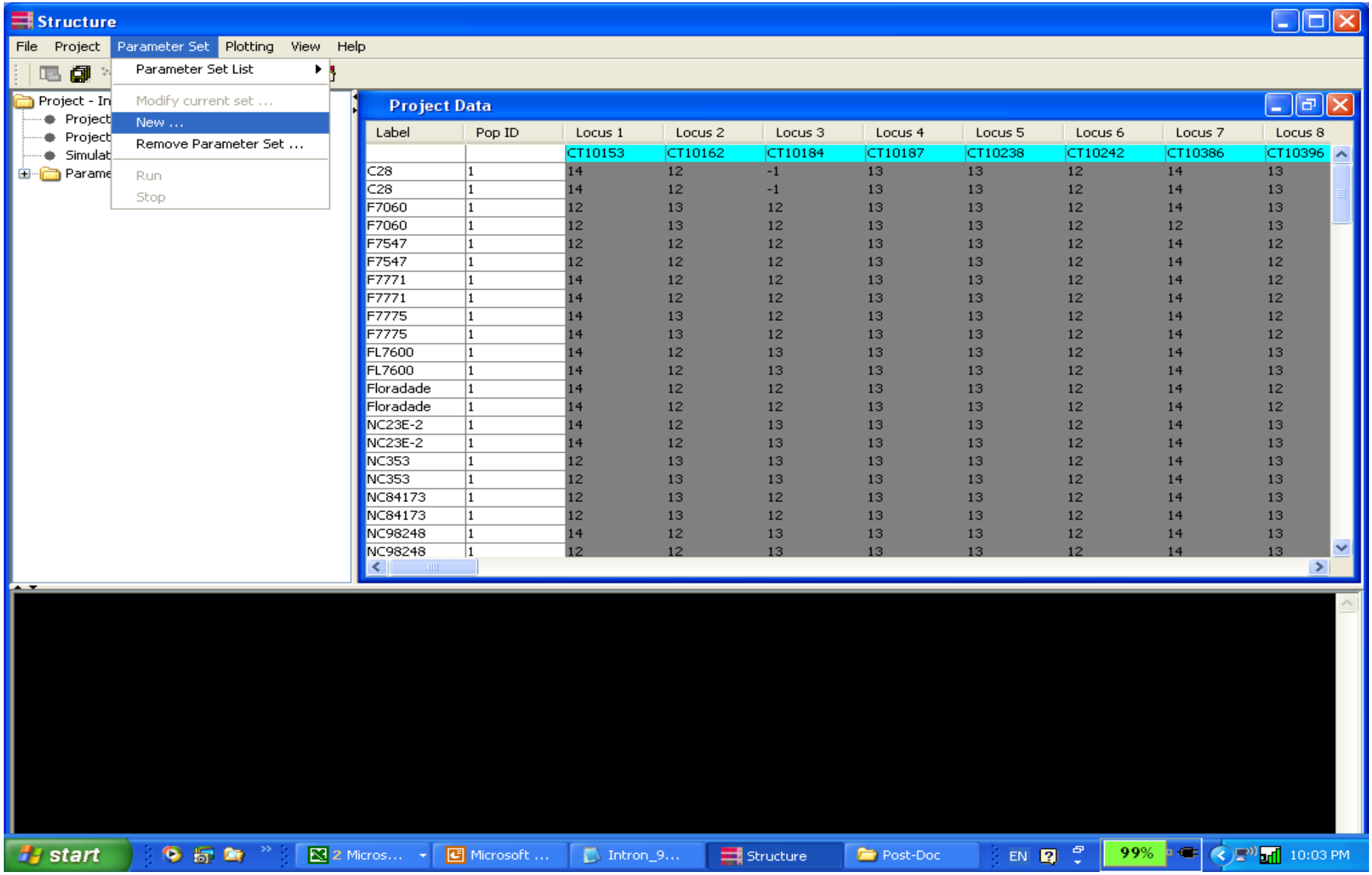
Importing input data into a project (cont.)



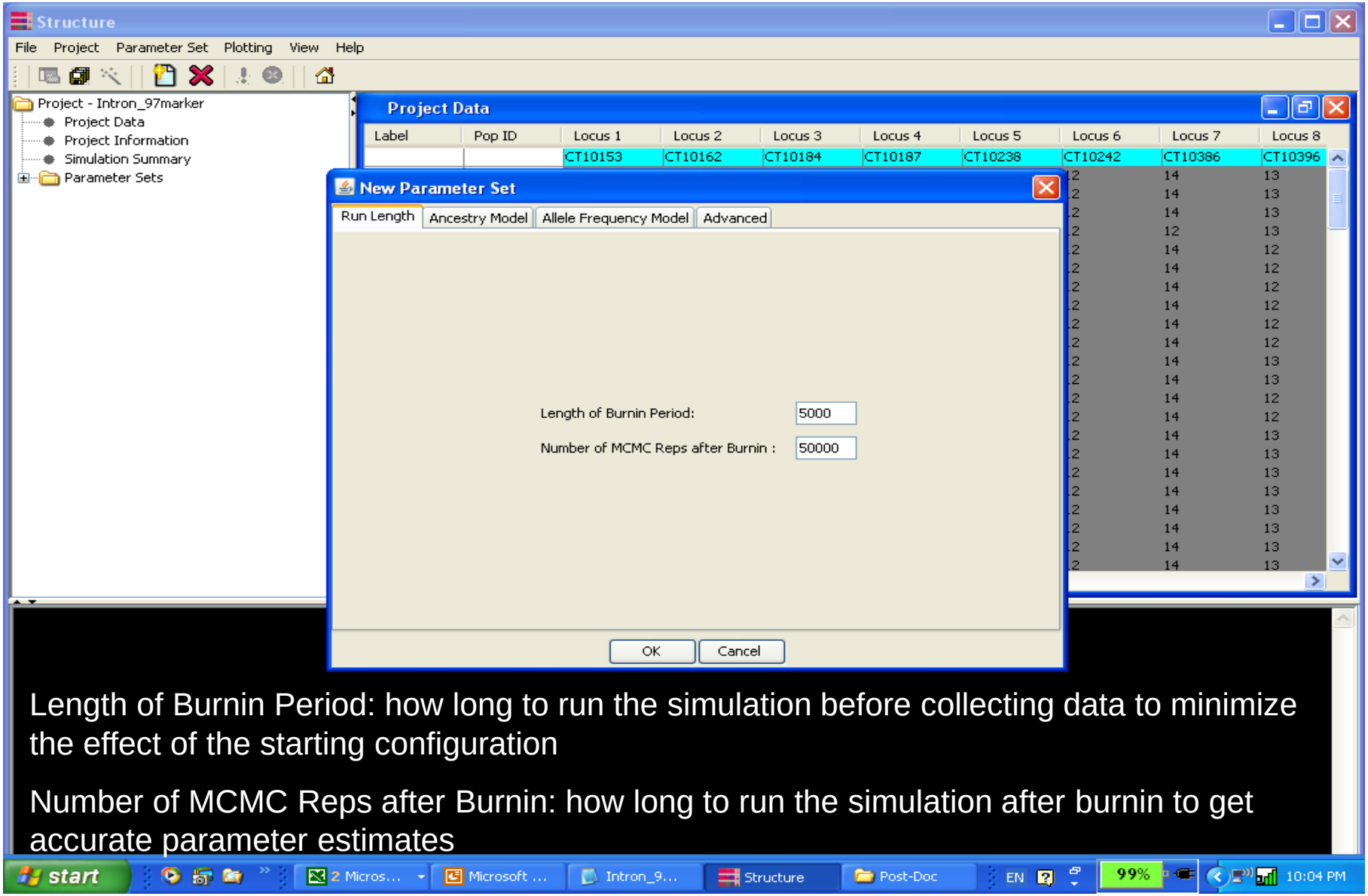
Importing input data into a project (cont.)



Configuring a parameter set



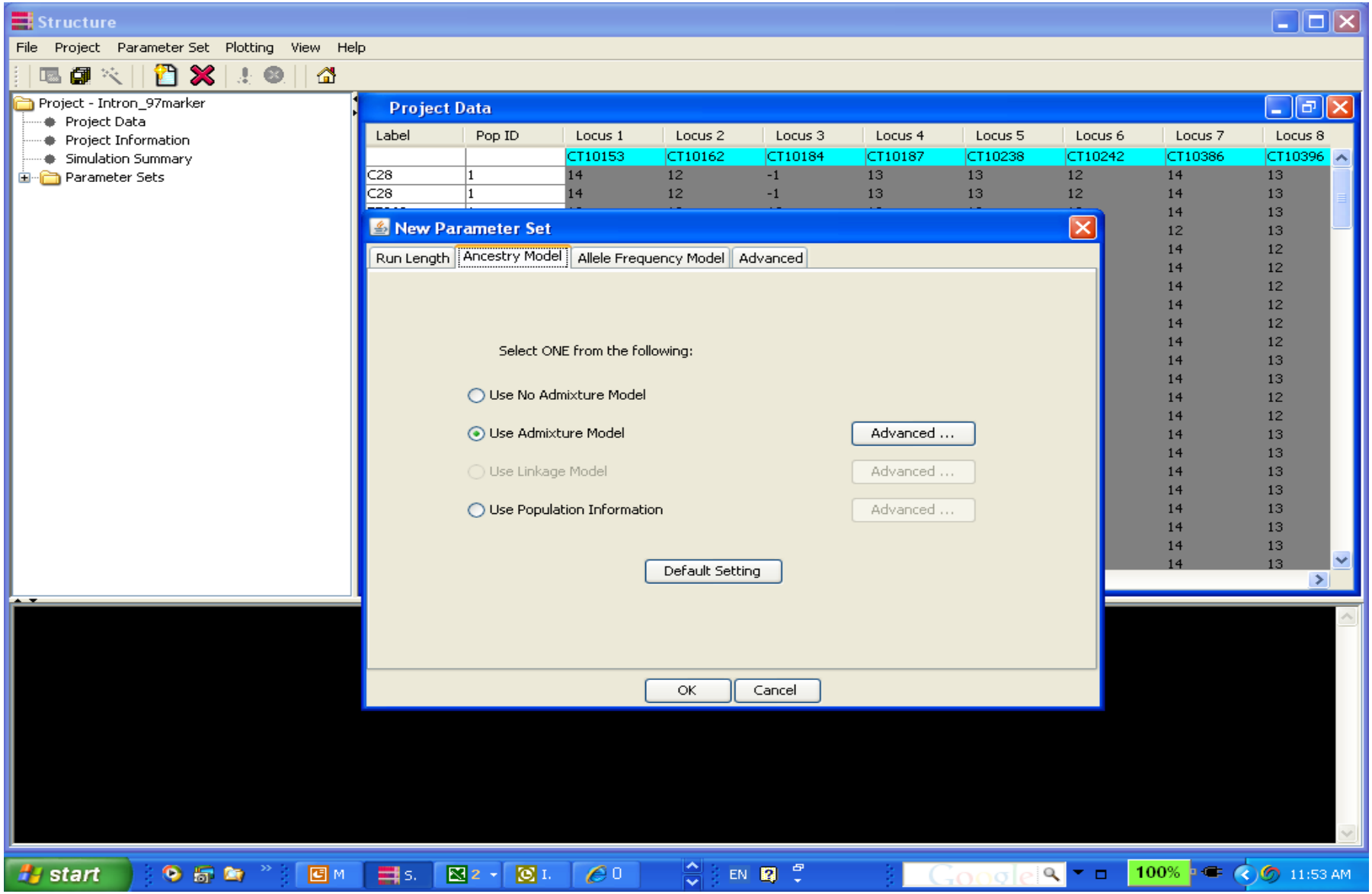
Configuring a parameter set (cont.)



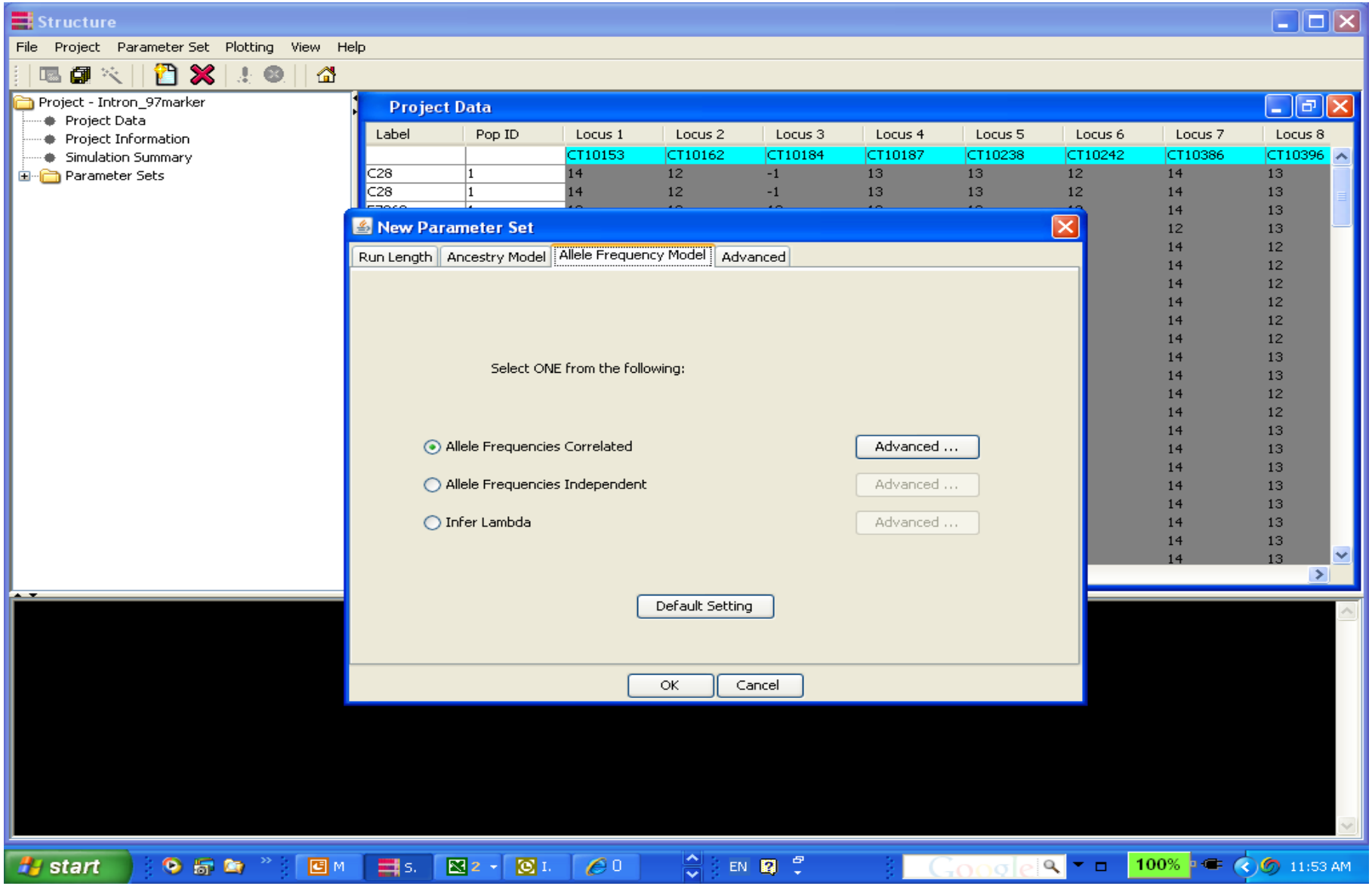
Length of Burnin Period: how long to run the simulation before collecting data to minimize the effect of the starting configuration

Number of MCMC Reps after Burnin: how long to run the simulation after burnin to get accurate parameter estimates

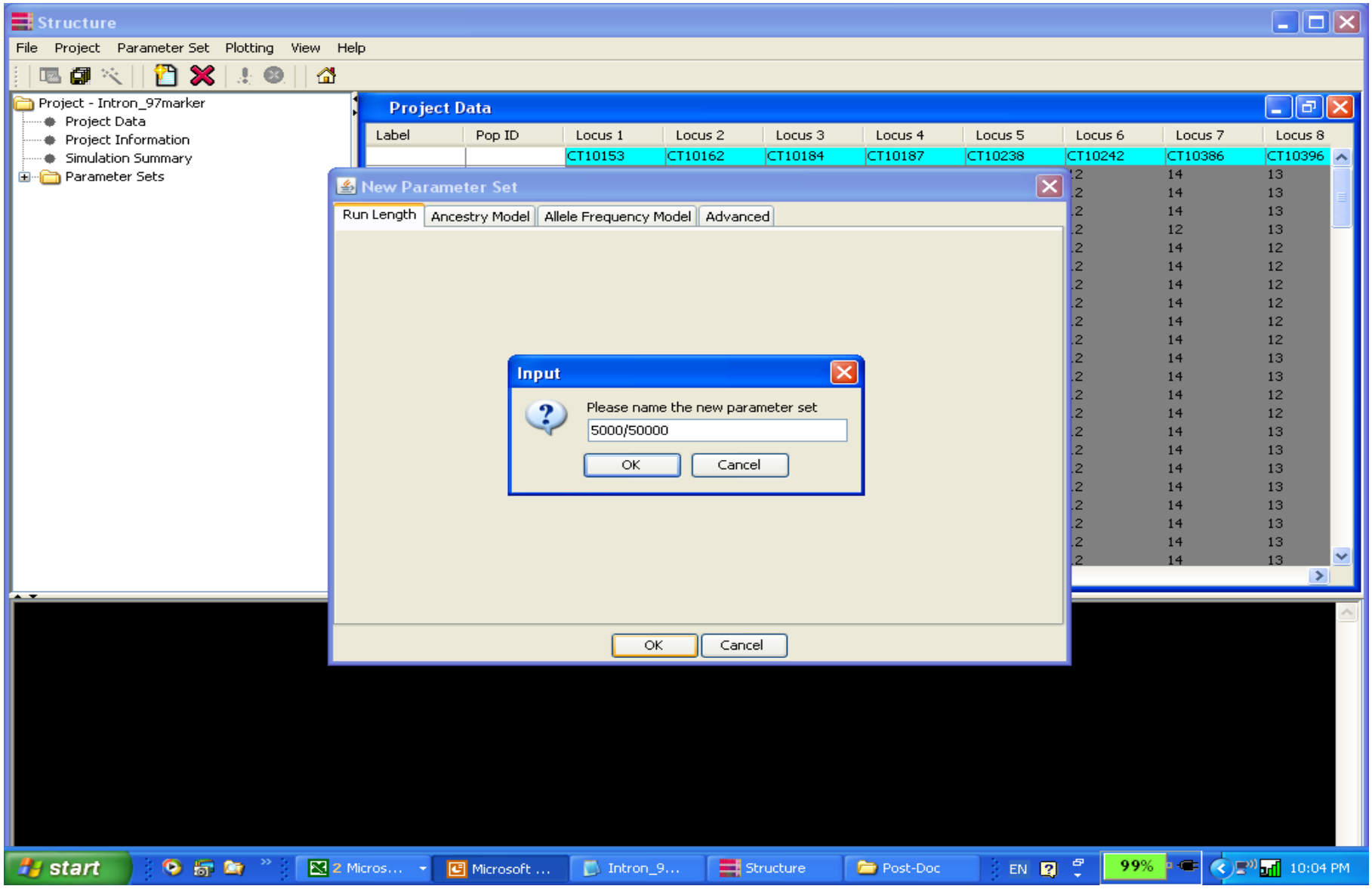
Configuring a parameter set (cont.)



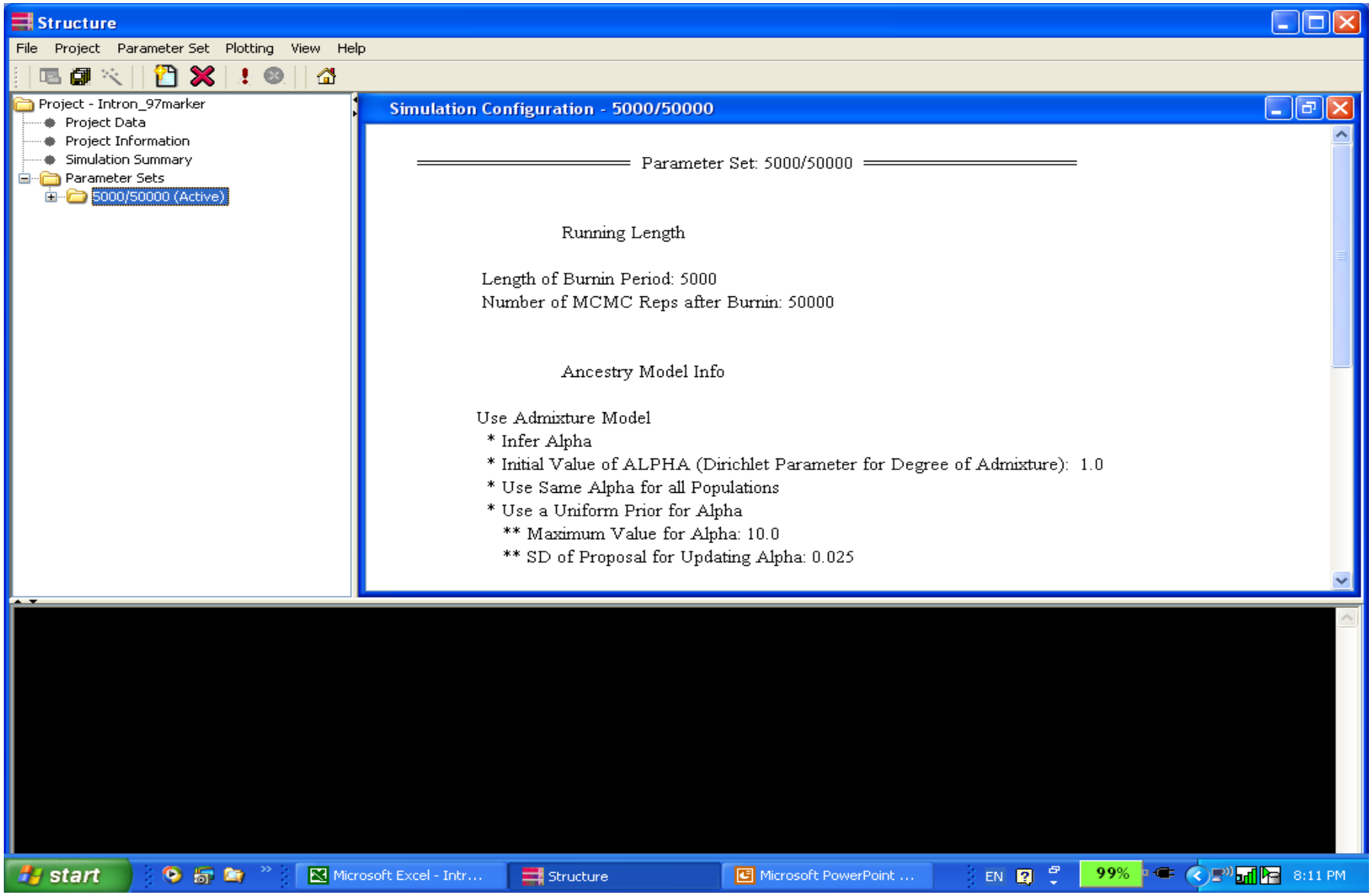
Configuring a parameter set (cont.)



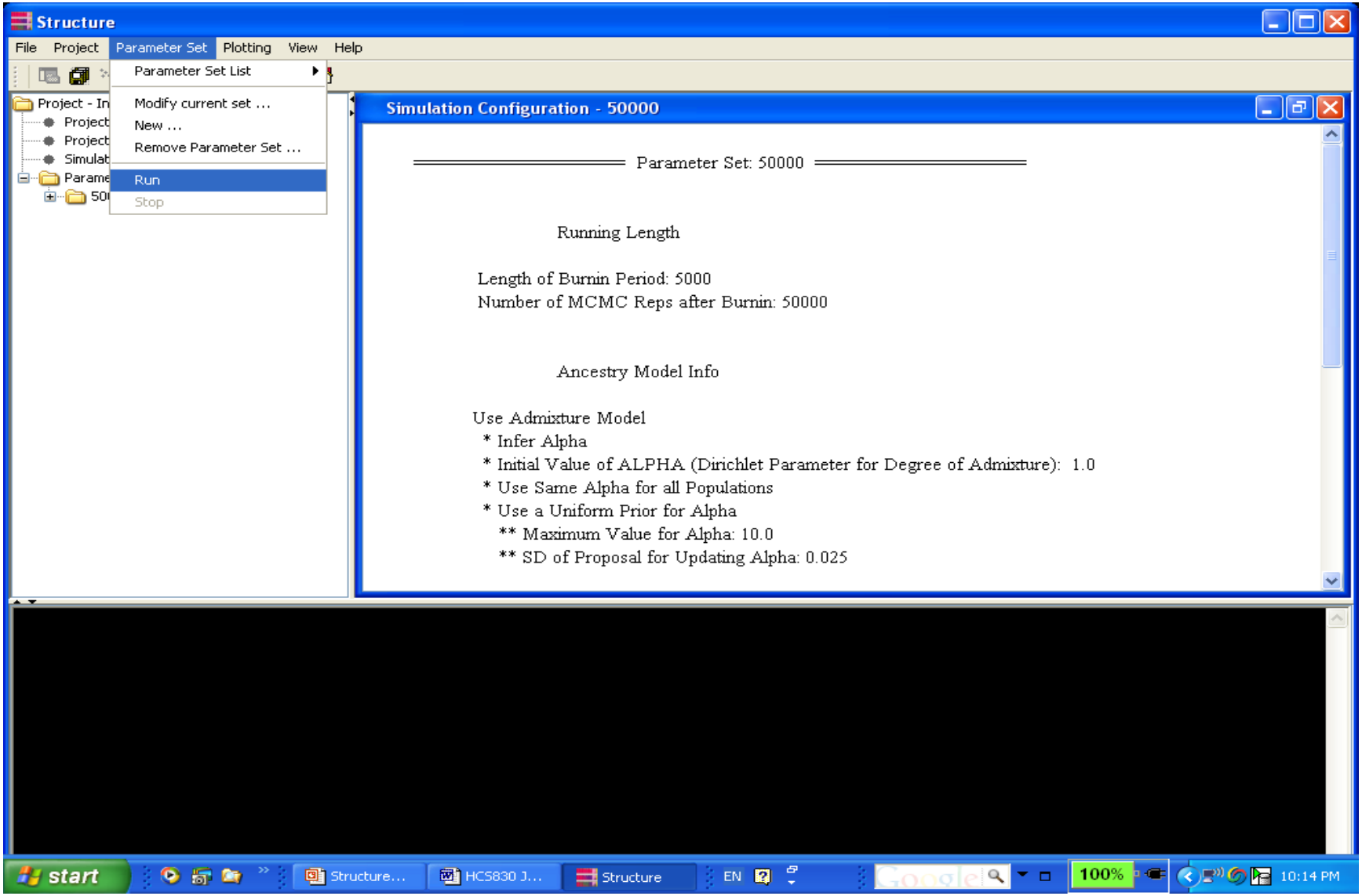
Configuring a parameter set (cont.)



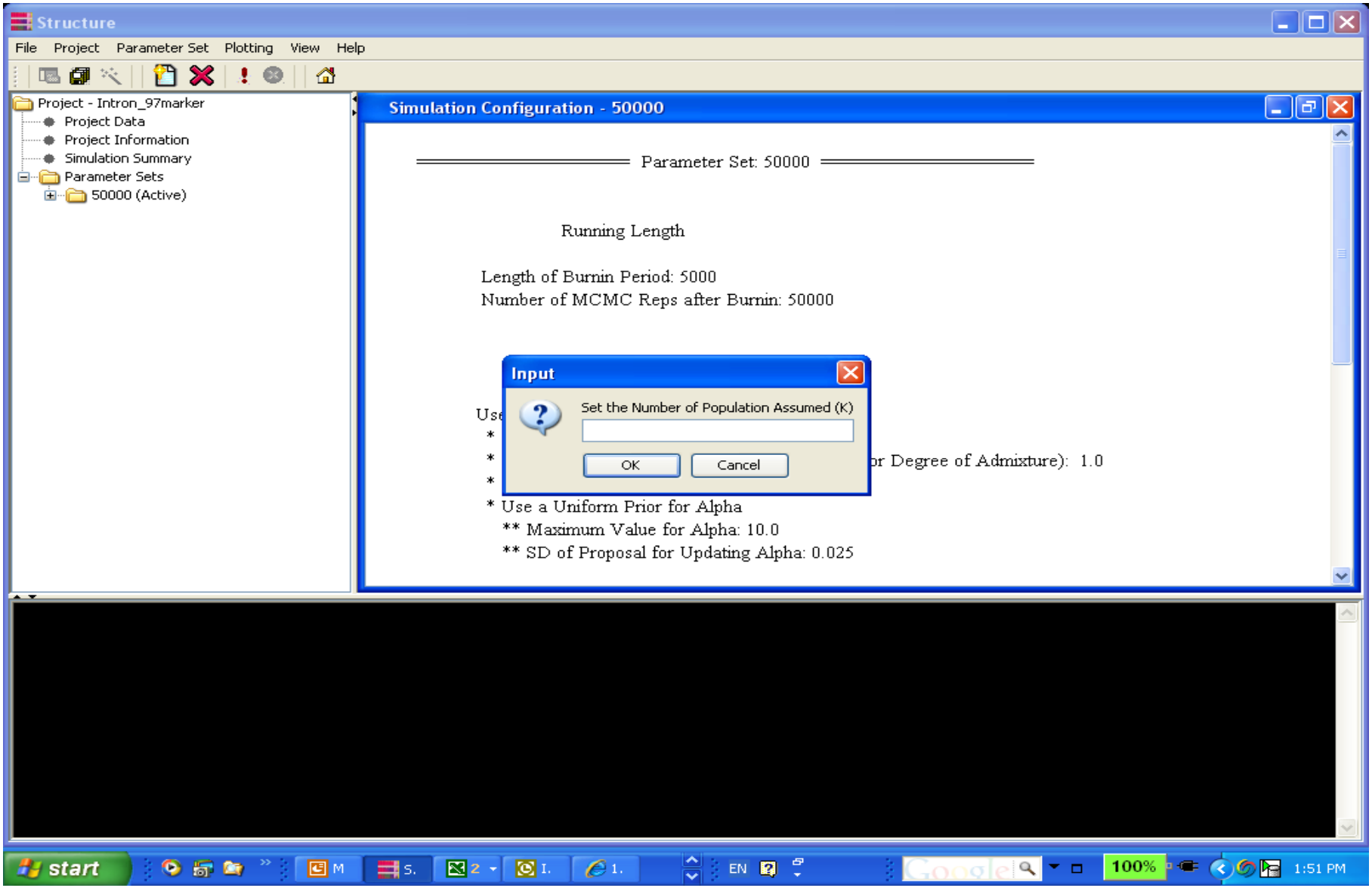
Configuring a parameter set (cont.)



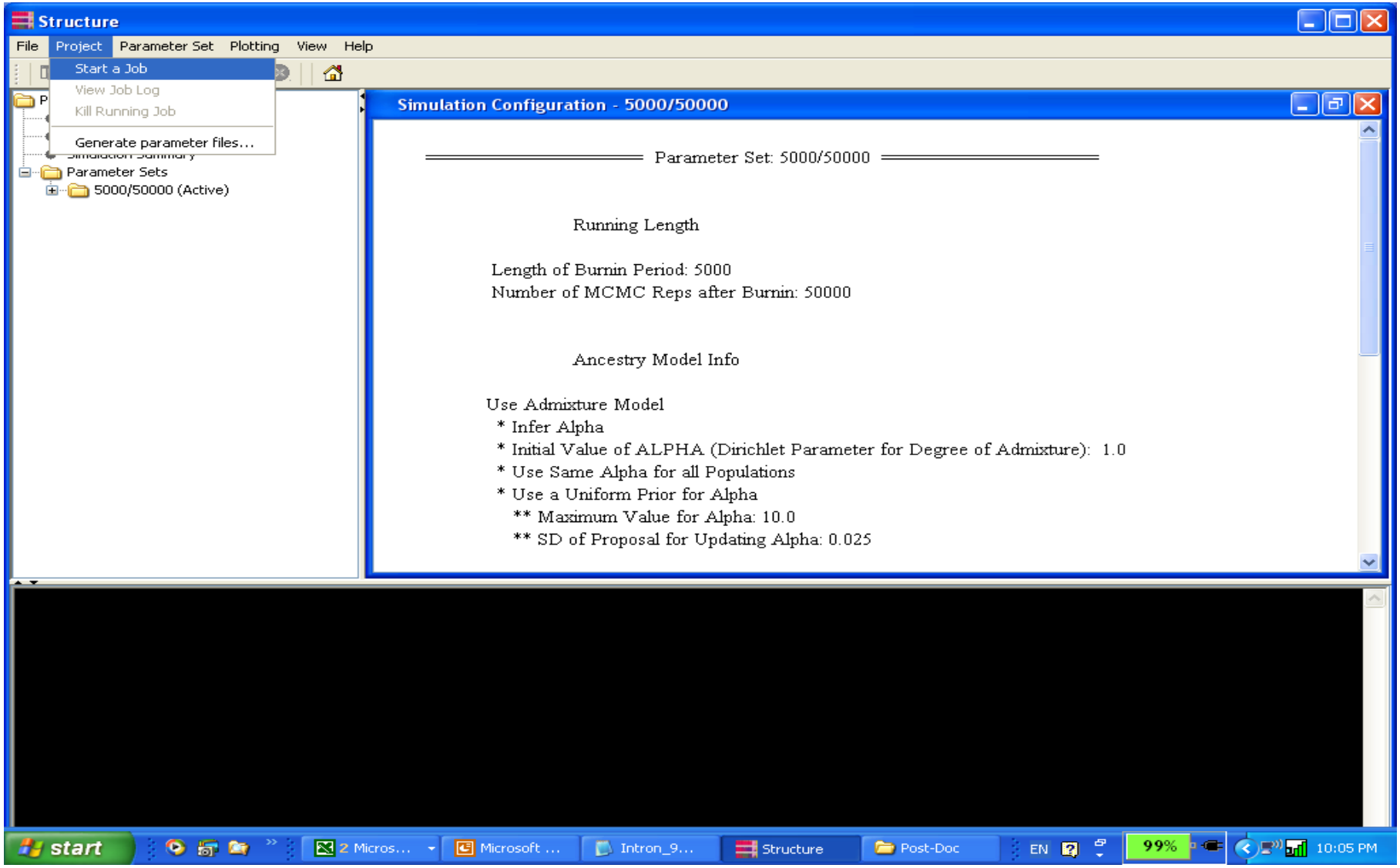
Running STRUCTURE: a single run



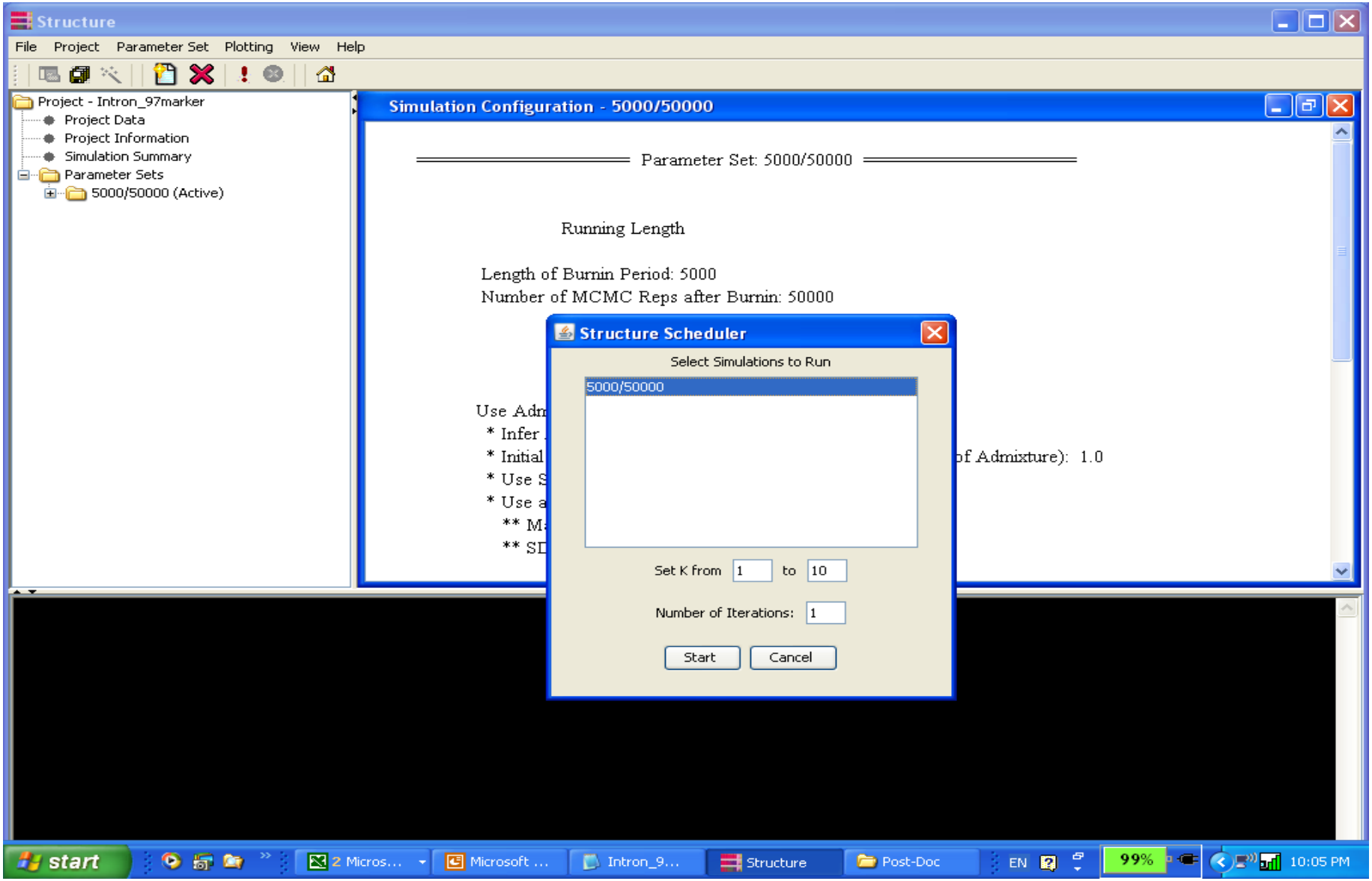
Running STRUCTURE: a single run (cont.)

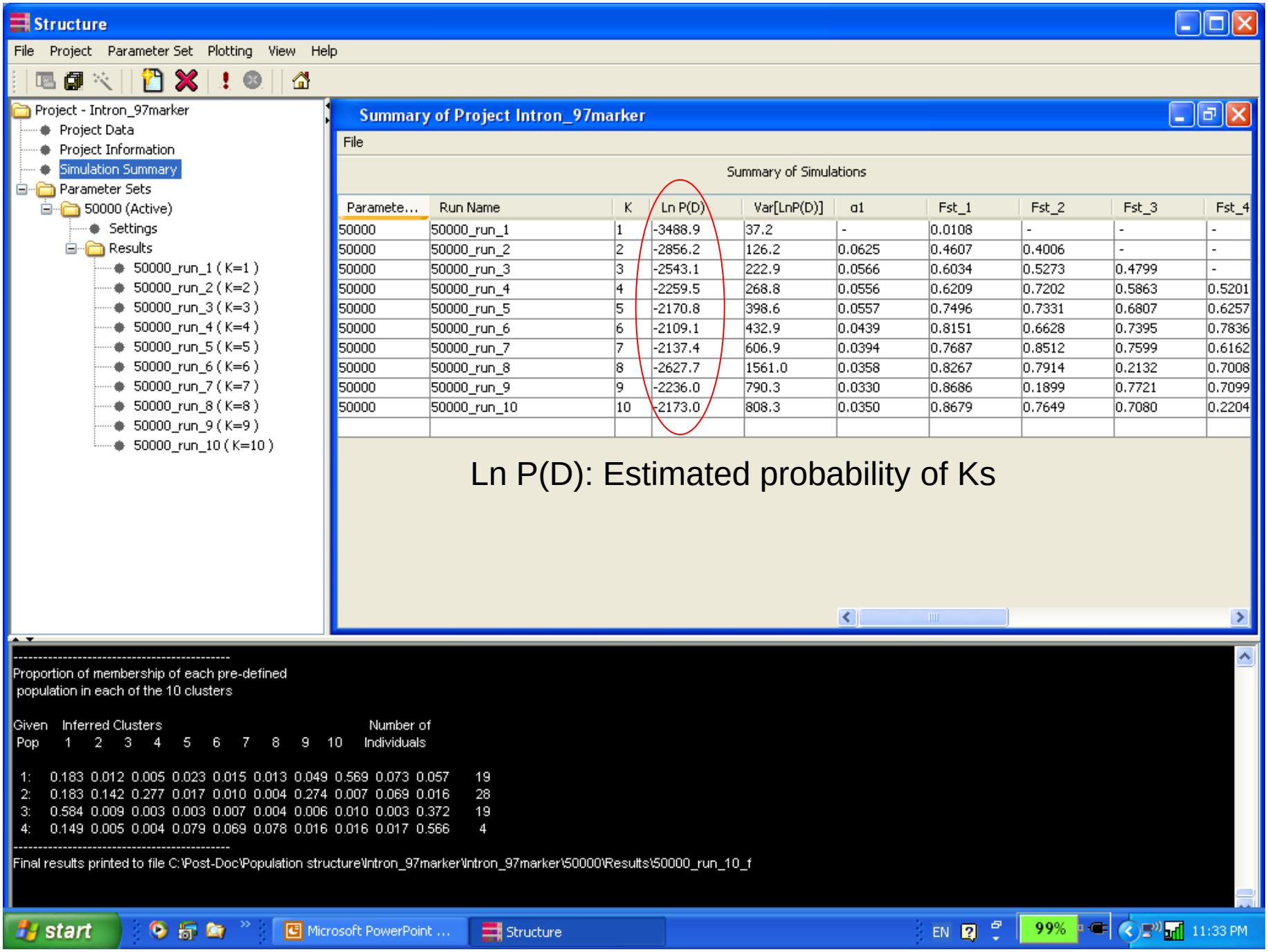


Running STRUCTURE: a batch run



Running STRUCTURE: a batch run (cont.)




Structure
File Project Parameter Set Plotting View Help
Project - Intron_97marker
Project Data
Project Information
Simulation Summary
Parameter Sets
50000 (Active)
Settings
Results
50000_run_1 (K=1)
50000_run_2 (K=2)
50000_run_3 (K=3)
50000_run_4 (K=4)
50000_run_5 (K=5)
50000_run_6 (K=6)
50000_run_7 (K=7)
50000_run_8 (K=8)
50000_run_9 (K=9)
50000_run_10 (K=10)
Summary of Project Intron_97marker
File
Summary of Simulations
Paramete... Run Name K Ln P(D) Var[LnP(D)] a1 Fst_1 Fst_2 Fst_3 Fst_4
50000 50000_run_1 1 -3488.9 37.2 - 0.0108 - - -
50000 50000_run_2 2 -2856.2 126.2 0.0625 0.4607 0.4006 - -
50000 50000_run_3 3 -2543.1 222.9 0.0566 0.6034 0.5273 0.4799 -
50000 50000_run_4 4 -2259.5 268.8 0.0556 0.6209 0.7202 0.5863 0.5201
50000 50000_run_5 5 -2170.8 398.6 0.0557 0.7496 0.7331 0.6807 0.6257
50000 50000_run_6 6 -2109.1 432.9 0.0439 0.8151 0.6628 0.7395 0.7836
50000 50000_run_7 7 -2137.4 606.9 0.0394 0.7687 0.8512 0.7599 0.6162
50000 50000_run_8 8 -2627.7 1561.0 0.0358 0.8267 0.7914 0.2132 0.7008
50000 50000_run_9 9 -2236.0 790.3 0.0330 0.8686 0.1899 0.7721 0.7099
50000 50000_run_10 10 -2173.0 808.3 0.0350 0.8679 0.7649 0.7080 0.2204
Ln P(D): Estimated probability of Ks
Proportion of membership of each pre-defined
population in each of the 10 clusters
Given Inferred Clusters Number of
Pop 1 2 3 4 5 6 7 8 9 10 Individuals
1: 0.183 0.012 0.005 0.023 0.015 0.013 0.049 0.569 0.073 0.057 19
2: 0.183 0.142 0.277 0.017 0.010 0.004 0.274 0.007 0.069 0.016 28
3: 0.584 0.009 0.003 0.003 0.007 0.004 0.006 0.010 0.003 0.372 19
4: 0.149 0.005 0.004 0.079 0.069 0.078 0.016 0.016 0.017 0.566 4
Final results printed to file C:\Post-Doc\Population structure\Intron_97marker\Intron_97marker\50000\Results\50000_run_10_f
start
Microsoft PowerPoint ...
Structure
EN
99%
11:33 PM

Inference of true K (number of populations)

- The log likelihood for each K, Ln P(D) = L(K)

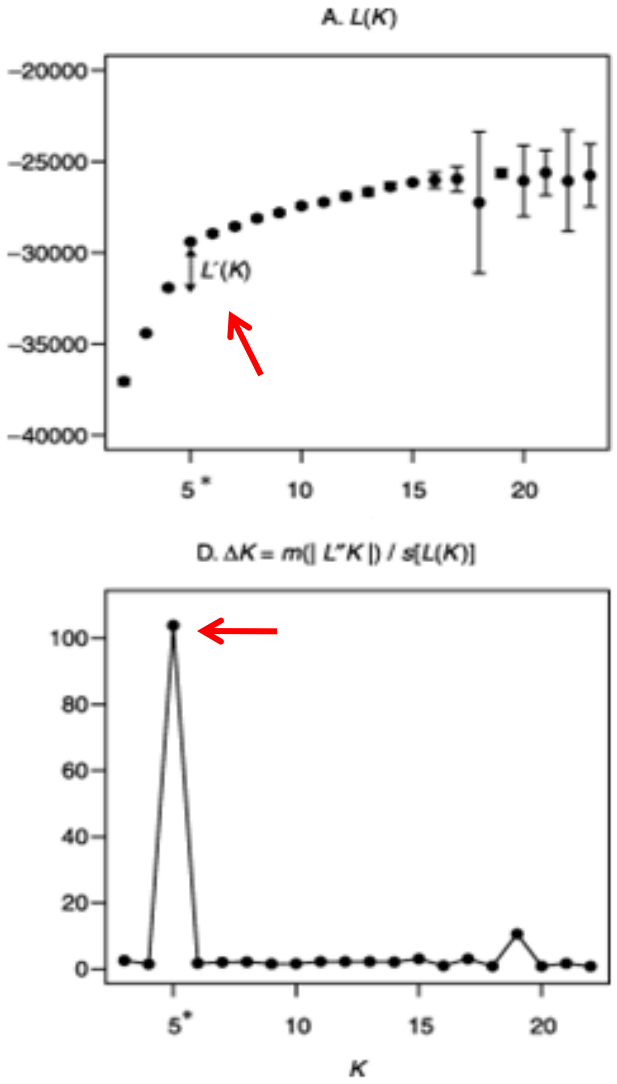
- Two approaches to determine the best K

1. Use of L(K): When K is approaching a true value, L(K) plateaus (or continues increasing slightly) and has high variance between runs (Rosenberg et al. 2001).

⇨ ***Nonparametric test (Wilcoxon test)***

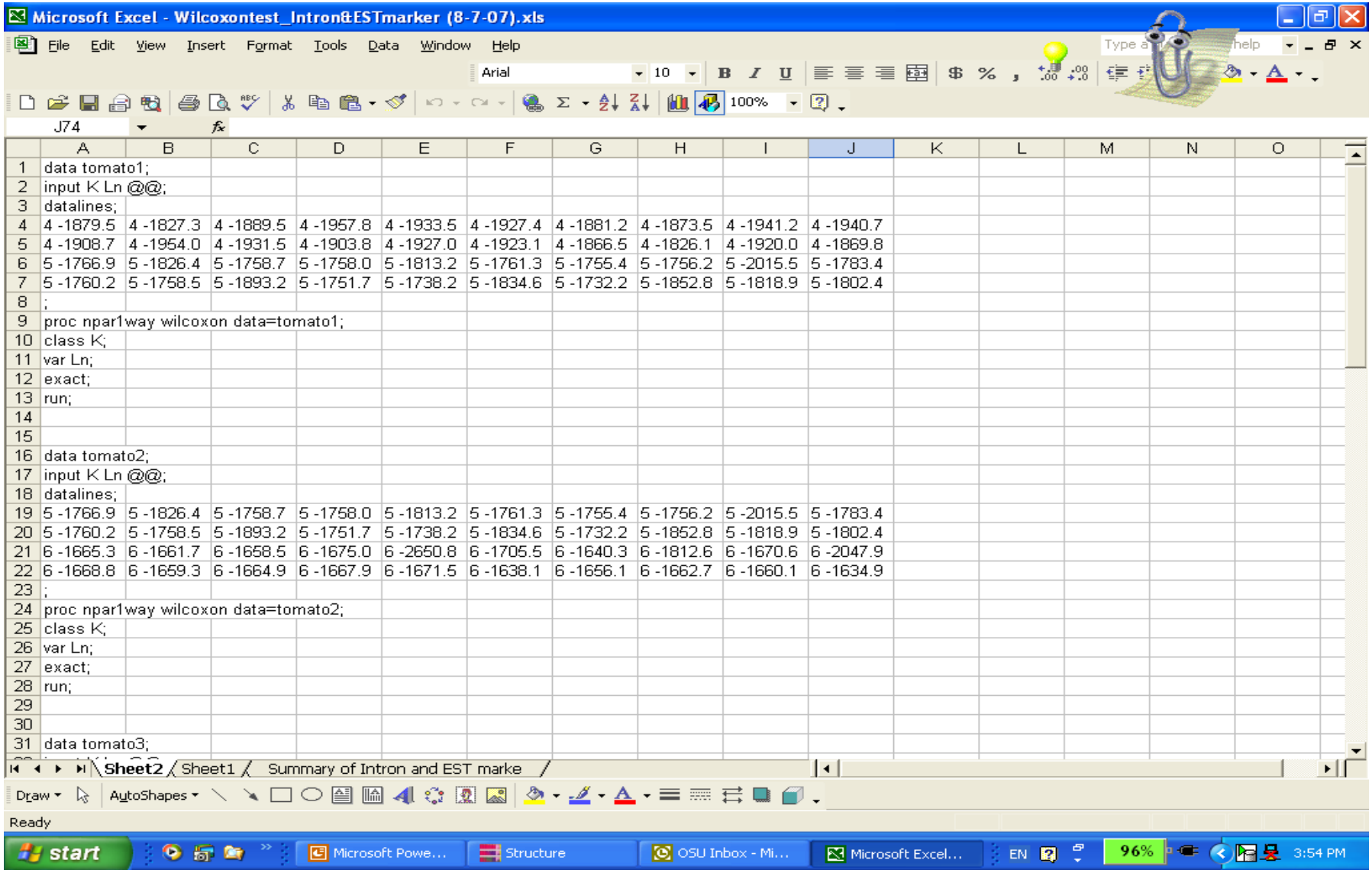
2. Use of an ad hoc quantity (ΔK): Calculated based on the second order rate of change of the likelihood (ΔK) (Evanno et al. 2005). The ΔK shows a clear peak at the true value of K.

⇨ ***ΔK = m([L"K])/s[L(K)]***



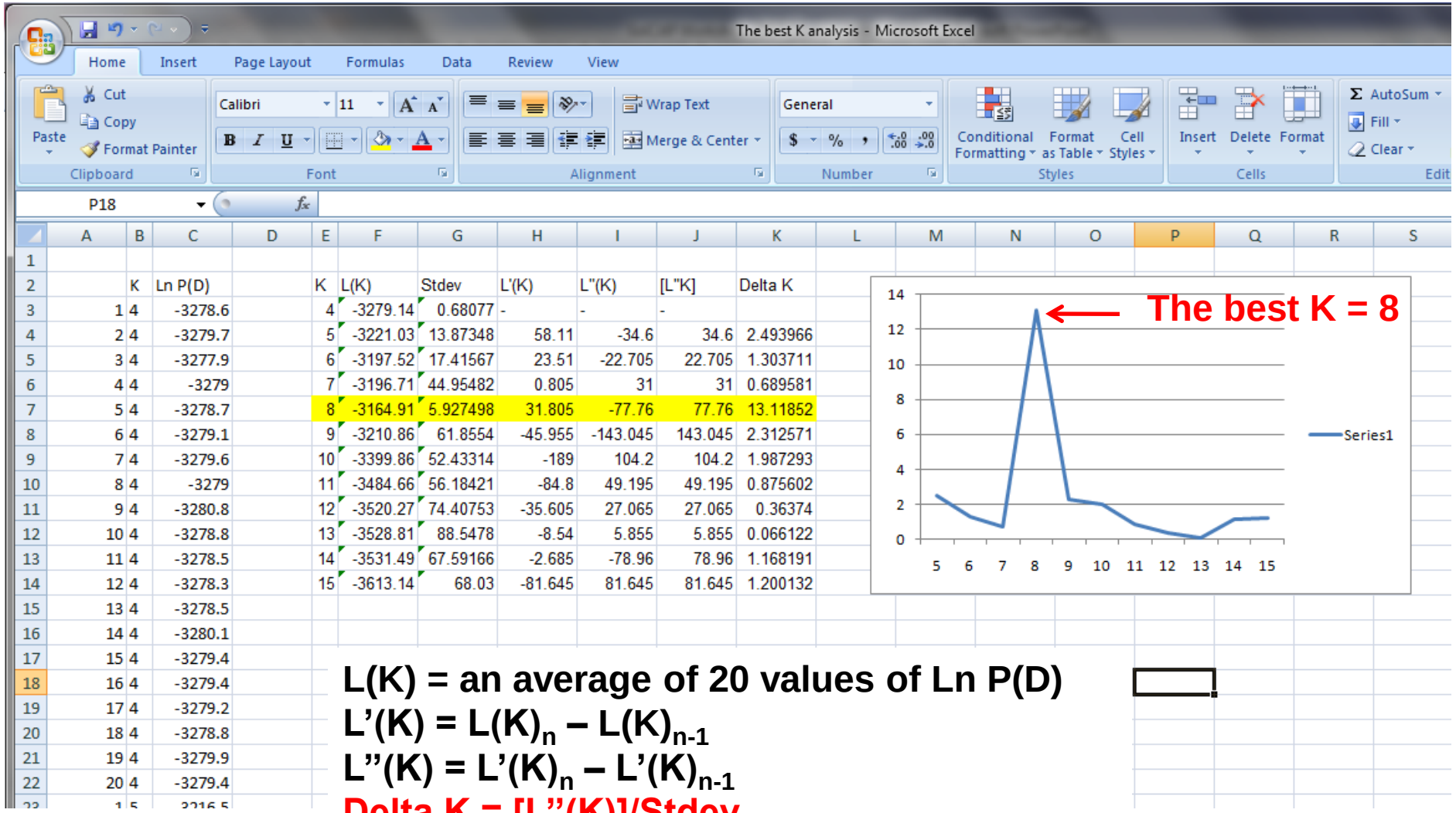
Evanno et al. 2005. Molecular Ecology 14: 2611-2620

SAS code for the nonparametric method



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	data tomato1;														
2	input K Ln @@;														
3	datalines;														
4	4 -1879.5	4 -1827.3	4 -1889.5	4 -1957.8	4 -1933.5	4 -1927.4	4 -1881.2	4 -1873.5	4 -1941.2	4 -1940.7					
5	4 -1908.7	4 -1954.0	4 -1931.5	4 -1903.8	4 -1927.0	4 -1923.1	4 -1866.5	4 -1826.1	4 -1920.0	4 -1869.8					
6	5 -1766.9	5 -1826.4	5 -1758.7	5 -1758.0	5 -1813.2	5 -1761.3	5 -1755.4	5 -1756.2	5 -2015.5	5 -1783.4					
7	5 -1760.2	5 -1758.5	5 -1893.2	5 -1751.7	5 -1738.2	5 -1834.6	5 -1732.2	5 -1852.8	5 -1818.9	5 -1802.4					
8	;														
9	proc npar1way wilcoxon data=tomato1;														
10	class K;														
11	var Ln;														
12	exact;														
13	run;														
14															
15															
16	data tomato2;														
17	input K Ln @@;														
18	datalines;														
19	5 -1766.9	5 -1826.4	5 -1758.7	5 -1758.0	5 -1813.2	5 -1761.3	5 -1755.4	5 -1756.2	5 -2015.5	5 -1783.4					
20	5 -1760.2	5 -1758.5	5 -1893.2	5 -1751.7	5 -1738.2	5 -1834.6	5 -1732.2	5 -1852.8	5 -1818.9	5 -1802.4					
21	6 -1665.3	6 -1661.7	6 -1658.5	6 -1675.0	6 -2650.8	6 -1705.5	6 -1640.3	6 -1812.6	6 -1670.6	6 -2047.9					
22	6 -1668.8	6 -1659.3	6 -1664.9	6 -1667.9	6 -1671.5	6 -1638.1	6 -1656.1	6 -1662.7	6 -1660.1	6 -1634.9					
23	;														
24	proc npar1way wilcoxon data=tomato2;														
25	class K;														
26	var Ln;														
27	exact;														
28	run;														
29															
30															
31	data tomato3;														

Inference of best K using the delta K method



	K	Ln P(D)	K	L(K)	Stdev	L'(K)	L"(K)	[L"K]	Delta K
1	4	-3278.6	4	-3279.14	0.68077	-	-	-	
2	4	-3279.7	5	-3221.03	13.87348	58.11	-34.6	34.6	2.493966
3	4	-3277.9	6	-3197.52	17.41567	23.51	-22.705	22.705	1.303711
4	4	-3279	7	-3196.71	44.95482	0.805	31	31	0.689581
5	4	-3278.7	8	-3164.91	5.927498	31.805	-77.76	77.76	13.11852
6	4	-3279.1	9	-3210.86	61.8554	-45.955	-143.045	143.045	2.312571
7	4	-3279.6	10	-3399.86	52.43314	-189	104.2	104.2	1.987293
8	4	-3279	11	-3484.66	56.18421	-84.8	49.195	49.195	0.875602
9	4	-3280.8	12	-3520.27	74.40753	-35.605	27.065	27.065	0.36374
10	4	-3278.8	13	-3528.81	88.5478	-8.54	5.855	5.855	0.066122
11	4	-3278.5	14	-3531.49	67.59166	-2.685	-78.96	78.96	1.168191
12	4	-3278.3	15	-3613.14	68.03	-81.645	81.645	81.645	1.200132
13	4	-3278.5							
14	4	-3280.1							
15	4	-3279.4							
16	4	-3279.4							
17	4	-3279.2							
18	4	-3278.8							
19	4	-3279.9							
20	4	-3279.4							

L(K) = an average of 20 values of Ln P(D)

L'(K) = $L(K)_n - L(K)_{n-1}$

L''(K) = $L'(K)_n - L'(K)_{n-1}$

Delta K = [L''(K)]/Stdev

Structure

File Project Parameter Set Plotting View Help

500000_run_76 (K=7)
500000_run_77 (K=7)
500000_run_78 (K=7)
500000_run_79 (K=7)
500000_run_7 (K=4)
500000_run_80 (K=7)
500000_run_81 (K=8)
500000_run_82 (K=8)
500000_run_83 (K=8)
500000_run_84 (K=8)
500000_run_85 (K=8)
500000_run_86 (K=8)
500000_run_87 (K=8)
500000_run_88 (K=8)
500000_run_89 (K=8)
500000_run_8 (K=4)
500000_run_90 (K=8)
500000_run_91 (K=8)
500000_run_92 (K=8)
500000_run_93 (K=8)
500000_run_94 (K=8)
500000_run_95 (K=8)
500000_run_96 (K=8)
500000_run_97 (K=8)
500000_run_98 (K=8)
500000_run_99 (K=8)
500000_run_9 (K=4)

Simulation Result: 500000(500000_run_85)

Bar plot Data plot Histogram Triangle plot Tree plot

Inferred ancestry of individuals:

Label (%Miss) : Inferred clusters

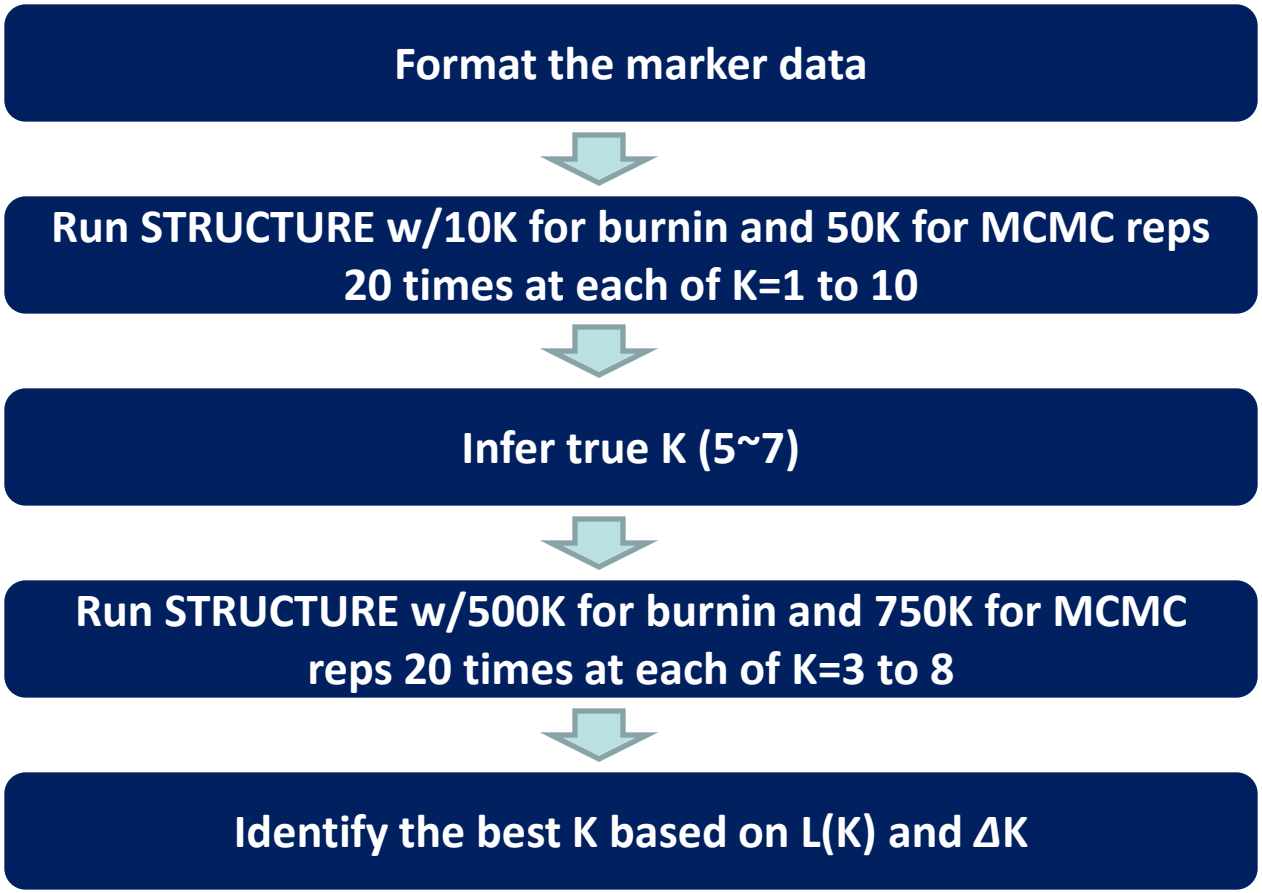
1	6111R1	(2)	:	0.035	0.010	0.012	0.143	0.524	0.252	0.013	0.011
2	6111R2	(0)	:	0.033	0.009	0.018	0.133	0.245	0.536	0.010	0.015
3	6111R3	(0)	:	0.008	0.005	0.015	0.124	0.150	0.674	0.012	0.011
4	6111S1	(2)	:	0.018	0.006	0.026	0.116	0.150	0.669	0.008	0.008
5	6111S2	(0)	:	0.145	0.012	0.042	0.126	0.234	0.384	0.014	0.043
6	6115S3	(2)	:	0.017	0.006	0.317	0.051	0.026	0.055	0.007	0.520
7	6115S4	(2)	:	0.013	0.006	0.023	0.019	0.018	0.014	0.007	0.899
8	6117R1	(0)	:	0.014	0.004	0.012	0.670	0.041	0.238	0.009	0.011
9	6117R2	(0)	:	0.012	0.007	0.272	0.285	0.030	0.124	0.019	0.251
10	6117S1	(0)	:	0.015	0.006	0.039	0.302	0.053	0.087	0.009	0.489
11	6117S2	(0)	:	0.036	0.009	0.150	0.265	0.207	0.088	0.008	0.238
12	6117S3	(0)	:	0.007	0.013	0.010	0.027	0.023	0.011	0.009	0.899
13	6117S4	(0)	:	0.015	0.007	0.085	0.376	0.061	0.095	0.008	0.353
14	6124R1	(0)	:	0.074	0.007	0.067	0.028	0.061	0.013	0.006	0.744
15	6124R2	(0)	:	0.022	0.010	0.014	0.110	0.426	0.012	0.005	0.401
16	6124R3	(0)	:	0.014	0.008	0.023	0.013	0.017	0.009	0.004	0.911
17	6124R4	(0)	:	0.021	0.016	0.014	0.030	0.079	0.014	0.008	0.818

← **Q-matrix**

Google

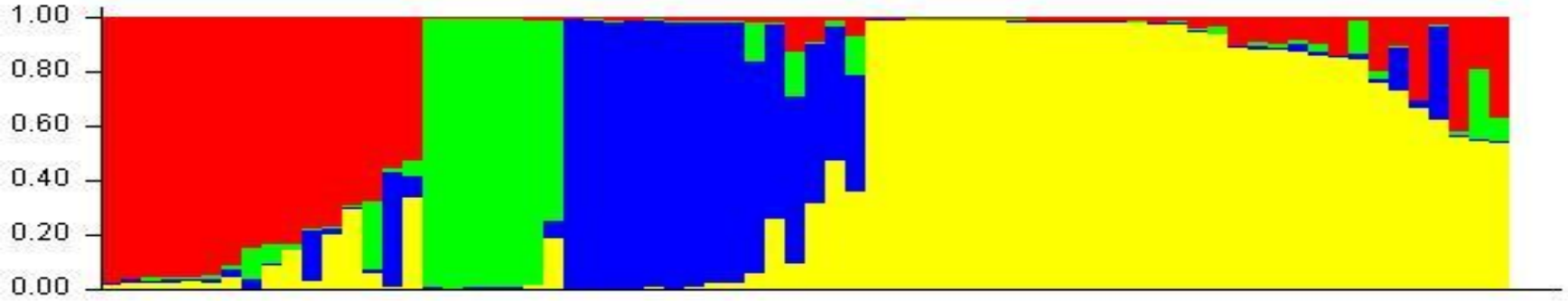
EN

An example of steps to identify the best K



We may not always be able to know the TRUE value of K, but we should aim for the smallest value of K that captures the major structure in the data

Pritchard et al. (2000)



Enjoy running STRUCTURE