Supplementary Figure 1:

```
baboon          ARATATGTGCACACTGATACACAGCAAACRTACACAGAGATCAGCACACACAAAGAGCCC 60
human           .G..........A.......................T....T.............T.

baboon          ACACCCACATGCACACACCCCTCAGGTGGGACGGATTCGACCACCACCACCTTTCCCCCA 120
human           ..G.....G..............T......A..G.T................T....

baboon          AACACATGGCTCTTGAACTGCCTTTCCTTGGATCAAGTTCAAGGGGATGGAGGAGCAGTG 180
human           ..........T.....................C.....................

baboon          AGAGTCAGCCGCCTTTCCACTCCAATTTCCCAGCACCTCCCTTATCTCTGCTTCACAAGT 240
human           ............C.................................C........

                                                 #
baboon          CACCCAGCCCCCCTCTTTTCCTTCCTTGTGCTTGRAGAGTCTCT---------------- 300
human           ..............C.....................A.....CCTTGCTGGAAAGCCC

baboon          ----------------------------------------ACTTGGTGGAAAACCCCCT 360
human           CCTGTTTTCTCAATCTCCCTTTCCACTTCGGTAAAATCTCT.....C......G......

baboon          GGTTTCTCAATCTCCTTTTCCACTTCGGTRAAATGCCCACTTTCTGGTCTCCACCTTTTT 420
human           .T............C.................................C..........

                                       *
baboon          CCTGCGTGCAGTCCCAACCAGTCAAATCYAACCTCAAAACAGGAAGCCCCGAGGCCAGTG 480
human           ....A...T............C.......................A...A.........

baboon          CCCCCCATAGACCTGAGGCTTGTGCAGGCAGTGGGCGTGGGGTGAGGCTTCCTGATGCCC 540
human           A.........G.............................A................

                                                          GATA1
baboon          CCTGTCCCTGCCCGGAACCTGATGGCCCTCATTAGTCCTTGGCTCTTATCTTGGAAGCAC 600
human           ............A..............................................

                                                    ↱
baboon          AGGCGCTGACAGCCGTCCCAGCCCTTCTGTCTG------TGAACCAAACGGTGCCATGGG 660
human           ..............................CGGGCC.....................

baboon          GAACTGTCTGCACCCGGTGAGTATGGGGCCAGGCCCCACAGTCCT 705
human           ............AG.....................G.....C
```

Figure S1. Alignment of baboon and human *FY cis*-regulatory sequences. The five variable sites identified in the Amboseli population are highlighted in gray. The *cis*-regulatory site associated with *Hepatocystis* infection is indicated with a hash, and the *cis*-regulatory site associated with allelic imbalance is indicated with an asterisk. The GATA1 binding motif implicated in *Plasmodium vivax* resistance in humans is boldfaced and labeled in italics, and the site of the T/C malaria-associated SNP in humans is underlined. The bent arrow indicates the transcription start site identified in reference 4.

Supplementary Figure 2:

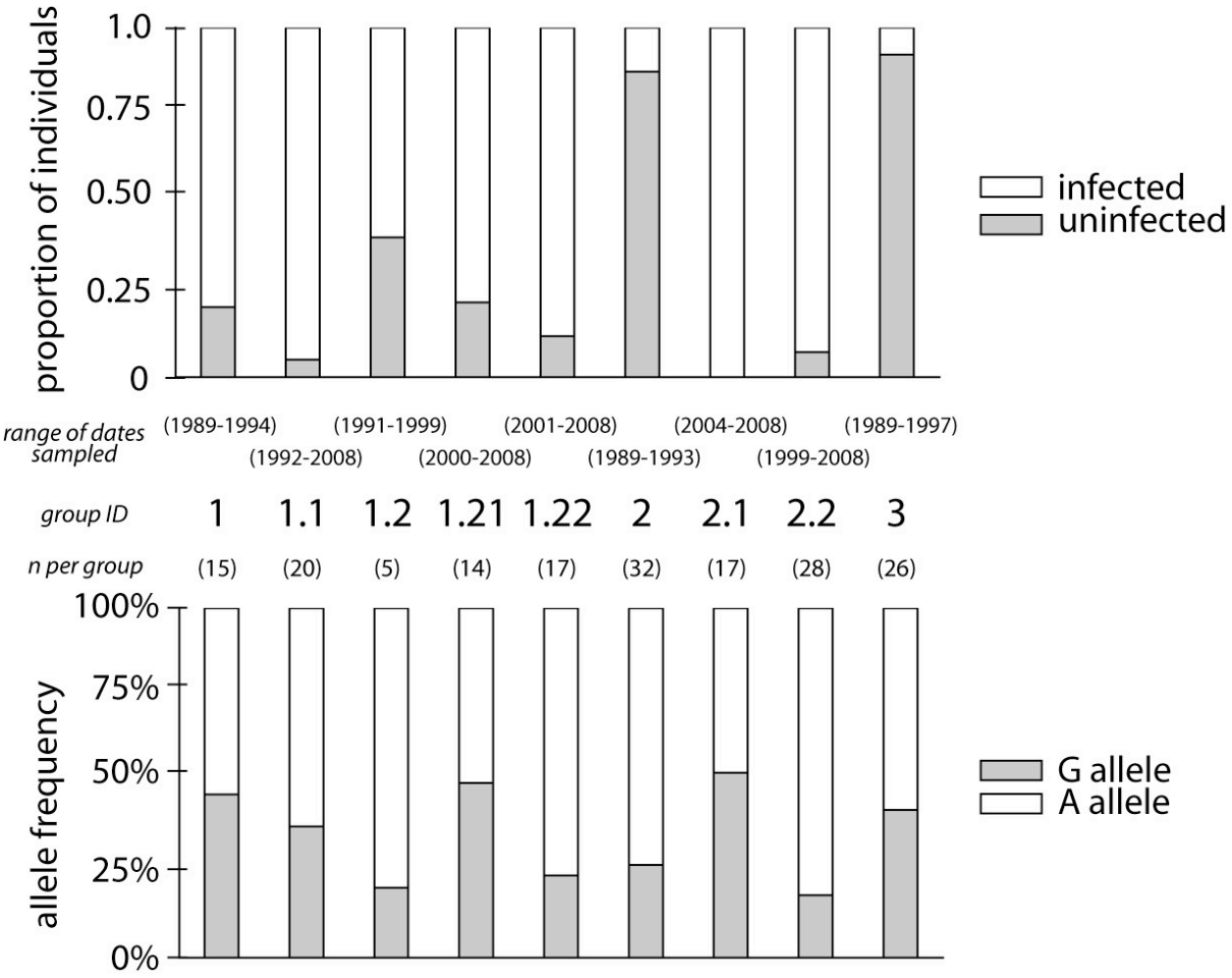

Figure S2. Differences by study group. The top bar graph shows differences in the proportion of individuals infected in each of the 9 study groups included in this study; the bottom bar graph shows the allele frequencies for the *Hepatocystis* associated *FY cis*-regulatory SNP. The x-axis label (study group) is given in the space between the two graphs, and is the same for both of them. Also given in parentheses are the range of years from which samples were obtained for each group (above the group ID), and the number of individuals sampled in each group (below the group ID).

Supplementary Figure 3

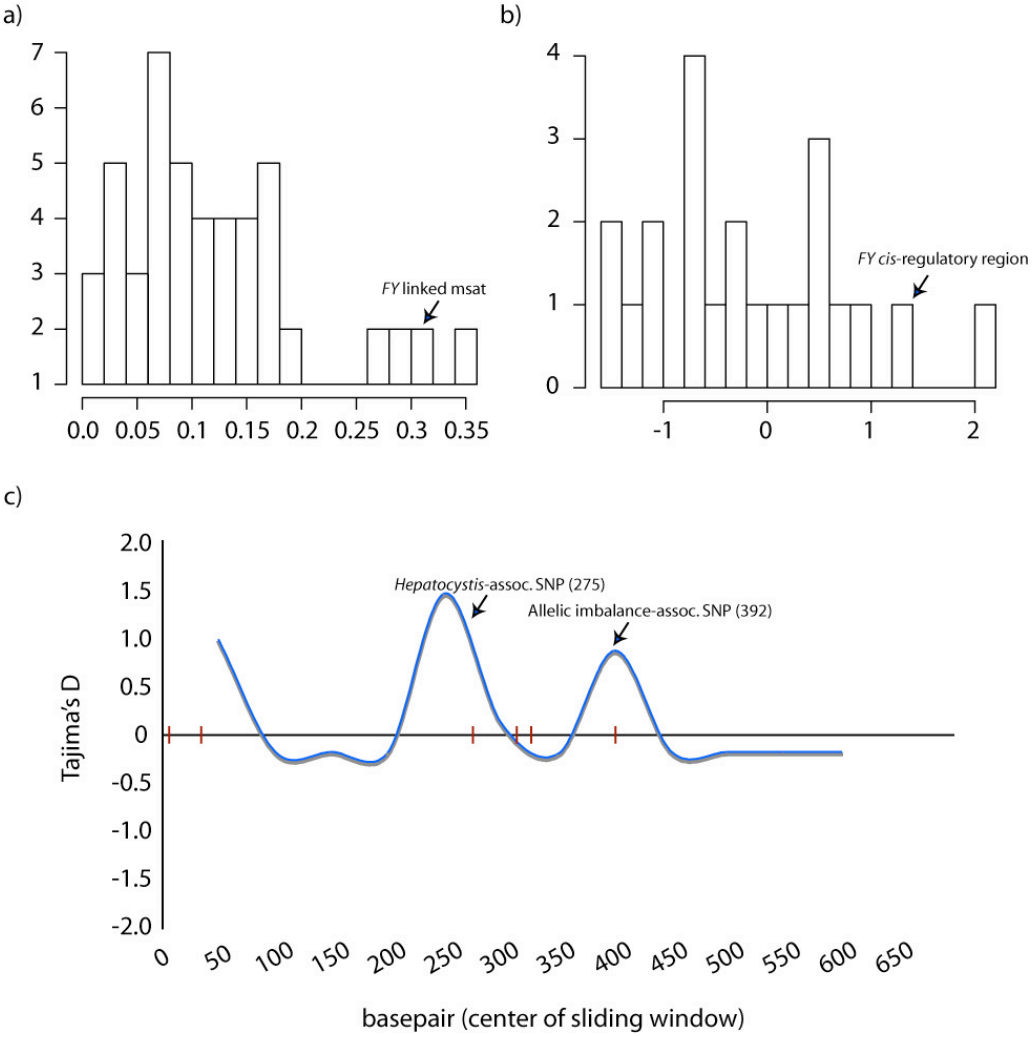

Figure S3. Comparison of genetic variation in and around the *FY cis*-regulatory region in relationship to other loci. Between population variation is shown in a) the distribution of F_{st} values between the Amboseli, Mikumi, and Masai Mara baboon populations for the *FY* linked microsatellite and 35 putatively neutral microsatellites around the genome. Within population variation in Amboseli is shown in b) the distribution of Tajima's D values within Amboseli from 22 loci in the baboon genome. Tajima's D was calculated using resequencing data for each locus and implemented using the DNAsp v. 4.9[27], assuming no recombination; and c) results of a sliding window analysis of Tajima's D for the *FY cis*-regulatory region within Amboseli, where window size = 100 bp and window interval = 50 bp. The locations of SNPs in the region are shown in red.

Supplementary Figure 4:

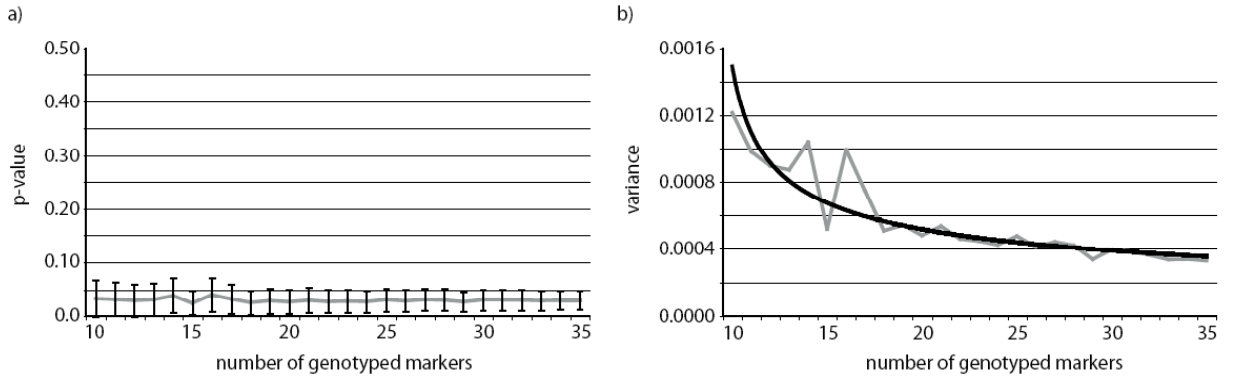

Figure S4. Stability and robustness of F_{st}-based inference. (a) The probability of observing the F_{st} value (or a more extreme value) for the *FY*-linked microsatellite marker under the null hypothesis of neutral evolution (see also the Supplementary Methods). Points on the gray line are given for different numbers of genotyped markers from 10 to 35, and represent the mean of the weighted averages of the *p*-value for 20 random subsamples. Error bars give the mean standard deviation of p over different subsamples. (b) The variance of the *p*-value decreases according to a power law distribution as the number of genotyped markers increases ($R^2 = 0.851$). The fit distribution is shown as a black line. Thus, confidence in the p-value also increases as the number of genotyped markers increases, but the added value of each additional marker decreases rapidly after a few dozen markers have been typed. Note however that the variance is quite small over the entire range of genotyped markers.

Supplementary Methods:

Robustness of the *Hepatocystis* screening assay

High rates of infection were found in groups sampled in 2004-8 as well as in groups sampled in the late 1980's and early 1990's, indicating that our ability to detect *Hepatocystis* infection did not markedly decrease with the age of the sample. However, to rule out the possibility that a failure to amplify *Hepatocystis* was due to poor quality DNA, we eliminated from subsequent analyses any individuals for whom we were not able to generate high quality genomic DNA sequence from other regions, including the *FY cis*-regulatory region.

We also used *Plasmodium*-genus specific primers[31] as a secondary confirmation of infection for 103 individuals (*Hepatocystis* is phylogenetically nested within the *Plasmodium* species that infect primates[15]). All individuals included in this study had concordant results with both the *Hepatocystis* mtDNA primers and the *Plasmodium* genus-specific primers.

Control for population structure in the association with *Hepatocystis*

When both phenotypic variation and genetic variation are structured and correlated with one another within a sample, false positive associations may result. We used two approaches to control for possible population structure in our sample. First, we included social group as a random effect when modeling *Hepatocystis* infection on genotype (infection rate was clearly structured by social group: see Figure S2, top). In baboons, sex-biased dispersal and philopatry predict that social group ('breeding group') will be the most important unit of population structure. We have also used this approach to take account of social group in previous studies of the Amboseli baboon population (e.g., ref 32). Second, we applied a principle components-based analysis to identify the major axes of population structure using genotype data from 47 unlinked loci from around the baboon genome. Estimates of population structure were obtained following the method of Price et al (2006)[33], using custom MATLAB code. Missing genotype data (9% of the overall data matrix) were imputed using two different methods: local least squares regression[34] and k nearest neighbors, with $k = 3$[35]; exploratory analyses with $k = 1 – 7$ produced very similar results. Results based on the k nearest neighbors approach are reported in the main text, but the results were qualitatively identical regardless of imputation method.

The *Hepatocystis* analysis was run using the *lmer* function in the R package *lme4*, version 0.99875-9[36]. In our final model, we incorporated projections from the first five eigenvectors obtained through PCA (these explained approximately 60% of the overall genetic variation in population).

Genotype matching in relatives

Because some of the individuals we sampled were related, we used simulations to test whether our result could be a spurious effect of elevated rates of genotype sharing. First-order ($r = 0.5$) relatives (as assessed by known pedigrees) were much more likely to share genotypes at the C/T SNP than random individuals, but this effect was considerably muted in second ($r = 0.25$) and third-order relatives ($r = 0.125$) and absent in unrelated individuals.

Specifically, we classified dyads of 166 individuals in the population as related at $r = 0.5$, $r = 0.25$, $r = 0.125$, or as unrelated, using maternal and paternal pedigree information available for

this population[10,11]. 1000 bootstrap resamplings of 406 dyads (the number of possible dyads for n = 29) of first-order (r = 0.5) relatives showed that these individuals shared the same genotype at the C/T SNP 34% (95% CI: 27.8% - 40.6%) more often than pairs of individuals drawn from the population at random. Dyads related at r < 0.5 were only about 10 – 14% more likely than random dyads to share genotypes, and unrelated dyads were about 1% less likely than random dyads to share genotypes (reflecting the outbred nature of the population and the accuracy of the pedigree data).

Assessing confidence in the F_{st} analysis

The F_{st} value for the *FY*-linked marker was significantly greater than expected under neutral evolution, based on comparing this marker to a set of 35 putatively neutral microsatellite markers around the baboon genome. Specifically, we modeled the distribution of F_{st} values for the 35 neutral markers as a gamma distribution, and asked about the likelihood of observing the value of F_{st} for the *FY*-linked marker or a more extreme value, given this model. The maximum likelihood parameter estimates for the gamma distribution yielded $p < 0.027$ as the probability of the observed F_{st} value for the FY-linked marker. However, because our inference was based on a modest number of markers, we formally tested the stability of our *p*-value estimate given uncertainty in the model (i.e., under other parameterizations of the gamma, including some parameter settings that might also be highly consistent with the data, but could potentially provide weaker support for the hypothesis of non-neutral evolution).

We therefore calculated the *p*-values for 10,000 other possible combinations of parameters for the gamma distribution and weighted these values by the likelihood of these parameter combinations, given the data. We sampled the values of the two parameters independently, in each case from a uniform distribution bounded by two standard deviations above or below the maximum likelihood estimate, where the standard deviations were based on the estimated marginal distribution for that parameter. This approach is equivalent to sampling from the posterior of a probability distribution, an approach that gives both the expectation of the true *p*-value (the mean of the posterior probability distribution), and the variance of the *p*-value across all the alternative parameterizations. We used random subsamplings of the data from n = 10 to n = 35 to examine how the variance decreases with increasing n (see Figure S3), and averaged the mean and Var(*p*) over multiple random subsamples of the same size.

The mean *p*-value from the posterior distribution on the parameters is very stable over the range of sample sizes we examined, and, as expected, the variance of *p* decreases with increasing n. Between n = 10 and n = 35 markers, the variance decreases almost an order of magnitude, from $Var_{n=10}(p) = 0.0012$ to $Var_{n=35}(p) = 0.00033$. In order to gauge whether increasing the sample size further would be useful, we fit a power law distribution to the decrease in Var(p) with increasing sample size ($R^2 = 0.851$). We found that increasing the sample size further is unlikely to have much of an effect on the mean or variance of p. For example, $Var_{n=50}(p)$ is expected to be around 0.00026, and $Var_{n=100}(p)$ around 0.0002.

All analyses were conducted using custom scripts in Ruby and R[7] by JT, and a modification of freely available code for estimating maximum likelihood parameters for gamma distributions[37].

Supplementary Tables:

IPBIR Repository #	Date of original sampling	Local Identification
BP00232	12 August 2004	York
BP00234	13 August 2004	Manda
BP00236	14 August 2004	Oscar
BP00237	16 August 2004	Stefano
BP00242	19 August 2004	Rocket
BP00243	21 August 2004	Leakey
BP00244	22 August 2004	Puck
BP00245	22 August 2004	Julius
BP00246	23 August 2004	Facko
BP00247	25 August 2004	Duke

Table S1: Masai Mara sample information. Ten *Papio anubis* samples from the Masai Mara National Reserve, Kenya, were obtained as extracted DNA from the Integrated Primate Biomaterial and Information Resource (IPBIR) courtesy of R. Sapolsky. IPBIR repository numbers for these samples, date of original sampling, and local identification for these individuals are provided here.

IPBIR Repository #	Local Identification
PR00883	2007
PR00884	2014
PR00888	3005
PR00889	3115
PR00890	3116
PR00893	3123
PR00895	3126
PR00897	3129
PR00898	3130
PR00899	3133
PR00901	4005
PR00902	5001
PR00903	5003
PR00908	5020
PR00911	5026
PR00916	IK01
PR00917	IK02
PR00922	KZ01
PR00924	KZ05
PR00925	KZ06

Table S2: Mikumi sample information. 20 *Papio cynocephalus* samples from Mikumi National Park, Tanzania, were obtained as extracted DNA from the Integrated Primate Biomaterial and Information Resource (IPBIR), courtesy of J. Rogers. IPBIR repository numbers for these samples and local identification for these individuals are provided here.

Primer names	Sequence (5' – 3')	Region amplified
FYprF1	TCATTATGCAGCCTCGACAG	5' *FY cis*-
FYprR1	GGGCATAGGGGTAAAGGACT	regulatory region
FYGF1	CCCCCTGCCACTCCTGTAA	*FY* transcribed
FYGR1	CCAAGGGTGTCCAGATGAGAA	region
HepatF3	CATTTACACGGTAGCACTAATCCTT	*Hepatocystis*
HepatR3	GGAATGGTTTTCAACATTGCAT	mtDNA
FYpyroF	CTTCAGGGAGGGGCAGAG	*FY*
FYpyroR	TTTTGCACTGTGTGGCTACG	pyrosequencing
FYpyroseq	GCCTGGTGGCAGAATA	assay 1
FYpyroF1	TCCTCTTCATGTTTTTCAGACC	*FY*
FYpyroR1	GCACCACAATGCTGAAGAGA	pyrosequencing
FYpyroseq1	CATGTTTTTCAGACCTCTC	assay 2

Table S3. Primers used in this study. Sequence details are provided for all primers designed by the authors. Pyrosequencing assays are given as a set of three primers: the forward and reverse primers for PCR, and an internal sequencing primer.

Locus	N	ungapped length	S	pi	theta/site	D
FY	344	636	6	0.00245	0.00147	1.260
CCL5	318	648	4	0.00036	0.00097	-1.046
CCR5	320	633	8	0.00130	0.00199	-0.722
CD58	304	161	2	0.00088	0.00197	-0.713
CD59	314	623	7	0.00217	0.00178	0.450
CXCR4	202	583	4	0.00070	0.00117	-0.709
CYP1A1	330	580	6	0.00166	0.00162	0.042
CYP1B1	302	682	15	0.00131	0.00350	-1.527
ESR1	242	415	1	0.00032	0.00040	-0.201
IFNGR1	332	488	5	0.00115	0.00161	-0.512
IL1A	342	514	6	0.00123	0.00182	-0.614
IL4R	332	461	5	0.00257	0.00170	0.924
IL6	286	602	6	0.00070	0.00160	-1.093
IL10	162	716	3	0.00015	0.00074	-1.312
IL12B1	322	589	6	0.00204	0.00160	0.515
IL19	338	272	5	0.00259	0.00287	-0.177
LTA	242	550	3	0.00070	0.00090	-0.339
MEFV	326	524	11	0.00381	0.00330	0.350
MPO	330	376	4	0.00006	0.00167	-1.600
MSR1	336	281	8	0.00901	0.00445	2.122
PHF11	332	318	4	0.00281	0.00197	0.711
TAP2	204	584	14	0.00478	0.00407	0.437

Table S4. Loci used for comparisons of Tajima's D. Short segments were resequenced within the transcribed region of each locus or 5' of the transcription start site in the putative *cis*-regulatory region for each locus given above. Haplotype was inferred using the program PHASE v. 2.1.1.[29]. In all cases, the individuals that were resequenced were included in the main set of individuals sequenced at the *FY cis*-regulatory region. N gives the number of alleles (2 per individual) in the

final sample, S is the number of segregating sites identified in each region, pi gives the mean pairwise distance between alleles, theta is Watterson's estimate of theta for nucleotide diversity, and D is the estimated value of Tajima's D for each locus.

Additional Works Cited:

31. Rougemont, M. *et al*. Detection of four *Plasmodium* species in blood from humans by 18S rRNA gene subunit-based and species-specific real-time PCR assays. J. Clin. Microbiol. **42**, 5636-5643 (2004).

32. Charpentier, M. J. E., Tung, J., Altmann, J. & Alberts, S. C. Age at maturity in wild baboons: genetic, environmental, and demographic influences. Mol. Ecol. **17**, 2026-2040 (2008).

33. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. **38**, 904-909 (2006).

34. Kim, H., Golub, G. H. & Park, H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. Bioinformatics **22**, 1410-1411 (2006).

35. Troyanskaya, O. et al. Missing value estimation methods for DNA microarrays. Bioinformatics **17**, 520-525 (2001).

36. R Development Core Team. R: A language and environment for statistical computing. (R Foundation for Statistical Computing, Vienna, Austria, 2007).

37. Wessa, P. Free statistics software. (Office for Research Development and Education, 2008).