	M	E	N		2	7	6	6	B	Dispatch: 24.9.09	Journal: MEN	CE: Anusha
	Journal Name			Manuscript No.						Author Received:	No. of pages: 4	PE: Gomathi

Molecular Ecology Resources (2009) doi: 10.1111/j.1755-0998.2009.02766.x

COMPUTER PROGRAM NOTE

2BAD: an application to estimate the parental contributions during two independent admixture events

J. BRAY,*1 V. C. SOUSA,†‡1 B. PARREIRA,† M. W. BRUFORD* and L. CHIKHI†§

*Cardiff School of Biosciences, Cardiff University, P.O. Box 915, Cardiff CF10 3TL, UK, †Instituto Gulbenkian de Ciência, Rua da Quinta Grande, N°6, P-2780-156 Oeiras, Portugal, ‡Faculdade de Ciências da Universidade de Lisboa, Centro de Biologia Ambiental, Campo Grande, Bloco C2-3°Piso, 1749-016 Lisboa, Portugal, §UMR 5174 CNRS/UPS Evolution et Diversité Biologique, Université Paul Sabatier, 118 Route de Narbonne, Bât. 4R3 b2, 31062 Toulouse cédex 09, France

Abstract

Several approaches have been developed to calculate the relative contributions of parental populations in single admixture event scenarios, including Bayesian methods. In many breeds and populations, it may be more realistic to consider multiple admixture events. However, no approach has been developed to date to estimate admixture in such cases. This report describes a program application, 2BAD (for 2-event Bayesian Admixture), which allows the consideration of up to two independent admixture events involving two or three parental populations and a single admixed population, depending on the number of populations sampled. For each of these models, it is possible to estimate several parameters (admixture, effective sizes, etc.) using an approximate Bayesian computation approach. In addition, the program allows comparing pairs of admixture models, determining which is the most likely given data. The application was tested through simulations and was found to provide good estimates for the contribution of the populations at the two admixture events. We were also able to determine whether an admixture model was more likely than a simple split model.

Keywords: approximate Bayesian computation, multiple admixture

Received 30 March 2009; revision received 30 June 2009; accepted 2 August 2009

Genetic data from present-day populations are increasingly being used to reconstruct the demographic history of populations. This history can be complex, involving population expansions, bottlenecks and admixture events. Genetic data have proven useful to infer parameters values for simple (Beaumont *et al.* 2002) or more complex (Fagundes *et al.* 2007) demographic models, including admixture models (Chikhi *et al.* 2001; Choisy *et al.* 2004; Excoffier *et al.* 2005; Sousa *et al.* 2009). Admixture occurs when two or more differentiated populations are brought into contact for a brief episode creating hybrid or admixed populations. For instance, admixture events occurred during the colonization of already occupied areas and during and after the domestication of animals and plants (e.g. the formation of new breeds through crossing; Blott *et al.* 1998). Several methods have been proposed to estimate admixture proportions based on genetic data, but only some of them try to explicitly model the demographic history of the populations sampled (e.g. Chikhi *et al.* 2001; Wang 2003). In general, these models assume that admixture took place during one unique event and that gene flow was negligible after that event; an assumption, which is particularly unrealistic for breed dynamics in some domestic species.

Here, we analyse several models, in which up to two independent admixture events may take place at different times, and we develop a method that estimates demographic parameters (the time since the admixture event, the relative contributions of the parental populations, etc.) taking into account the sampling procedure, genetic drift and mutations for microsatellite loci data. Fig. 1 shows the demographic models considered. It is assumed that an ancestral population of size N_A splits t_{split} generations ago into two or three parental populations (P_1, P_2, P_3), with effective sizes N_1, N_2, N_3. The first admixture event occurred t_{adm1} generations ago and the second admixture event occurred t_{adm2} generations ago. In the first model (Fig. 1A), admixture first occurs between P_1 and P_2 giving rise to the hybrid population (H), with

Correspondence: L. Chikhi, Fax: +351 21 440 40 79;
E-mail: chikhi@igc.gulbenkian.pt

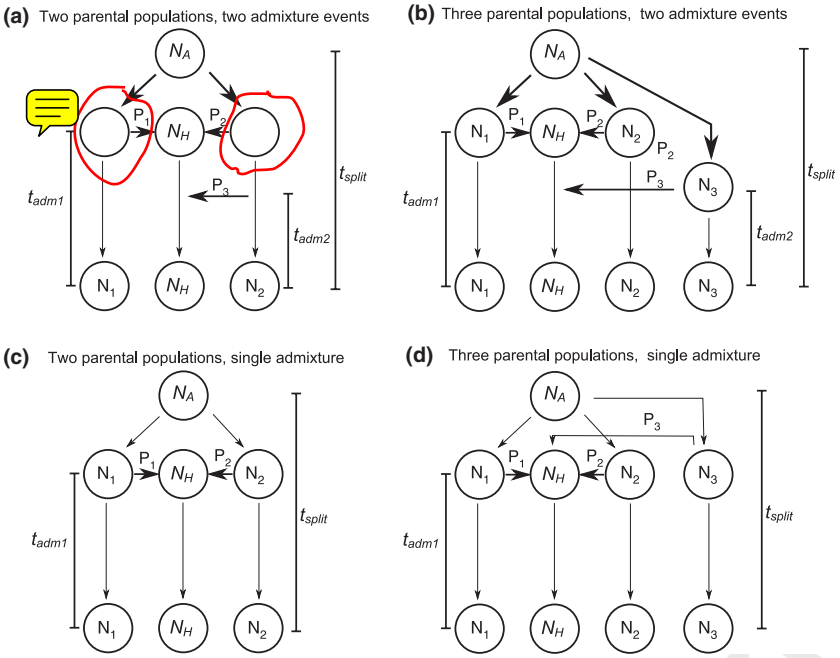


Fig. 1 The four admixture models considered.

effective size N_H. Then the second admixture event takes place, involving only P_2. In the second model (Fig. 1b), the only difference is that the second admixture event involves a third population, P_3. The last two models (Fig. 1c and d) assume a single admixture event. They can thus be seen as special cases of the previous models by fixing $t_{adm1} = t_{adm2}$. The model assumes that all loci have the same mutation rate and that the markers evolve according to the stepwise-mutation model (SMM).

The flexibility of the 2BAD program should allow its application for many biological situations, where two or three populations are thought to have potentially contributed to the genetic pool of potential admixed populations, and where the dating of these events is not clearly identified. Admixture events involving more than two parental populations are common in humans (e.g. Latin American Mestizos, Wang *et al.* 2008) and breeds (Bray *et al.* 2009). They are less documented in natural populations, but the situation could be common in freshwater fish species, when restocking is carried out from more than one source population (Kelly *et al.* 2006), and in plants that were put into contact from more than two refugia. Also, the fact that 2BAD allows testing alternative models should prove important to identify such cases where there is uncertainty on the number of admixture events and on their timing.

Recently, approximate Bayesian computation (ABC) methods (Beaumont *et al.* 2002) have become popular as an alternative to full-likelihood methods because of their flexibility and ability to be applied to complex demographic models at a relative low computational cost (e.g. Excoffier *et al.* 2005; Fagundes *et al.* 2007). ABC algorithms are based on a rejection scheme to obtain an approximate sample from the joint posterior distribution. Briefly, this involves five steps: (i) definition of a demographic model, including the prior distributions of the parameters of the model; (ii) simulation of datasets with different parameter values drawn from the prior distributions; (iii) computation of a set of summary statistics (e.g. number of alleles, expected heterozygosity, etc.) for each dataset; (iv) comparison of the observed and simulated summary statistics using a distance metric (e.g. Euclidean distance, but see Sousa *et al.* 2009 for the use of different distances); and (v) rejection of the parameters that generate datasets that are distant from the observed data.

In this study, we show that it is possible to apply an ABC approach to the admixture models described above and estimate the different parameters using reasonably large microsatellite data sets, similar to those commonly used for livestock breeds and increasingly available in endangered species. The method was implemented in a user-friendly program named 2BAD. The user provides an input file with the allele frequencies for each locus for one admixed and two or three parental populations. The user can then either estimate the parameters within one of the appropriate models, or compare two demographic models (e.g. one admixture vs. a split model, or one admixture event vs. two admixture models) using Beaumont (2008) approach. In both cases, the user selects and defines the prior distribution for each parameter (mutation rate, effective sizes, time of admixture and contribution of parental populations). Depending on the parameter, the user can select uniform, gamma, lognormal or beta prior distributions. For each parameter set,

genetic data are simulated using the coalescent with the ms program of Hudson (2002). In practice, the program uses MATLAB and C code to build the interface, run ms and perform the ABC inference step. For each locus, a set of predefined summary statistics are computed, namely: (i) expected heterozygosity for each population and over all populations; (ii) number of alleles in each population and overall populations: (iii) number of private alleles in each population; (iv) number of gaps in the allelic distribution in each population; (v) pairwise F_{ST} and overall F_{ST}. For each of these statistics, we considered the mean across loci and standardized them according to the mean and standard deviation computed using a set of 10 000 simulations. The distance between the standardized summary statistics for the simulated data and the observed data is computed with a Euclidean distance. The parameter sets that generated the simulated data with the smallest distances are then accepted. The user specifies the tolerance level defined as the proportion of simulations to be kept. The program outputs the point estimates of the different parameters and a histogram to represent the posterior distribution. Several text files are produced saving the point estimates and 95% credible intervals for each parameter, the accepted parameter values, the accepted summary statistics and the corresponding distances.

The performance of the ABC methodology was assessed using a simulation study. Datasets simulated with known parameter values were analysed as pseudo-observed datasets, and the estimates obtained using 2BAD were then compared with the known parameter values. We simulated data under an admixture model with three parental populations and two admixture events (Fig. 1B). To assess the effect of genetic drift on the quality of the estimates, we simulated data assuming a scenario with limited drift and another one with strong drift. The low and strong drift scenarios correspond to effective sizes sampled from U[1000, 15000] and U[100, 1000] respectively and to t_{split} values sampled from U[1000, 15000] and U[100, 1000] respectively. For the other parameters, we used the same priors: t_{adm1} and t_{adm2} were sampled from U[0, 100] in generations, the mutation rates (per locus per generation) from U[10^{-5}, 10^{-3}] and p_1 and p_3 from U[0,1]. For each of these two scenarios, five hundred independent datasets of twenty independent microsatellite loci each were simulated and analysed with 2BAD. The tolerance value was set as 1% (1000 accepted simulations out of 10^6). The effect of the number of simulations was assessed by repeating the analysis with 10^6 and 10^7 simulations.

The results show that 2BAD returned point estimates close to the true parameter values for all parameters (Fig. 2). As expected, the estimates obtained under the strong drift have higher error (seen in Fig. 2). It is noteworthy that the method was able to accurately estimate p_1 and p_3, showing, for the first time that ABC methods are able to quantify the contribution of parental popula-

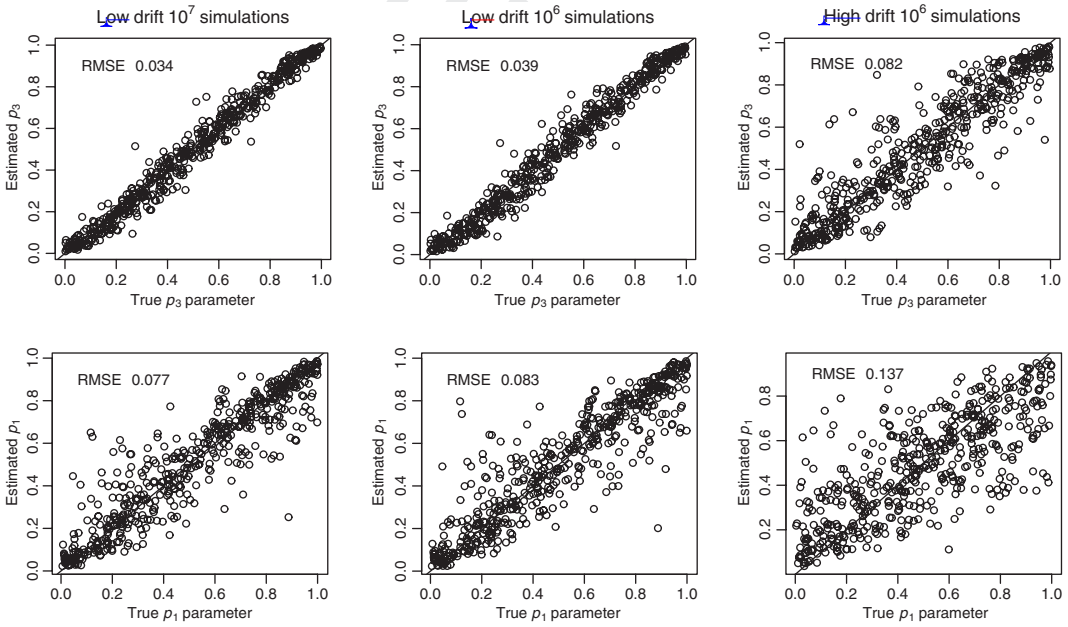


Fig. 2 True vs. point estimates of the admixture parameters. In this figure, the *x* axis represents the true value for p_1 and p_3, whereas the y axis represents the corresponding point estimates obtained using 2BAD. The different panels represent different amounts of drift and different numbers of simulations. The Root Mean Square Error is shown in each panel as a measure of precision.

tions under two admixture events. No major differences were found between the estimates obtained with 10^6 and 10^7 simulations, suggesting that one million simulations should be sufficient to obtain good estimates. While this is in agreement with our results on a simpler admixture model (Sousa *et al*. 2009), larger simulations may provide better estimates. Overall, our results show that good estimates are obtained. We also found that the method is robust to some extent to bottlenecks taking place after the admixture event, as may have been the case in some rare breeds (e.g. Bray *et al*. 2009, Sousa *et al*., in preparation).

To conclude, we have developed an easy-to-use program, which implements a method allowing population genetics inference for an admixture model involving up to two independent admixture events and an easy-to-use procedure for model choice. It is important to add as a final note that the models implemented in 2BAD do not take into account events such as bottlenecks, expansions and migration, which might all affect estimates provided by 2BAD. Testing the robustness of 2BAD to all these factors would be beyond the scope of this study. However, we are currently performing a simulation study to assess the effect of bottlenecks and the performance of the model choice procedure (Sousa *et al*., in prep). Our preliminary results suggest that recent bottlenecks do not lead to biased estimates. They also show that it is possible to separate a pure population split model from an admixture model. Finally, we found that it is also possible to determine whether a single admixture event is more likely than a model with two admixture events.

Acknowledgements

We thank the Rare Breeds Survival Trust, Dexter Cattle Society, the Instituto Gulbenkian de Ciência, the Université Paul Sabatier and Cardiff University for funding and infrastructural support for this research. Thanks go to A. Coutinho and B. Crouau-Roy for their continuous support. This work was supported by the 'Fundação Ciência e Tecnologia' (FCT PhD studentship to V. Sousa SFRH/BD/22224/2005), the Rare Breeds Survival Trust, the Dexter Cattle Society and Cardiff University (PhD studentship to T. Bray). Calculations were performed using the High Performance Computing resource at the 'Instituto Gulbenkian de Ciência' (IGC) with the help of P. Fernandes (FCT H200741/re-equip/2005). LC was partly funded by the FCT grant PTDC/BIA-BDE/71299/2006. The program 2BAD is freely available for research use from the authors and from the Rare Breeds Survival Trust (http://downloads.igc.gulbenkian.pt/program2bad/), to whom applications for licenses for commercial use should be addressed. We finally would like to thank the Subject Editor (V. Castric) for his critical comments which led us to implement the model choice procedure and stimulated a more rigorous simulation study, altogether improving 2BAD.

References

Beaumont MA (2008) Joint determination of topology, divergence time and immigration in population trees. In: *Simulations, Genetics and Human Prehistory, (McDonald Institute Monographs)* (eds Matsumura S, Forster P & Renfrew C), pp 134–154. McDonald Institute for Archaeological Research, Cambridge.

Beaumont MA, Zhang W, Balding DJ (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Berniell-Lee G, Plaza S, Bosch E *et al*. (2008) Admixture and sexual bias in the population settlement of La Réunion Island (Indian Ocean). *American Journal of Physical Anthropology*, **136**, 100–107.

Blott SC, Williams JL, Haley CS (1998) Genetic relationships among European cattle breeds. *Animal Genetics*, **29**, 273–282.

Bray TC, Chikhi L, Sheppy AJ, Bruford MW (2009) The population genetic effects of ancestry and admixture in a subdivided cattle breed. *Animal Genetics*. DOI; 10.1111/j.1365-2052.2009.01850.x.

Chikhi L, Bruford MW, Beaumont MA (2001) Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics*, **158**, 1347–1362.

Choisy MP, Franck P, Cornuet JM (2004) Estimating admixture proportions with microsatellites: comparison of methods based on simulated data. *Molecular Ecology*, **13**, 955–968.

Excoffier L, Estoup A, Cornuet J-M (2005) Bayesian analysis of an admixture model with mutations and arbitrarily linked markers. *Genetics*, **169**, 1727–1738.

Fagundes NJR, Ray N, Beaumont M *et al*. (2007) Statistical evaluation of alternative models of human evolution. *Proceedings of the National Academy of Sciences USA*, **104**, 17614–17619.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Kelly DW, Muirhead JR, Heath DD, Macisaac HJ (2006) Contrasting patterns in genetic diversity following multiple invasions of fresh and brackish waters. *Molecular Ecology*, **15**(12), 3641–3653.

Sousa VC, Fritz M, Beaumont MA, Chikhi L (2009) Approximate Bayesian computation without summary statistics: the case of Admixture. *Genetics*, **181**.

Wang J (2003) Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, **164**, 747–765.

Wang S, Ray N, Rojas W *et al*. (2008) Geographic patterns of genome admixture in Latin American Mestizos. *PLoS Genetics*, **4**.

Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Relative Root Mean Square Error (RRMSE) for the different parameters estimated for the three parental, two admixture events model.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Author Query Form

Journal: MEN

Article: 2766

Dear Author,
During the copy-editing of your paper, the following queries arose. Please respond to these by marking up your proofs with the necessary changes/additions. Please write your answers on the query sheet if there is insufficient space on the page proofs. Please write clearly and follow the conventions shown on the attached corrections sheet. If returning the proof by fax do not write too close to the paper's edge. Please remember that illegible mark-ups may delay publication.

Many thanks for your assistance.

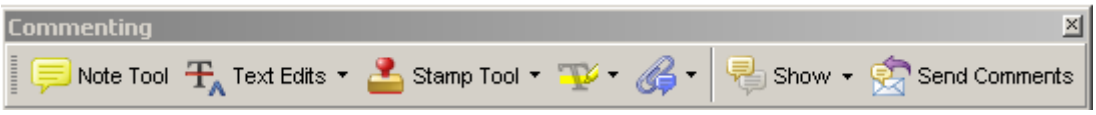
Query reference	Query	Remarks
Q1	**AUTHOR: Beaumont, 2002 has been changed to Beaumont *et al.* 2002 so that this citation matches the Reference List. Please confirm that this is correct.**	
Q2	**AUTHOR: Please provide all author names with initials for this 'in preparation'.**	
Q3	**AUTHOR: Berniell-Lee *et al.* (2008) has not been cited in the text. Please indicate where it should be cited; or delete from the Reference List.**	
Q4	**AUTHOR: Please provide the volume number, page range for reference Bray *et al.* (2009).**	
Q5	**AUTHOR: Please provide the page range for reference Sousa *et al.* (2009).**	
Q6	**AUTHOR: Please provide the page range for reference Wang *et al.* (2008).**	
Q7	**AUTHOR: Table S1 has not been mentioned in the text. Please cite the table in the relevant place in the text.**	

USING E-ANNOTATION TOOLS FOR ELECTRONIC PROOF CORRECTION

Required Software

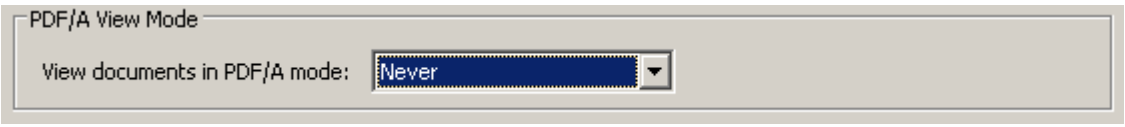
Adobe Acrobat Professional or Acrobat Reader (version 7.0 or above) is required to e-annotate PDFs. Acrobat 8 Reader is a free download: http://www.adobe.com/products/acrobat/readstep2.html

Once you have Acrobat Reader 8 on your PC and open the proof, you will see the Commenting Toolbar (if it does not appear automatically go to Tools>Commenting>Commenting Toolbar). The Commenting Toolbar looks like this:



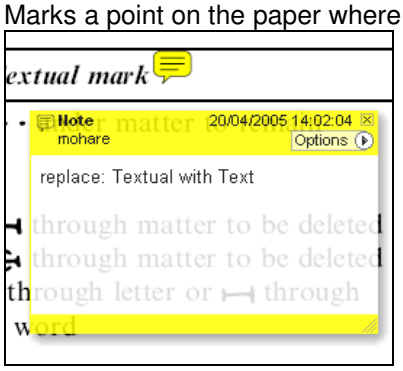
If you experience problems annotating files in Adobe Acrobat Reader 9 then you may need to change a preference setting in order to edit.

In the "Documents" category under "Edit – Preferences", please select the category 'Documents' and change the setting "PDF/A mode:" to "Never".



Note Tool — For making notes at specific points in the text

Marks a point on the paper where a note or question needs to be addressed.

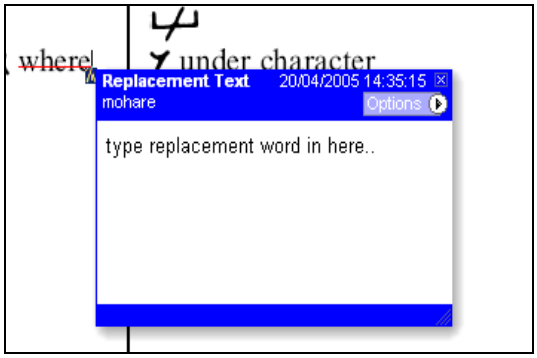


How to use it:

1. Right click into area of either inserted text or relevance to note
2. Select Add Note and a yellow speech bubble symbol and text box will appear
3. Type comment into the text box
4. Click the X in the top right hand corner of the note box to close.

Replacement text tool — For deleting one word/section of text and replacing it

Strikes red line through text and opens up a replacement text box.

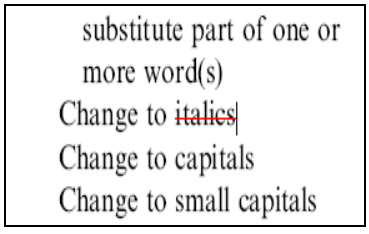


How to use it:

1. Select cursor from toolbar
2. Highlight word or sentence
3. Right click
4. Select Replace Text (Comment) option
5. Type replacement text in blue box
6. Click outside of the blue box to close

Cross out text tool — For deleting text when there is nothing to replace selection

Strikes through text in a red line.



How to use it:

1. Select cursor from toolbar
2. Highlight word or sentence
3. Right click
4. Select Cross Out Text

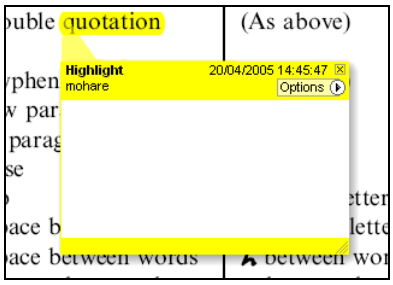
Approved tool — For approving a proof and that no corrections at all are required.

How to use it:

1. Click on the Stamp Tool in the toolbar
2. Select the Approved rubber stamp from the 'standard business' selection
3. Click on the text where you want to rubber stamp to appear (usually first page)

Highlight tool — For highlighting selection that should be changed to bold or italic.

Highlights text in yellow and opens up a text box.

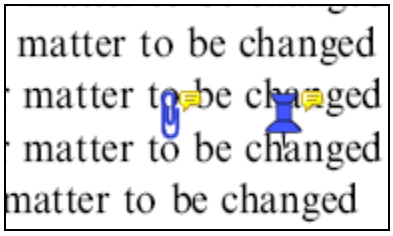


How to use it:

1. Select Highlighter Tool from the commenting toolbar
2. Highlight the desired text
3. Add a note detailing the required change

Attach File Tool — For inserting large amounts of text or replacement figures as a files.

Inserts symbol and speech bubble where a file has been inserted.

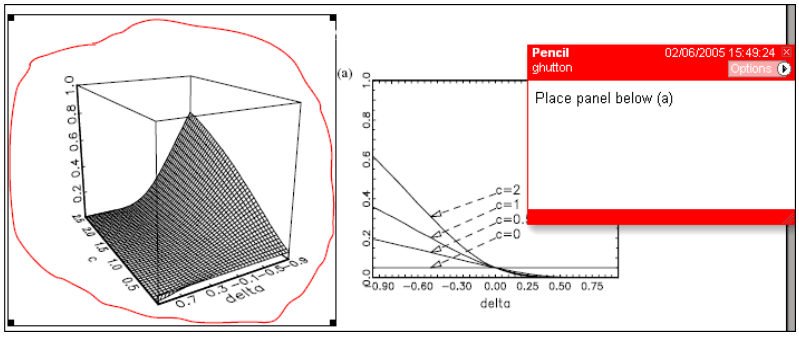


How to use it:

1. Click on paperclip icon in the commenting toolbar
2. Click where you want to insert the attachment
3. Select the saved file from your PC/network
4. Select appearance of icon (paperclip, graph, attachment or tag) and close

Pencil tool — For circling parts of figures or making freeform marks

Creates freeform shapes with a pencil tool. Particularly with graphics within the proof it may be useful to use the Drawing Markups toolbar. These tools allow you to draw circles, lines and comment on these marks.



How to use it:

1. Select Tools > Drawing Markups > Pencil Tool
2. Draw with the cursor
3. Multiple pieces of pencil annotation can be grouped together
4. Once finished, move the cursor over the shape until an arrowhead appears and right click
5. Select Open Pop-Up Note and type in a details of required change
6. Click the X in the top right hand corner of the note box to close.

Help

For further information on how to annotate proofs click on the Help button to activate a list of instructions: