Supplementary results for: A meta-analysis of computational biology benchmarks reveals publication bias affects on speed and accuracy

Paul P. Gardner[1,2]*, James Paterson[1,2], Fatemeh Ashari Ghomi[1,2], Sinan Uğur Umu[1,2], Stephanie McGimpsey[1,2]

Abstract
In the below we provide additional results for our investigation of computational biology benchmarks.

Keywords
computational biology — accuracy — benchmarks — meta-analysis — software development

[1] *School of Biological Sciences, University of Canterbury, Christchurch, New Zealand.*
[2] *Biomolecular Interaction Centre and the Bio-Protection Research Centre, University of Canterbury, Christchurch, New Zealand.*
*__Corresponding author__: paul.gardner@canterbury.ac.nz

Literature mining

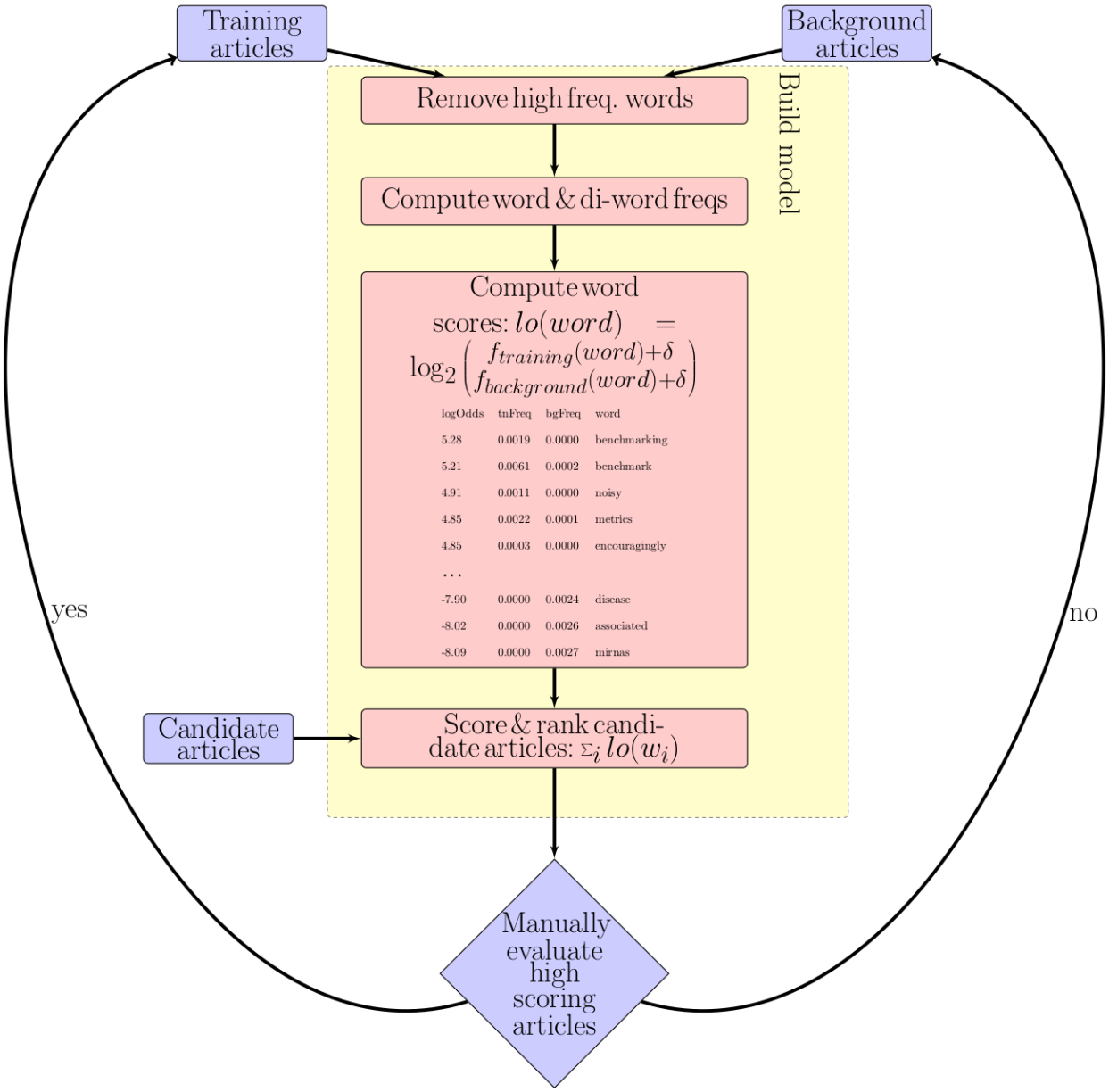

Figure S1. In order to improve the identification of benchmark articles that rank both accuracy and speed we developed a tool for ranking PubMed articles based upon word association scores (measured in 'bits'). In brief, keywords were extracted from titles and abstracts for both training (in this case benchmark articles) and background articles (articles published between 2013 and 2015 with "bioinformatics" in the title or abstract). Log-odds ratios were computed for each keyword (measured in 'bits'). Candidate articles that matched a hand-selected list of keywords associated with benchmarks were then scored and ranked with a "sum of bits" score. High ranking articles were then inspected, those that met our criteria were added to the training set, those that didn't were added to the background set of articles.

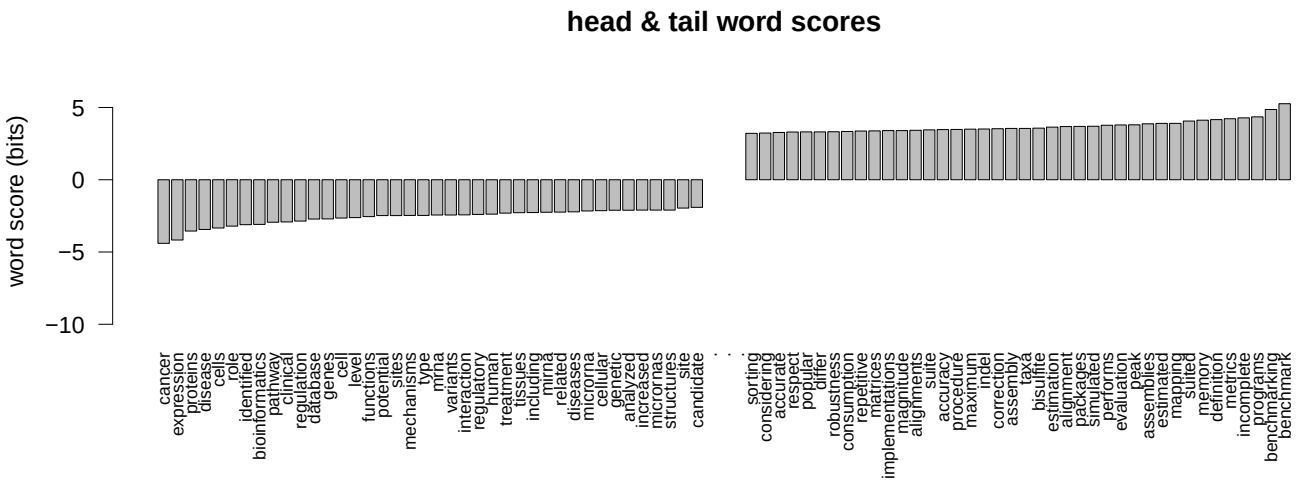

Figure S2. The 40 highest and lowest scoring words that are associated with bioinformatic benchmark articles from the training benchmark articles, compared to the background articles. The log-odds ratios (measured in bits) are indicated on the y-axis.

Data mining

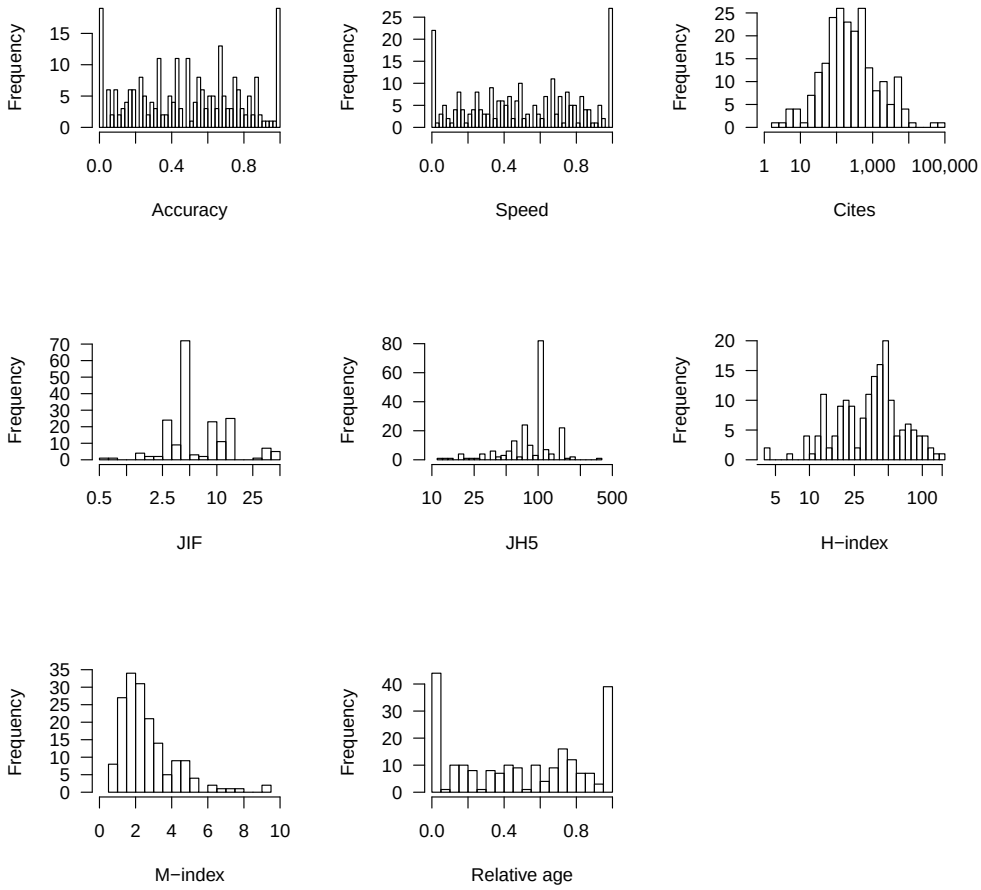

Figure S3. The distributions for the measures we expected to be predictive of software quality used in this study. These are, reading from left to right, top to bottom: Accuracy – the mean normalised accuracy rank for each benchmarked method; Speed – the mean normalised speed rank for each benchmarked method; Cites – the number of citations to the most cited manuscript describing a method, data from GoogleScholar; JIF – the Journal Impact Factor to the highest impact journal that has published a manuscript describing a method, data from 2014 Thompson-Reuters Citation Reports; JH5 – the Journal H5 index to the highest impact journal that has published a manuscript describing a method, data from GoogleScholar 2015 Metrics; H-index – the H-index for the highest profile corresponding author from the manuscripts describing a method, data from GoogleScholar User Profiles; M-index – the M-index (H-index/(#years since first publication)) for the highest profile corresponding author from the manuscripts describing a method, data from GoogleScholar User Profiles;

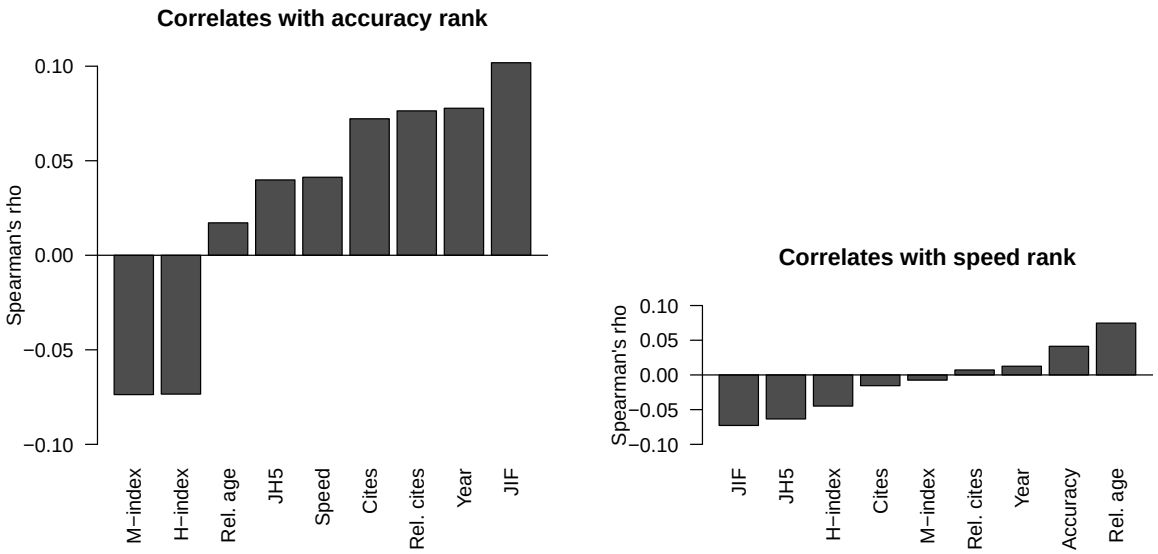

Figure S4. The correlation between method accuracy (on the left) and method speed (on the right) and measures we expected to be predictive of software quality. E.g. author reputation measures (H-index, M-index), journal reputation (JH5 and JIF), number of users (citations and relative citations) or the recency of methods (year and relative age). The correlations were estimated using Spearman's ρ. The signficant relationships are indicated with a "*".

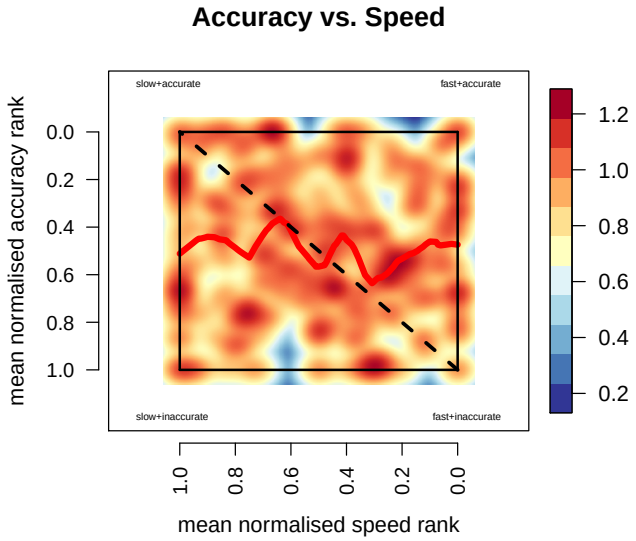

Figure S5. A smoothed color density representation of a scatterplot of normalised speed ranks and normalised accuracy ranks. Dark red regions are indicative of of a high density of points, blue shades indicate the reverse. The expected inverse relationship between speed and accuracy is indicated with a dashed black line, points above this line could be considered comparatively fast and accurate, points below are the reverse. A locally weighted regression (lowess) curve is shown in red.

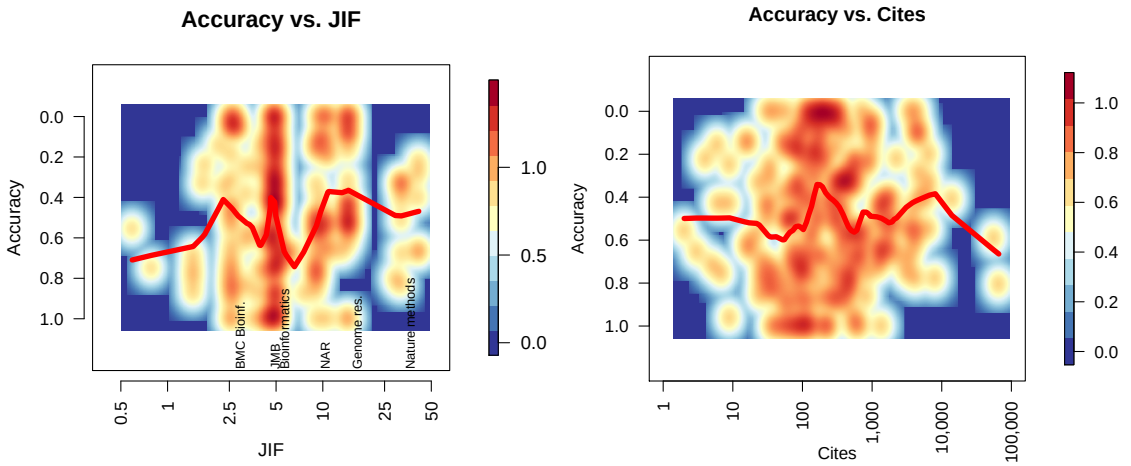

Figure S6. A smoothed color density representation of a scatterplot of journal impact factors (JIF) on the left and number of citations on the right vs normalised accuracy ranks. As above, dark red regions are indicative of of a high density of points, blue shades indicate the reverse. The JIF for journals that are major publishers of bioinformatic software are indicated on the left.

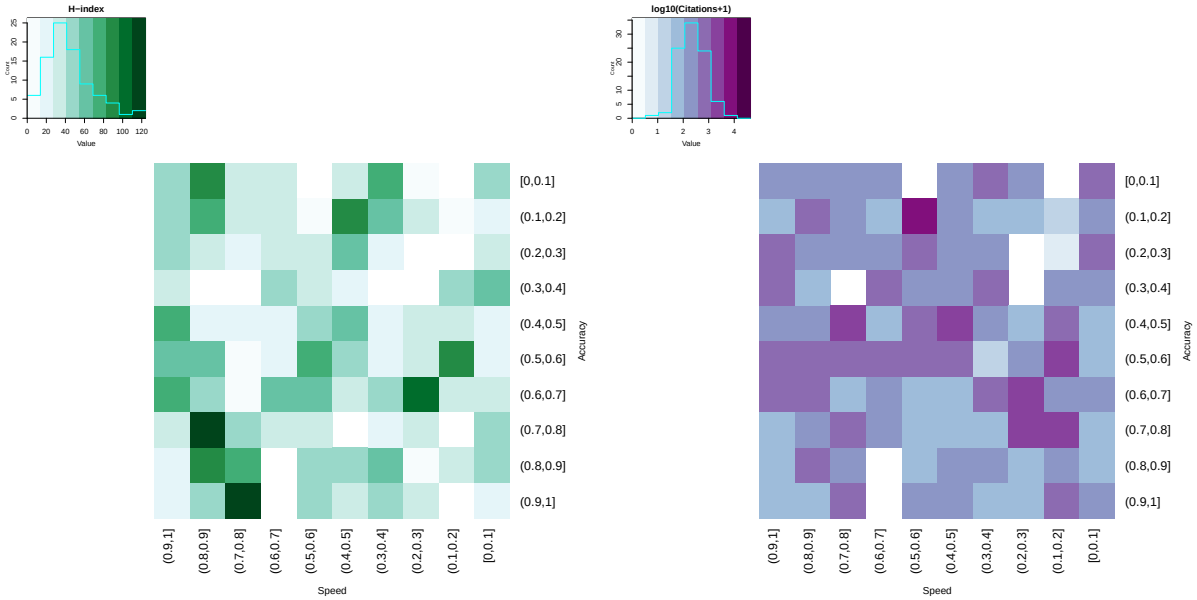

Figure S7. Heatmaps of normalised speed ranks and normalised accuracy ranks, both x and y dimensions are discretised into a 10×10 matrix. The shading indicates median H-indices for corresponding authors (darker shade indicates a higher H) in the figure on the left. The shading indicates median citations (on a log scale), for software tools in the figure on the right.

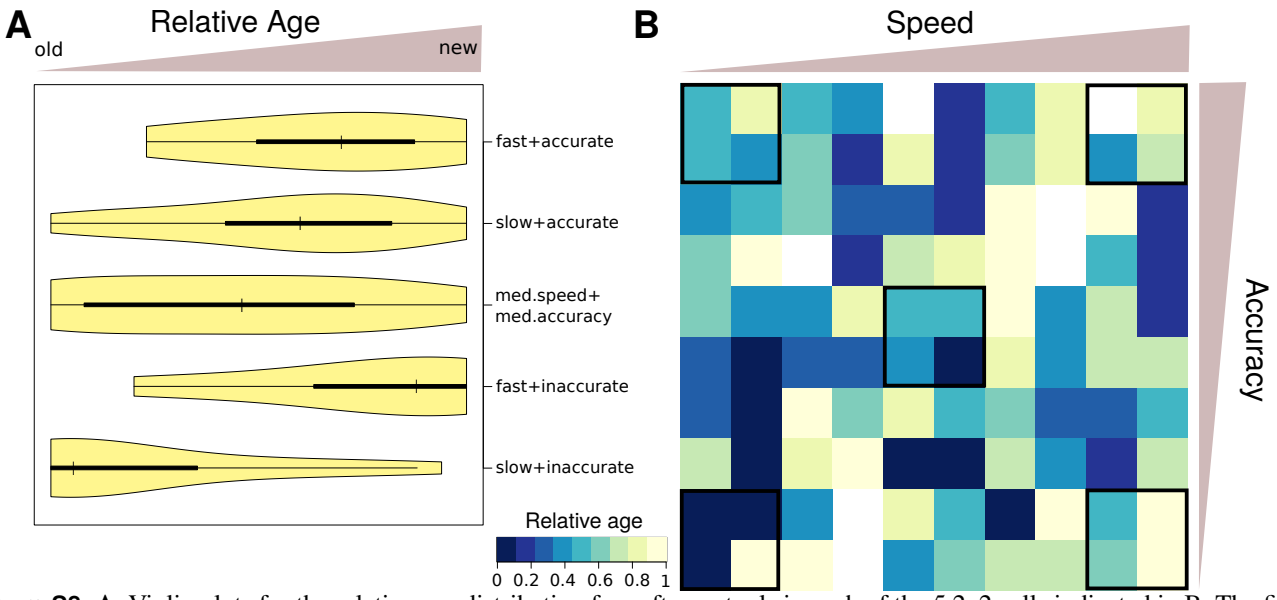

Figure S8. A. Violin plots for the relative age distribution for software tools in each of the 5 2x2 cells indicated in B. The five boxes correspond to the four extreme corners of the speed vs accuracy spectrum (i.e. slow and inaccurate, slow and accurate, fast and inaccurate, fast and accurate) and the central box (medium speed and medium accuracy). **B.** A heatmap indicating the relative age of software in a range of relative accuracy and speed rankings. Blue colours indicate an abundance of older software tools in an accuracy and speed category, while light colours indicate younger software in an accuracy and speed category.