Text mining of 15 million full-text scientific articles

David Westergaard[1,2], Hans-Henrik Stærfeldt[2], Christian Tønsberg[3], Lars Juhl Jensen[2] [†], Søren Brunak[1†]

[1]Center for Biological Sequence Analysis, Department of Bio and Health Informatics, Technical University of Denmark, DK-2800 Lyngby, Denmark

[2]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, DK-2200 Copenhagen, Denmark

[3]Office for Innovation and Sector Services, Technical Information Center of Denmark, Technical University of Denmark, DK-2800 Lyngby, Denmark

[†]To whom correspondence should be addressed: brunak@cbs.dtu.dk or lars.juhl.jensen@cpr.ku.dk

Abstract

Across academia and industry, text mining has become a popular strategy for keeping up with the rapid growth of the scientific literature. Text mining of the scientific literature has mostly been carried out on collections of abstracts, due to their availability. Here we present an analysis of 15 million English scientific full-text articles published during the period 1823–2016. We describe the

23   development in article length and publication sub-topics during these nearly 250

24   years. We showcase the potential of text mining by extracting published protein–

25   protein, disease–gene, and protein subcellular associations using a named entity

26   recognition system, and quantitatively report on their accuracy using gold

27   standard benchmark data sets. We subsequently compare the findings to

28   corresponding results obtained on 16.5 million abstracts included in MEDLINE

29   and show that text mining of full-text articles consistently outperforms using

30   abstracts only.

31

32

33   **Introduction**

34   Text mining has become a widespread approach to identify and extract

35   information from unstructured text. Text mining is used to extract facts and

36   relationships in a structured form that can be used to annotate specialized

37   databases, to transfer knowledge between domains and more generally within

38   business intelligence to support operational and strategic decision-making [1–3].

39   Biomedical text mining is concerned with the extraction of information

40   regarding biological entities, such as genes and proteins, phenotypes, or even

41   more broadly biological pathways (reviewed extensively in [3–9]) from sources

42   like scientific literature, electronic patient records, and most recently patents

43   [10–13]. Furthermore, the extracted information has been used as annotation of

44   specialized databases and tools (reviewed in [3,14]). In addition, text mining is

45   routinely used to support manual curation of biological databases [15,16]. Thus,

2

46 text mining has become an integral part of many resources serving a wide

47 audience of scientists. The main text source for scientific literature has been the

48 MEDLINE corpus of abstracts, essentially due to the restricted availability of full-

49 text articles. However, full-text articles are becoming more accessible and there

50 is a growing interest in text mining of complete articles. Nevertheless, to date no

51 studies have presented a systematic comparison of the performance comparing

52 abstracts and full-texts in corpora that are similar in size to MEDLINE.

53

54 Full-text articles and abstracts are structurally different [17]. Abstracts are

55 comprised of shorter sentences and very succinct text presenting only the most

56 important findings. By comparison, full-text articles contain complex tables,

57 display items and references. Moreover, they present existing and generally

58 accepted knowledge in the introduction (often presented in the context of

59 summaries of the findings), and move on to reporting more in-depth results,

60 while discussion sections put the results in perspective and mention limitations

61 and concerns. The latter is often considered to be more speculative compared to

62 the abstract [3].

63

64 While text-mining results from accessible full-text articles have already become

65 an integral part of some databases (reviewed recently for protein-protein

66 intaractions [18]), very few studies to date have compared text mining of

67 abstracts and full-text articles. Using a corpus consisting of ~20,000 articles from

68 the PubMed Central (PMC) open-access subset and Directory of Open Access

69 Journals (DOAJ), it was found that many explicit protein–protein interactions

3

70   only are mentioned in the full text [19]. Additionally, in a corpus of 1,025 full-text

71   articles it was noticed that some pharmacogenomics associations are only found

72   in the full text [20]. One study using a corpus of 3,800 articles with focus on

73   *Caenorhabditis elegans* noted an increase in recall from 45% to 95% when

74   including the full text [21]. Other studies have worked with even smaller corpora

75   [17,22,23]. One study have even noted that the majority of claims within an

76   article is not reported in the abstract [24]. Whilst these studies have been of

77   significant interest, the number of full-text articles and abstracts used for

78   comparison are nowhere near the magnitude of the actual number of scientific

79   articles published to date, and it is thus unclear if the results can be generalized

80   to the scientific literature as a whole. The earlier studies have mostly used

81   articles retrieved from PMC in a structured XML file. However, full-text articles

82   received or downloaded directly from the publishers often come in the PDF

83   format, which must be converted to a raw unformatted text file. This presents a

84   challenge, as the quality of the text mining will depend on the proper extraction

85   and filtering of the unformatted text. A previous study dealt with this by writing

86   custom software taking into account the structure and font of each journal at

87   that time [21]. More recent studies typically provide algorithms that

88   automatically determines the layout of the articles [25–27].

89

90   In this work, we describe a corpus of 15 million full-text scientific articles from

91   Elsevier, Springer, and the open-access subset of PMC. The articles were

92   published during the period 1823–2016. We highlight the possibilities by

93   extracting protein–protein associations, disease–gene associations, and protein

4

94   subcellular localization from the large collection of full-text articles using a

95   Named Entity Recognition (NER) system combined with a scoring of co-

96   mentions. We quantitatively report the accuracy and performance using gold

97   standard benchmark data sets. Lastly, we compare the findings to corresponding

98   results obtained on the matching set of abstracts included in MEDLINE as well as

99   the full set of 16.5 million MEDLINE abstracts.

100

101

102   **Results**

103

104   **Growth and temporal development in full text corpora**

105   The growth of the data set over time is of general interest in itself, however, it is

106   also important to secure that the concepts used in the benchmarks are likely to

107   be present in a large part of the corpus. We found that the number of full-text

108   articles has grown exponentially over a long period (Fig 1a, a log-transformed

109   version is provided in Supplementary Fig 1). We also observed that the growth

110   represents a mixture of two components: one from 1823–1944, and another

111   from 1945–2016. Fitting an exponential curve to the years 1945–2016 we found

112   that the growth rate is 0.103 ($p < 2 * 10^{-16}, R^2 = 0.95$). Thus, the doubling time

113   for the full-text corpus is 9.7 years. In comparison, MEDLINE had a growth rate

114   of 0.195 ($p < 2 * 10^{-16}, R^2 = 0.91$) and a doubling time of 5.1 years. We noticed

115   that there was a drop in the number of full-text publications around the years

116   1914–1918 and 1940–1945. Likewise, we see a decrease in the number of

117   publications indexed by MEDLINE in the entire period 1930–1948.

118   **Fig 1**: Temporal corpus statistics. **(a)** Number of publications per year in the

119   period 1823-2016. The growth in publications was found to fit an exponential

120   model. **(b)** Temporal development in the distribution of six different topical

121   categories in the period 1823-2016. Publications from health science journals

122   made up nearly 75% of all publications until 1950, at which point it started to

123   decrease rapidly. To date, it makes up approximately 25% of the publications in

124   the full-text corpus. **(c)** Development in the number of pages per article in the

125   period 1823-2016. The range of pages varies from 1-1,572 pages. Until year

126   1900 the number of one-page articles were increasing, at one point making up

127   75% of all articles. At the end of the 19th century, the number of one-page

128   articles started to decrease, and by the start of the 21th century they made up

129   less than 20%. Conversely, the number of articles with 11+ pages has been

130   increasing, and by the start of the 21th century made up more than 20% of all

131   articles.

132

133   We binned the full-text articles into four categories based on the number of

134   pages (see Methods). The average length of articles has increased considerably

135   during the almost 250 years studied (Fig 1b). Whereas 75% of the articles were

136   1–3 pages long at the end of the 20th century, less than 25% of the articles

137   published after year 2000 are that short. Conversely, articles with ten or more

138   pages only made up between 0.7%-7% in the 19th century, a level that had grown

139   to 20% by the start of the 21st century.

140

141 In the full-text corpora we found a total of 12,781 unique journal titles. The most

142 prevalent journals are tied to health or life sciences, such as *The Lancet,*

143 *Tetrahedron Letters,* and *Biochemical and Biophysical Research Communications,*

144 or the more broad journals such as *PLoS ONE* (see Supplementary Table 1 for the

145 top-15 journals). *The Lancet* publishes only very few articles per issue, it was

146 established in 1823 and has been active in publishing since then, thus explaining

147 why it so far has nearly published 400,000 articles. In contrast, *PLoS ONE* was

148 launched in 2006, and has published more than 172,000 articles. Of the 12,781

149 journal titles, 6,900 had one or more category labels assigned by librarians at the

150 Technical University of Denmark. The vast majority of the full-texts, 13,343,040,

151 were published in journals with one or more category labels. The frequency of

152 each category within the corpus can be seen in Supplementary Fig 2. We

153 observed that before the 1950's health science dominated and made up almost

154 75% of all publications (Fig 1c). At the start of the 1950's the fraction started to

155 decrease, and to date health science makes up approximately 25% of all

156 publications in the full-text corpus. Inspecting the remaining eleven categories in

157 a separate plot we found that there was no single category that was responsible

158 for the growth (Supplementary Fig 3).

159

160 **Evaluating information extraction across corpora**

161 We analyzed and compared four different corpora comprising all full-text

162 articles (14,549,483 articles, All Full-texts), full-text articles that had a separate

163 abstract (10,376,626 articles, Core Full-texts), the abstract from the full-text

164   articles (10,376,626 abstracts, Core Abstracts), and the MEDLINE corpus

165   (16,544,511 abstracts, MEDLINE).

166

167   We have used quite difficult, but still well established benchmarks, to illustrate

168   the differences in performance when comparing text mining of abstracts to full-

169   text articles. Within biology, and specifically in the area of systems biology,

170   macromolecular interactions and the relationships between genes, tissues and

171   diseases are key data that drive modeling and the analysis of causal biochemical

172   mechanisms. Knowledge of interactions between proteins is extremely useful

173   when revealing the components, which contribute to mechanisms in both health

174   and disease. As many biological species from evolution share protein orthologs,

175   their mutual interactions can often be transferred, for example from an

176   experiment in another organism to the corresponding pair of human proteins

177   where the experiment has not yet been performed. Such correspondences can

178   typically be revealed by text mining as researchers in one area often will not

179   follow the literature in the other and *vice versa*.

180

181   We ran the text mining pipeline on the two full-text and two abstract corpora. In

182   all cases we found that the AUC-value was far greater than 0.5, from which we

183   conclude that the results were substantially better than random (Fig 2). The

184   biggest gain in performance when using full-text was seen in finding associations

185   between diseases and genes (Supplementary Table 2). Compared to MEDLINE,

186   the traditional corpus used for biomedical text mining, there was an increase in

187   the AUC from 0.85 to 0.91. The smallest gain was associations between proteins,

188   which increased from 0.70 to 0.73. Likewise, the Core Full-texts always

189   performed better than Core Abstracts, signifying that some associations are only

190   reported in the main body of the text. Consequently, traditional text mining of

191   abstracts will never be able to find this information.

192

193   **Fig 2**: Benchmarking the four different corpora. In all cases the AUC is far greater

194   than 0.5, indicating that the results obtained are better than random. The biggest

195   gain in AUC is seen for disease-gene associations **(a)**, followed by protein-

196   compartment associations **(c)** and protein-protein associations **(b)**.

197

198   It has previously been speculated if text mining of full-text articles may be more

199   difficult and lead to an increased rate of false positives [3]. To investigate this we

200   altered the weights of the scoring system. The scoring scheme used here has

201   weights for within sentence, within paragraph and within document co-

202   occurrences (see Methods). When setting the document weight to zero versus

203   using the previously calibrated value we found that having a non-zero small

204   value does indeed improve extraction of known facts in all cases (Supplementary

205   Fig 4). Inspecting the gain in AUC we found that it is lower, compared to having a

206   document weight (Supplementary Table 2). In one case, protein–protein

207   associations, the MEDLINE abstract corpus outperforms the full-text articles.

208   Abstracts are generally unaffected by the document weight, mainly because

209   abstracts are almost always one paragraph. Overall, the difference in

210   performance gain is largest for full-texts and lowest for abstracts and MEDLINE.

211   Hence, all the full-text information is indeed valuable and necessary.

9

212

213   For practical applications, it is often necessary to have a low False Positive Rate

214   (FPR). Accordingly, we evaluated the True Positive Rate (TPR) of the different

215   corpora at the 10% FPR (TPR@10%FPR) (Fig 3). We found that full-texts have

216   the highest TPR@10%FPR for disease-gene associations (Supplementary Table

217   3). When considering protein–protein associations and protein-compartment

218   associations, full-texts perform equivalently to Core Abstracts and Core Full-

219   texts. The result was similar to when we evaluated the AUC across the full range,

220   removing the document weight has the biggest impact on the full-texts

221   (Supplementary Fig 5), while abstracts remain unaffected.

222

223   **Fig 3**: Benchmarking the four different corpora at low false positive rates. At a

224   false positive rate of 10%, relevant to practical applications, the full-text corpus

225   still outperforms the collection of MEDLINE abstracts for the extraction of

226   disease-gene associations. Conversely, the performance is the same for protein-

227   protein associations and protein-compartment associations.

228

229   **Discussion**

230   We have investigated a unique corpus consisting of 15 million full-text articles

231   and compared the results to the most commonly used corpus for biomedical text

232   mining, MEDLINE. We found that the full-text corpus outperforms the MEDLINE

233   abstracts in all benchmarked cases. To our knowledge, this is the largest

234   comparative study to date of abstracts and full-text articles. We envision that the

10

235 results presented here can be used in future applications for discovering novel

236 associations from mining of full-text articles, and as a motivation to always

237 include full-text articles when available and to improve the techniques used for

238 this purpose.

239

240 The corpus consisted of 15,032,496 full-text documents, mainly in PDF format.

241 1,504,674 documents had to be discarded for technical reasons, primarily

242 because they were not in English. Further, a large number of documents were

243 also found to be duplicates or subsets of each other. On manual inspection we

244 found that these were often conference proceedings, collections of articles etc.,

245 which were not easily separable without manual curation. We also managed to

246 identify the list of references in the majority of the articles thereby reducing

247 some repetition of knowledge that could otherwise lead to an increase in the

248 false positive rate.

249

250 We have encountered and described a number of problems when working with

251 full-text articles converted from PDF to TXT from a large corpus. However, the

252 majority of the problems did not stem from the PDF to TXT conversion, which

253 could potentially be solved using a layout aware conversion tool. Examples

254 include LA-PDFText [27], SectLabel [26] of PDFX [25], of which the latter is not

255 practical for very large corpora as it only exists as an online tool. Nonetheless, to

256 make use of the large volume of existing articles it is necessary to solve these

257 problems. Having all the articles in a structured XML format, such as the one

258 provided by PubMed Central, would with no doubt produce a higher quality

11

259   corpus. This may in turn further increase the benchmark results for full-text

260   articles. Nevertheless, the reality is that many articles are not served that way.

261   Consequently, the performance gain we report here should be viewed as a lower

262   limit as we have sacrificed quality in favor of a larger volume of articles. The

263   solutions we have outlined here will serve as a guideline and baseline for future

264   studies.

265

266   The increasing article length may have different underlying causes, but one of

267   the main contributors is most likely increased funding to science worldwide

268   [28,29].   Experiments and protocols are consequently getting increasingly

269   complex and interdisciplinary – aspects that also contribute to driving

270   meaningful publication lengths upward. The increased complexity has also been

271   found to affect the language of the articles, as it is becoming more

272   specialized[30]. It was outside the scope of this paper to go further into socio-

273   economic impact. We have limited this to presenting the trends from what could

274   be computed from the meta-data.

275

276   Previous papers are – in terms of benchmarking – only making qualitative

277   statements about the value of full-text articles as compared to text in abstracts.

278   In one paper a single statement is made on the potential for extracting

279   information, but no quantitative evidence is presented [31]. In a paper targeting

280   pharmacogenomics it is similarly stated that that there are associations that only

281   are found in the full-text, but no quantitative estimates are presented [20]. In a

282   paper analyzing around 20,000 full-text papers a search for physical protein

283   interactions was made, concluding that these contain considerable higher levels

284   of interaction [19]. Again, no quantitative benchmarks were made comparing

285   different sources. In this paper, we have made a detailed comparison of four

286   different corpora that provides a strong basis for estimating the added value of

287   using full-text articles in text mining workflows.

288

289   The results presented here are purely associational. Through rigorous

290   benchmarking and comparison of a variety of biologically relevant associations,

291   we have demonstrated that a substantial amount of relevant information is only

292   found in the full body of text. Additionally, by modifying the document weight we

293   found that it was important to take into account the whole document and not

294   just individual paragraphs. Consequently, as text mining methods improve and

295   become more sophisticated, the quantitative benchmarks will improve. Event-

296   based text mining will be the next step for a deeper interpretation and extending

297   the applicability of the results [5]. With more development it may also be

298   possible to extract quantitative values, as has been demonstrated for

299   pharmacokinetics [32]. However, this was outside the scope of this article.

300

301   The Named Entity Recognition (NER) system used depends heavily on the

302   dictionaries and stop word lists. A NER system is also very sensitive to

303   ambiguous words. To combat this we have used dictionaries from well-known

304   and peer-reviewed databases, and we have included other dictionaries to avoid

305   ambiguous terms. Other approaches to text mining have previously been

306   extensively reviewed [10,14,32].

307

308   The full-text corpus presented here consists of articles from Springer, Elsevier

309   and PubMed. However, we still believe that the results presented here are valid

310   and can be generalized across publishers, to even bigger corpora. Preprocessing

311   of corpora is an ongoing research project, and it can be difficult to weed out the

312   rubbish when dealing with millions of documents. We have tried to use a process

313   where we evaluate the quality of a subset of randomly selected articles

314   repeatedly and manually, until it no longer improves.

315

316

317   **Methods**

318   **MEDLINE Corpus**

319   The MEDLINE corpus consists of 26,385,631 citations. We removed empty

320   citations, corrections and duplicate PubMed IDs. For duplicate PubMed IDs we

321   kept only the newest entry. This led to a total of 16,544,511 abstracts for text

322   mining.

323

324   **PMC Corpus**

325   The PubMed Central corpus comprises 1,488,927 freely available scientific

326   articles (downloaded 27th January 2017). Each article was retrieved in XML

327   format. The XML file contains the article divided into paragraphs, article category

328   and meta-information such as journal, year published, etc. Articles that had a

329   category matching Addendum, Corrigendum, Erratum or Retraction were

330   discarded. A total of 5,807 documents were discarded due to this, yielding a total

331   of 1,483,120 articles for text mining. The article paragraphs were extracted for

332   text mining. No further pre-processing of the text was done. The journals were

333   categorized according to categories (described in the following section) by

334   matching the ISSN number. The number of pages for each article was also

335   extracted from the XML, if possible. Permission for use of the PMC corpus was

336   obtained by the Technical Information Center of Denmark (DTU Library).

337

338   **TDM Corpus**

339   The Technical Information Center of Denmark (DTU Library) TDM corpus is a

340   collection of full-text articles from the publishers Springer and Elsevier, where

341   the library has obtained permission for use in the context of text mining. The

342   corpus covers the period from 1823 to 2016. The corpus comprises 3,335,400

343   and 11,697,096 full-text articles in PDF format, respectively. An XML file

344   containing meta-data such as publication date, journal, etc. accompanies each

345   full-text article. PDF to TXT conversion was done using pdftotext v0.47.0, part of

346   the Poppler suite (poppler.freedesktop.org). 192 articles could not be converted

347   to text due to errors in the PDF file. The article length, counted as the number of

348   pages, was extracted from the XML file. If not recorded in the XML file we

349   counted the number of pages in the PDF file using the Unix tool pdfinfo v0.26.5.

350   Articles were grouped into four bins, determined from the 25%, 50%, and 75%

351   quantiles, respectively. These were found to be 1-4 pages (0-25%), 5-7 pages

352   (25-50%), 8-10 pages (50-75%) and 11+ pages (75%-100%). Each article was,

15

353 based on the journal where it was published, assigned to one or more of the

354 following seventeen categories: Health Sciences, Chemistry, Life Sciences,

355 Engineering, Physics, Agriculture Sciences, Material Science and Metallurgy,

356 Earth Sciences, Mathematical Sciences, Environmental Sciences, Information

357 Technology, Social Sciences, Business and Economy and Management, Arts and

358 Humanities, Law, Telecommunications Technology, Library and Information

359 Sciences. Due to the large number of categories, we condensed anything not in

360 the top-6 into the category "Other". The top-six categories *health science*,

361 *chemistry*, *life sciences*, *engineering*, *physics* and *agricultural sciences* make up

362 74.8% of the data (Supplementary Fig 2). The assignment of categories used in

363 this study was taken from the existing index for the journal made by the

364 librarians at the DTU Library. For the temporal statistics, the years 1823-1900

365 were condensed into one.

366

367 **Pre-processing of PDF-to-text converted documents**

368 Following the PDF-to-text conversion of the Springer and Elsevier articles we ran

369 a language detection algorithm implemented in the python package langdetect

370 v1.0.7 (https://pypi.python.org/pypi/langdetect). We discarded 902,415 articles

371 that were not identified as English. We pre-processed the remaining raw text

372 from the articles as follows:

373   1. Non-printable characters were removed using the POSIX filter [[:^print:]].

374   2. A line of text was removed if digits make up more than 10% of the text, or

375   symbols make up more than 10% of the text, or lowercase text was less

376   than 50%. Symbols are anything not matching [0-9A-Za-z].

16

377   3. Removal of acknowledgements and reference- or bibliography-lists using

378   a rule-based system explained below.

379   4. Text was split into sentences and paragraphs using a rule-based system

380   described below.

381

382   We assumed that acknowledgements and reference lists are always at the end of

383   the article. Upon encountering either of the terms: "acknowledgement",

384   "bibliography", "literature cited", "literature", "references", and the following

385   misspellings thereof: "refirences", "literatur", "références", "referesces". In some

386   cases the articles had no heading indicating the start of a bibliography. We tried

387   to take these cases into account by constructing a RegEx that matches the typical

388   way of listing references (e.g. [1] Westergaard, …). Such a pattern can be

389   matched by the RegEx "^\[\d+\]\s[A-Za-z]". The other commonly used pattern,

390   "1. Westergaard, …", was avoided since it may also indicate a new heading.

391   Keywords were identified based on several rounds of manual inspection. In each

392   round, 100 articles in which the reference list had not been found was randomly

393   selected and inspected. We were unable to find references in 286,287 and

394   2,896,144 Springer and Elsevier articles, respectively. Manual inspection of 100

395   randomly selected articles revealed that these articles indeed did not have a

396   reference list or that the pattern was not easily describable with simple metrics,

397   such as keywords and RegEx. Articles without references were not discarded.

398

399   The PDF to text conversion often breaks up paragraphs and sentences, due to

400   new page, new column, etc. Paragraph and sentence splitting was performed

401   using a ruled-based system. If the previous line of text does not end with a ".!?",

402   and the current line does not start with a lower-case letter, it is assumed that the

403   line is part of the previous sentence. Otherwise, the line of text is assumed to be a

404   new paragraph.

405

406   **Text article filtering**

407   A number of Springer and Elsevier documents were removed due to technical

408   issues post pre-processing. An article was removed if:

409   1. Article contained no text post-preprocessing (51,399 documents).

410   2. Average word length was below the 2% quantile (263,902 documents).

411   3. Article contained specific keywords, described below (286,958

412   documents).

413

414   Some PDF files without texts are scans of the original article (point 1). We did

415   not attempt to make an optical character recognition conversion (OCR) as the old

416   typesetting fonts often are less compatible with present day OCR programs, and

417   this can lead to text recognition errors [33,34]. For any discarded document, we

418   still used the meta-data to calculate summary statistics. In some cases the PDF to

419   text conversion failed, and produced non-sense data with a white space between

420   the characters of a majority of the words (point 2). To empirically determine a

421   cutoff we gradually increased the cutoff and repeatedly inspected 100 randomly

422   selected articles. At the 2% quantile we saw no evidence of broken text.

423

424   Articles with the following keywords in the article were discarded: Author Index,

425   Key Word Index, Erratum, Editorial Board, Corrigendum, Announcement, Books

426   received, Product news, and Business news (point 3). These keywords were

427   found as part of the process of identifying acknowledgements and reference lists.

428   Further, any article that was available through PubMed Central was

429   preferentially selected by matching doi identifiers. This left a total of 14,549,483

430   full-text articles for further analysis.

431

432   Some articles were not separable, or were subsets of others. For instance,

433   conference proceedings may contain many individual articles in the same PDF.

434   We found 1,911,365 articles in which this was the case. In these cases we

435   removed the duplicates, or the shorter texts, but kept one copy for text mining. In

436   total, we removed 898,048 duplicate text files.

437

438   The majority of articles had a separate abstract. We matched articles from

439   PubMed Central to their respective MEDLINE abstract using the PMCID to

440   PubMed ID conversion file available from PMC. Articles from Springer and

441   Elsevier typically had a separate abstract in the meta-data. Any abstract from an

442   article that was part of the 1,911,365 articles that could not be separated was

443   removed. This led to a total of 10,376,626 abstracts for which the corresponding

444   full-text was also included downstream, facilitating a comparative analysis.

445

446 **Text mining of articles**

447 We performed text mining of the articles using a Named Entity Recognition

448 (NER) system, described earlier[35–38]. The software is open source and can be

449 downloaded from https://bitbucket.org/larsjuhljensen/tagger. The NER

450 approach is dictionary based, and thus depends on well-constructed dictionaries

451 and stop word lists. We used the gene names from the STRING dictionary v10.0

452 [35], disease names from the Disease Ontology (DO) [39] and compartment

453 names from the Gene Ontology branch cellular component [40]. Stop word lists

454 were all created and maintained in-house. Pure NER based approaches often

455 struggles with ambiguity of words. Therefore, we included additional

456 dictionaries that we do not report the results from. If any identified term was

457 found in multiple dictionaries, it was discarded due to ambiguity. The additional

458 dictionaries include small molecule names from STITCH [41], tissue names from

459 the Brenda Tissue Ontology [42], Gene Ontology biological process and

460 molecular function [40], and the mammalian phenotype ontology [43]. The latter

461 is a modified version made to avoid clashes with the disease ontology. The

462 dictionaries can be downloaded from http://download.jensenlab.org/.

463

464 In the cases where the dictionary was constructed from an ontology co-

465 occurrences were backtracked through all parents. E.g. the term type 1 diabetes

466 mellitus from the Disease Ontology is backtracked to its parent, diabetes

467 mellitus, then to glucose metabolism disease, etc.

468

469 Co-occurrences were scored using the scoring system described in [44]. In short,

470 a weighted count for each pair of entities (e.g. disease-gene) was calculated using

471 the formula,

472
$$C(i,j) = \sum_{k=1}^{n} w_d \delta_{dk}(i,j) + w_p \delta_{pk}(i,j) + w_s \delta_{sk}(i,j)$$

473   (1)

474 where δ is an indicator function taking into account whether the terms i,j co-

475 occur within the same document (d), paragraph (p), or sentence (s). w is the co-

476 occurrence weight here set to 1.0, 2.0, and 0.2, respectively. Based on the

477 weighted count, the score $S(i,j)$ was calculated as,

478
$$S(i,j) = C_{ij}^{\alpha} \left(\frac{C_{ij}C_{..}}{C_{i.}C_{.j}}\right)^{1-\alpha}$$

479   (2)

480 where α is set to 0.6. All weights were optimized using the KEGG pathway maps

481 as benchmark (described further below). The S scores were converted to Z

482 scores, as described earlier [45].

483

484 **Benchmarking of associations**

485 PPIs were benchmarked using pathway maps from the KEGG database [46]. Any

486 two proteins in the same pathway were set to be a positive example, and any two

487 proteins present in at least one pathway, but not the same, were set as a negative

21

488 example. This approach assumes that the pathways are near complete and

489 includes all relevant proteins. The same approach has been used for the STRING

490 database [44]. The disease–gene benchmarking set was created by setting the

491 disease-gene associations from UniProt [47] and Genetics Home Reference

492 (https://ghr.nlm.nih.gov/, accessed 23th March 2017) as positive examples. The

493 positive examples were then shuffled, and the shuffled examples were set as

494 negative examples. Shuffled examples that ended up overlapping with the

495 positive examples were discarded. This approach has previously been described

496 [36]. The protein–compartment benchmark set was created by extracting the

497 compartment information for each protein from UniProt and counting these as

498 positive examples. For every protein found in at least one compartment, all

499 compartments where it was not found were set as negative examples. The same

500 approach has been used previously [38].

501

502 Receiver Operating Characteristic (ROC) curves were created by gradually

503 increasing the *Z*-score and calculating the True Positive Rate (TPR) and False

504 Positive Rate (FPR), as described in eqs. (3) and (4).

505

506
$$TPR = \frac{True\ Positives}{Positives}$$

507 (3)

508

$$FPR = \frac{True\ Negatives}{Negatives}$$

510   (4)

511   We compare the ROC curves by the Area Under the Curve (AUC), a metric

512   ranging from 0 to 1.

513

514

515   **Acknowlegements**

519 **References**

520

521 1.   Azevedo A. Integration of Data Mining in Business Intelligence Systems.
522   1st Editio. Azevedo A, Santos MF, editors. Integration of Data Mining in
523   Business Intelligence Systems. IGI Publishing Hershey, PA, USA; 2014. 314
524   p.
525 2.   Krallinger M, Valencia A. Text-mining and information-retrieval services
526   for molecular biology. Vol. 6, Genome biology. 2005.6(7):224.
527 3.   Fleuren WWM, Alkema W. Application of text mining in the biomedical
528   domain. Vol. 74, Methods. 2015.74:97–106.
529 4.   Luo Y, Riedlinger G, Szolovits P. Text Mining in Cancer Gene and Pathway
530   Prioritization. Vol. 13, Cancer Informatics. 2014.13(Suppl 1):69–79.
531 5.   Ananiadou S, Thompson P, Nawaz R, McNaught J, Kell DB. Event-based text
532   mining for biology and functional genomics. Vol. 14, Briefings in functional
533   genomics. 2015.14(3):213–30.
534 6.   Hoffmann R, Krallinger M, Andres E, Tamames J, Blaschke C, Valencia A.
535   Text mining for metabolic pathways, signaling cascades, and protein
536   networks. Vol. 283/pe21, Sci. STKE. 2005.283/pe21:e21.
537 7.   Liu F, Chen J, Jagannatha A, Yu H. Learning for Biomedical Information
538   Extraction: Methodological Review of Recent Advances. arXiv:1606.07993
539   [cs]. 2016. Cited 20 June 2017.
540 8.   Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text
541   mining, information extraction, and retrieval applications for biology. Vol.
542   9 Suppl 2, Genome biology. 2008.9 Suppl 2(Suppl 2):S8.
543 9.   Gonzalez GH, Tahsin T, Goodale BC, Greene AC, Greene CS. Recent
544   advances and emerging applications in text and data mining for
545   biomedical discovery. Vol. 17, Briefings in Bioinformatics. 2016.17(1):33–
546   42.
547 10. Rebholz-Schuhmann D, Oellrich A, Hoehndorf R. Text-mining solutions for
548   biomedical research: enabling integrative biology. Vol. 13, Nature Reviews
549   Genetics. 2012.13(12):829–39.
550 11. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards
551   better research applications and clinical care. Vol. 13, Nature Reviews
552   Genetics. 2012.13(6):395–405.
553 12. Rodriguez-Esteban R, Bundschus M. Text mining patents for biomedical
554   knowledge. Vol. 21, Drug Discovery Today. 2016.21(6):997–1002.
555 13. Simmons M, Singhal A, Lu Z. Text mining for precision medicine: Bringing
556   structure to ehrs and biomedical literature to understand genes and
557   health. In: Vol. 939, Advances in Experimental Medicine and Biology.
558   Springer Singapore; 2016. p. 139–66.
559 14. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from
560   information retrieval to biological discovery. Vol. 7, Nature reviews.
561   Genetics. 2006.7(2):119–29.
562 15. Winnenburg R, Wächter T, Plake C, Doms A, Schroeder M. Facts from text:
563   Can text mining help to scale-up high-quality manual curation of gene
564   products with ontologies? Vol. 9, Briefings in Bioinformatics.

565   2008.9(6):466–78.

566 16. Wei C-H, Kao H-Y, Lu Z. Text mining tools for assisting literature curation.
567   In: Proceedings of the 5th ACM Conference on Bioinformatics,
568   Computational Biology, and Health Informatics - BCB '14 [Internet]. New
569   York, New York, USA: ACM Press; 2014. p. 590–1.

570 17. Cohen KB, Johnson HL, Verspoor K, Roeder C, Hunter LE, Verspoor K, et al.
571   The structural and content aspects of abstracts versus bodies of full text
572   journal articles are different. Vol. 11, BMC Bioinformatics. 2010.11(1):492.

573 18. Papanikolaou N, Pavlopoulos GA, Theodosiou T, Iliopoulos I. Protein-
574   protein interaction predictions using text mining methods. Vol. 74,
575   Methods. 2015.74:47–53.

576 19. Samuel J, Yuan X, Yuan X, Walton B. Mining online full-text literature for
577   novel protein interaction discovery. In: 2010 IEEE International
578   Conference on Bioinformatics and Biomedicine Workshops, BIBMW 2010
579   [Internet]. IEEE; 2010. p. 277–82.

580 20. Garten Y, Altman R. Pharmspresso: a text mining tool for extraction of
581   pharmacogenomic concepts and relationships from full text. Vol. 10, BMC
582   bioinformatics. 2009.10(Suppl 2):S6.

583 21. Müller HM, Kenny EE, Sternberg PW. Textpresso: An ontology-based
584   information retrieval and extraction system for biological literature. Vol. 2,
585   PLoS Biology. 2004.2(11):e309.

586 22. Martin EPG, Bremer EG, Guerin M-C, DeSesa C, Jouve O. Analysis of
587   protein/protein interactions through biomedical literature: Text mining of
588   abstracts vs. text mining of full text articles. In: Vol. 3303, Knowledge
589   Exploration in Life Science Informatics. Springer, Berlin, Heidelberg; 2004.
590   p. 96–108.

591 23. Corney DPA, Buxton BF, Langdon WB, Jones DT. BioRAT: extracting
592   biological information from full-length papers. Vol. 20, Bioinformatics.
593   2004.20(17):3206–13.

594 24. Blake C. Beyond genes, proteins, and abstracts: Identifying scientific claims
595   from full-text biomedical articles. Vol. 43, Journal of Biomedical
596   Informatics. 2010.43(2):173–89.

597 25. Constantin A, Pettifer S, Voronkov A. Pdfx. Proceedings of the 2013 ACM
598   symposium on Document engineering - DocEng '13. 2013.:177.

599 26. Luong M-T, Nguyen TD, Kan M-Y. Logical Structure Recovery in Scholarly
600   Articles with Rich Document Features. Vol. 1, International Journal of
601   Digital Library Systems. 2012.1(4):1–23.

602 27. Ramakrishnan C, Patnia A, Hovy E, Burns GAPC. Layout-aware text
603   extraction from full-text PDF of scientific articles. Vol. 7, Source Code for
604   Biology and Medicine. 2012.7(1):7.

605 28. Adams J. Collaborations: The rise of research networks. Vol. 490, Nature.
606   2012.490(7420):335–6.

607 29. Eckhouse S, Lewison G, Sullivan R. Trends in the global funding and
608   activity of cancer research. Vol. 2, Molecular Oncology. 2008.2(1):20–32.

609 30. Plaven-Sigray P, Matheson GJ, Schiffler BC, Thompson WH. The Readability
610   Of Scientific Texts Is Decreasing Over Time. bioRxiv. 2017.:119370.

611 31. Mallory EK, Zhang C, Ré C, Altman RB. Large-scale extraction of gene
612   interactions from full-text literature using DeepDive. Vol. 32,
613   Bioinformatics. 2015.32(1):106–13.

614   32.   Fluck J, Hofmann-Apitius M. Text mining for systems biology. Vol. 19, Drug
615   Discovery Today. 2014.19(2):140–4.
616   33.   Thompson P, Batista-Navarro RT, Kontonatsios G, Carter J, Toon E,
617   McNaught J, et al. Text mining the history of medicine. Rocha LM, editor.
618   Vol. 11, PLoS ONE. 2016.11(1):e0144717.
619   34.   Lopresti D. Optical character recognition errors and their effects on
620   natural language processing. Vol. 12, International Journal on Document
621   Analysis and Recognition. 2009.12(3):141–51.
622   35.   Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas
623   J, et al. STRING v10: Protein-protein interaction networks, integrated over
624   the tree of life. Vol. 43, Nucleic Acids Research. 2015.43(D1):D447–52.
625   36.   Pletscher-Frankild S, Pallej?? A, Tsafou K, Binder JX, Jensen LJ. DISEASES:
626   Text mining and data integration of disease-gene associations. Vol. 74,
627   Methods. 2015.74:83–9.
628   37.   Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ.
629   Comprehensive comparison of large-scale tissue expression datasets. Vol.
630   3, PeerJ. 2015.3:e1054.
631   38.   Binder JX, Pletscher-Frankild S, Tsafou K, Stolte C, O'Donoghue SI,
632   Schneider R, et al. COMPARTMENTS: Unification and visualization of
633   protein subcellular localization evidence. Vol. 2014, Database.
634   2014.2014(0):bau012-bau012.
635   39.   Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, et al.
636   Disease ontology: A backbone for disease semantic integration. Vol. 40,
637   Nucleic Acids Research. 2012.40(D1):D940-6.
638   40.   Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
639   Ontology: Tool for The Unification of Biology. Vol. 25, Nature Genetics.
640   2000.25(1):25–9.
641   41.   Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5:
642   Augmenting protein-chemical interaction networks with tissue and
643   affinity data. Vol. 44, Nucleic Acids Research. 2016.44(D1):D380–4.
644   42.   Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The
645   BRENDA Tissue Ontology (BTO): The first all-integrating ontology of all
646   organisms for enzyme sources. Vol. 39, Nucleic Acids Research.
647   2011.39(SUPPL. 1):D507–13.
648   43.   Smith CL, Eppig JT. The mammalian phenotype ontology: Enabling robust
649   annotation and comparative analysis. Vol. 1, Wiley Interdisciplinary
650   Reviews: Systems Biology and Medicine. 2009.1(3):390–9.
651   44.   Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, et
652   al. STRING v9.1: Protein-protein interaction networks, with increased
653   coverage and integration. Vol. 41, Nucleic Acids Research.
654   2013.41(D1):D808-15.
655   45.   Mørk S, Pletscher-Frankild S, Caro AP, Gorodkin J, Jensen LJ. Protein-
656   driven inference of miRNA-disease associations. Vol. 30, Bioinformatics.
657   2014.30(3):392–7.
658   46.   Wasmuth E V, Lima CD. OUP accepted manuscript. Vol. 45, Nucleic Acids
659   Research. 2016.45(D1):1–15.
660   47.   Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, et al.
661   UniProt: A hub for protein information. Vol. 43, Nucleic Acids Research.
662   2015.43(D1):D204–12.

Supporting Information Captions

663

664 **S1 Fig**: Number of publications per year on the log scale.

665

666 **S2 Fig**: Category overview across all journals and years. The bar chart indicates

667 the frequency, whilst the line is the cumulative sum. The first six categories

668 contribute 74,8%. Due to the large number of categories, anything outside the

669 top-6 was condensed into the joint category "Other".

670

671 **S3 Fig**: Temporal trend for the categories embedded in the "Other" category. We

672 note that the category has grown as a whole, but that the growth is not tied to

673 one category.

674

675 **S4 Fig**: Benchmarking the four different corpora not using a document weight.

676 (a-c) The increase in performance has fallen, compared to including a document

677 weight. In one case, protein-protein associations, the MEDLINE corpus

678 outperforms the full text articles.

679

680 **S5 Fig**: Benchmarking the four different corpora at a low false positive rate not

681 using a document weight. The increase in performance has fallen. In one case, for

682 protein-protein associations, the MEDLINE corpus outperforms the full text

683 articles.

684

685   **S1 Table**: The top 15 journals in the corpora.

686

687   **S2 Table:** Area under the curve (AUC) for the four different corpora, with and

688   without document weight for scoring co-occurrences.

689

690   **S3 Table:** True Positive Rate at 10% False Positive Rate (TPR@10%FPR) for the

691   four different corpora, with and without document weight for scoring co-

692   occurrences.

693
694

Fig 1
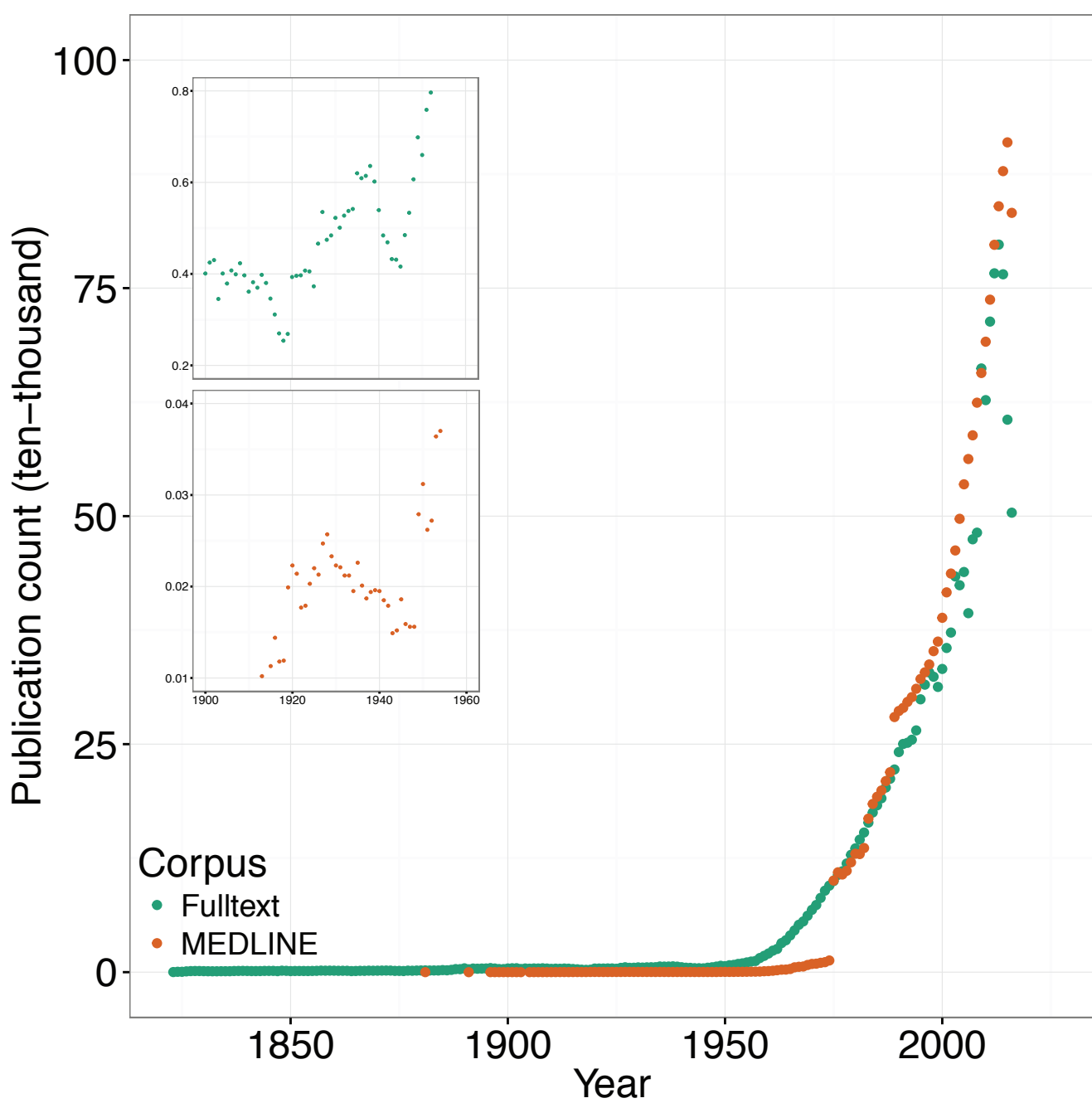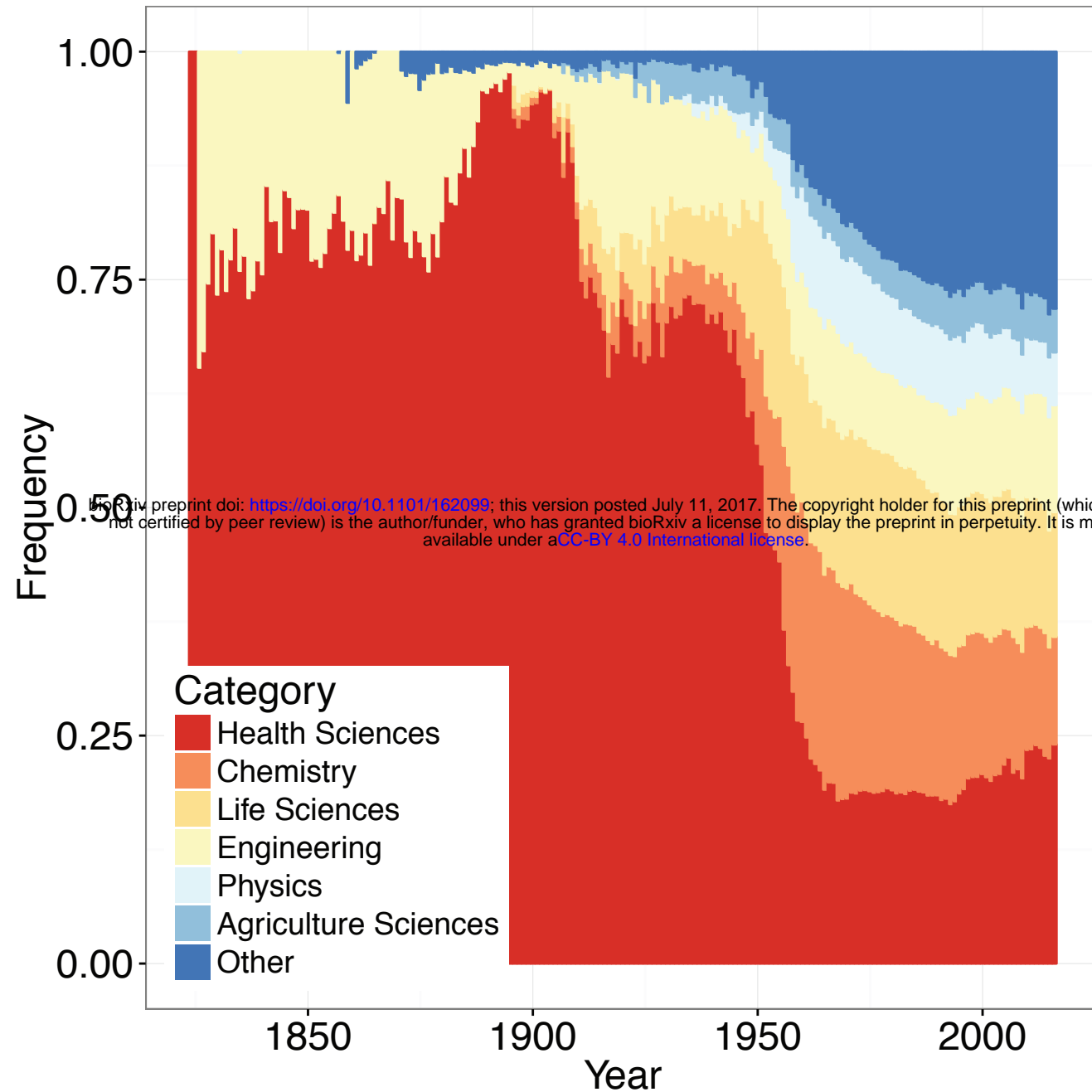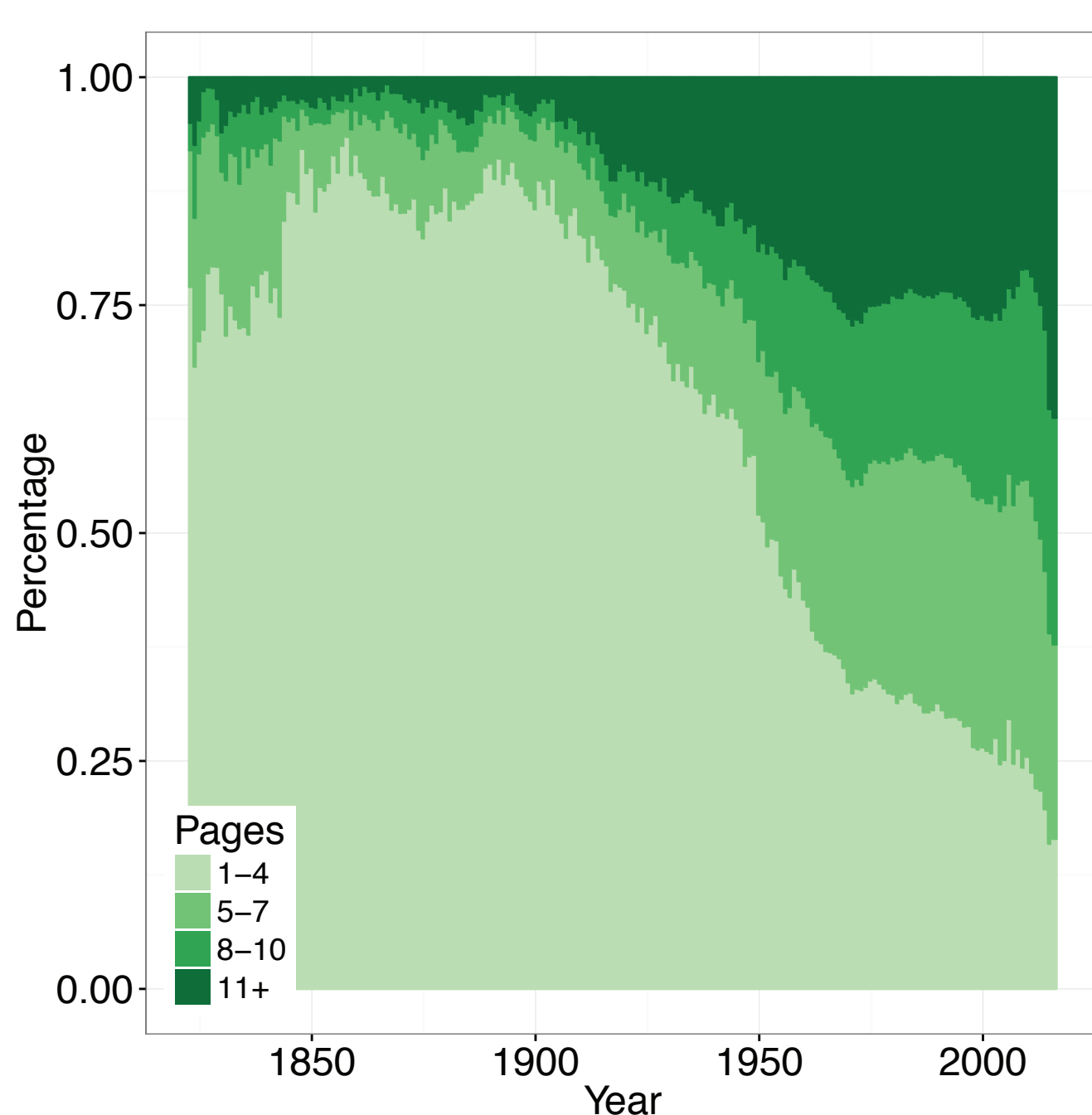
Fig 2

a

Disease-Gene Associations

- All Fulltexts
- MEDLINE
- Core Fulltexts
- Core Abstracts

b

Protein-Protein Associations

c
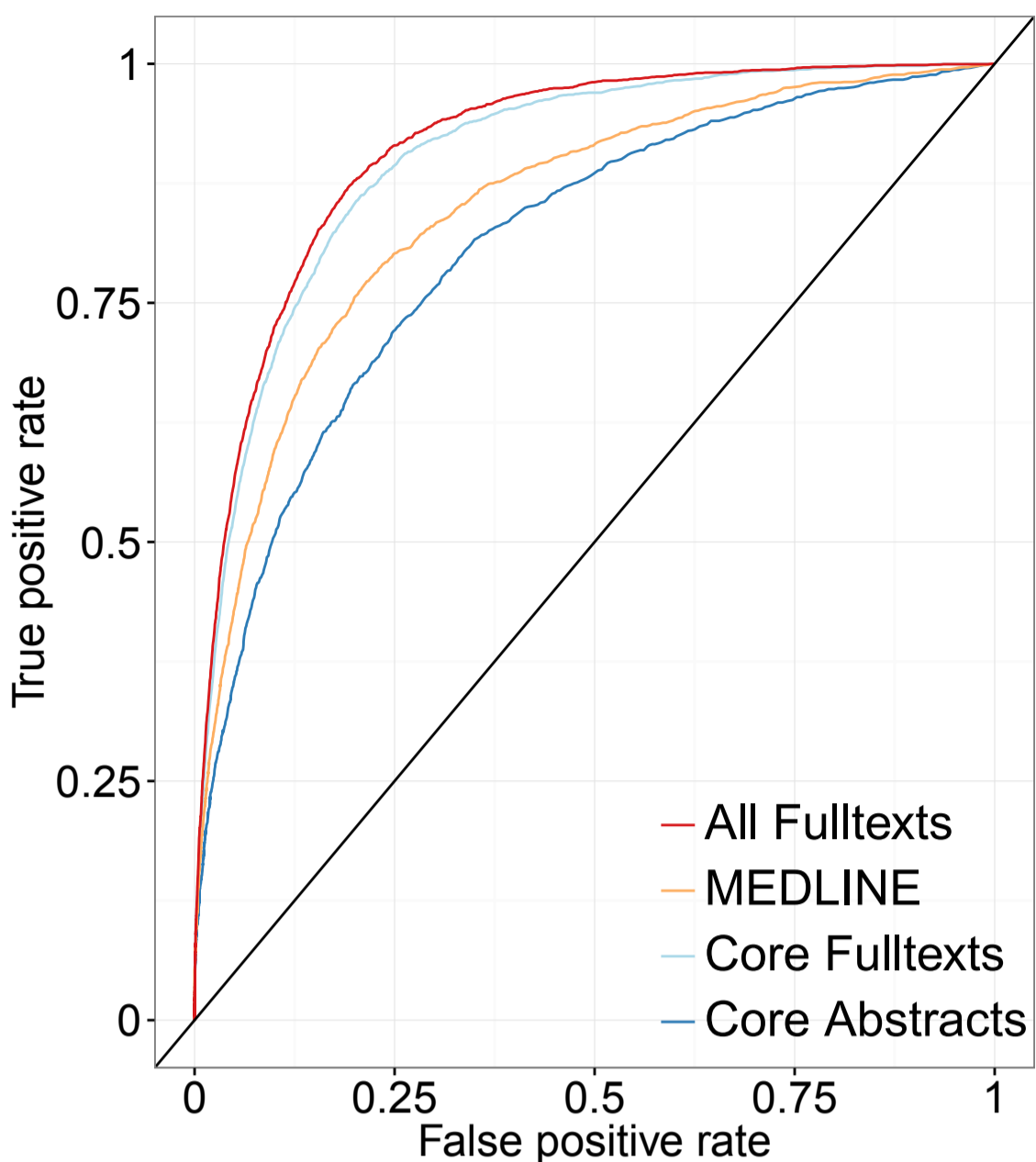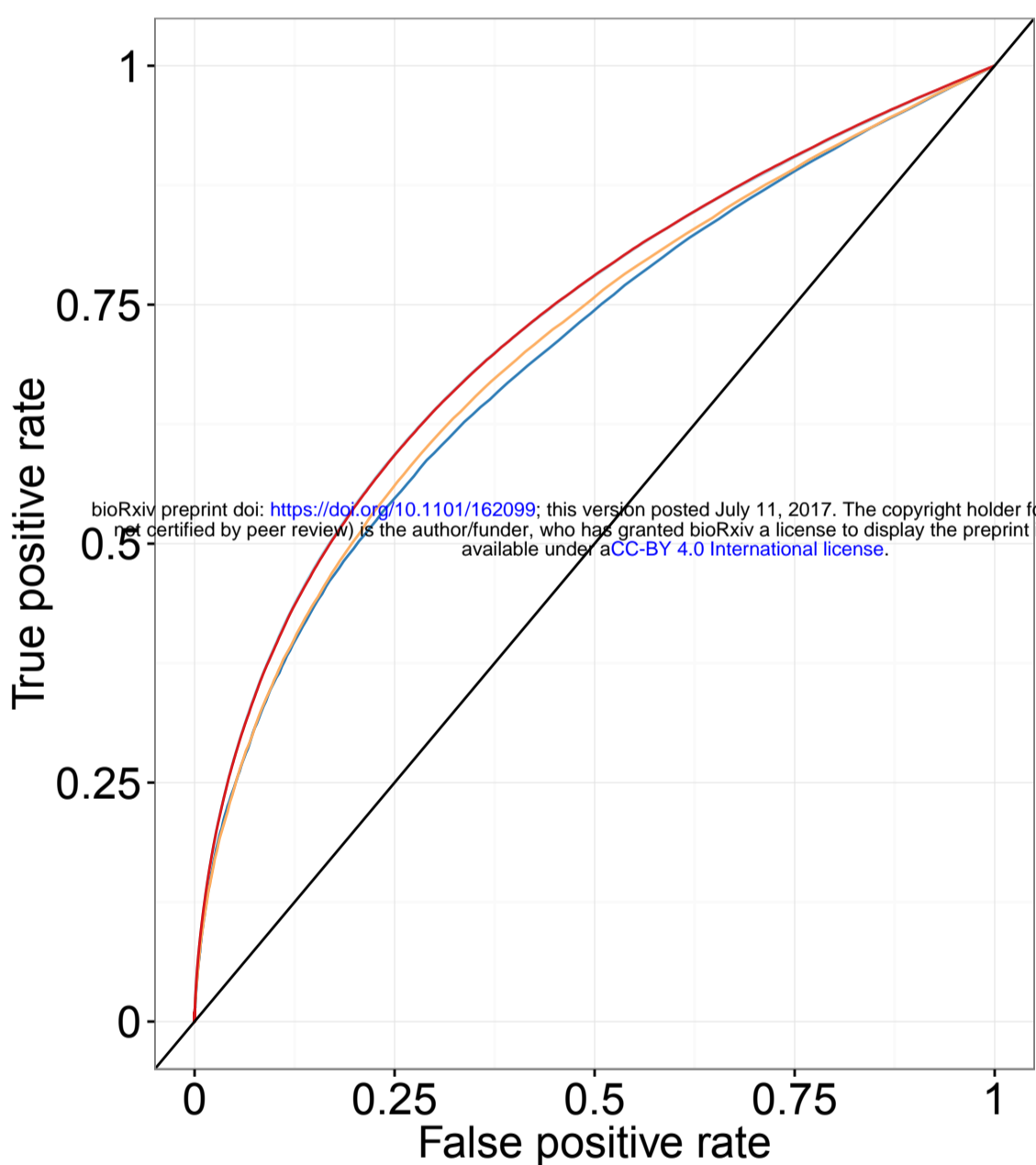
Protein-Compartment Associations

Fig 3

a

Disease-Gene Associations

- All Fulltexts
- MEDLINE
- Core Fulltexts
- Core Abstracts

b

Protein-Protein Associations

c
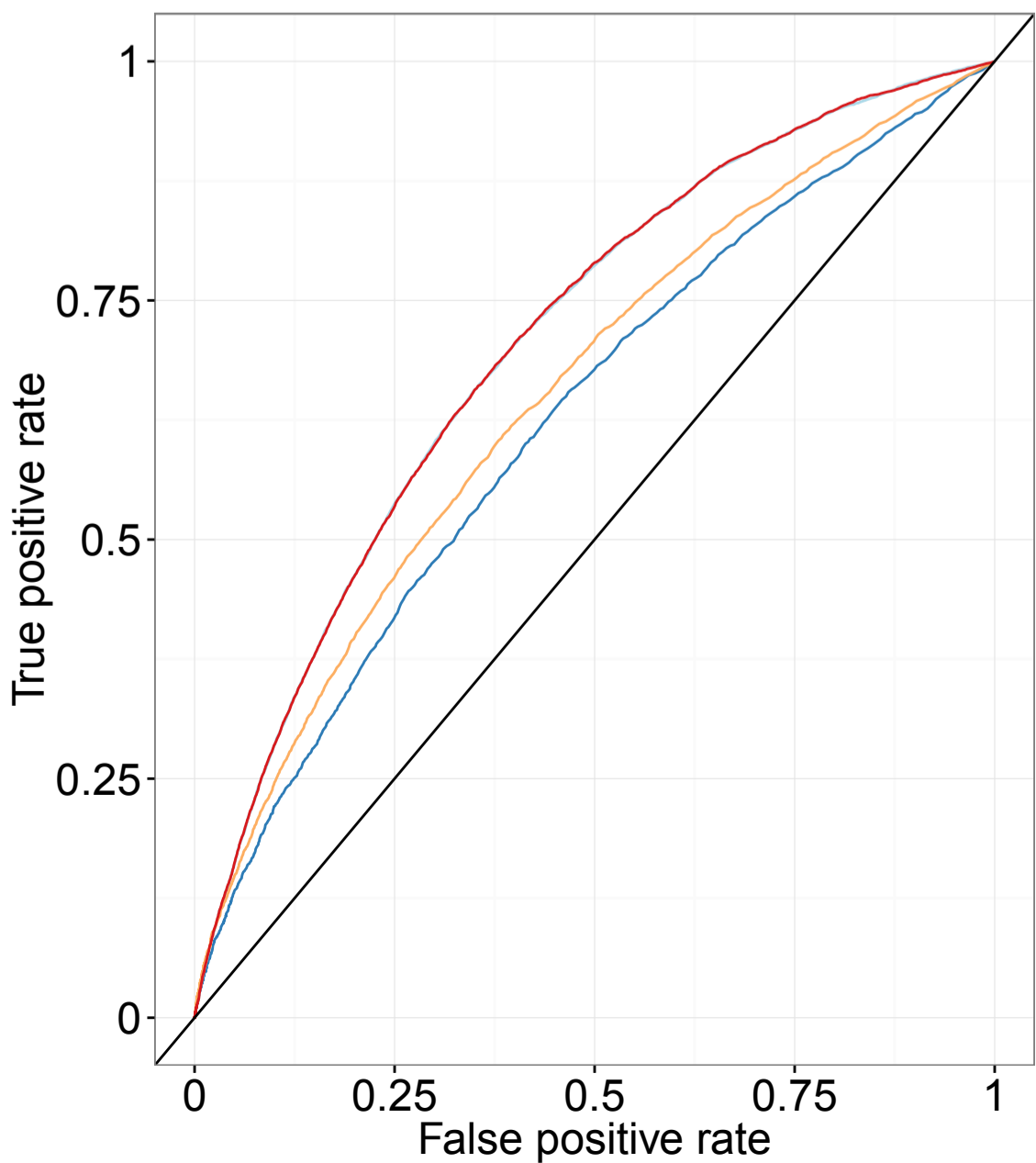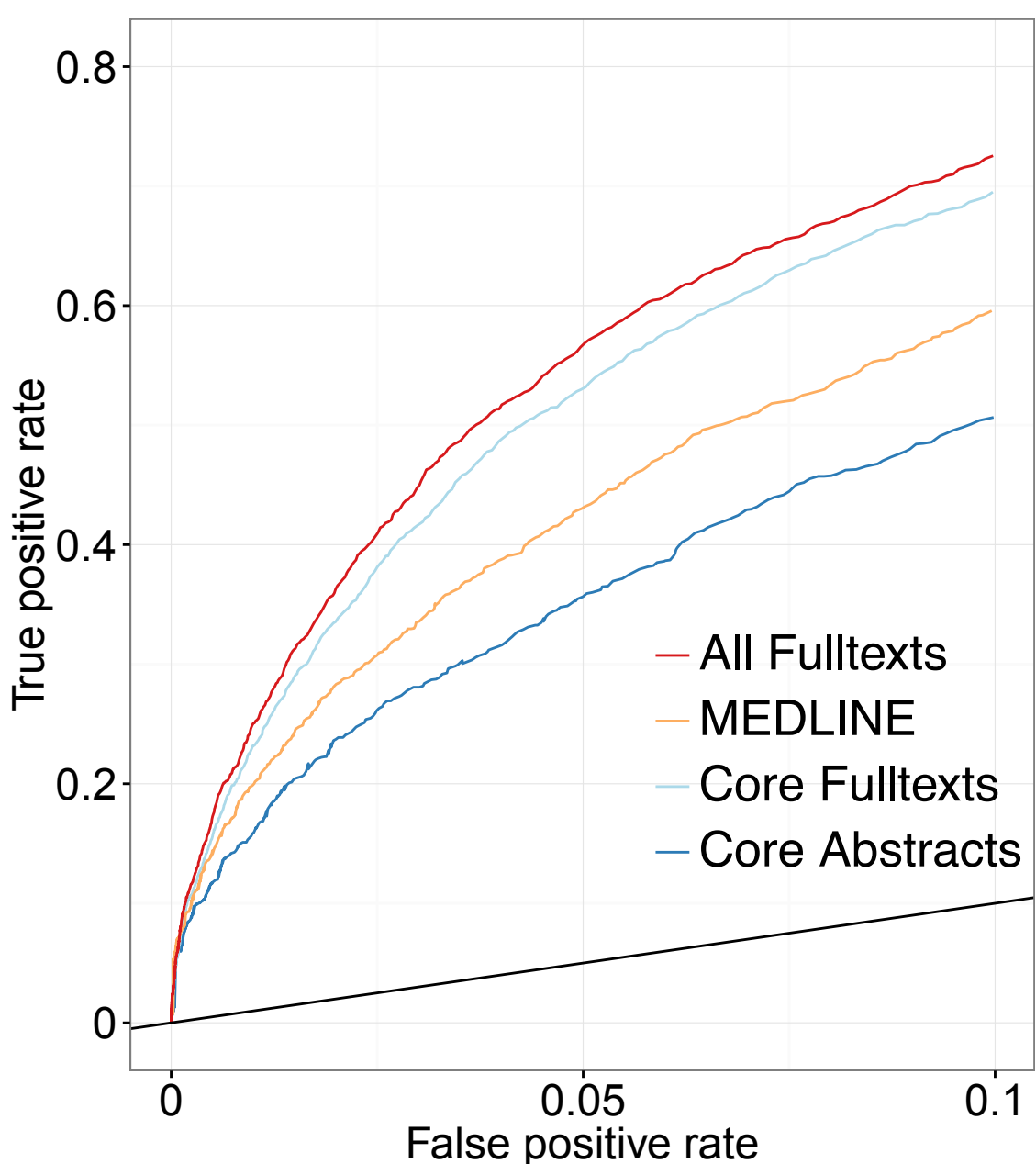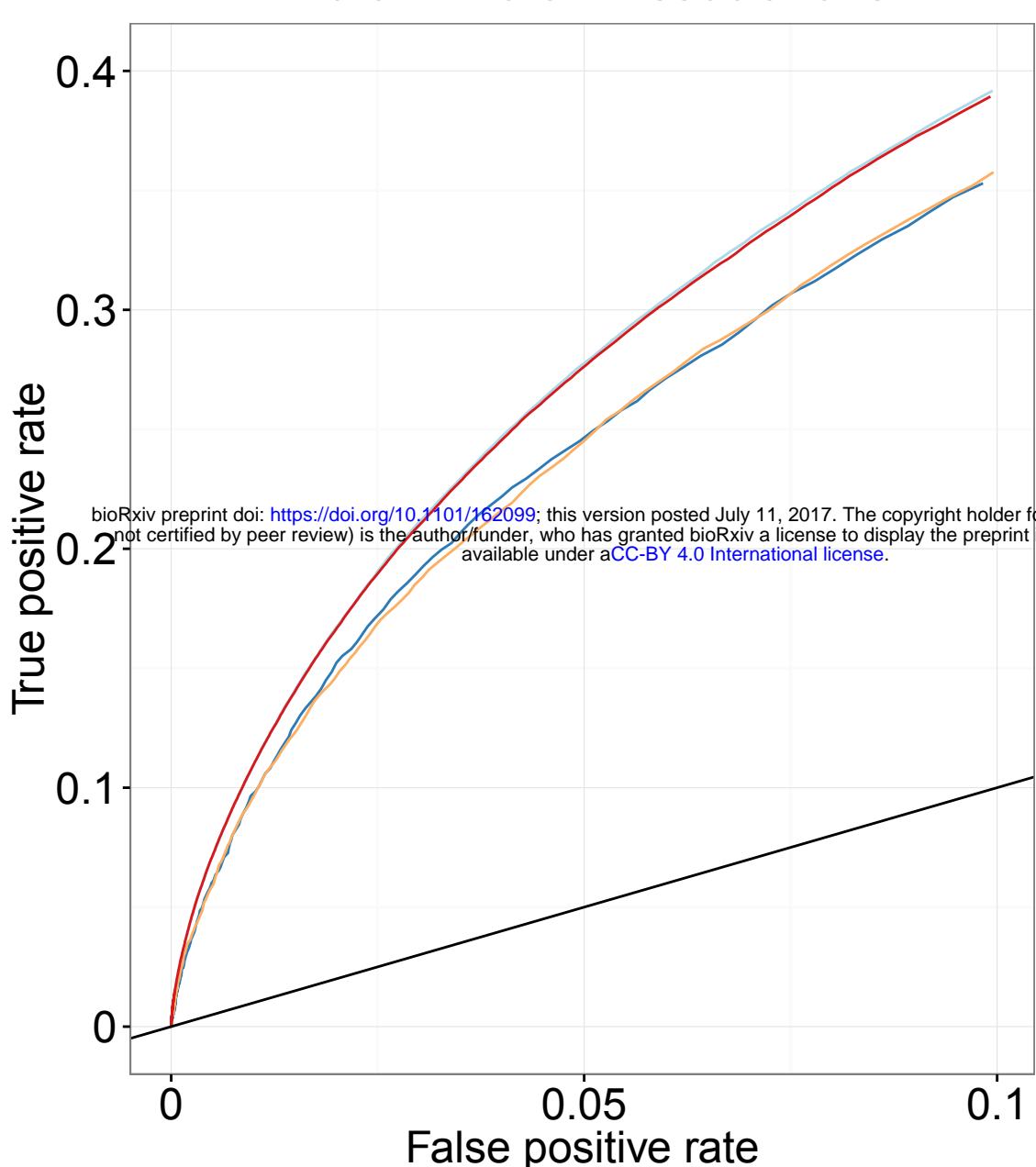
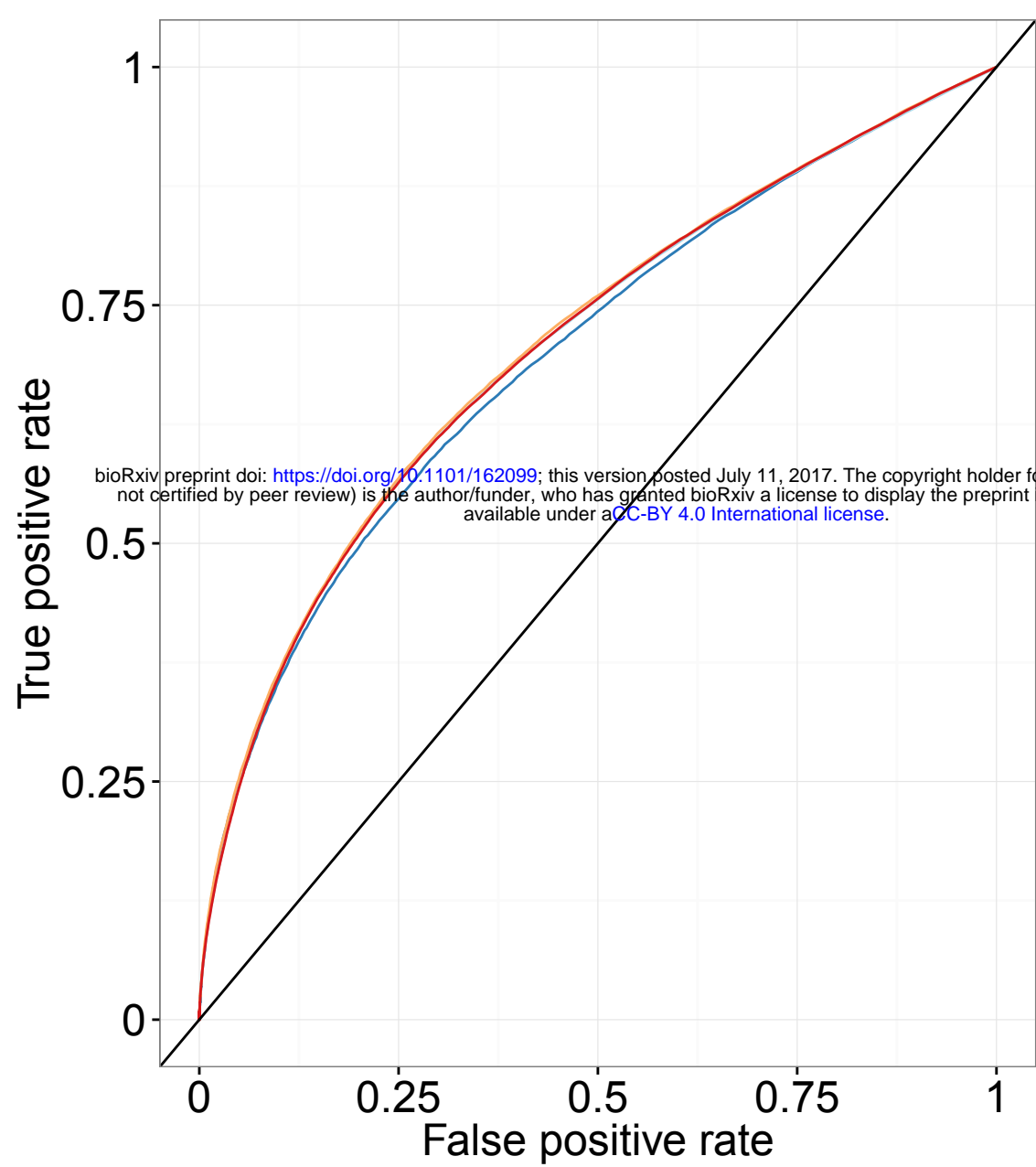Protein-Compartment Associations
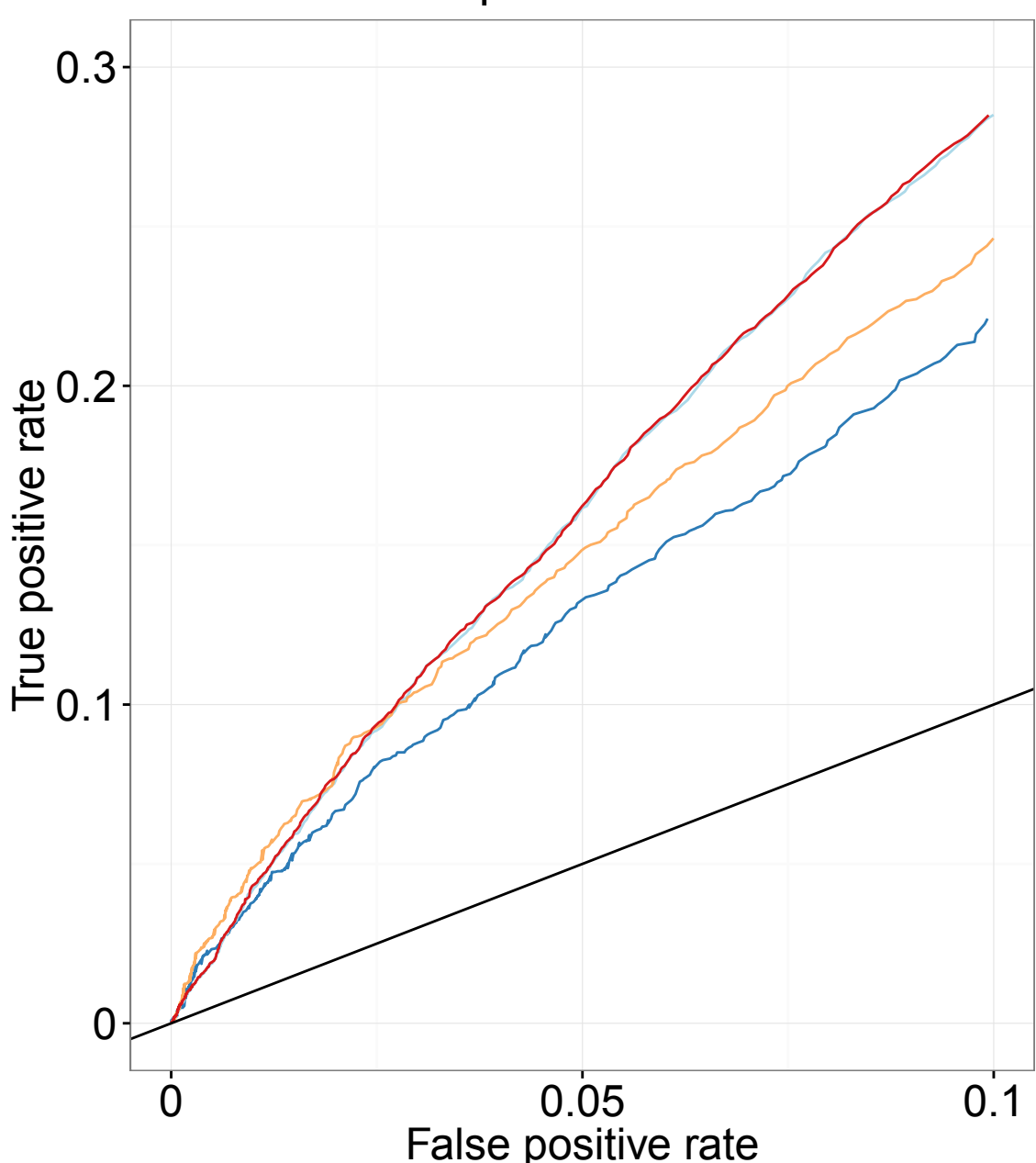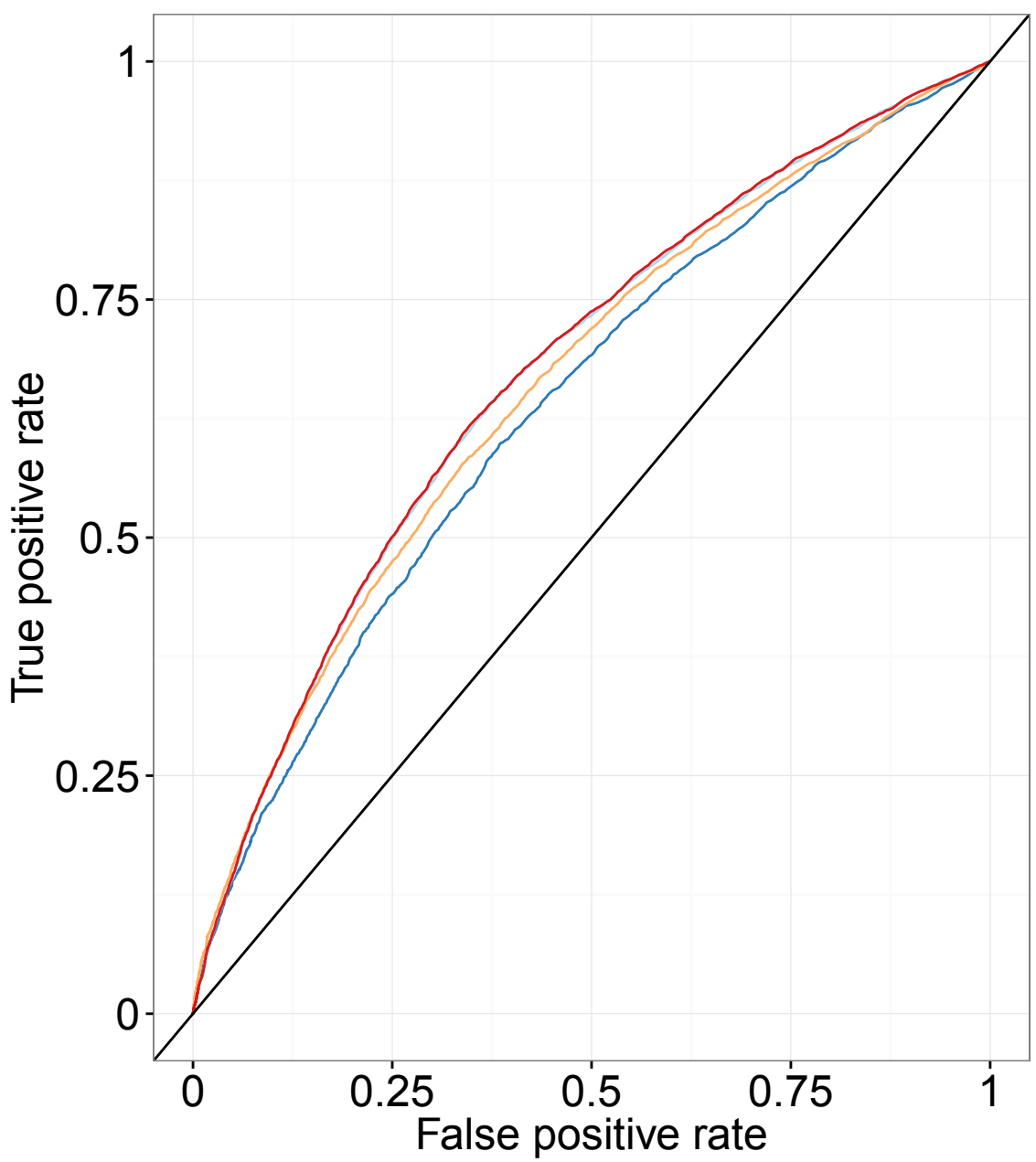
S1 Fig
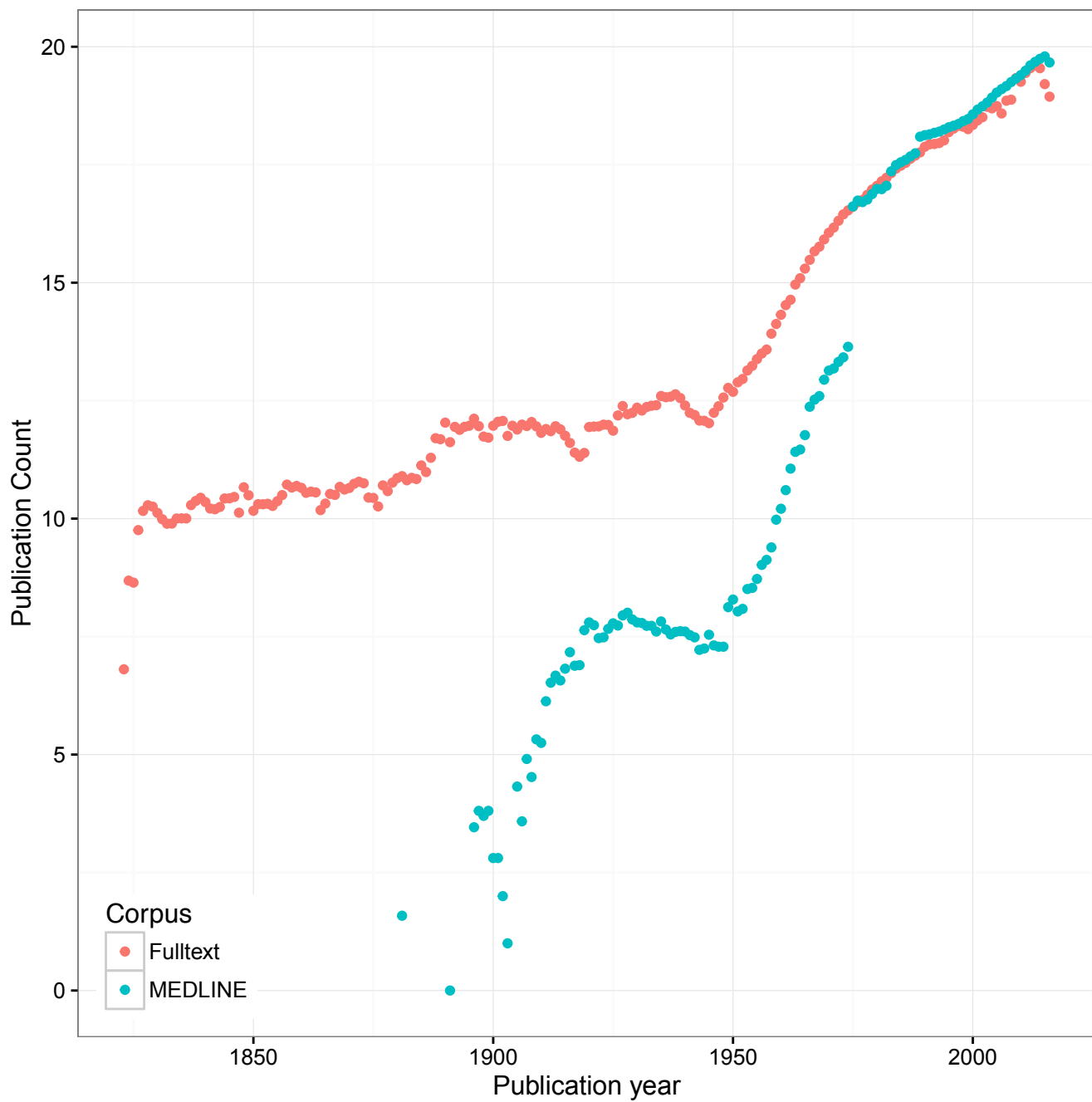
S2 Fig
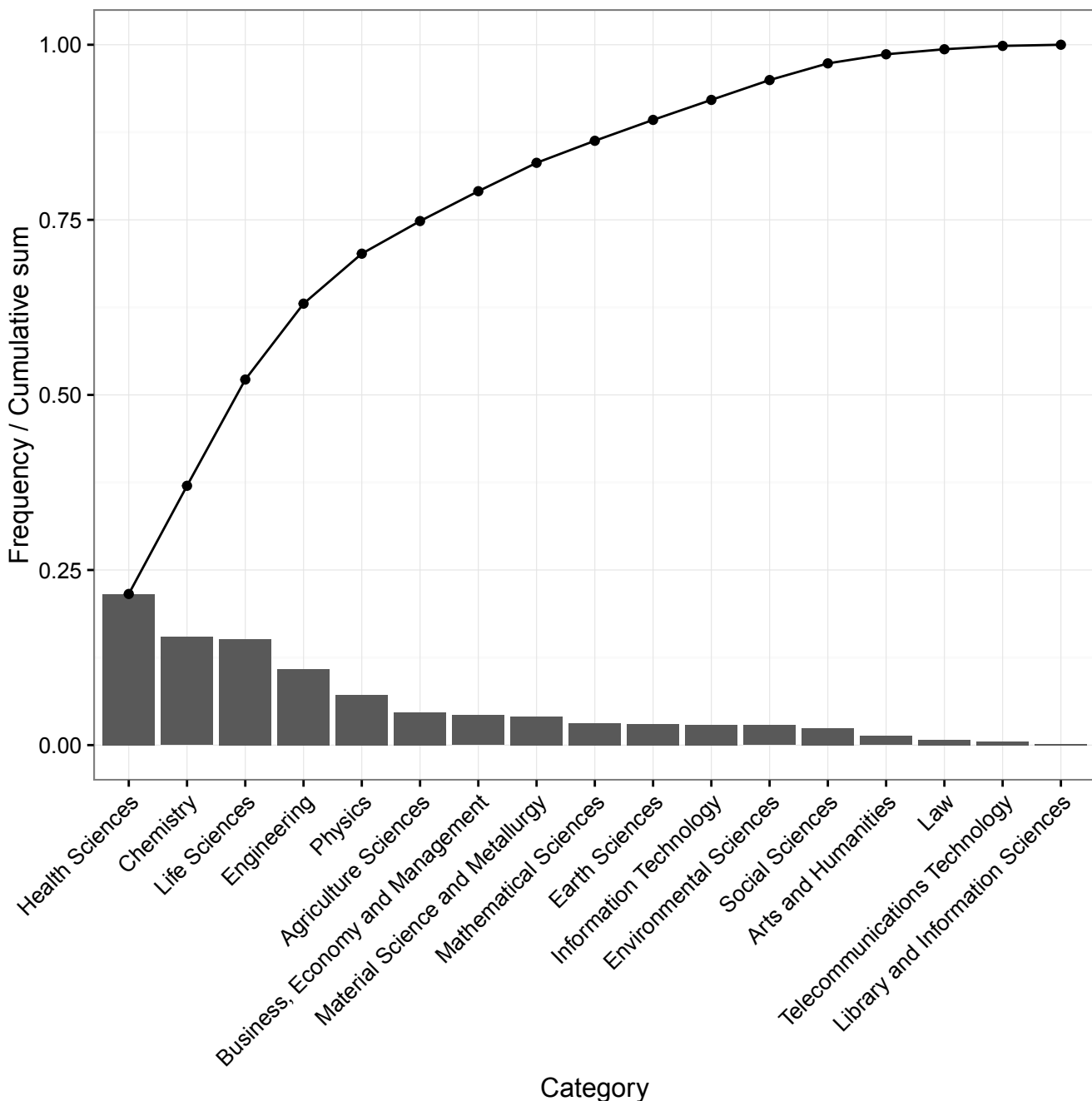
Supplementary Figure 3

S4 Fig

a

b

c

S5 Fig

a

b

c