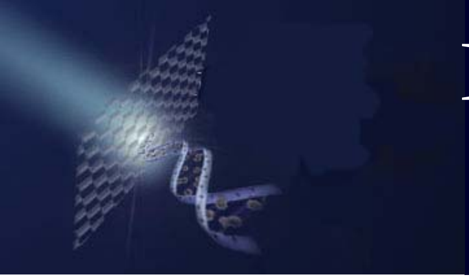
The Characteristics of Highly Similar Scientific Publications

Skip Garner reporting for the team of quantitative ethics researchers on work supported by ORI/NLM.

Virginia Bioinformatics Institute

Virginia Tech

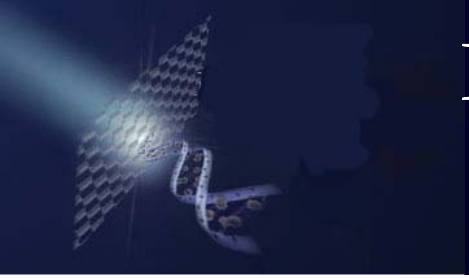


If you remember only one thing from this talk, remember ‘etblast.org’

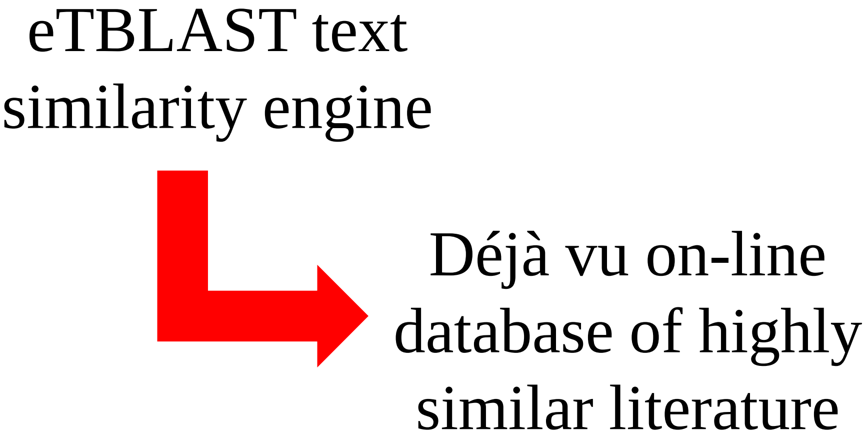
eTBLAST text similarity engine

Alternative approach to accessing the literature

Freely available for 10 years…
used by scientists, editors, reviewers 1,000s of times a day



If you remember only one thing from this talk, remember ‘etblast.org’

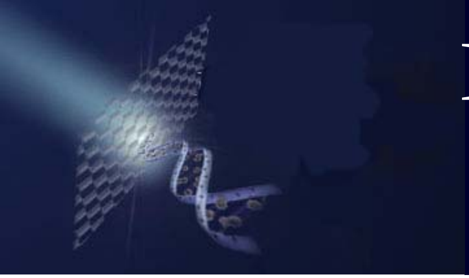


eTBLAST text similarity engine

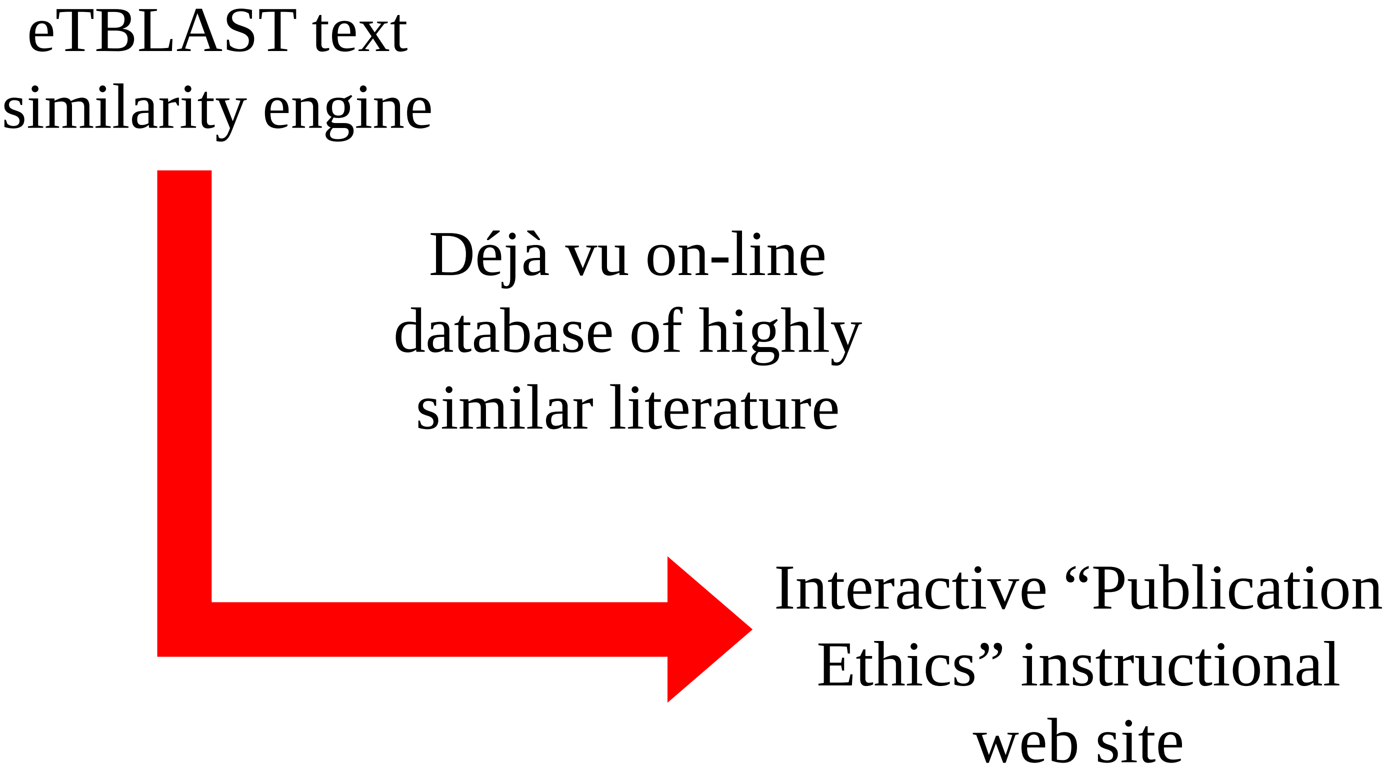
Déjà vu on-line database of highly similar literature

~80,000 pairs of entries

Dynamic and accessed 100 to 15,000 times a day



If you remember only one thing from this talk, remember ‘etblast.org’

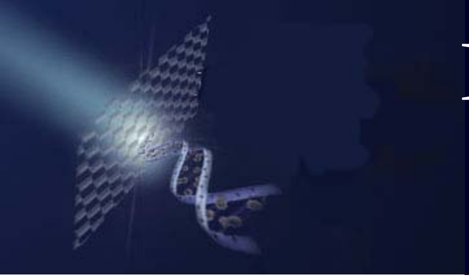


eTBLAST text similarity engine

Déjà vu on-line database of highly similar literature

Interactive “Publication Ethics” instructional web site

Targets professionals, editors, reviewers



If you remember only one thing from this talk, remember ‘etblast.org’

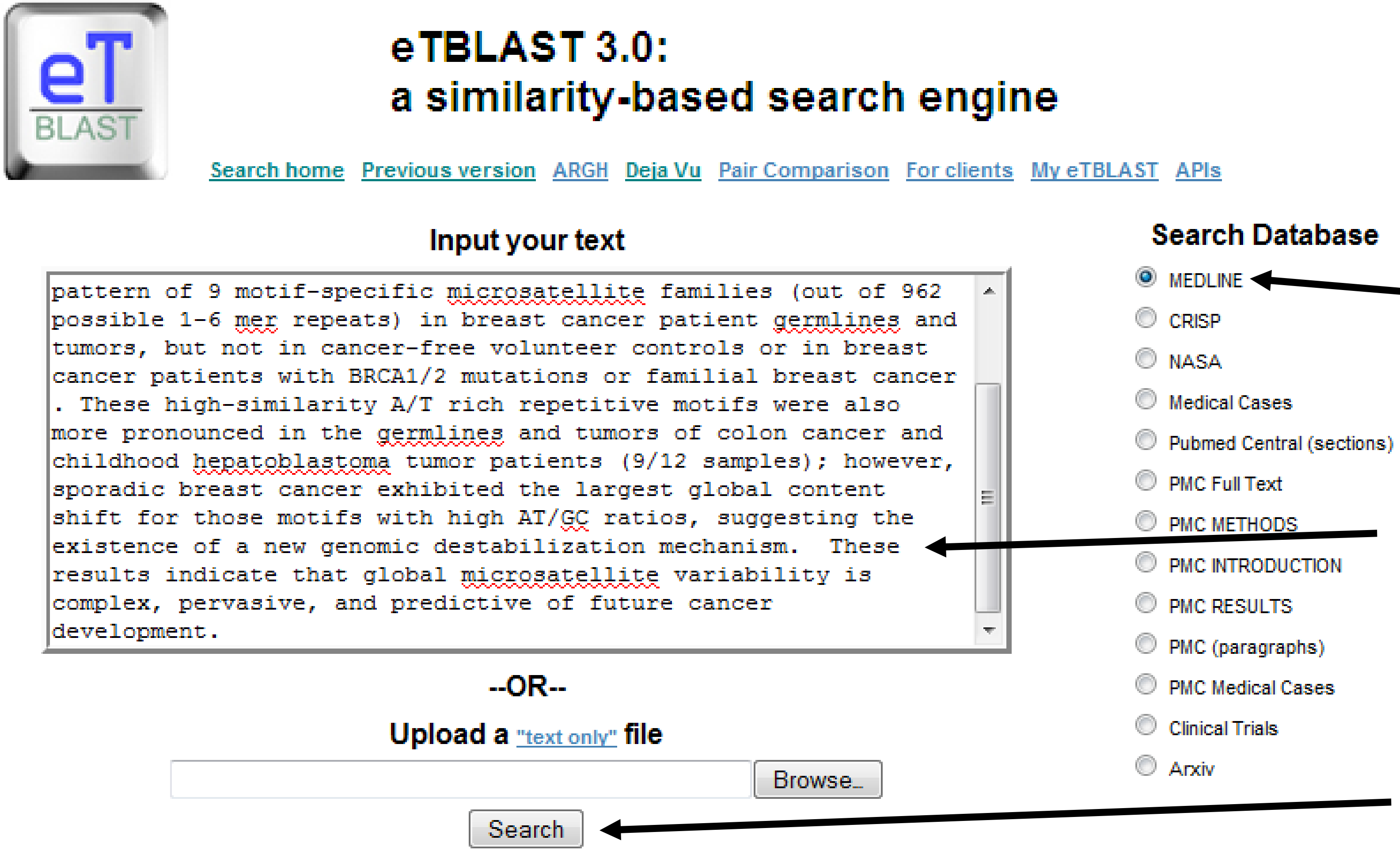
eTBLAST text similarity engine

Heliotext’s ultra-secure turnkey text analytics implementations for business intelligence, marketing, contract/grant evaluation, meeting organization

Déjà vu on-line database of highly similar literature

Interactive “Publication Ethics” instructional web site

eTBLAST, compares a query to a number of text databases.



eTBLAST 3.0:
a similarity-based search engine

Input your text

pattern of 9 motif-specific microsatellite families (out of 962 possible 1-6 mer repeats) in breast cancer patient germlines and tumors, but not in cancer-free volunteer controls or in breast cancer patients with BRCA1/2 mutations or familial breast cancer . These high-similarity A/T rich repetitive motifs were also more pronounced in the germlines and tumors of colon cancer and childhood hepatoblastoma tumor patients (9/12 samples); however, sporadic breast cancer exhibited the largest global content shift for those motifs with high AT/GC ratios, suggesting the existence of a new genomic destabilization mechanism. These results indicate that global microsatellite variability is complex, pervasive, and predictive of future cancer development.

--OR--

Upload a "text only" file

Search Database

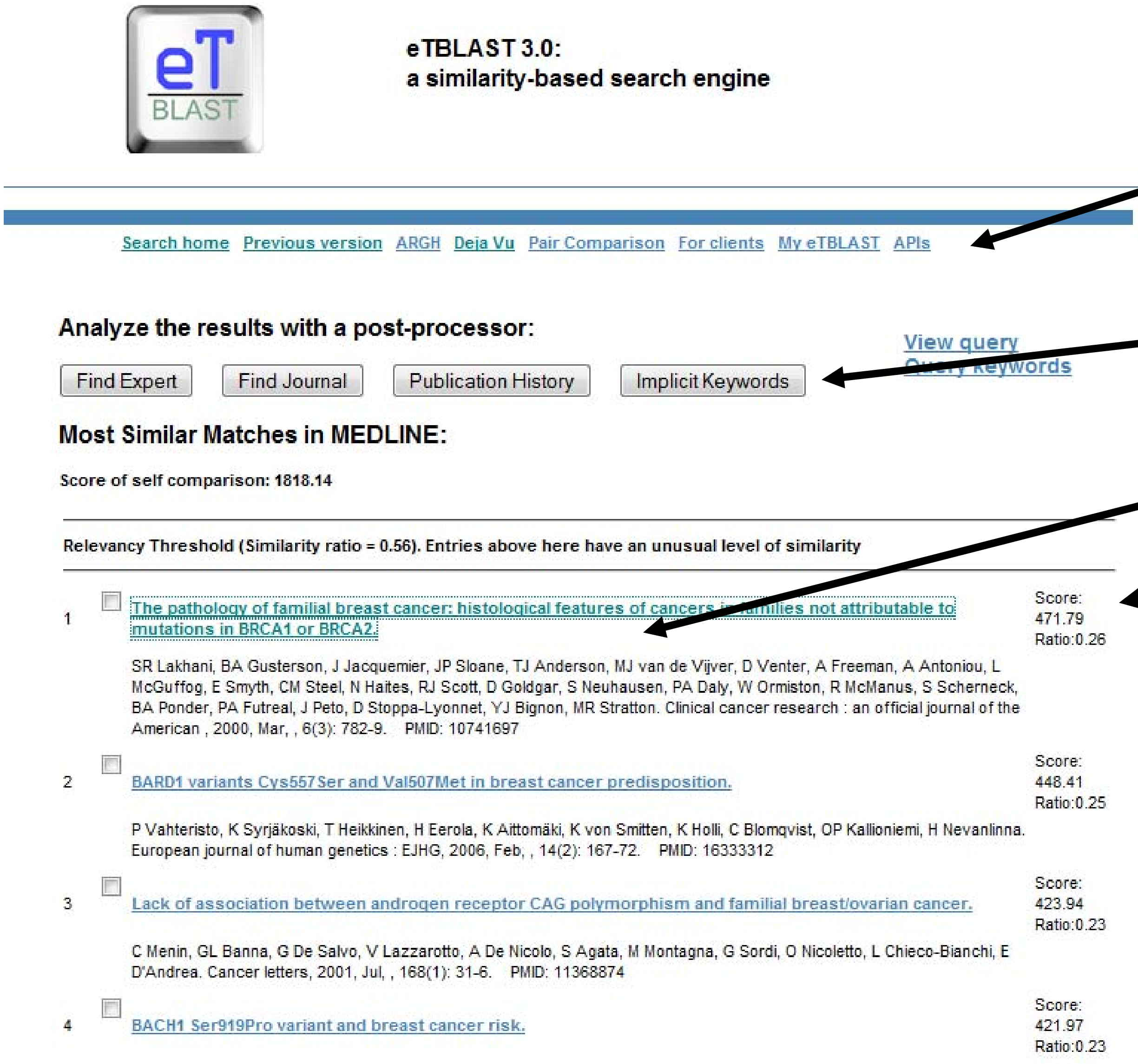
- ● MEDLINE
- ○ CRISP
- ○ NASA
- ○ Medical Cases
- ○ Pubmed Central (sections)
- ○ PMC Full Text
- ○ PMC METHODS
- ○ PMC INTRODUCTION
- ○ PMC RESULTS
- ○ PMC (paragraphs)
- ○ PMC Medical Cases
- ○ Clinical Trials
- ○ Arxiv

Select database to search

Paste your text in here

And search

eTBLAST results are linked to the abstract and other tools, of value while writing, reviewing or studying



eTBLAST 3.0:
a similarity-based search engine

Search home   Previous version   ARGH   Deja Vu   Pair Comparison   For clients   My eTBLAST   APIs

Analyze the results with a post-processor:

View query
Query keywords

Find Expert | Find Journal | Publication History | Implicit Keywords

Most Similar Matches in MEDLINE:

Score of self comparison: 1818.14

Relevancy Threshold (Similarity ratio = 0.56). Entries above here have an unusual level of similarity

1 □ The pathology of familial breast cancer: histological features of cancers in families not attributable to mutations in BRCA1 or BRCA2.
Score: 471.79 Ratio:0.26
SR Lakhani, BA Gusterson, J Jacquemier, JP Sloane, TJ Anderson, MJ van de Vijver, D Venter, A Freeman, A Antoniou, L McGuffog, E Smyth, CM Steel, N Haites, RJ Scott, D Goldgar, S Neuhausen, PA Daly, W Ormiston, R McManus, S Scherneck, BA Ponder, PA Futreal, J Peto, D Stoppa-Lyonnet, YJ Bignon, MR Stratton. Clinical cancer research : an official journal of the American , 2000, Mar, , 6(3): 782-9.   PMID: 10741697

2 □ BARD1 variants Cys557Ser and Val507Met in breast cancer predisposition.
Score: 448.41 Ratio:0.25
P Vahteristo, K Syrjäkoski, T Heikkinen, H Eerola, K Aittomäki, K von Smitten, K Holli, C Blomqvist, OP Kallioniemi, H Nevanlinna. European journal of human genetics : EJHG, 2006, Feb, , 14(2): 167-72.   PMID: 16333312

3 □ Lack of association between androgen receptor CAG polymorphism and familial breast/ovarian cancer.
Score: 423.94 Ratio:0.23
C Menin, GL Banna, G De Salvo, V Lazzarotto, A De Nicolo, S Agata, M Montagna, G Sordi, O Nicoletto, L Chieco-Bianchi, E D'Andrea. Cancer letters, 2001, Jul, , 168(1): 31-6.   PMID: 11368874

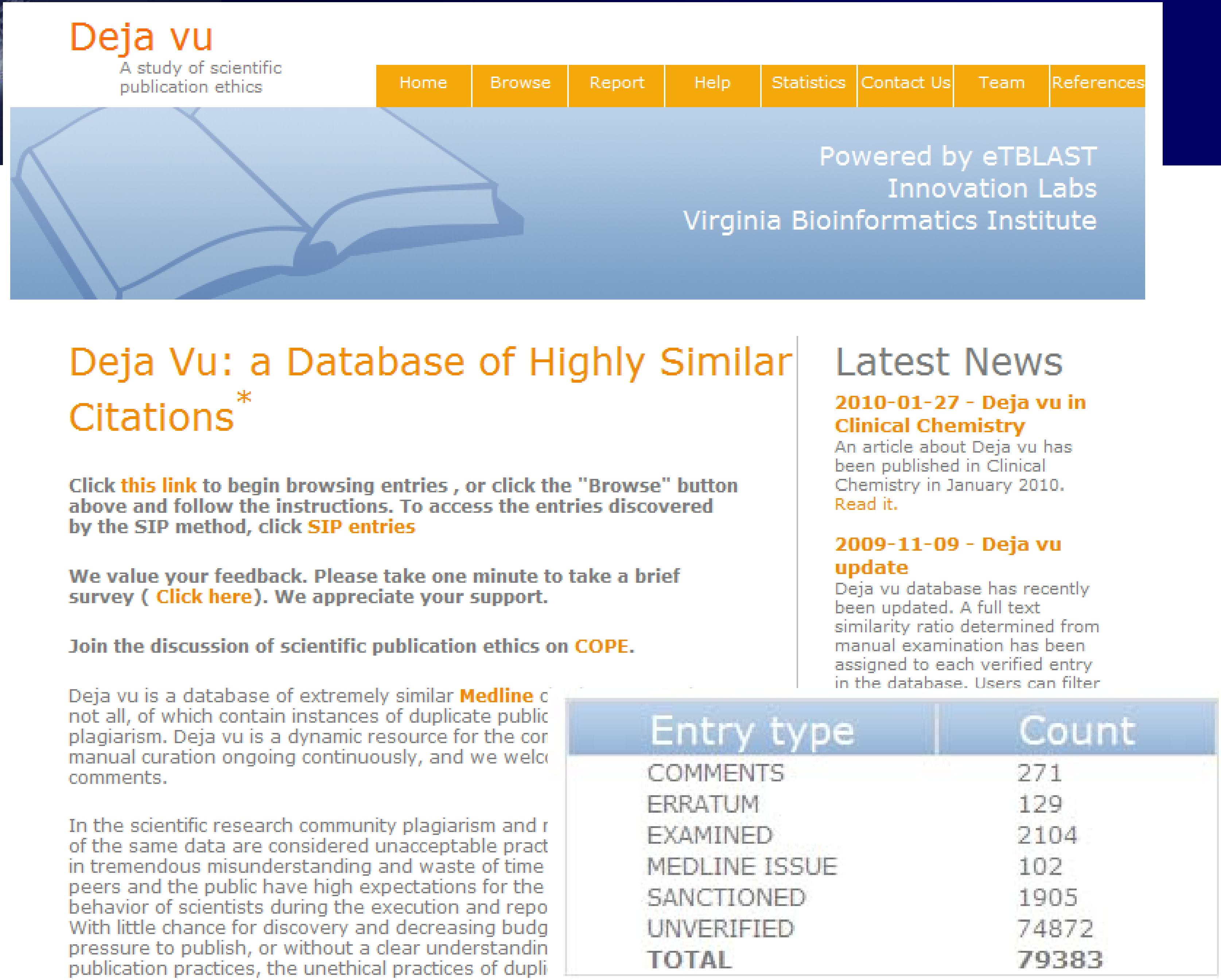
4 □ BACH1 Ser919Pro variant and breast cancer risk.
Score: 421.97 Ratio:0.23

Links to Déjà vu, etc.

Post-processors that analyze all ‘hits’ as a set

Ranked records

Raw similarity score



Deja vu

A study of scientific
publication ethics

Powered by eTBLAST
Innovation Labs
Virginia Bioinformatics Institute

Deja Vu: a Database of Highly Similar Citations*

Click this link to begin browsing entries , or click the "Browse" button above and follow the instructions. To access the entries discovered by the SIP method, click SIP entries

We value your feedback. Please take one minute to take a brief survey (Click here). We appreciate your support.

Join the discussion of scientific publication ethics on COPE.

Deja vu is a database of extremely similar **Medline** c
not all, of which contain instances of duplicate public
plagiarism. Deja vu is a dynamic resource for the cor
manual curation ongoing continuously, and we welc
comments.

In the scientific research community plagiarism and r
of the same data are considered unacceptable pract
in tremendous misunderstanding and waste of time
peers and the public have high expectations for the
behavior of scientists during the execution and repo
With little chance for discovery and decreasing budg
pressure to publish, or without a clear understandin
publication practices, the unethical practices of dupli

Latest News

2010-01-27 - Deja vu in Clinical Chemistry

An article about Deja vu has been published in Clinical Chemistry in January 2010. Read it.

2009-11-09 - Deja vu update

Deja vu database has recently been updated. A full text similarity ratio determined from manual examination has been assigned to each verified entry in the database. Users can filter

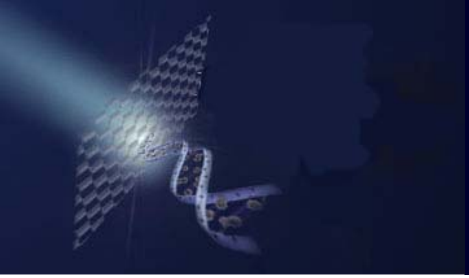
Entry type	Count
COMMENTS	271
ERRATUM	129
EXAMINED	2104
MEDLINE ISSUE	102
SANCTIONED	1905
UNVERIFIED	74872
TOTAL	**79383**

There are a large number of potentially plagiarized papers in Medline

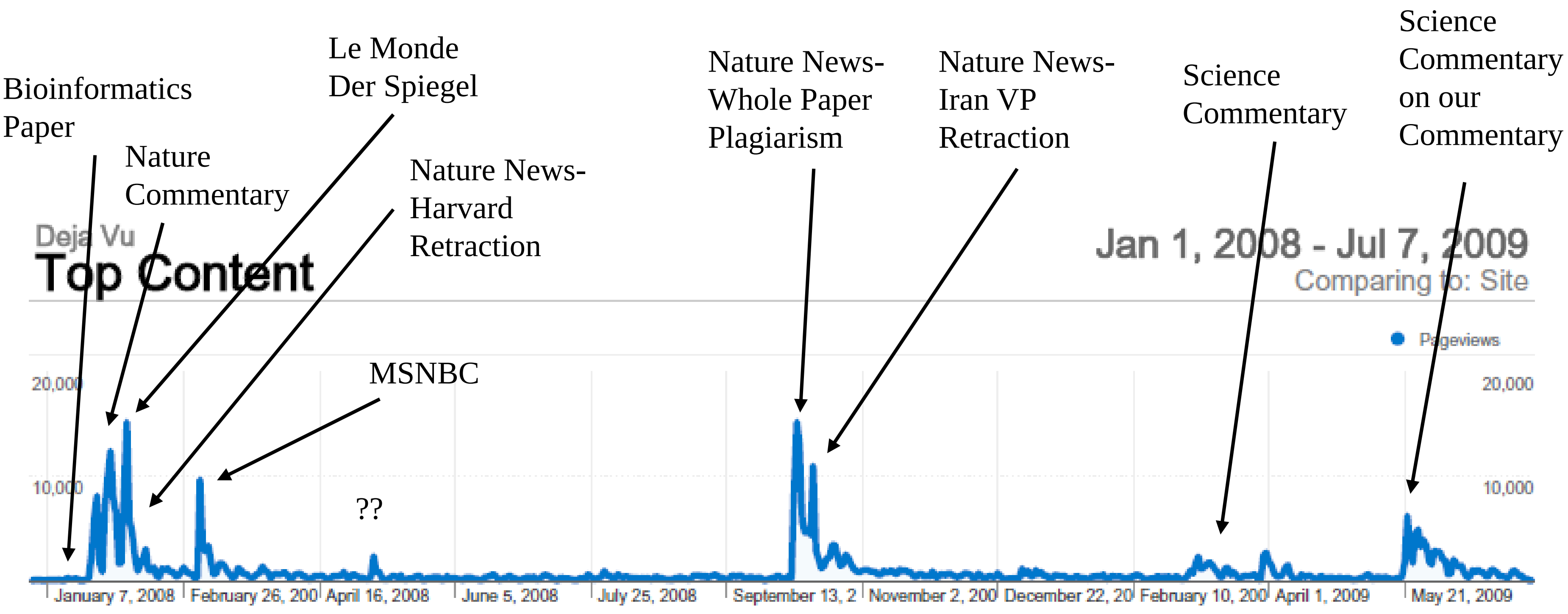
- Entries in Déjà vu with no overlapping authors   7,947
- Stakeholders surveyed for 206 pairs of articles
- Average full text similarity 86%
- Pairs with similar table/figure 72%
- Overall survey response rate 90.8% found:

- 93% of authors unaware they were duplicated
- 26% of duplicate authors denied wrongdoing,
- 35% admitted and apologized,
- 16% co-authors claiming no involvement in
- writing manuscript
- 13% were not aware that they were 'authors'

- Total investigations initiated 90+
- Retractions 50+ (+~72)



Déjà vu access statistics confirms interest in publishing ethics



Bioinformatics Paper

Nature Commentary

Le Monde
Der Spiegel

Nature News-
Harvard
Retraction

MSNBC

??

Nature News-
Whole Paper
Plagiarism

Nature News-
Iran VP
Retraction

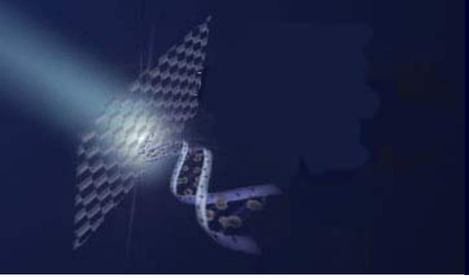
Science Commentary

Science
Commentary
on our
Commentary

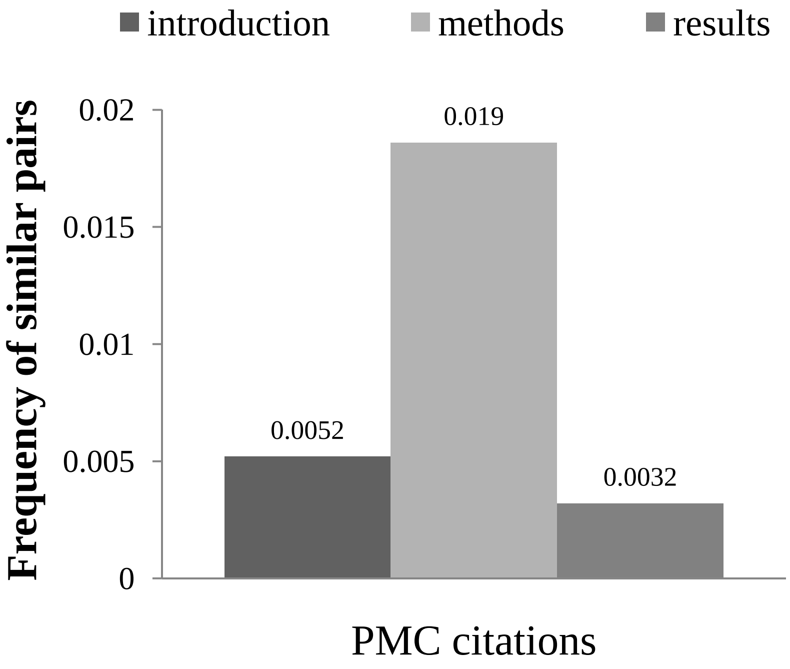
91,418 pages were viewed a total of 492,754 times

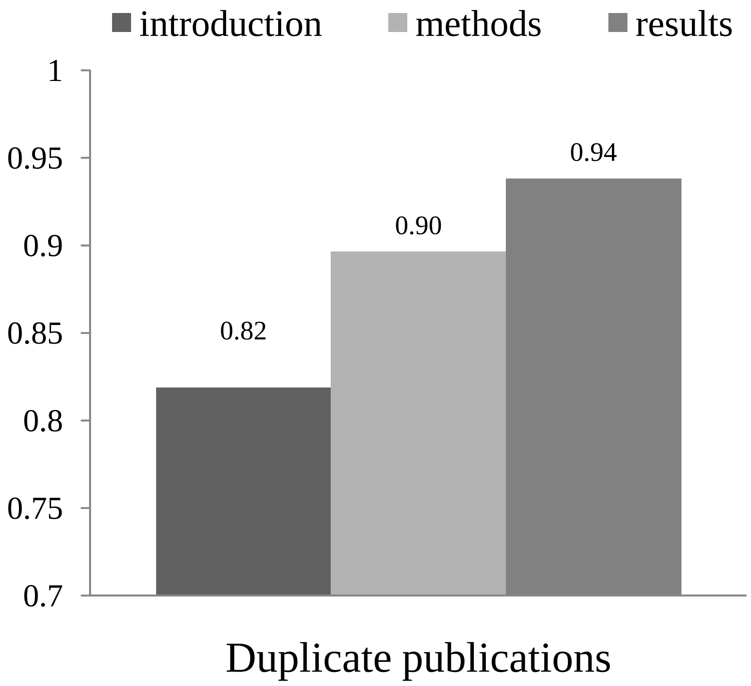
80,079 unique visitors

- •Identifying duplicate content using Statistically Improbable Phrases, Bioinformatics, 2010
- •Quaere verum: Responding to the editorial, "Primum non Nocere“, Clinical Chemistry, 2010
- •Characterizations of the text similarity in full text biomedical citations, submitted
- • “Are there too many review articles?”, in preparation



Striking full text difference between all publications and a subset of 'duplicate' publications



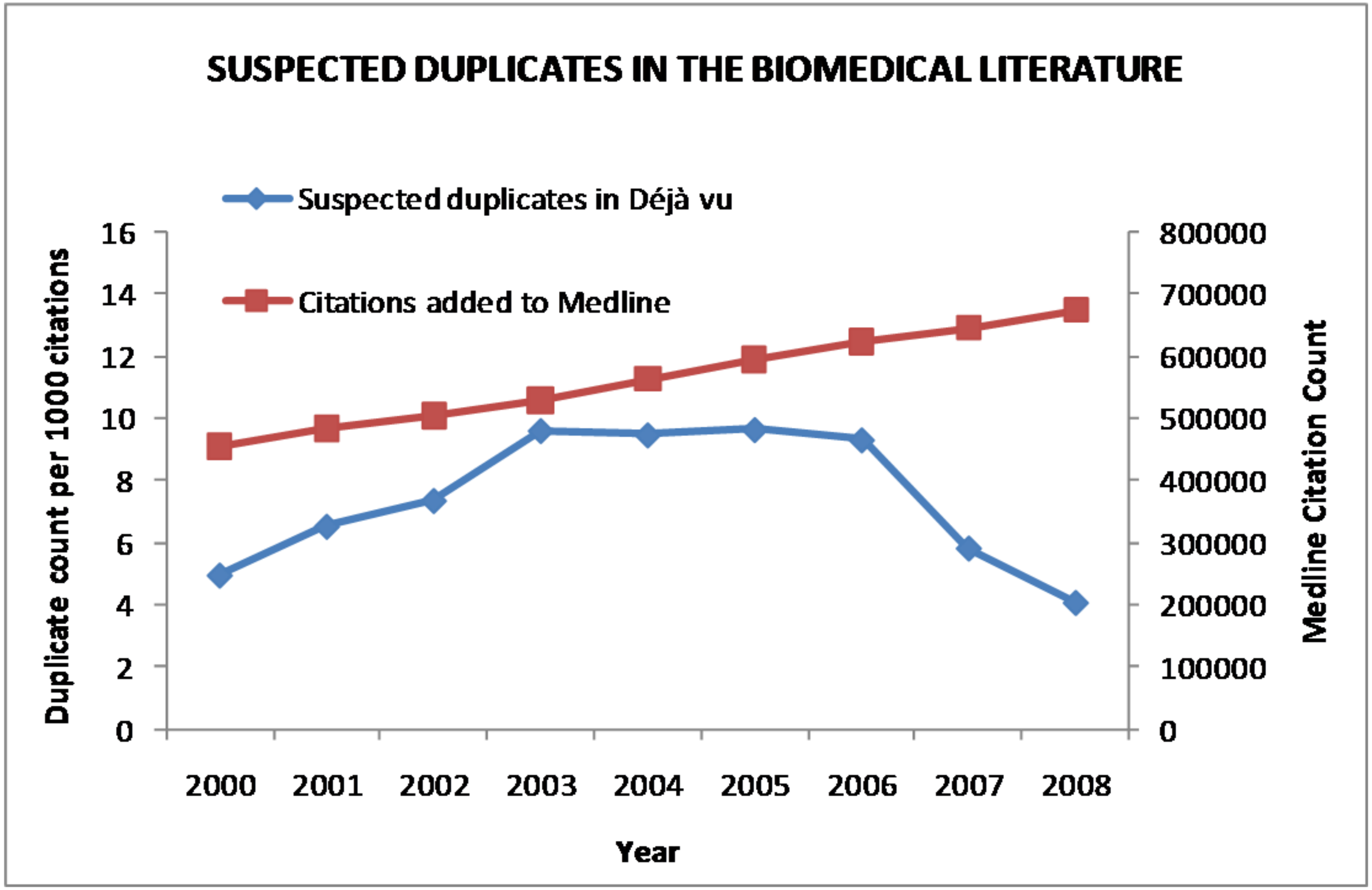


72,011 full text papers

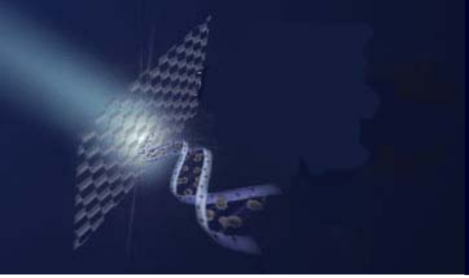
~400 duplicate papers

High similarity in the methods section is a good thing, but one would expect the results section to be unique and add to scientific knowledge.

There is good news, the duplicate rate is decreasing.

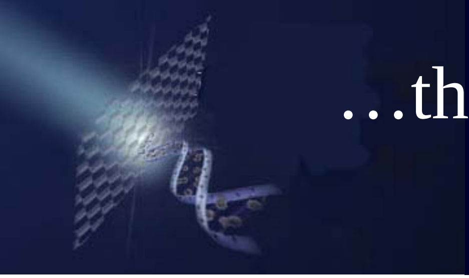


…but, there are ~3,000 highly similar pairs of papers added annually…



Funding is difficult because reviewers find “ the approach, tactics and findings to be too controversial”, “the research is done”, and it “should be handled by the national databases” … or perhaps a contract…

Regardless….



…there is much to be investigated and resolved, a sampling includes:

- Why do so many highly similar articles also contain falsified/fabricated data, inappropriate authorship, inappropriate changes in experimental design?

- What is a “retraction”? Researchers and clinicians continue to use “retracted” papers because only ~10% of the official retractions propagate back to Medline.

- We found 3 journals whose editorial staff are engaged in “plagiaristic” activities. There is no policy for “de-indexing” compromised journals.

- The journals that primarily publish ethically questionable articles have low impact factors (~1) , are small with limited resources, so their editorial and review staff need a free public service.

- There are 76,000 more pairs of questionable manuscripts that need to be inspected, and this number is growing at ~3,000/year.

But mainly this has to be pursued, because as one
of those whose work was ‘reused’ put it…

“[My] major concern is that false data will lead to changes in surgical practice regarding procedures.”

Or as one of the authors of a “plagiarized” paper responded

“There are probably only "x" amount of word combinations that could lead to "y" amount of statements. … I have no idea why the pieces are similar, except that I am sure I do not have a good enough memory and it is certainly not photographic, to have allowed me to have "copied" his piece. ... I did in fact review it [the original article] for whatever journal it was published in.”

(Paper was retracted and author has since resigned chairmanship of his clinical department at Harvard)