Global Tech Mining, Atlanta, September 2015

Underreporting research relevant to local needs in the global south

Database biases in the representation of knowledge on rice

Ismael Rafols, Tommaso Ciarli, and Diego Chavarro

Ingenio (CSIC-UPV), Universitat Politècnica de València

SPRU (Science Policy Research Unit), University of Sussex, Brighton, UK
Observatoire des Sciences et des Téchniques (OST-HCERES), Paris

Introduction

- Increasing demand for science to help societal problems
- Local knowledge important for:
  - Supporting local communities in specific contexts
    - Agriculture, health
  - Global challenges need local knowledge
    - Climate change, pandemics…understanding local conditions is crucial to explaining global effects and trends.
- Mapping research landscape of a topic (science supply)
  - We need a representation of the knowledge on research topics relevant for a problem.
- Conventional databases (WoS, Scopus) only have limited local literature coverage.
  - How can this effect the representationof the knowledge landscape?

Bias in bibliometric databases

- Web of Science is biased towards English-speaking publications and biomedical publications (Archambault et al., 2006).
- Scopus has a broader coverage, but similar ranks regarding country production over different fields,
  - ‘indicators of scientific production and citation at the country level are stable and largely independent of the database’(Archambault et al., 2009, p. 1320).

- In **international benchmarking,** major int’l organisations continue to use the main databases WoS (e.g UNESCO, 2010) and Scopus (e.g Royal Society, 2011).
- Recommendations have been made on the need to improve scientometric indicators in order to "properly evaluate global science" (Royal Society, 2011, p. 107).

What is the extent of bias?

Country and Topic bias

Databases compared

Publications on rice were downloaded from:

- **WoS** (including SCI-Expanded, SSCI, A&HCI, CPCI-S and CPCI-SSH) searching “rice” or “oryza” in the field “topic”.
  - 99,500 records
- Scopus searching in **title, abstract or keywords**, i.e. TIT-ABS-KEY ("rice" OR "oryza").
  - 95,701 records.
- Database CAB Abstracts, documents with “rice” or “oryza” were searched in **title and abstract**.
  - **227,873 records**!

CAB Abstracts (http://www.cabdirect.org/) is a database focused on **environment and agriculture**. http://www.cabi.org/

CABI (CAB abstracts): **environment and agriculture**

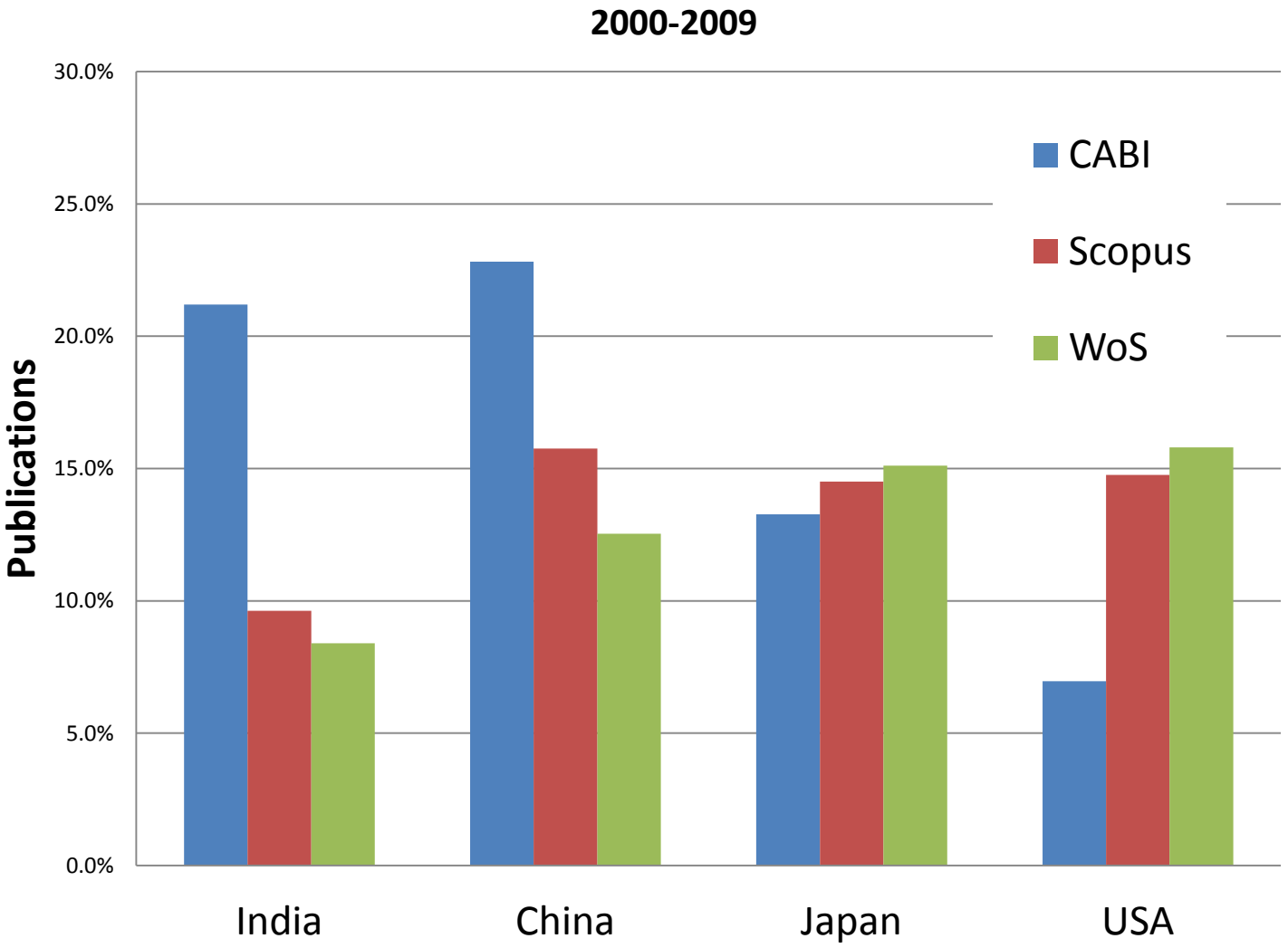
CAB Abstracts (http://www.cabdirect.org/) is a database focused on **environment and agriculture**. http://www.cabi.org/

- CABI, an inter-governmental, not-for-profit organization that was set up by a United Nations treaty, with 48 member countries (many **Commonwealth**)
- Mission :“providing information and applying scientific expertise to solve problems in **agriculture and the environment**”.
- Both **CAB Abstracts** (for agriculture and environment) and **Global Health** (for public health) are aimed at facilitating the retrieval of relevant information for practitioners, very much as MEDLINE for medical research.

Comparison of document types

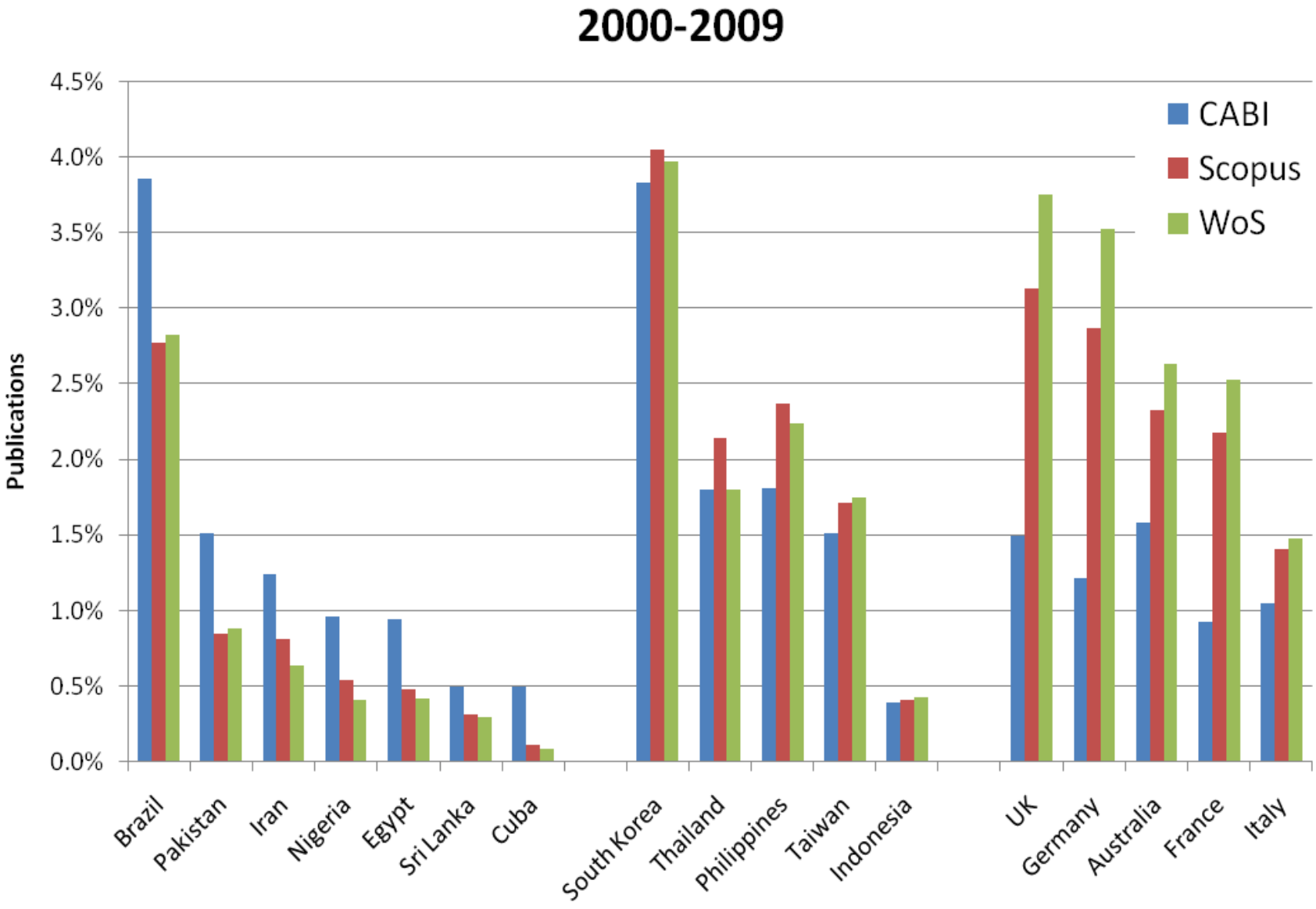
WoS			Scopus			CAB Abstracts		
Doc type	**%**	**Cum %**	**Doc type**	**%**	**Cum%**	**Doc type**	**%**	**Cum%**
Article	**81%**	**81.2%**	**JOUR**	**93.7%**	**93.7%**	**Journal article**	**85%**	**84.8%**
Proceedings Paper	**7.1%**	**88.3%**	**CONF**	**3.5%**	**97.2%**	**Conference paper**	**6.8%**	**91.6%**
Review	3.4%	91.7%	SER	1.6%	98.8%	Miscellan-eous	4.7%	96.3%
Meeting Abstract	2.7%	94.4%	INPR	0.9%	99.7%	Book chapter	2.0%	98.3%
Note	2.4%	96.8%	CHAP	0.3%	99.9%	Book	1.9%	100.2%
Book Review	1.6%	98.4%	BOOK	0.1%	100.0%	Annual report	0.9%	101.1%
Editorial Material	0.7%	99.1%				Bulletin	0.6%	101.7%
Letter	0.6%	99.6%				Conference proceedings	0.5%	102.2%
Correction	0.3%	99.9%				Bulletin article	0.4%	102.7%

Coverage bias against developing countries (rice pubs)



Ciarli and Rafols (2014)

Coverage comparison in other countries

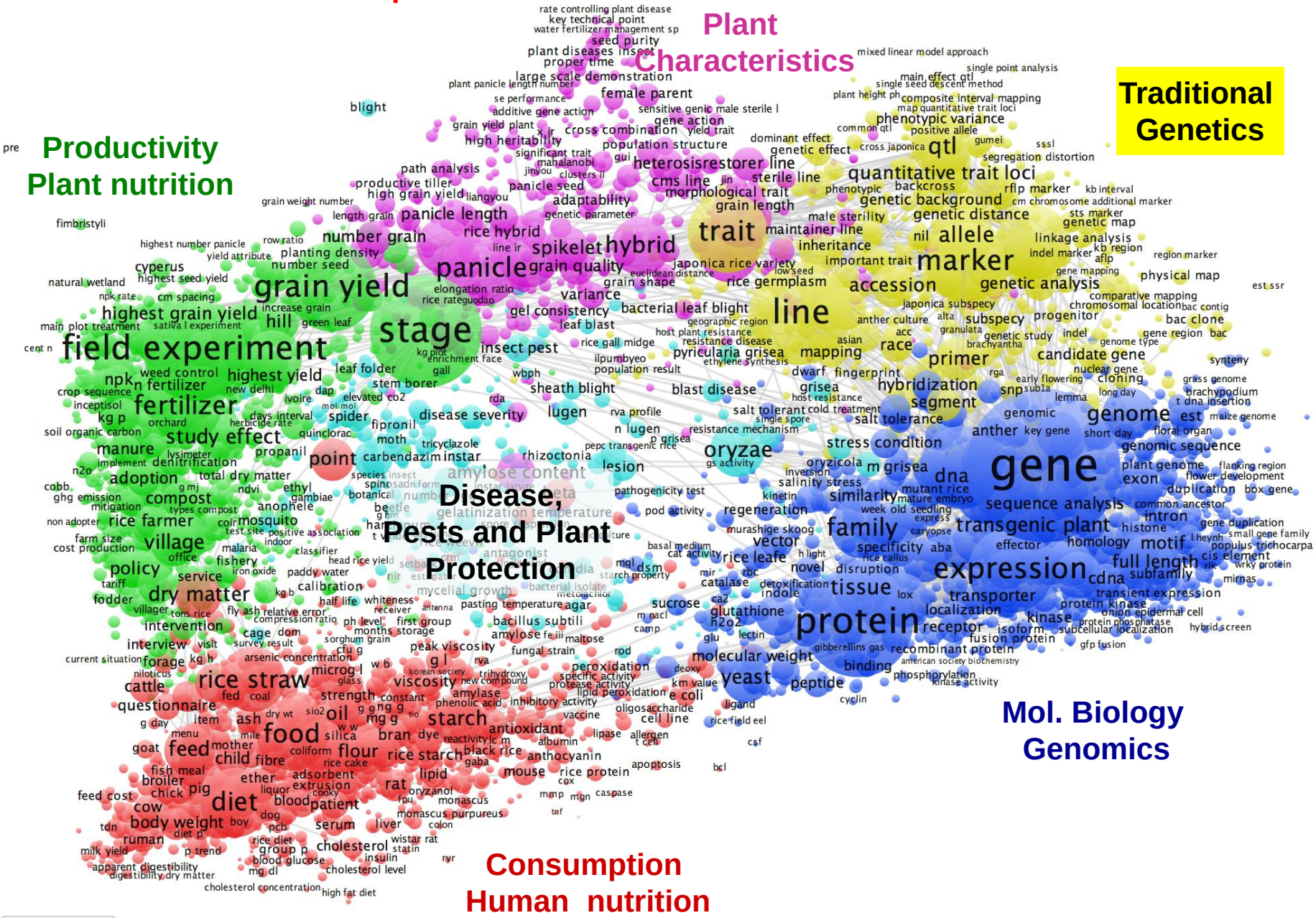


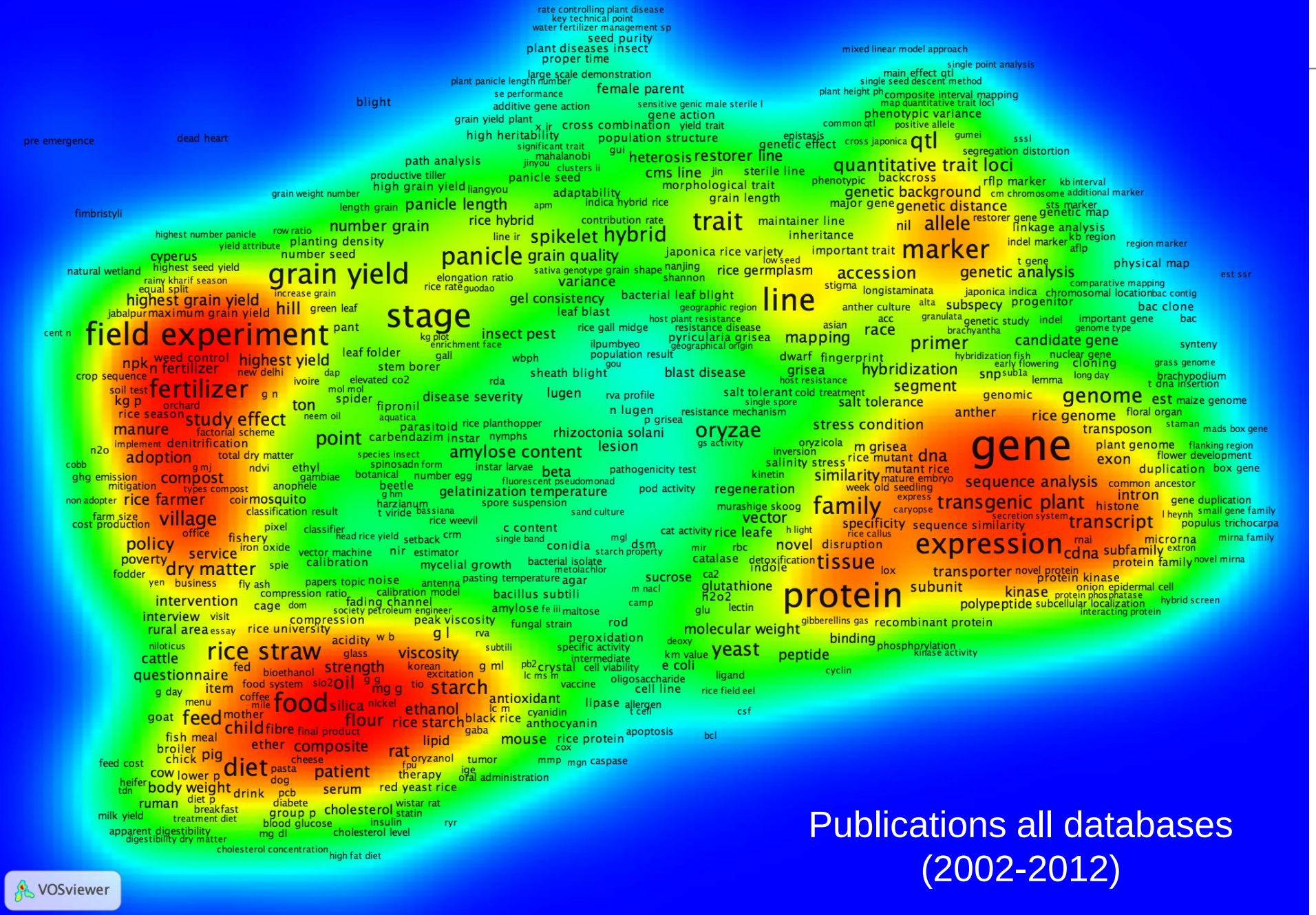
Method for clustering and paper assignation

- Match databases (use various fields)
- Use words from abstracts in VOSviewer co-word
  - Relevance algorithm
- Word clustering in VOSviewer → find topics
- Assign fractionally papers to clusters of words.

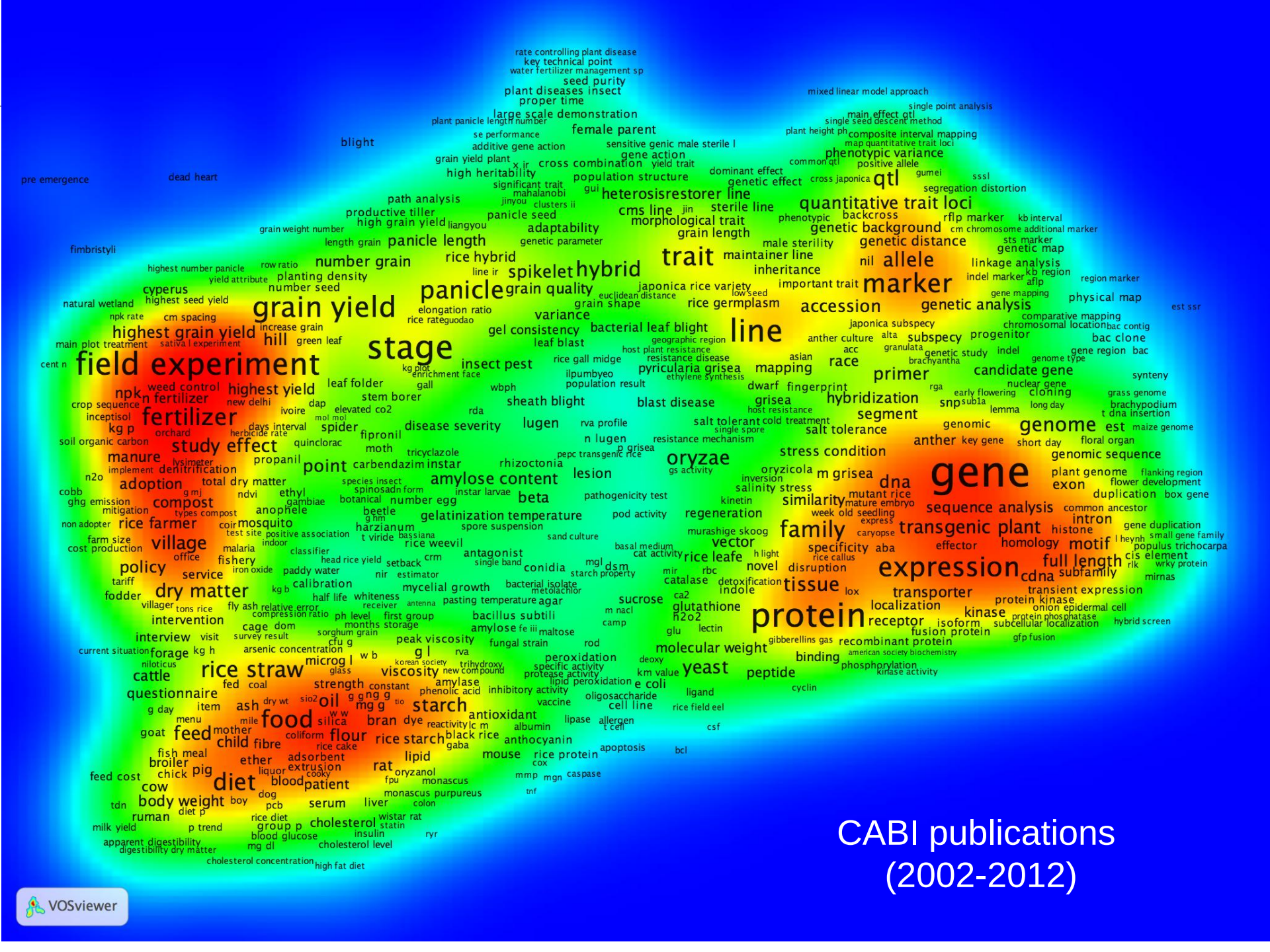
- Robustness of clustering was confirmed using clustering method directly based on papers
  - Similarity on word co-occurrence, with Blondel et al. algorithm)

Rice research topics



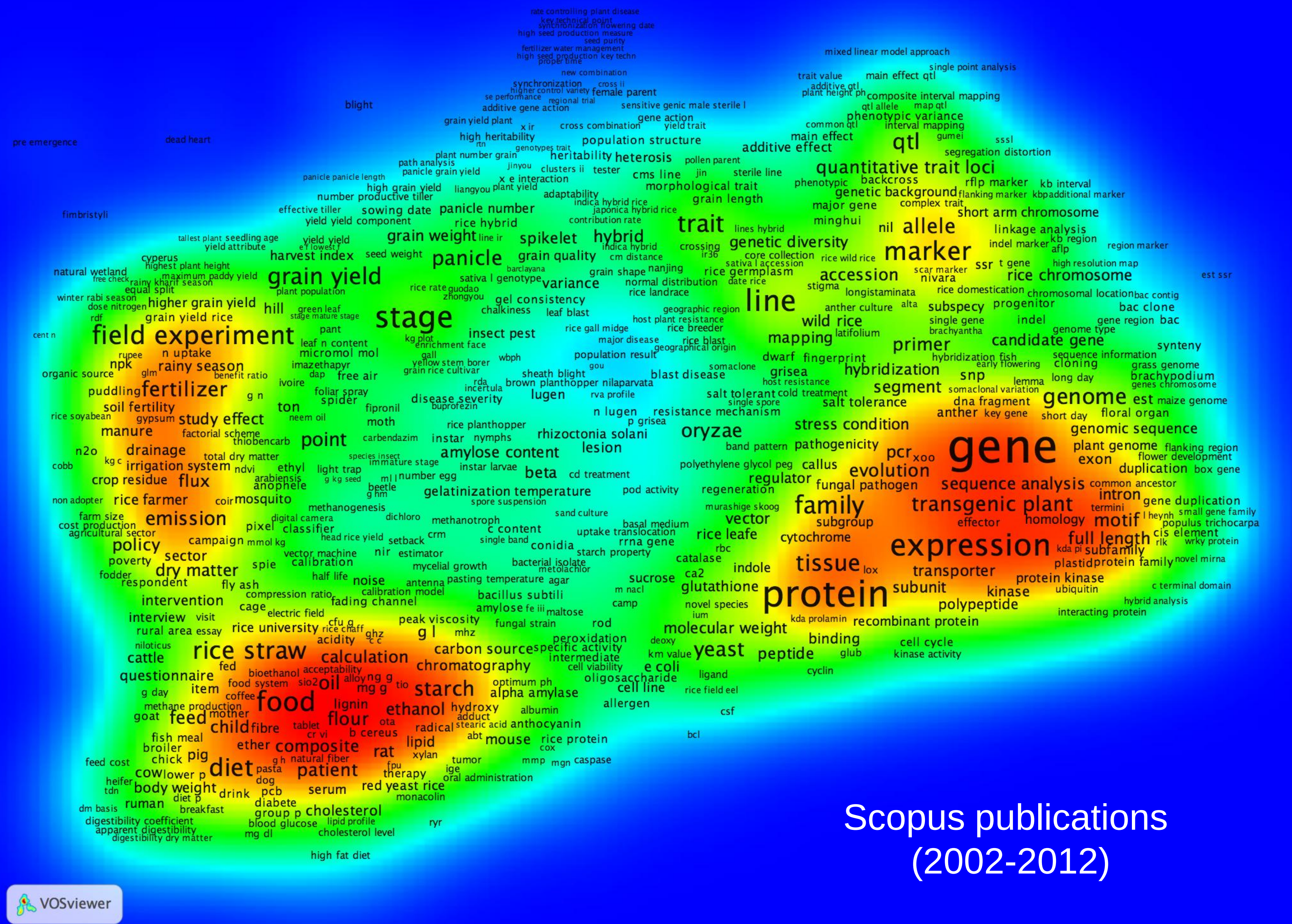


Publications all databases
(2002-2012)



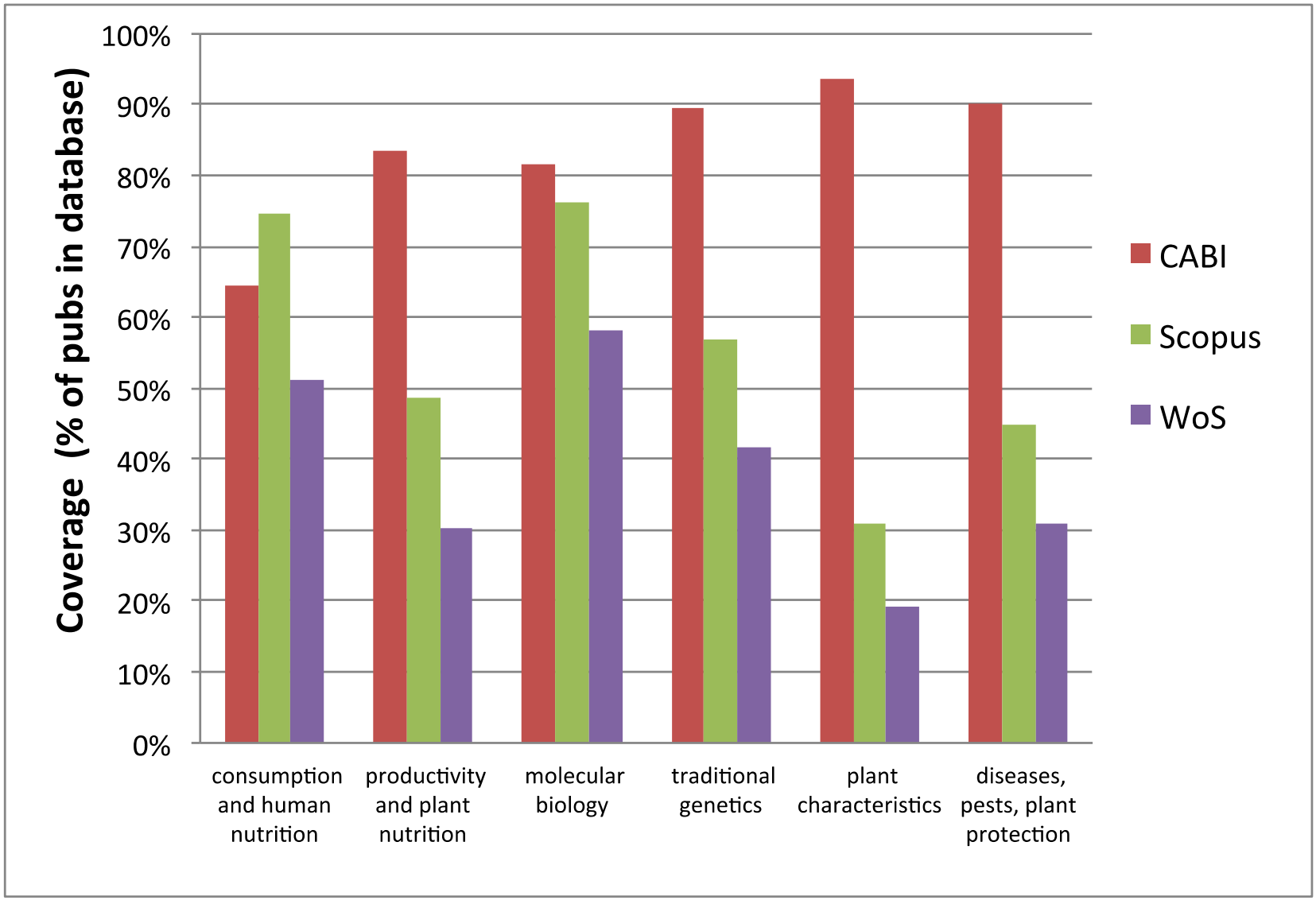
CABI publications
(2002-2012)

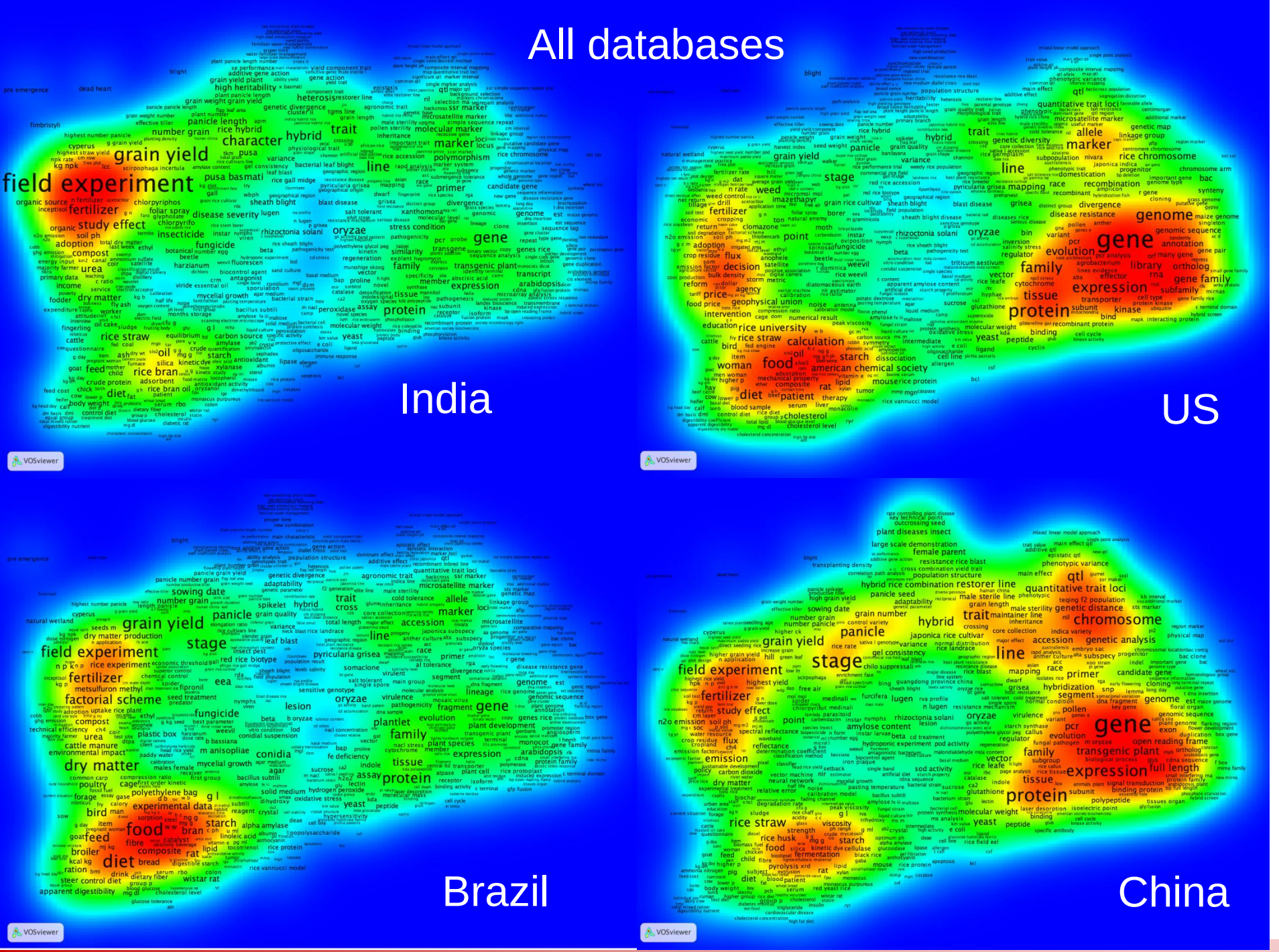
VOSviewer

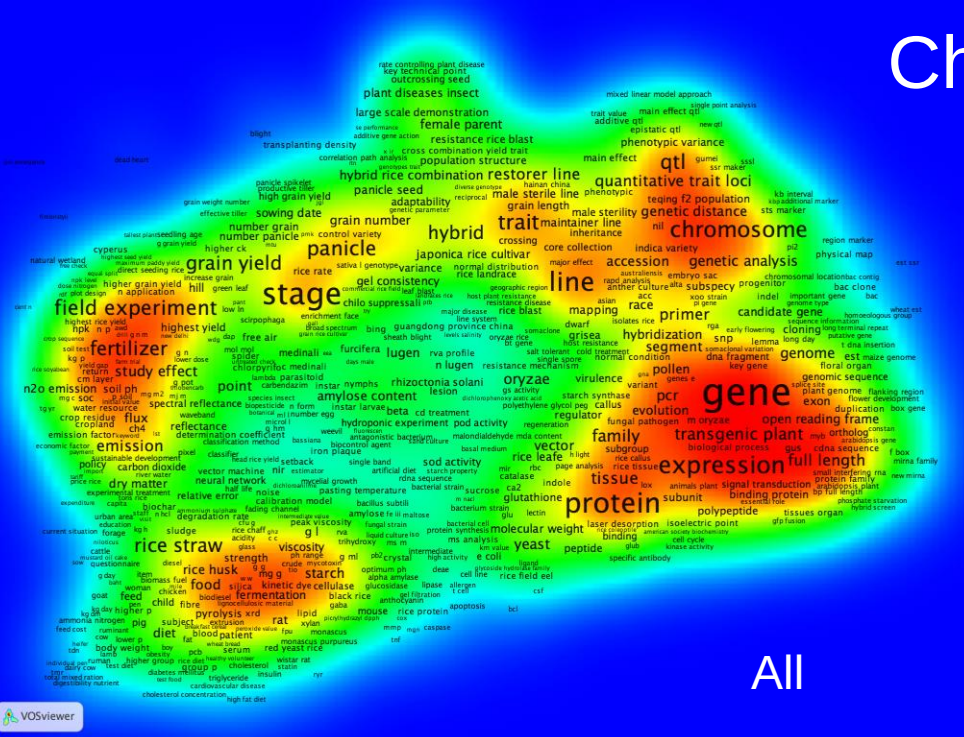


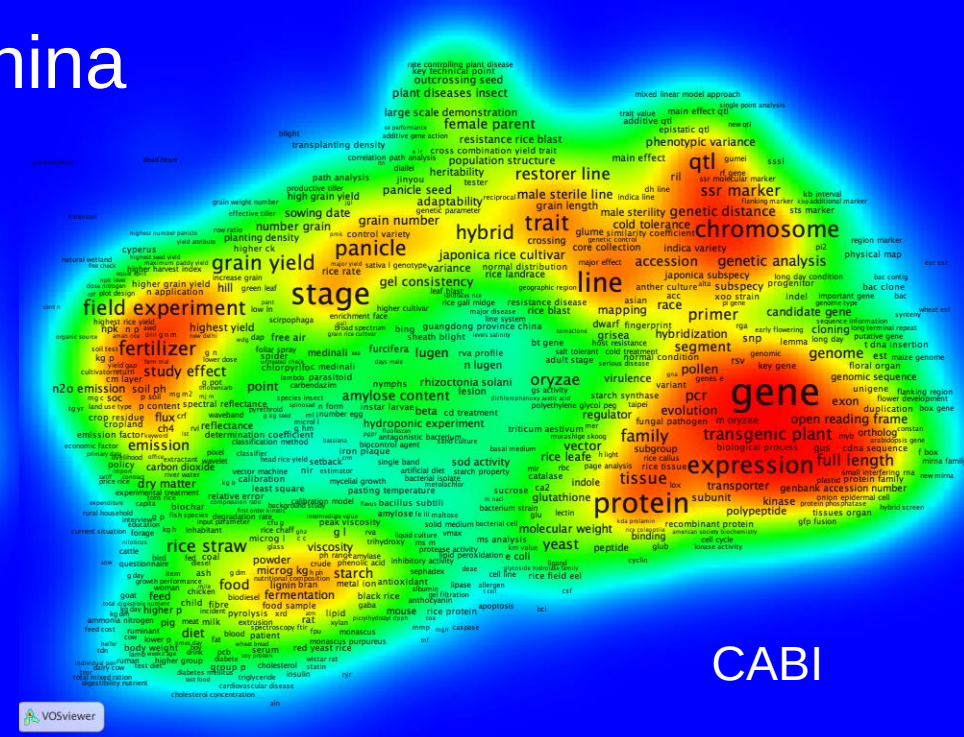
Scopus publications
(2002-2012)

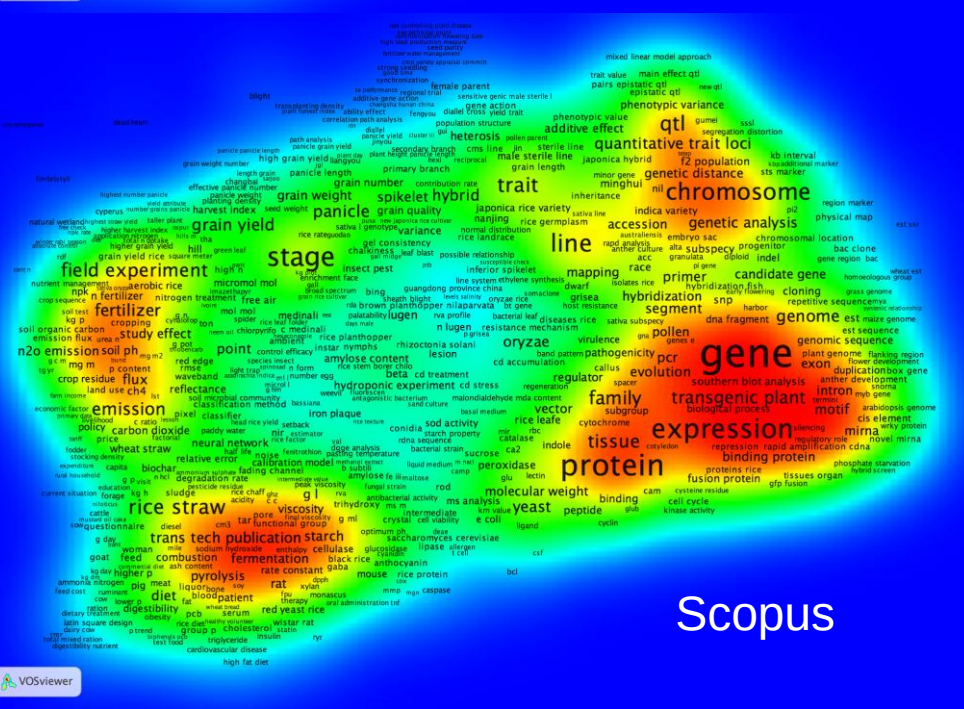
Coverage by database and research topics (2002-2012)

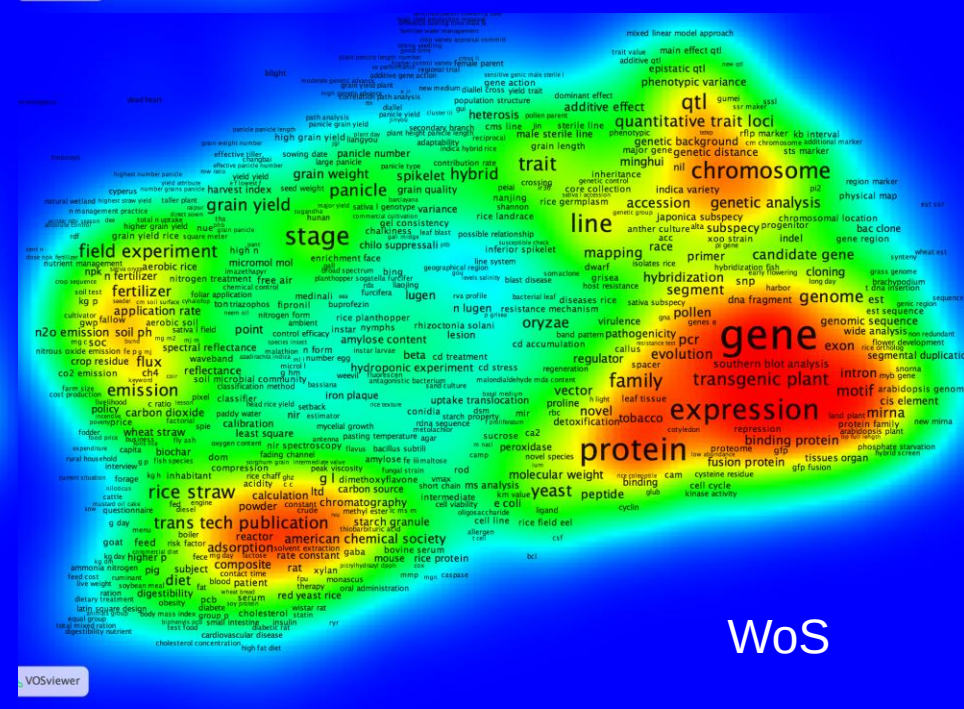



All databases
India
US
Brazil
China
VOSviewer
VOSviewer
VOSviewer
VOSviewer

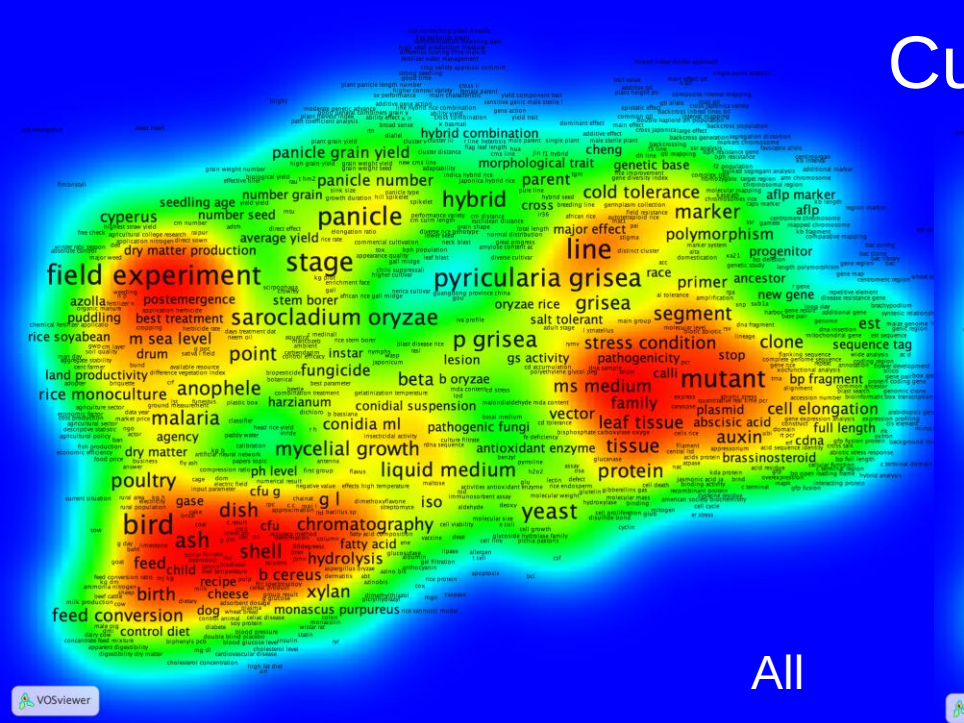

China
All


CABI

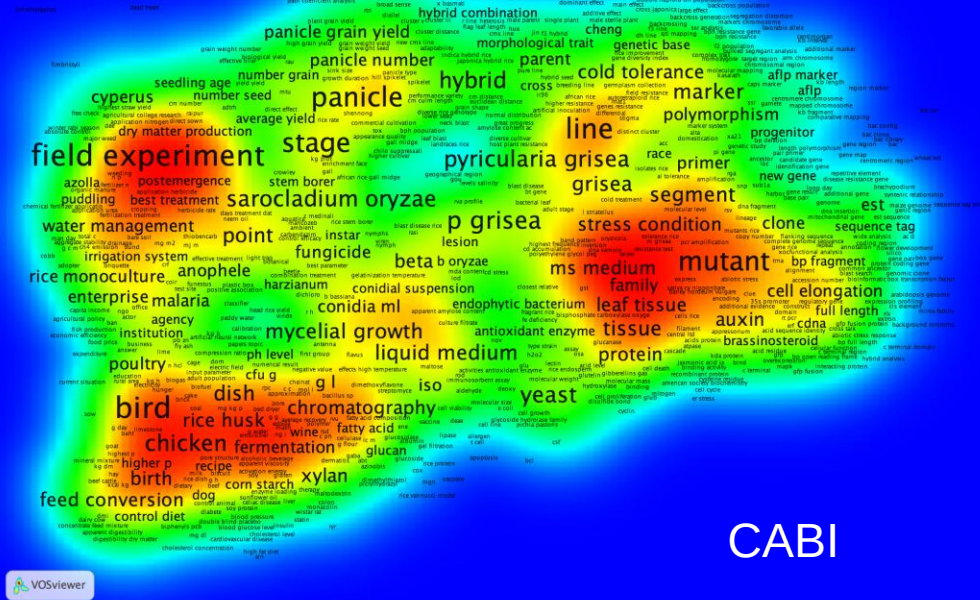

Scopus


WoS

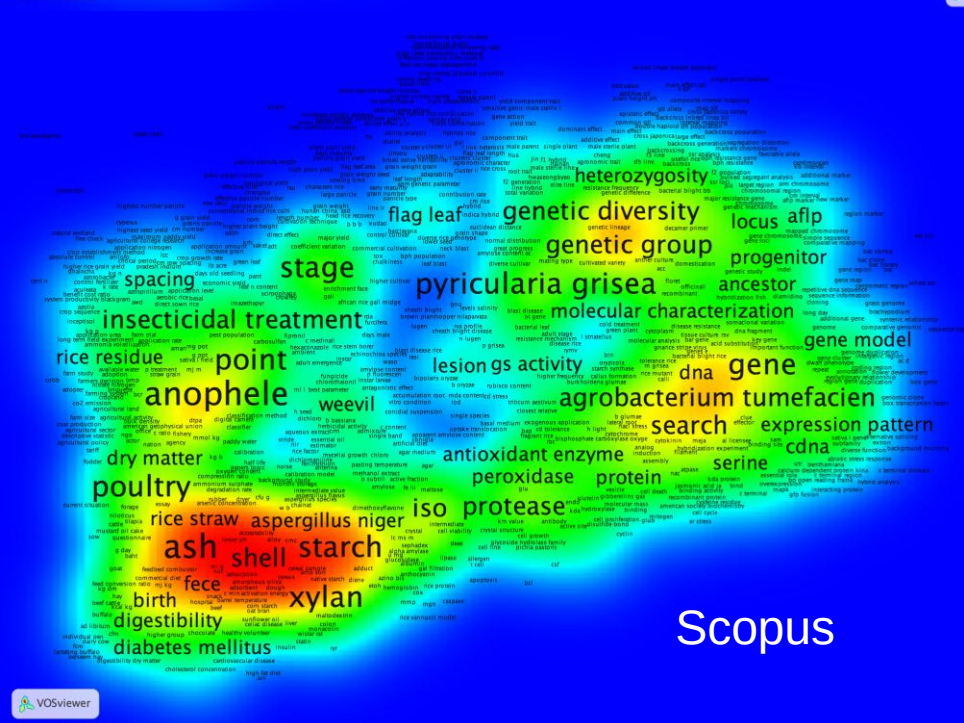
Cuba



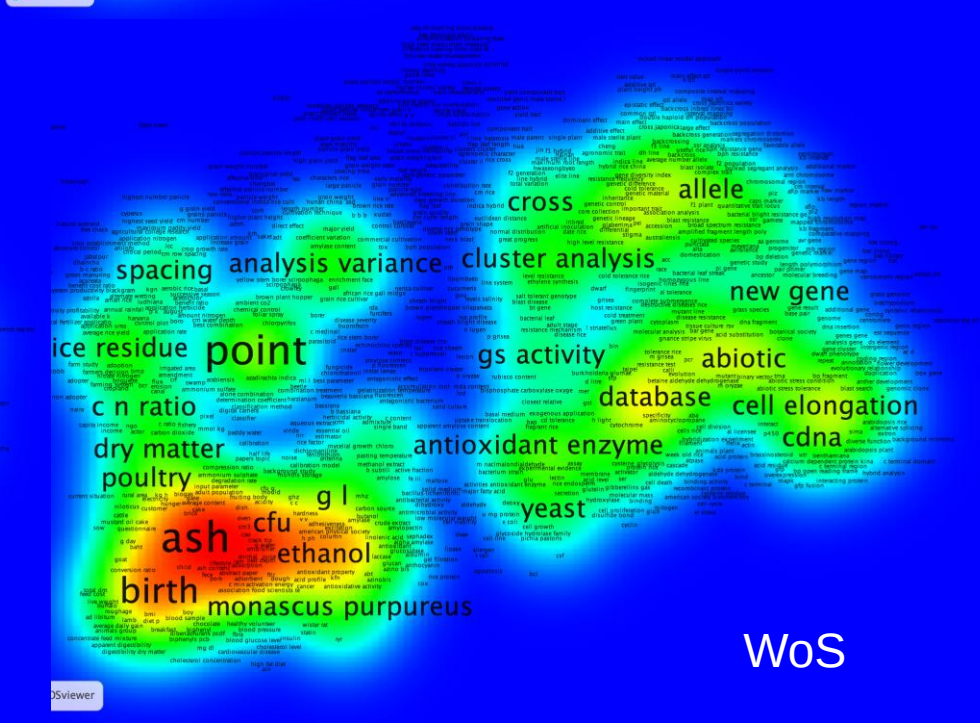
All



CABI



Scopus



WoS

Findings regarding country coverage

- Assumption on the stability of indicators of scientific production are incorrect (Archambault et al., 2009).

- Number of publications is very dependent on the database when one analyses low and middle income countries.

- Important result for international organisations such as FAO, IFRI or UNESCO (UNESCO, 2010) that aim to work on human development.

- Proliferation in the last two decades of journal indexing systems at the regional level, such as Scielo or Redalyc that aim to provide visibility to local journals, often in languages other than English (Chavarro, 2013)

Findings regarding topic coverage

- Significant differences of coverage between research topics by database
  - Conventional databases (WoS and Scopus) have a larger coverage of molecular biology, traditional genetics and consumption
  - CABI has much better coverage about productivity, plant nutrition, plant characteristics and plant protection.

- High coverage appears to be related to
  - Resarch interests of actors in developed countries such as seed companies, food & industry
- Lower coverage appears to be related to
  - Potential interests of small farmers, local contexts.
  - Exception – nutrition? (to be confirmed)

- Need to contrast results with stakeholders.

Uneven coverage of databases

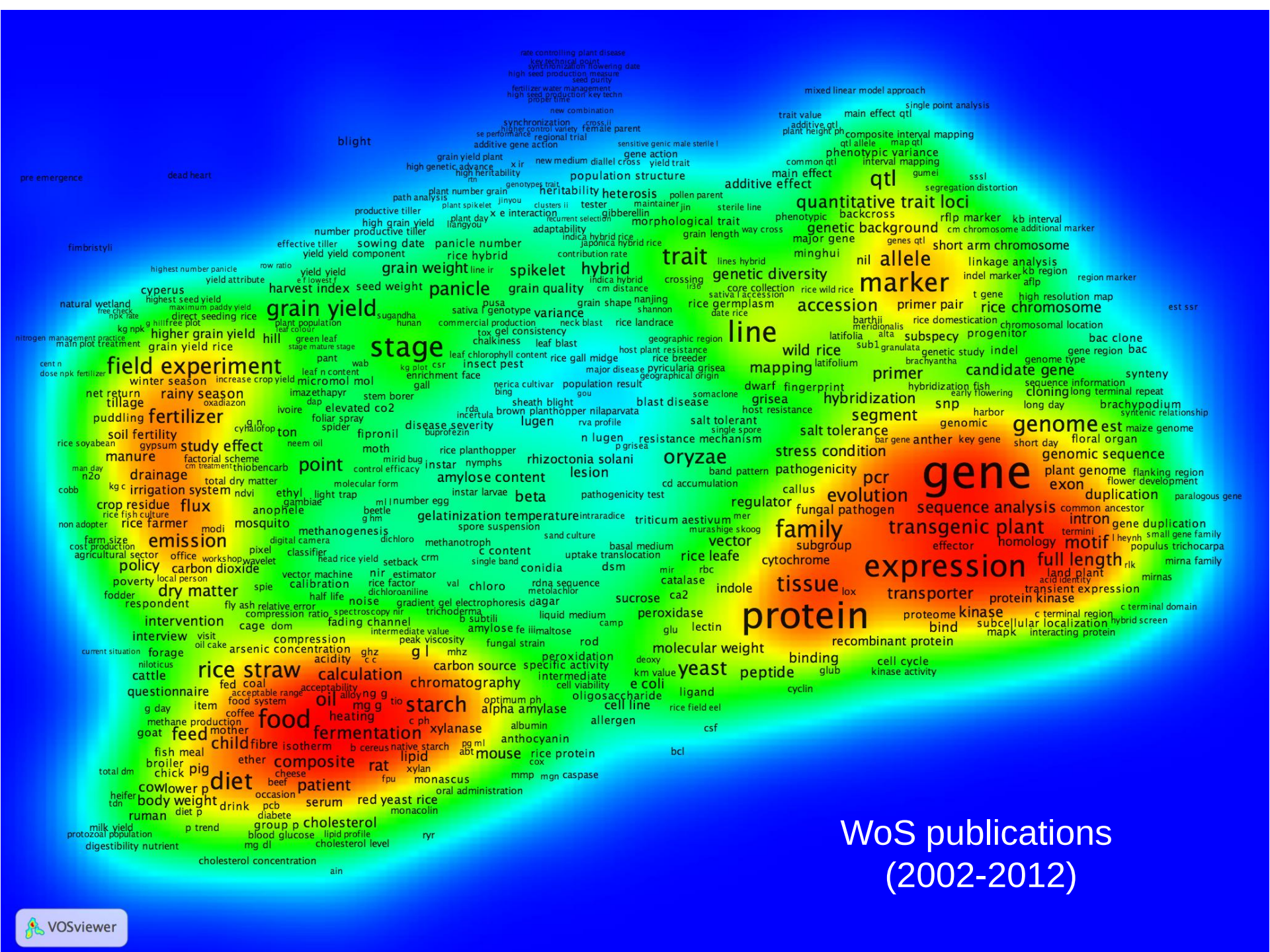
"When comparing databases one easily forgets that each database has a different purpose.

Thus, most of the subject specific databases (including CAB) aim for data completeness, whereas others like Web of Science, following Garfield's original idea, consider only the "core" journals, which are responsible for 80% of the citations in each discipline.

Thus, it is obvious that the coverage is biased in favour of journals published in industrialised countries, because these normally have a higher impact. (…)

Considering the conclusions, **it is alarming to see how often scientometric analyses are performed without even the correct choice of adequate data sources for the required purpose**. "

Reviewer of an earlier version of this paper


WoS publications
(2002-2012)
VOSviewer

Comparison of document types

	CABI		WoS	
Language	# docs	%	# docs	%
English	**148577**	**71.84%**	**92554**	**94.93%**
Chinese	20544	9.93%	490	0.50%
Japanese	13844	6.69%	2032	2.08%
Portuguese	5356	2.59%	1015	1.04%
French	3942	1.91%	560	0.57%
Spanish	3320	1.61%	307	0.31%
Korean	3018	1.46%	31	0.03%
Russian	2396	1.16%	162	0.17%
Italian	1546	0.75%	22	0.02%
German	1462	0.71%	214	0.22%