The ContentMine Scraping Stack

Richard Smith-Unna **Peter Murray-Rust**

University of Cambridge

CONTENT
MINE

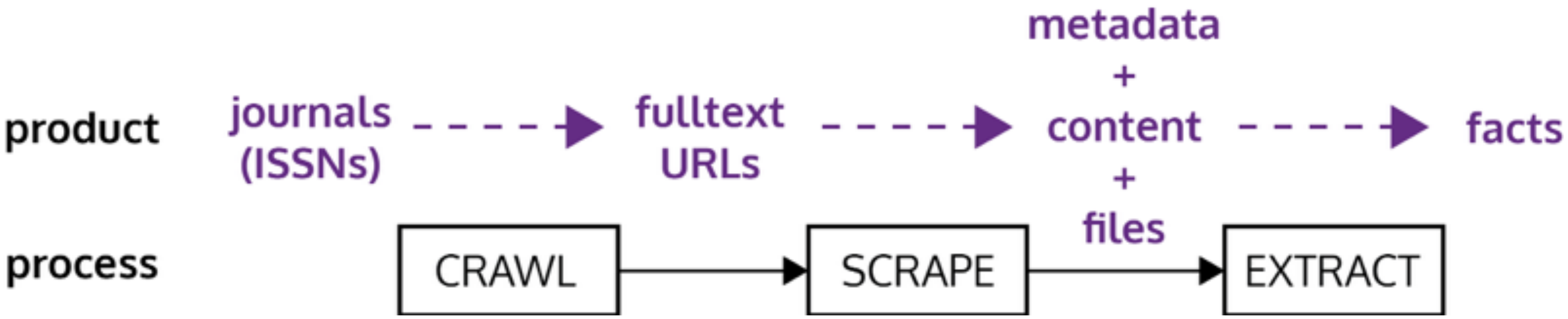
“*make 100,000,000 facts from the scholarly literature open, accessible and reusable*”

our mission

The scale of the task

- ~ 27,000 peer reviewed journals (Ulrich's)
- > 5,000 publishers
- new papers every day

The pipeline



product

journals (ISSNs)

fulltext URLs

metadata + content + files

facts

process

CRAWL

SCRAPE

EXTRACT

scraperJSON

- scrapers all have the same plumbing
- ignore the plumbing, just configure
- **benefits**
  - supports large collections of scrapers
  - no programming required
  - not limited to one piece of software

Basic scraperJSON

name of the scraper

the URL(s) it applies to

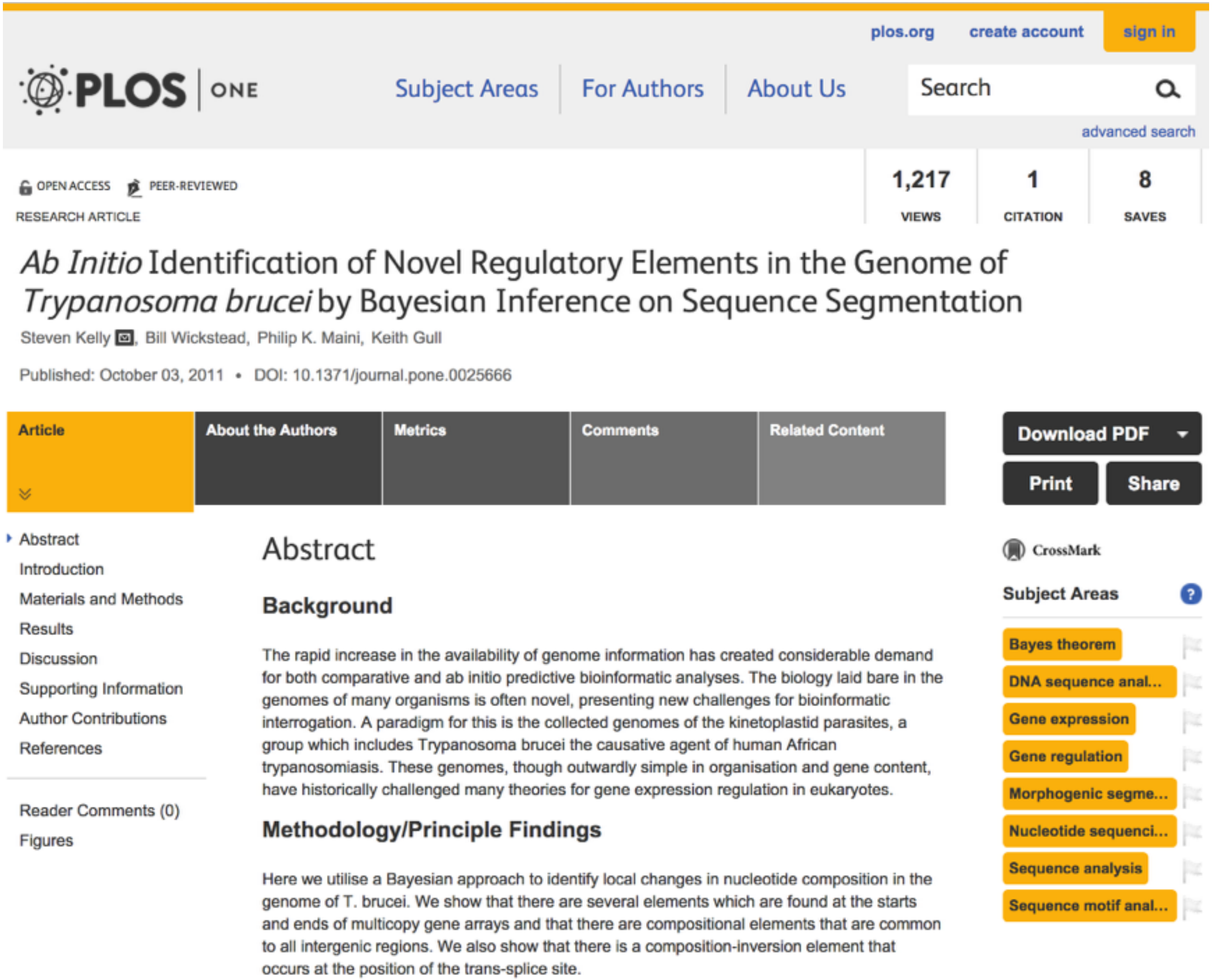
the elements to capture

element name

where to find it

```
{
  "name": "PLOS",
  "url": "plos\\w*.org",
  "elements": {
    "title": {
      "selector": “//h1[@property=‘dc:title’]”,
    }
  }
}
```

http://github.com/ContentMine/scraperJSON



RESEARCH ARTICLE

Ab Initio Identification of Novel Regulatory Elements in the Genome of *Trypanosoma brucei* by Bayesian Inference on Sequence Segmentation

Steven Kelly, Bill Wickstead, Philip K. Maini, Keith Gull

Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

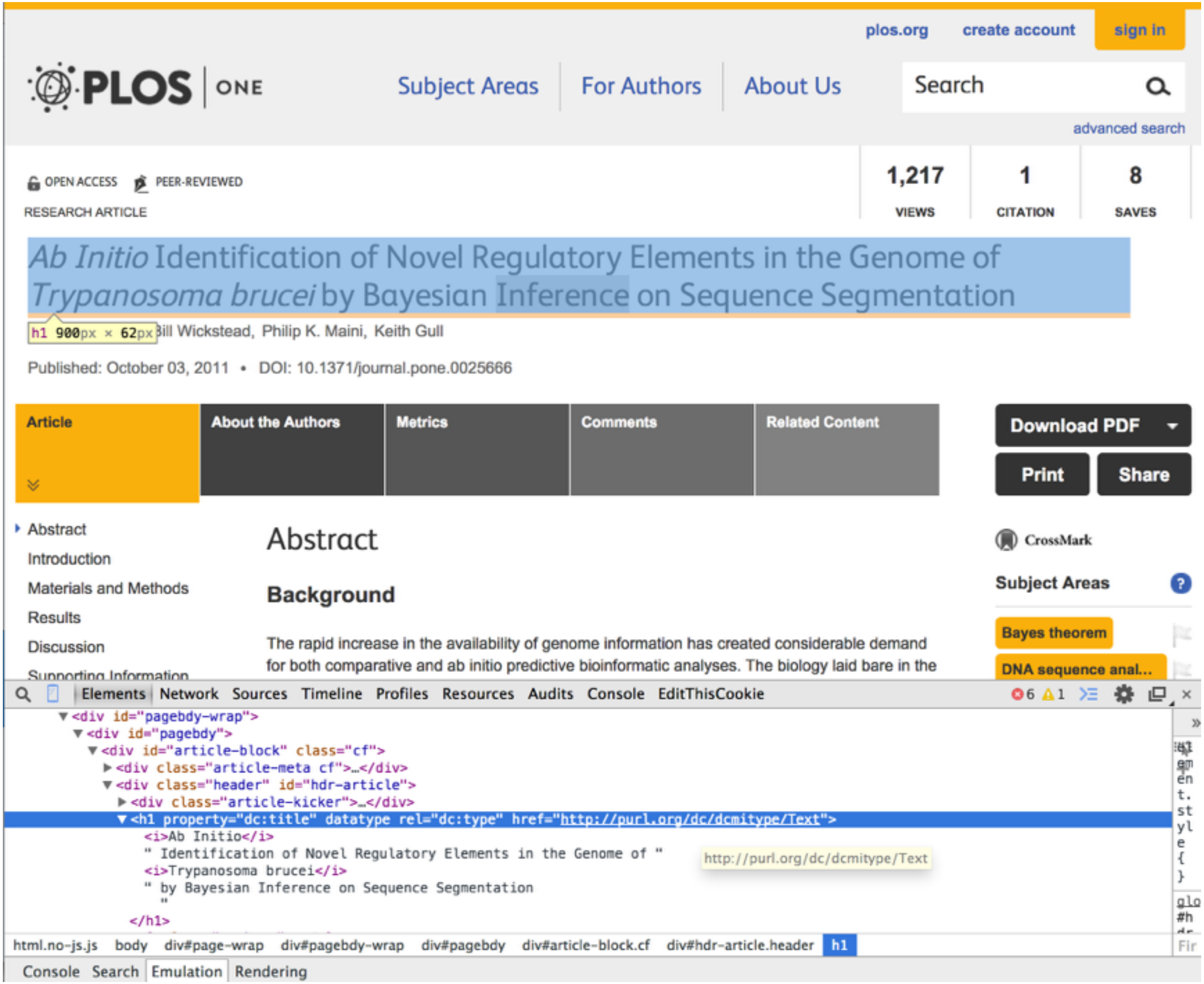
Abstract

Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the genomes of many organisms is often novel, presenting new challenges for bioinformatic interrogation. A paradigm for this is the collected genomes of the kinetoplastid parasites, a group which includes Trypanosoma brucei the causative agent of human African trypanosomiasis. These genomes, though outwardly simple in organisation and gene content, have historically challenged many theories for gene expression regulation in eukaryotes.

Methodology/Principle Findings

Here we utilise a Bayesian approach to identify local changes in nucleotide composition in the genome of T. brucei. We show that there are several elements which are found at the starts and ends of multicopy gene arrays and that there are compositional elements that are common to all intergenic regions. We also show that there is a composition-inversion element that occurs at the position of the trans-splice site.



Ab Initio Identification of Novel Regulatory Elements in the Genome of *Trypanosoma brucei* by Bayesian Inference on Sequence Segmentation

Bill Wickstead, Philip K. Maini, Keith Gull

Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

Abstract

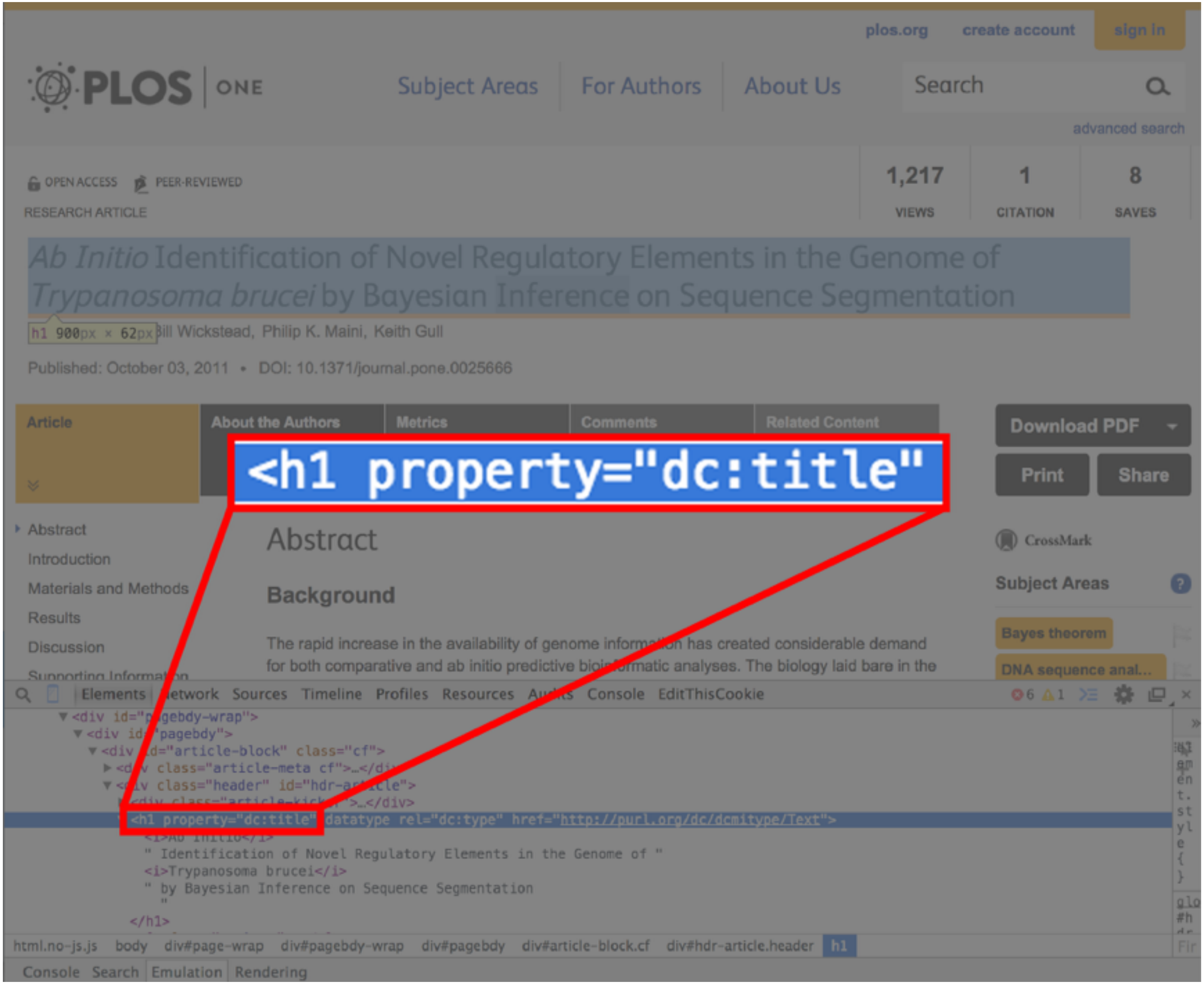
Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the

```
▼<div id="pagebdy-wrap">
  ▼<div id="pagebdy">
    ▼<div id="article-block" class="cf">
      ▶<div class="article-meta cf">…</div>
      ▼<div class="header" id="hdr-article">
        ▶<div class="article-kicker">…</div>
        ▼<h1 property="dc:title" datatype rel="dc:type" href="http://purl.org/dc/dcmitype/Text">
            <i>Ab Initio</i>
            " Identification of Novel Regulatory Elements in the Genome of "
            <i>Trypanosoma brucei</i>
            " by Bayesian Inference on Sequence Segmentation
            "
          </h1>
```

http://purl.org/dc/dcmitype/Text

html.no-js.js body div#page-wrap div#pagebdy-wrap div#pagebdy div#article-block.cf div#hdr-article.header h1



Ab Initio Identification of Novel Regulatory Elements in the Genome of *Trypanosoma brucei* by Bayesian Inference on Sequence Segmentation

Bill Wickstead, Philip K. Maini, Keith Gull

Published: October 03, 2011 • DOI: 10.1371/journal.pone.0025666

<h1 property="dc:title"

Abstract

Background

The rapid increase in the availability of genome information has created considerable demand for both comparative and ab initio predictive bioinformatic analyses. The biology laid bare in the

```
▼<div id="pagebdy-wrap">
  ▼<div id="pagebdy">
    ▼<div id="article-block" class="cf">
      ▶<div class="article-meta cf">…</div>
      ▼<div class="header" id="hdr-article">
        ▶<div class="article-kicker">…</div>
        ▼<h1 property="dc:title" datatype rel="dc:type" href="http://purl.org/dc/dcmitype/Text">
            <i>Ab Initio</i>
            " Identification of Novel Regulatory Elements in the Genome of "
            <i>Trypanosoma brucei</i>
            " by Bayesian Inference on Sequence Segmentation
            "
          </h1>
```

Basic scraperJSON

name of the scraper

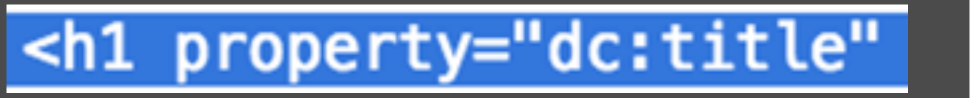
the URL(s) it applies to

the elements to capture

element name

where to find it

```
{
  "name": "PLoS",
  "url": "plos\\w*.org",
  "elements": {
    "title": {
      "selector": “//h1[@property=‘dc:title’]”,
    }
  }
}
```



<h1 property="dc:title"

http://github.com/ContentMine/scraperJSON

bibJSON output

```
{

  "title": "Ab Initio Identification of Novel
Regulatory Elements in the Genome of Trypanosoma
brucei by Bayesian Inference on Sequence
Segmentation"

}
```

thresher & quickscrape

- reference implementation of scraperJSON
- **thresher** is the scraping library
  - http://github.com/ContentMine/thresher
- **quickscrape** is the command-line tool
  - http://github.com/ContentMine/quickscrape
- Node.js, **MIT licensed**

journal-scrapers

http://github.com/ContentMine/journal-scrapers

a self-testing collection of scraperJSON scrapers for academic journals

- PLOS
- MDPI
- PeerJ
- Wiley
- ScienceDirect
- Springer
- Taylor & Francis
- NPG, AAAS, RSC, ACS, …

Future work

- GUI (browser plugin) for creating scrapers
- Standalone GUI for scraping

Acknowledgements

SHUTTLEWORTH
FELLOW

- Peter Murray-Rust
- Michelle Brook
- Mark MacGillivray
- Emanuil Tolev
- Ross Mounce
- Jenny Molloy
- Our volunteer community and collaborators
- **Funding: Shuttleworth Foundation**

http://contentmine.org
http://github.com/ContentMine