

# The LAILAPS plant science search engine: Explore plant genome databases

**Jinbo Chen**, Christian Colmsee, Uwe Scholz and Matthias Lange  
(Research Group BIT, IPK Gatersleben, Germany)  
14 October 2015 – NETTAB & IB 2015



# Outline

- Motivation
- Overview
- Implementation and Example
- Technology Details

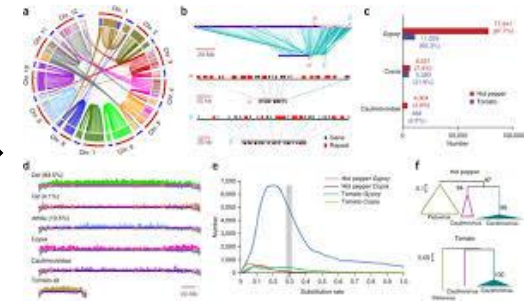
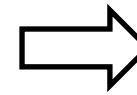
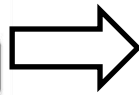
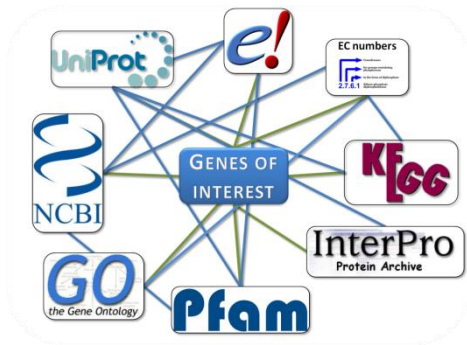
# Search Engines in Life Science

Information search for for Hordeum vulgare (barley)



Information search gets an difficult and time consuming task in a heterogeneous ecosystem of life science

# LAILAPS Integrated Search



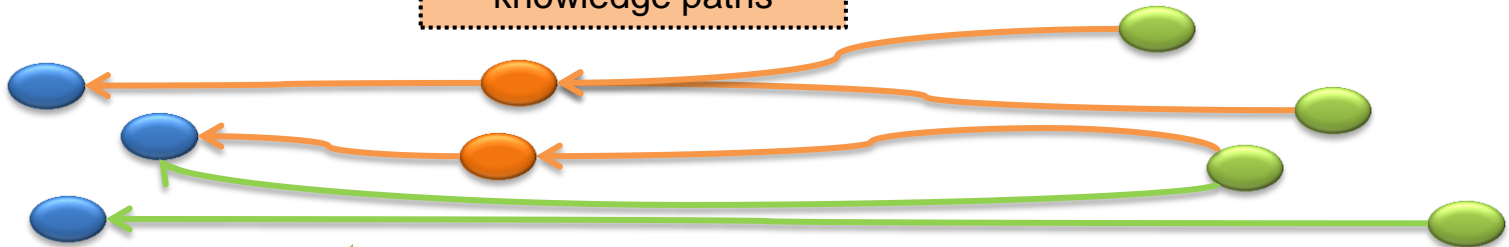
textual notated knowledge

- literature
- ontologies
- curated literature excerpts (proteins, pathways)

none textual facts databases

- genome
- metabolome
- phenome

indirect / transitive knowledge paths



direct functional annotation

- homologies, wet lab, ...

# Information Retrieval Components

## 1. Index

- text / data decomposition
- language processing
- synonyms, homonyms

## 2. Query

- efficient search in content

### 3. Ranking

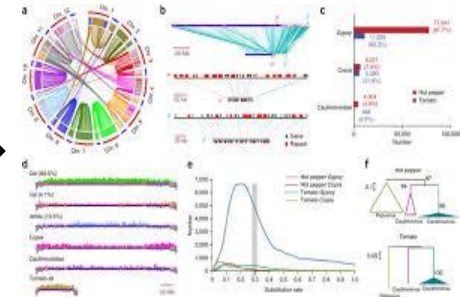
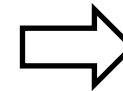
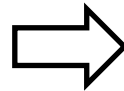
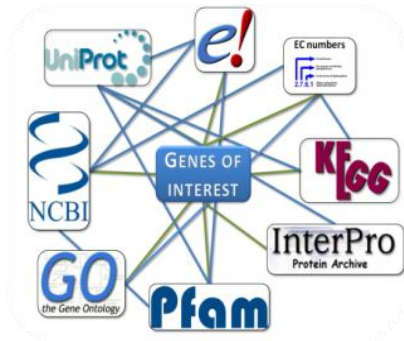
- feature extraction
- ranking functions
- pertinence (subjective user relevance profiles)

## 4. Presentation

- intuitive user interface
- related entries („page like this“)
- query suggestion („did you mean“)



# Indexing: Running Instances\*



indexed  
documents

linked  
records

transPlant, EU/UK

gene models, protein, ontologies,  
literature

$63.5 \cdot 10^6$

$50 \cdot 10^6$

IPK, Germany

plant genomes,  
germplasm collections

$62.7 \cdot 10^6$

$3 \cdot 10^6$

Helmholz Center  
Munich, Germany

barley and wheat genome, literature

$60.5 \cdot 10^6$

$0.6 \cdot 10^6$

Barcelona Supercomp.  
Center, Spain

IPK mirror

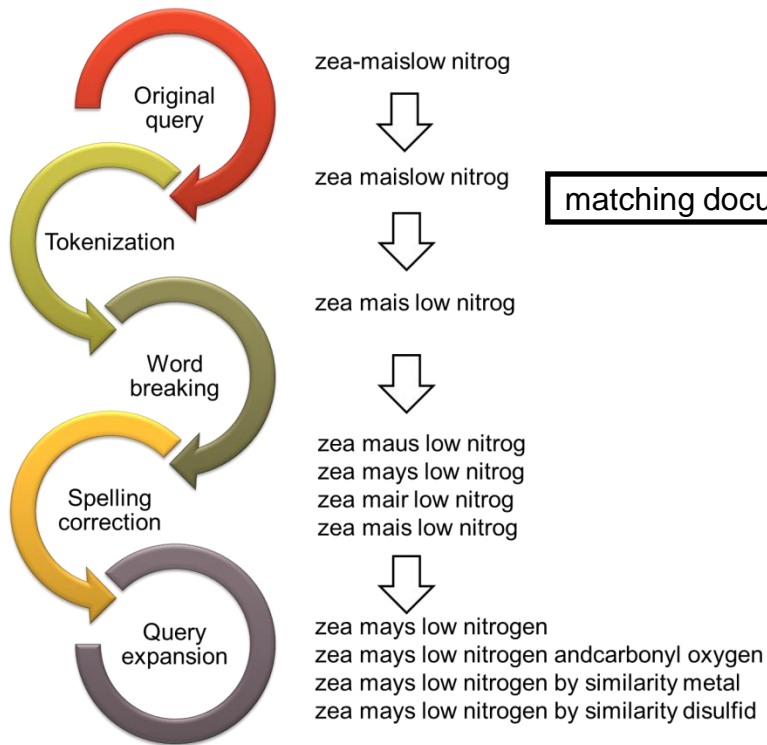
$63.5 \cdot 10^6$

$50 \cdot 10^6$

\*[http://lailaps.ipk-gatersleben.de/products/running\\_instance.html](http://lailaps.ipk-gatersleben.de/products/running_instance.html)

# Query and Ranking

query workflow

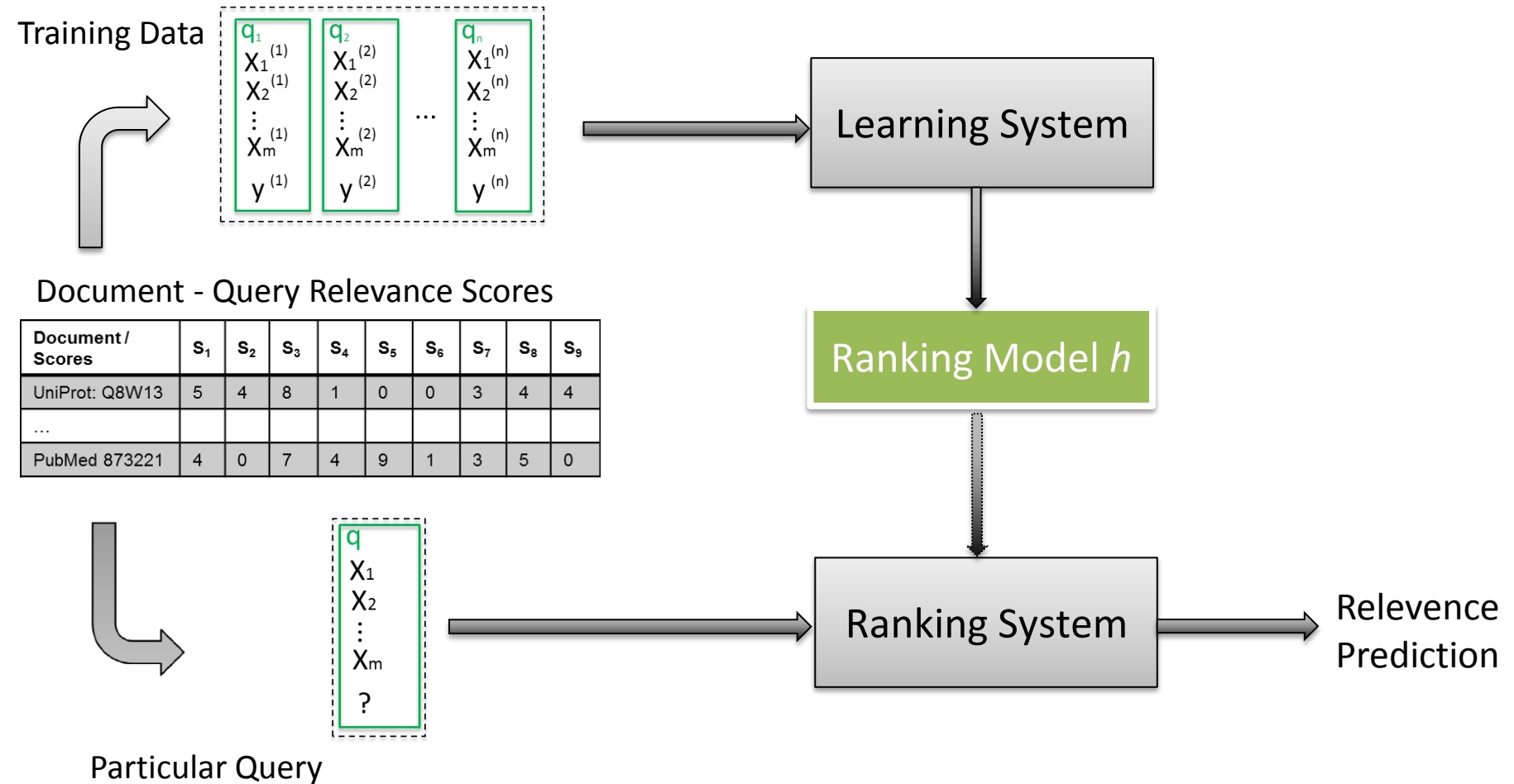


matching documents

query result

Document / Scores	S <sub>1</sub>	S <sub>2</sub>	S <sub>3</sub>	S <sub>4</sub>	S <sub>5</sub>	S <sub>6</sub>	S <sub>7</sub>	S <sub>8</sub>	S <sub>9</sub>
UniProt: Q8W13	5	4	8	1	0	0	3	4	4
...									
PubMed 873221	4	0	7	4	9	1	3	5	0

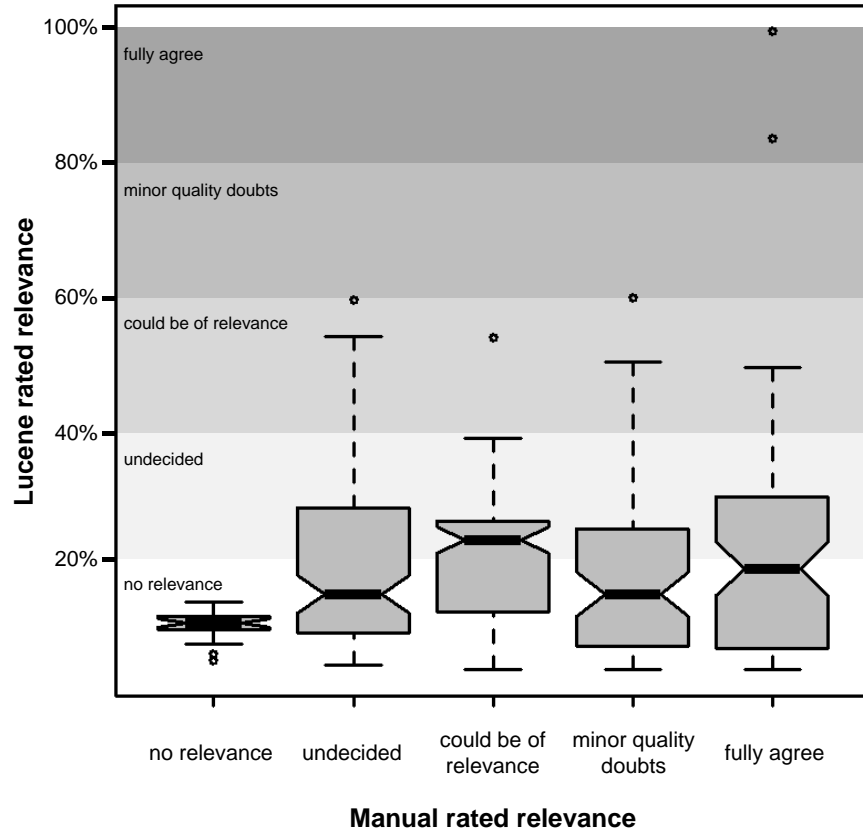
# Query and Ranking



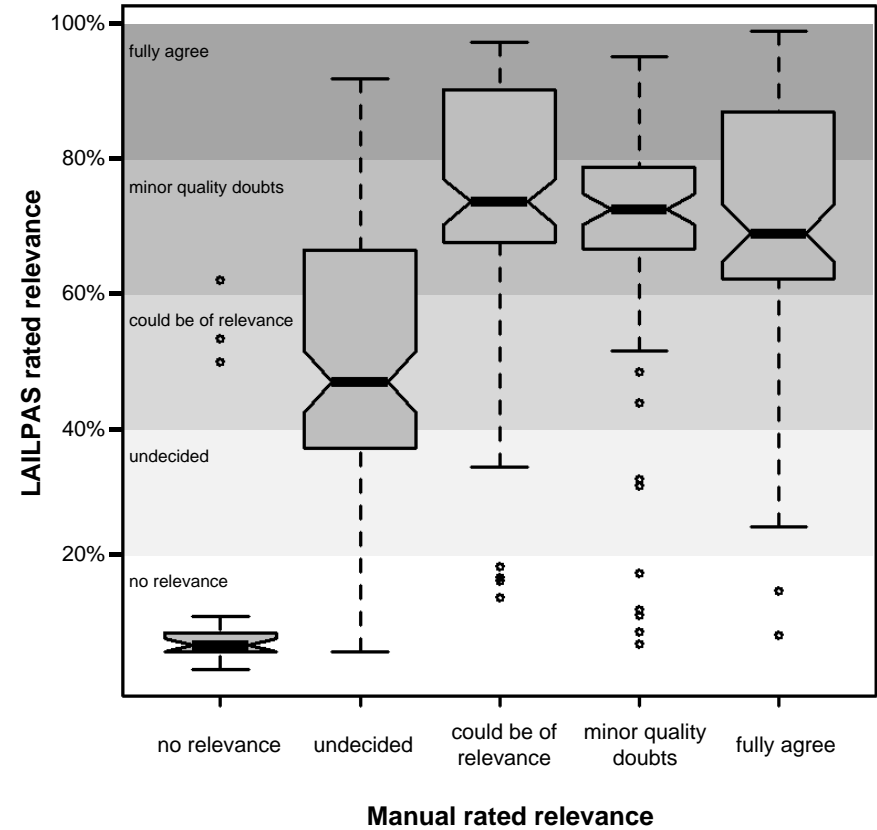


# LAILAPS Ranking Results

Relevance scoring - Expert vs. Lucene ranking



Relevance scoring - Expert vs. LAILAPS ranking



Esch et al. (*Plant and Cell Physiology*, Database Issue 2015)

# LAILAPS Example

**Query:**  
"salt stres barley"

## Spelling correction & synonym expansion:

(salt OR "CG2196") AND (stress) AND (barley OR "HORDEUM VULGARE" OR ...)

## Get most relevant trait data:

1. D0V4H8 (Score: 51.9% - uniprot\_trembl)  
'response to salt stress ...'
2. P28524 (Score: 48.9% - uniprot\_sprot)  
'... increase in roots during salt stress.'

## Link & rank genome annotations:

1. D0V4H8: INRA(21) EBI(3) MIPS(29) IPK(22)
2. P28524: INRA(4) MIPS(5) IPK(4)  
PubMed(1) BioModels(2)

The screenshot displays the LAILAPS web interface. At the top, the 'Barley project' is highlighted. The 'BAC Network Database' section shows a network of BAC clones (BAC\_39, BAC\_40, BAC\_41, BAC\_42) connected by lines. Below this, the 'Element Report' for AK355761 is shown, detailing its name, version, description (transcript\_IBSC), and references. The 'Gene family information' section shows the orthologous gene family as 'Gene family'. The 'Sequence report for AK355761' section provides a view of the sequence (CDS, Protein). The 'Protein domain informations from SIMAP' section shows Simap homologs in Vitis vinifera (12x) and a genome browser view. The 'Feature Infos' section for Vv50004g07950.t01 shows positions, length, source, and other features. The 'Germin-Like Polypeptides Increase in Barley Roots during Salt Stress' section includes the author information (Hurkman W.J., Tao H.P., Tanaka C.K.) and an abstract describing the isolation and characterization of Gs1 and Gs2 polypeptides.

**Barley project**

**BAC Network Database**

Home BAC List FPC View BLAST Results BAC Details BAC Dotplot Gene List

**Information Restricted**

More information are currently only available for members of the International Barley Sequencing Consortium (IBSC).

**Morex Contig Details**

Contig Name: morex\_contig\_6255  
Contig length: 4101  
Chromosome: 2H  
cM: 135.623229461756

**Related Genes**

Morex Contig	Gene
morex_contig_6255	AK355761

1 - 1

**Element Report**

**Name:** AK355761 **Version:** 1

**Element type:** transcript\_IBSC

**Description:** Superoxide dismutase

**Comment:** IPR001189 (Manganese/iron superoxide dismutase)

**Contig:** morex\_contig\_6255

**References**

**Confidence:** High-confidence (HC) barley gene

**References:** Crowsnest\_SyntenyToRice  
LINK to ENSEMBL plants barley instance

**Gene family information**

Show orthologous gene family: [Gene family](#)

**Sequence report for AK355761**

**View sequence:** [CDS](#) [Protein](#)

**Protein domain informations from SIMAP**

Simap homologs in: [Vitis vinifera](#)

**Vitis vinifera (12x) Genome Browser: 1.209 kbp from chr6:8,6**

**Feature Infos:**

Feature	Value
Positions	8686646 ... 8687855
Length	1210
Source	V1
trnshmm type	OUT
trnshmm coordnones	NULL
hmm type	OUT
hmm coord	NULL
psort score	3.5, 6
psort location	chloroplast/mitochondria, mitochondria
targetp location	M
targetp length	86
kegg ec	ec:1.15.1.1
kegg description	superoxide dismutase

**Germin-Like Polypeptides Increase in Barley Roots during Salt Stress.**

Hurkman W.J., Tao H.P., Tanaka C.K.

**Author information**

**Abstract**

The 26 kilodalton, isoelectric point 6.3 and 6.5 (Gs1 and Gs2) polypeptides that increase in barley (*Hordeum vulgare* L.) roots during salt stress were isolated and identified. Both Gs1 and Gs2 had high sequence similarity to germin, a protein that increases significantly in germinating wheat seeds. Like germin, Gs1 and Gs2 were resistant to proteases and were glycosylated. Immunoblots were probed with antibodies to Gs1 and Gs2 to determine the distribution of these polypeptides among organs and cell-free fractions. Gs1 and Gs2 were present in roots and coleoptiles, but absent from leaves. In roots, Gs1 and Gs2 were present in the mature region, but not the tip. Gs1 and Gs2 increased in roots, but decreased in coleoptiles in response to salt stress. Gs1 and Gs2 were distributed among the soluble, microsomal, and cell wall fractions of roots, but the majority of Gs1 and Gs2 was present in the soluble fraction. Although Gs1 and Gs2 were heat stable, their synthesis was not affected by abscisic acid treatment. Gs2 accumulated during abscisic acid treatment, whereas Gs1 did not. However, a 25.5 kilodalton, isoelectric point 6.1 polypeptide that was immunologically related to Gs1 did accumulate with abscisic acid treatment.

# Personal data and customization

## CSV format

metadata repository	metadata id	genomics datarecord ID	evidence
Pfam	PF04998	morex_contig_46112	1.0
InterPro	IPR007081	morex_contig_46112	1.0
gene_ontology	GO:0003677	morex_contig_46112	1.0
gene_ontology	GO:0003899	morex_contig_46112	1.0
gene_ontology	GO:0006351	morex_contig_46112	1.0

## GFF3 format

1	2	3	4	5	6	7	8	9
Arad1A.1	EMBL	gene	75	1268	.	-	0	note=conserved hypothetical protein;Dbxref=UniProtKB/TrEMBL:A0A060T284;...
Arad1A.1	EMBL	exon	75	1268	.	-	0	Parent=HG937691.3
Arad1A.1	EMBL	gene	1779	2441	.	-	0	note=no similarity;Dbxref=UniProtKB/TrEMBL:A0A060SWA0;...
Arad1A.1	EMBL	exon	1779	2441	.	-	0	Parent=HG937691.5
Arad1A.1	EMBL	gene	3702	4043	.	+	0	note=similar to uniprotIP39534 Saccharomyces cerevisiae SACE0J00586g MBB1_YEAST...

# Personal data and customization

Filter

Indexed Databases

☒ Biomedical Literature Databases

☒ PubMed (243)

☒ Ontologies

☒ Gene Ontology Consortium (6)

☒ Plant Ontology (0)

☒ Plant Diversity Resources

☒ GBIS-I Genebank Information System of the IPK Gatersleben

☒ Protein Sequence Databases

☒ InterPro (5)

☒ Pfam (2)

☒ RCSB PDB (2)

☒ UniProtKB(Swiss-Prot) (184)

☒ UniProtKB(TrEMBL) (1013)

Linked Databases

☒ Public Database

☒ BARLEX/Barley/IPK (10)

☒ EnsemblPlants (501)

☒ PlantsDB/Barley/PGSB (10)

☒ PlantsDB/Wheat/PGSB (11)

☒ PubMed (1233)

☒ Private Database

☒ EnsemblPlants(wheat) (2)

Link Types

☐ indirectlink (1243)

☐ allhits (6527)

☒ directlink (1455)

apply

**F8T7T0; SubName: Full=FTa1 {ECO:0000313|EMBL:AEI99551.1}; SubName: Full=Flowering I**  
Score:57,2% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:* Regulator of **Flowering Time**.; *'Fine Mapping Links the FTa1 Flowering Time; DESCRIPTION:=Regulator of flowering time {ECO:0000313|EMBL:AGC60097.1}; SubName: Full;*  
EnsemblPlants (1) PubMed (5)

**Q7X9X7; 'Analysis of the molecular basis of flowering time variation in Arabidopsis**  
Score:57,1% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:*'Analysis of the molecular basis of **flowering time** variation;  
PubMed (3)

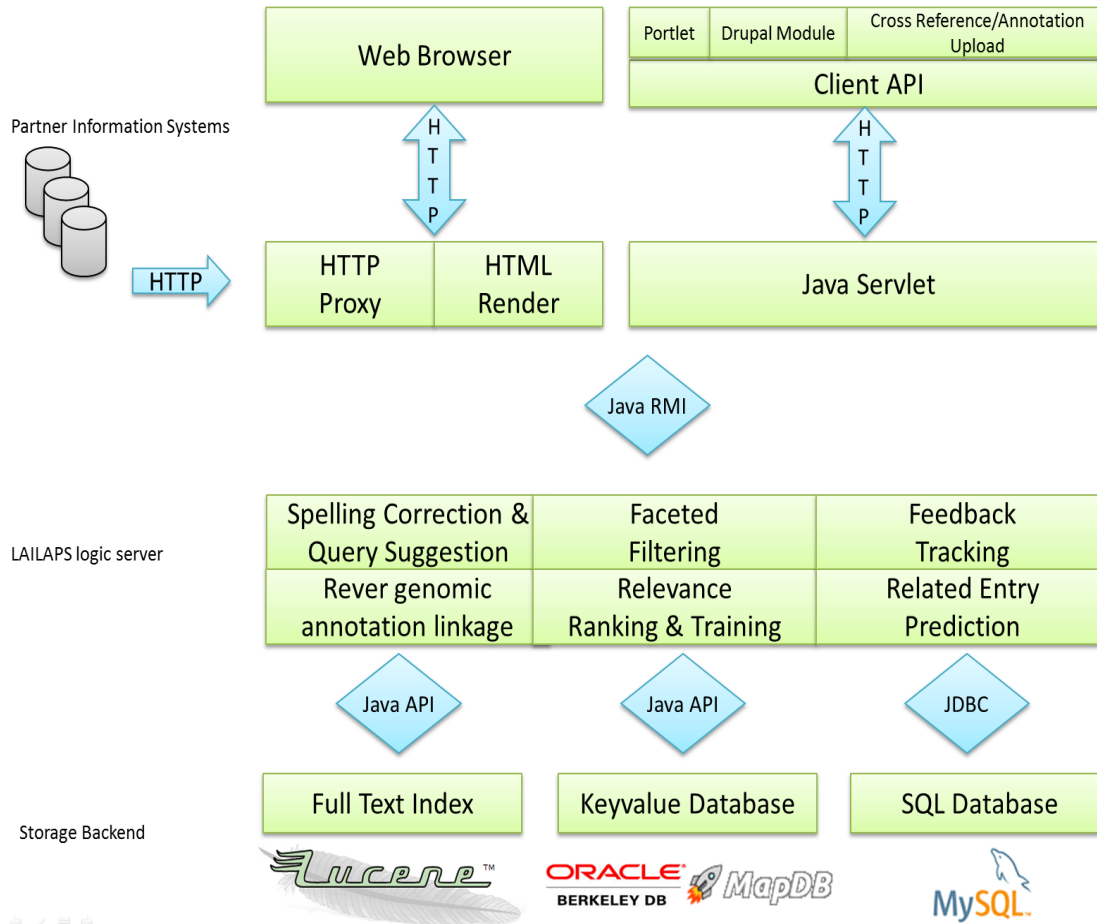
**C0SSC4; 'Diversification in flowering time due to tandem FT-like gene duplication,**  
Score:57% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:*'Diversification in **flowering time** due to tandem FT-like gene;  
PubMed (2)

**Q5Q9J1; SubName: Full=At5g10140 {ECO:0000313|EMBL:ABR46217.1}; SubName: Full=FLC {E**  
Score:57% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:* FRI and FLC **flowering time** genes generates a latitudinal cline;  
*DESCRIPTION:* protein {ECO:0000313|EMBL:AAV51219.1}; SubName: Full=**Flowering time**;  
EnsemblPlants (1) PubMed (4)

**Q58T28; 'Analysis of the molecular basis of flowering time variation in Arabidopsis**  
Score:57% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:*'Analysis of the molecular basis of **flowering time** variation;  
PubMed (4)

**B2XAJ0; SubName: Full=FLC {ECO:0000313|EMBL:AFP49589.1}; SubName: Full=Flowering lo**  
Score:57% Source:UniProtKB(TrEMBL)  
*REFERENCE TITLE:* C homeolog is associated with **flowering time** variation;  
*DESCRIPTION:=Flowering locus C {ECO:0000313|EMBL:ABX25246.1}; Flags: Fragment;;*  
PubMed (3)

# Implementation



## Efficient Algorithms and Data Structures:

- In-memory data structures
  - compressed HashMaps
  - bit-efficient, dynamic number types
- In-memory pre-filter
  - Bloom filter
  - Tries
- Hybrid of relational and key-value databases
- Off-heap memory (avoid garbage collection)

## Hardware Requirements:

- RAM: 16GB
- CPU: 4 core, 4GHz
- storage 800GB SSD
- cost < \$1500
- performance:
  - max. response time: 20 sec. (broad query: e.g. „gene“ etc.)
  - 25 parallel queries



# Acknowledgements



Member of the



Matthias Lange  
Daniel Arend  
Christian Colmsee  
Uwe Scholz  
Thomas Münch  
Jens Bauernfeind



[lailaps.ipk-gatersleben.de](http://lailaps.ipk-gatersleben.de)



from March 2015

de.NBI

*Thank you for your attention  
and see you at the Poster*