



# Extraction d'information pour la sélection du blé par marqueur génétique



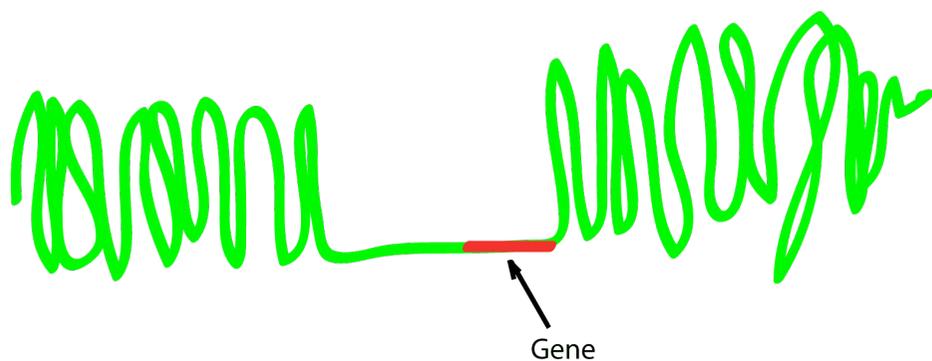
*Dialekti Valsamou, Robert Bossy, Marion Ranoux,  
Wiktorja Golik, Pierre Sourdille, Claire Nédellec*

**Clermont-Ferrand – 13 mai 2014**

## Sélection du blé assistée par marqueur génétique (*SAM*)

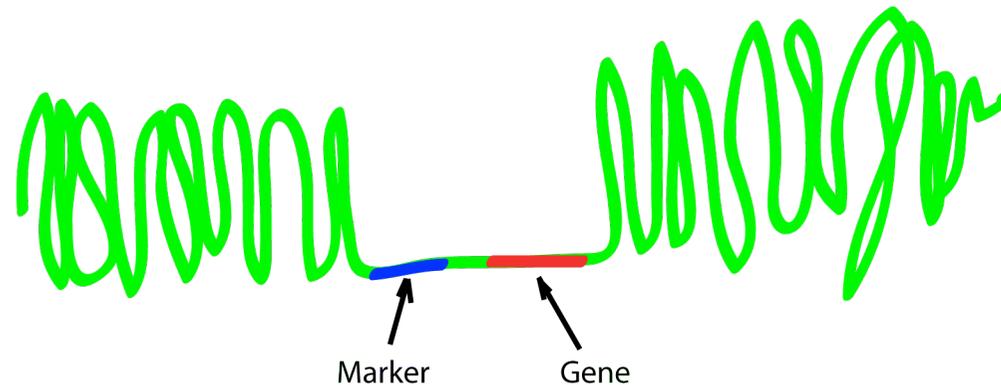
Les gènes déterminent les traits des variétés de blé (ex. la taille).

Les allèles des gènes déterminent les phénotypes (ex. grand)





## Sélection du blé assistée par marqueur génétique



Les marqueurs génétiques indiquent la proximité de l'allèle du gène.

Ils sont plus facile à identifier expérimentalement que le gène.

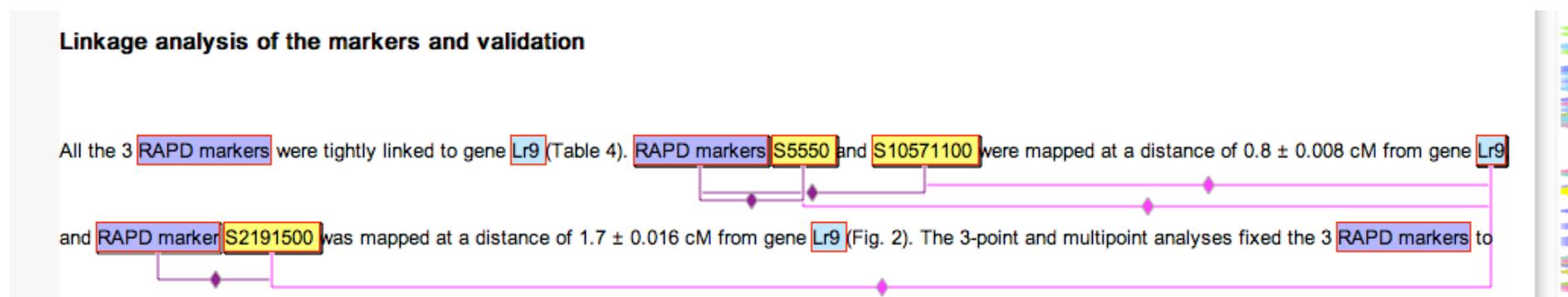
Ils sont utilisés pour sélectionner les variétés qui présentent un phénotype d'intérêt agronomique.

# Projet FSOV Sélection du Blé par Marqueur, *SamBlé*

Les marqueurs sont majoritairement décrits dans des articles scientifiques

## Objectif

Extraire automatiquement à partir d'articles, les relations entre les marqueurs moléculaires, la méthode mise en œuvre, les gènes, les traits observables, les phénotypes et les variétés.





## Projet FSOV Sélection du Blé par Marqueur, *SamBlé*

### Résultats attendus

- Moteur de recherche sémantique de la bibliographie scientifique
- Base de données des informations extraites de la bibliographie interrogeable en ligne

### Basés sur

- Des méthodes d'annotation sémantiques de texte
- Une ontologie des traits, phénotypes et facteurs biotiques et abiotiques pour le blé
- Un modèle de connaissance des gènes, marqueurs, traits, phénotypes, variétés, *etc.*

01



# *Moteur de recherche*

# Moteur de recherche sémantique *SamBlé*

**Alvis** Search Engine

Search: (resistance to a fungal pathogen) sr2

Results | Query details | Ontology navigation | Refinement | +

Query terms : [resistance to a fungal pathogen](#) (pathogen resistance) [sr2](#) (gene) [more details...](#)

1-10 among 46 results.

- BAC-derived markers for assaying the [stem rust resistance](#) gene, [Sr2](#), in wheat breeding programs**  
Journal: MOLECULAR BREEDING Date: 2008  
BAC-derived markers for assaying the [stem rust resistance](#) gene, [Sr2](#), in wheat breeding programs 123 [More...](#)
- Fine genetic mapping fails to dissociate durable [stem rust resistance](#) gene [Sr2](#) from pseudo-black chaff in common wheat (*Triticum aestivum* L.)**  
Journal: THEORETICAL AND APPLIED GENETICS Date: 2006  
Fine genetic mapping fails to dissociate durable [stem rust resistance](#) gene [Sr2](#) from pseudo-black chaff in common wheat (*Triticum aestivum* L.) wheat. *Phytopathology* 82:835-838 Sorrells ME, La Rota M, Bermudez-Kandianis CE et al. (2003) Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res* 13:1818-1827 Spielmeier W, Sharp PJ, Lagudah ES (2003) Identification and validation of markers linked to broad-spectrum stem rust resistance gene [Sr2](#) in wheat (*Triticum aestivum* L.). *Crop Sci* 43:333-336 Yan L, Loukoianov A, Tranquilli G, Helguera M, Fahima T, Dubcovsky J (2003) Positional cloning of the wheat vernalization gene VRN1. *Proc Natl Acad Sci USA* 100:6263-6268 [More...](#)
- BAC-derived markers for assaying the [stem rust resistance](#) gene, [Sr2](#), in wheat breeding programs**  
Journal: MOLECULAR BREEDING Date: 2008  
BAC-derived markers for assaying the [stem rust resistance](#) gene, [Sr2](#), in wheat breeding programs [rust resistance](#) genes that can mask the effect of this gene (McIntosh et al. 1995). Moreover, the resistance phenotype is only expressed at the adult plant stage, which delays the classification of progeny (Roelfs 1988). The phenotypic trait pseudo-black chaff

**Refinement shortcuts**

**Concepts**

- rust resistance
- stem rust resistance
- black chaff

**Taxa**

- Triticum aestivum*
- Oryza sativa*
- Embryophyta

**Genes**

- Sr2
- Lr34
- Yr18

**Varieties**

- Halberd
- Cranbrook
- Kota

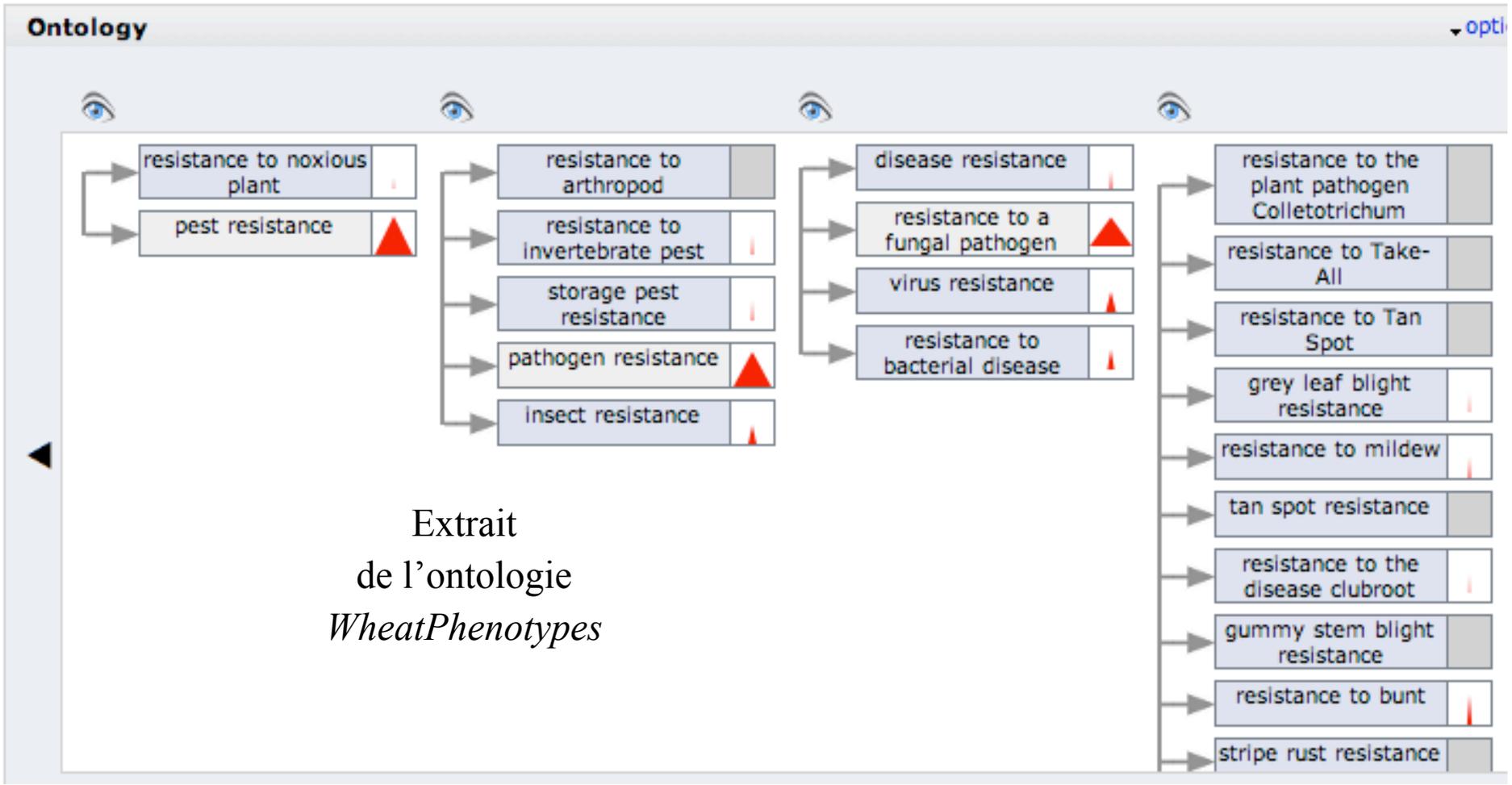
**Markers**

- Yr18
- Sun2
- wmc291

**Journals**

- MOLECULAR BREEDING

<http://bibliome.jouy.inra.fr/test/alvisir/FSOV/>





## Collection d'articles pour le moteur de recherche *SamBlé*

- Références sélectionnées dans WoS par la requête « Wheat marker gene aestivum » parmi les journaux les plus pertinents.
- 3 170 articles scientifiques complets téléchargés par *AlvisCrawler*
- Annotation sémantique par la *Suite Alvis*.
- Indexation et IHM de *AlvisIR*

\_02



# *Extraction d'Information*

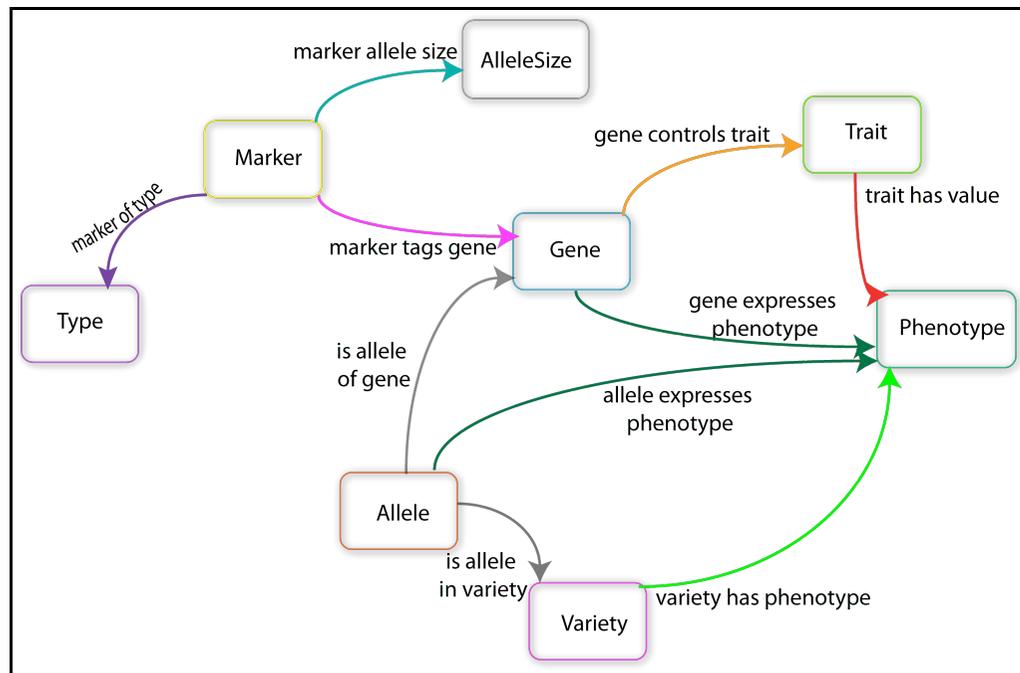
# Annotation sémantique du texte (*AlvisAE*)

**Linkage analysis of the markers and validation**

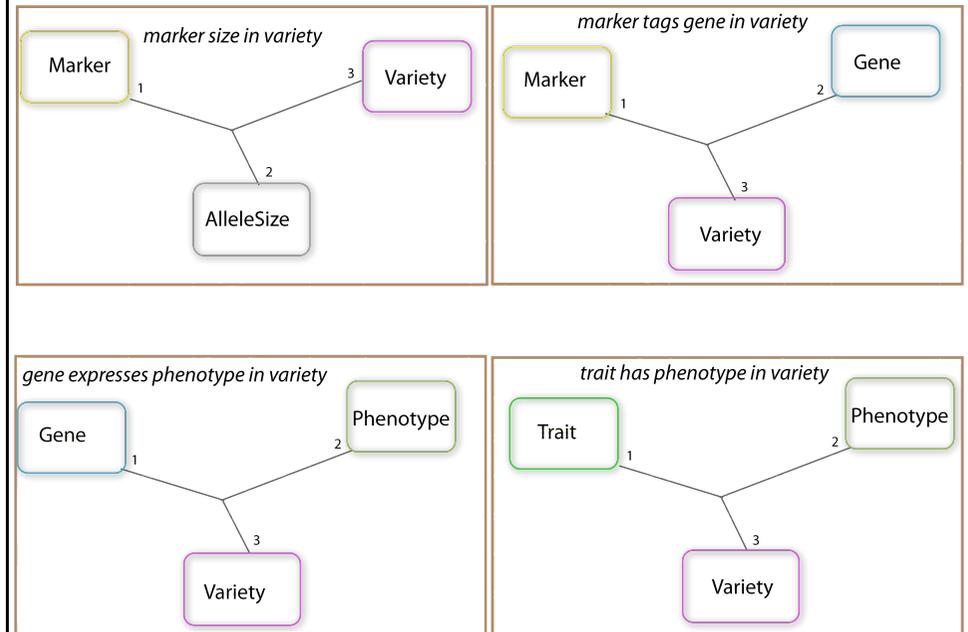
All the 3 RAPD markers were tightly linked to gene Lr9 (Table 4). RAPD markers S5550 and S10571100 were mapped at a distance of  $0.8 \pm 0.008$  cM from gene Lr9 and RAPD marker S2191500 was mapped at a distance of  $1.7 \pm 0.016$  cM from gene Lr9 (Fig. 2). The 3-point and multipoint analyses fixed the 3 RAPD markers to within a distance of 1.7 cM from gene Lr9, and to 1 side of the Lr9 locus (Fig. 2). On the basis of the sequence of cloned RAPD-marker fragment S5550 a pair of forward and reverse primers were synthesized, which specifically amplified the single 550-bp band of SCAR marker SCS5550 in resistant F2 plants including the 81 heterozygous F2 plants and the parent HW 2055 carrying gene Lr9. The sequences of the 2 primers of SCAR marker SCS5550 were 5'-TGCGCCCTCAAAGGAAG-3' (forward primer) and 5'-TGCGCCCTTCTGAACTGTAT-3' (reverse primer). The underlining identifies the original sequence of RAPD primer S5. The SCS5550 marker fragment was also amplified in 10 leaf-rust-resistant wheat lines carrying gene Lr9 in 10 different genetic backgrounds, and was completely absent in the

# Modèle de Connaissance

Modèle conçu pour l'annotation sémantique, manuelle et automatique et la base de données.  
C'est une représentation formelle des connaissances biologiques à extraire du texte



Relations binaires



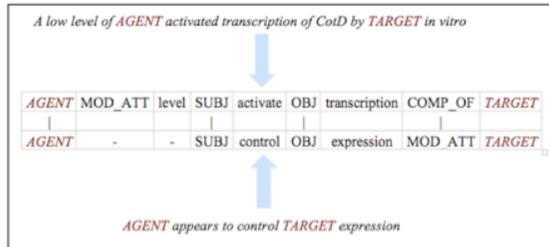
Relations ternaires



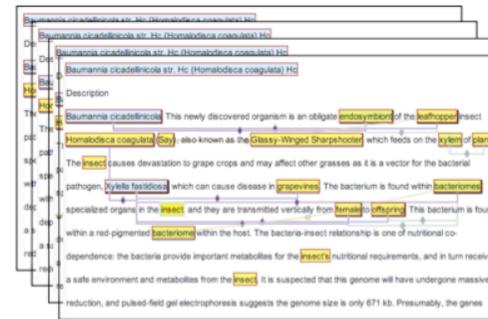
# Principe de l'extraction automatique d'information par la Suite Alvis

1. Annotation manuelle d'exemples dans les texte, suivant le modèle (avec *AlvisAE*)
2. Analyse sémantique des textes par *AlvisNLP*
3. Induction de règles par apprentissage automatique (*AlvisML*)
4. Prédiction des entites et des relations dans de nouveaux textes (*AlvisNLP*)

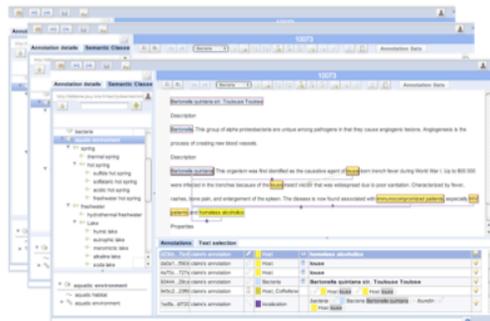
## Machine learning for entity and relation prediction



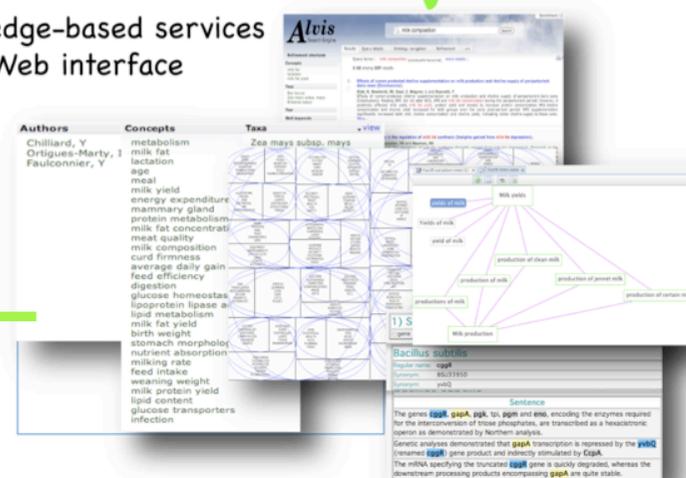
Fine-grained text-bound semantic annotation



Semi-automatic annotation of examples of entities and relations. Web interface



Knowledge-based services Web interface



## Annotation d'exemples d'apprentissage

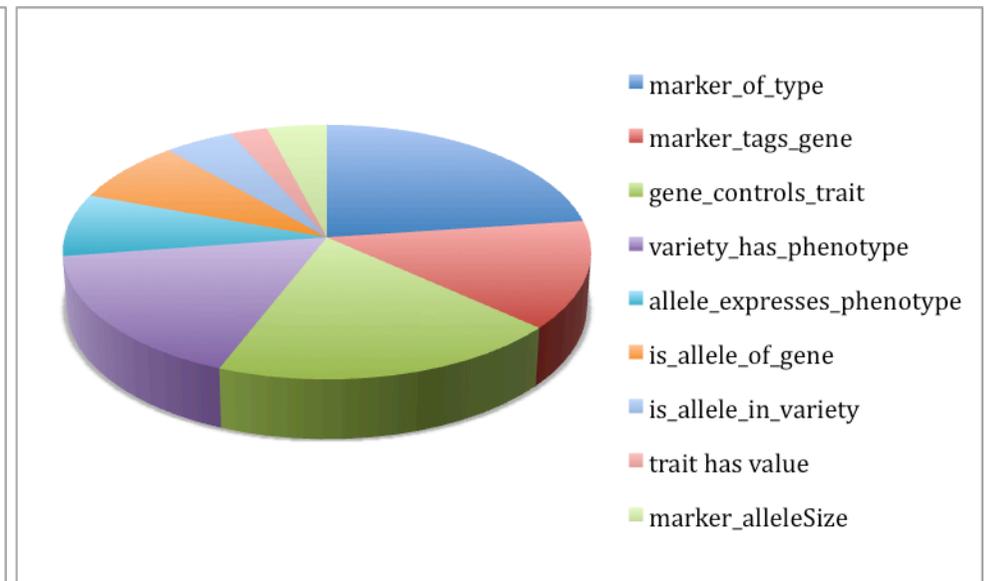
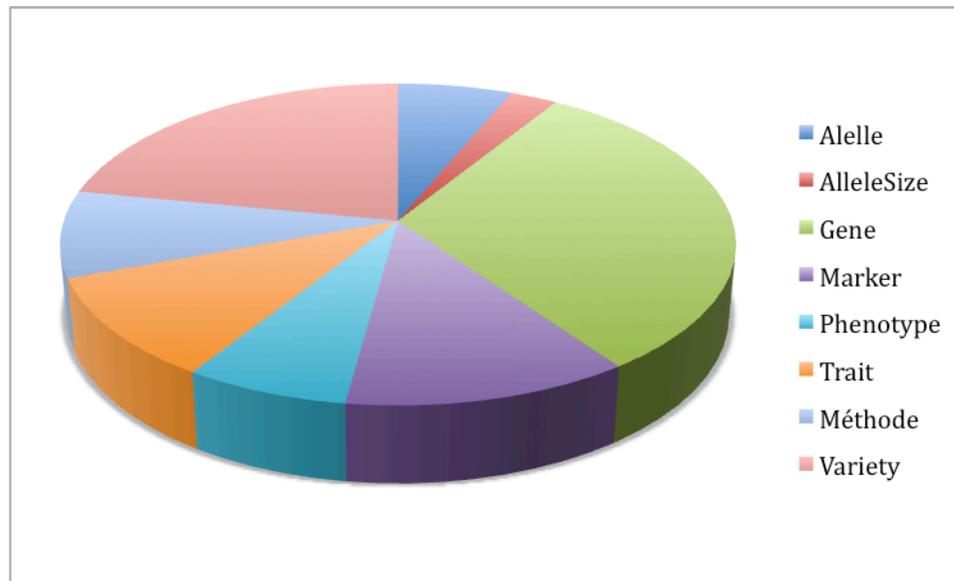
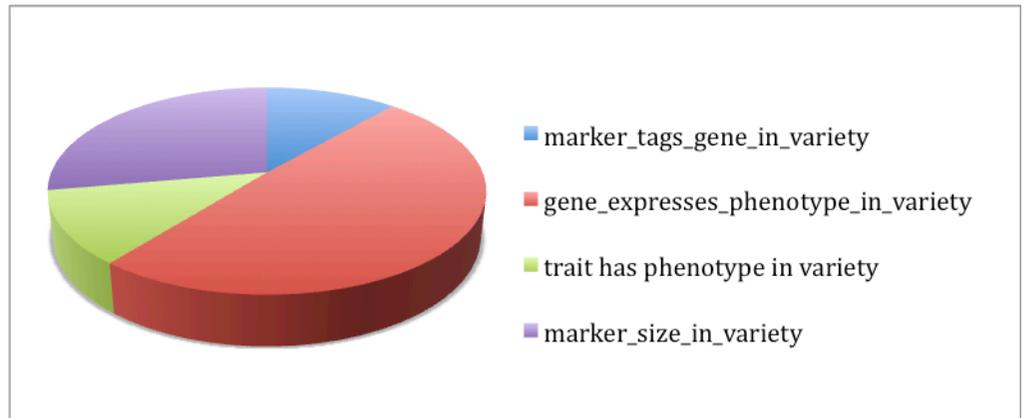
### 2 campagnes d'annotation (*en cours*)

13 partenaires participants (sélectionneurs)

72 articles, 292 sections

5 848 entités, 1 342 relations binaires,

207 relations ternaires



03



# *Reconnaissance automatique d'entités*



## Types d'entités

### Deux types d'entités

#### Entités figées

Gènes : *Lr9*

Allèles : *Ppd-D1a*

Tailles d'allèles : *64 bp*

Variétés : *HW 2055*

Marqueurs : *S5550*

Types de méthodes : *RAPD*

#### Entités non figées

Traits et phénotypes : grande diversité de termes

*superior fusarium head blight resistance,*      *quality of wheat flour,*  
*resistance to leaf rust infection,*                      *low grain protein content trait*

La projection de termes connus sur le texte ne suffit pas, il faut pouvoir prédire des termes nouveaux.



## Méthode de prédiction des entités figées

- Projection de nomenclatures (GRIN, MasWheat, GenBank, Gramene, *etc.*)
- Expressions régulières pour désambigüiser à l'aide du contexte (ex. variété *Leeds*)
- Expressions régulières pour traiter les variations typographiques  
*Puccinia graminis f. sp. tritici (Pgt)*



## Prédiction des entités figées, premiers résultats

Mesurés en comparant prédiction et annotation manuelle

	Comparaison exacte			Recouvrement partiel		
	Rappel	Précision	F1	Rappel	Précision	F1
<b>Gene</b>	0.61	0.49	0.54	0.73	0.61	0.66
<b>Marker</b>	0.58	0.65	0.61	0.59	0.66	0.62
<b>Type</b>	0.54	0.62	0.58	0.56	0.64	0.60
<b>AlleleSize</b>	0.39	0.49	0.43	0.46	0.50	0.48

- Le rappel augmente significativement avec le relâchement des bornes
- Gènes et marqueurs ne sont pas toujours bien distingués les uns des autres



# Prédiction des traits et phénotypes

## Principe

- Extraire automatiquement *tous* les termes des documents
- Filtrer les termes qui *ressemblent* à des termes de l'ontologie *WheatPhenotypes*

## Analyse linguistique du texte par AlvisNLP

- Segmentation par **SegMig**. Analyse syntaxique par **TreeTagger** [Schmid, 1994]
- Extraction de termes par **BioYaTeA** [Golik et al., 2012]
- Filtrage des termes désignant des traits ou des phénotypes par **ToMap** [Golik et al., 2011] à l'aide de l'ontologie *WheatPhenotypes*

## Extraction des termes par *BioYaTeA*

Stem rust caused by *Puccinia graminis f. sp. tritici Eriks and Henn* and leaf rust caused by *Puccinia triticina Rob. ex Desm.* are major constraints to wheat production worldwide. In the present study, F4-derived SSD population, developed from a cross between Australian cultivars 'Schomburgk' and 'Yarralinka', was used to identify molecular markers linked to rust resistance genes Lr3a and Sr22. A total of 1,330 RAPD and 100 ISSR primers and 33 SSR primer pairs selected on the basis of chromosomal locations of these genes were used. The ISSR marker UBC 840540 was found to be linked with Lr3a in repulsion at a distance of 6.0 cM. Markers cfa2019 and cfa2123 flanked Sr22 at a distance of 5.9 cM (distal) and 6.0 cM (proximal), respectively. The use of these markers in combination would predict the presence or absence of Sr22 in breeding populations. A previously identified PCR-based diagnostic marker STS638 linked to Lr20 was validated in this population. This marker showed a recombination value of 7.1 cM with Lr20.

### Termes

Groupes nominaux (en jaune) entre les frontières prédéfinies (en rouge)

Sous-termes extraits récursivement en fonction de leurs occurrences dans les textes

Exemple : *identified [PCR-based diagnostic[ marker]] STS638*



## Principe de ToMap (1)

En trois étapes.

Un terme du texte désigne un phénotype, s'il désigne exactement un concept de phénotype de l'ontologie *WheatPhenotypes*

Texte	<i>cultivars that originally were <b>resistant to leaf rust</b>, and races</i>	
Concept	<i>resistant to leaf rust</i>	



## Principe de ToMap (2)

Un terme du texte désigne un phénotype, si sa tête syntaxique (mot principal) est égale à une tête de terme de l'ontologie *WheatPhenotypes*

Texte	<i>Prevalence of <b>ToxA-sensitive</b> alleles of the wheat gene Tsn1</i>	
Concepts	<i>toxin <u>sensitive</u>    <u>sensitive</u> to photoperiod    drought <u>sensitive</u></i>	
Texte	<i><b>polymorphism information <u>content</u></b></i>	
Concept	<i>Grain Protein <u>Content</u></i>	

La tête *Content* est ambiguë. Un patron contextuel permet de désambigüiser.



## Principe de ToMap (3)

Un terme du texte désigne un phénotype si

- Sa tête appartient à la liste des têtes ambiguës, définie manuellement préalablement (*number, level, plant, size, time, index*)
- Et que la tête de son sous-terme est une tête d'un terme de l'ontologie *WheatPhenotypes*

Texte *indicating a **higher level of polymorphism*** → Non retenu



Texte *genes conferring a **high level of stripe rust resistance** are*

Concept *stripe rust resistance*



Il est intéressant de conserver aussi *high level of stripe rust resistance* (cas 2)  
en plus de *stripe rust resistance* (cas 3)

## Résultats préliminaires

Les textes d'articles annotés ne sont pas utilisables pour l'évaluation.  
L'annotation manuelle des phénotypes dans n'est pas terminée.

La précision est mesurée sur un ensemble de 1 789 résumés de 300 000 mots.

	Sans désambiguïsation		Avec désambiguïsation	
Validation	Nb termes	Proportion	Nb de termes	Proportion
<b>Exemples positifs</b>	<b>245</b>	<b>81 %</b>	<b>212</b>	<b>95 %</b>
Corrects et précis	227	76 %	176	79 %
Corrects et généraux	18	5 %	36	16 %
<b>Exemples négatifs</b>	<b>54</b>	<b>19 %</b>	<b>11</b>	<b>5%</b>
Erreur d'analyse linguistique	5	1,7%	4	2 %
Erreur de la méthode	16	5,4%	7	3 %
Erreur de paramètre (têtes ambiguës)	33	11 %	0	0 %



## Conclusion

**Le modèle de connaissance** pertinent pour l'annotation des textes et la représentation des connaissances

Une fois formés, les sélectionneurs produisent les **annotations manuelles** nécessaires à l'apprentissage.

**L'ontologie WheatPhenotypes** couvre tous les traits d'intérêt agronomique.

Elle indexe la collection d'article du **moteur de recherche sémantique *SamBlé***. Il donne aux sélectionneurs l'accès à la connaissance bibliographique

**Extraction d'information** à systématiser et à exporter dans une base de donnée intégrée (projet *BreedWheat*)

- La méthode OntoMap produit des résultats encourageants malgré la forte polysémie