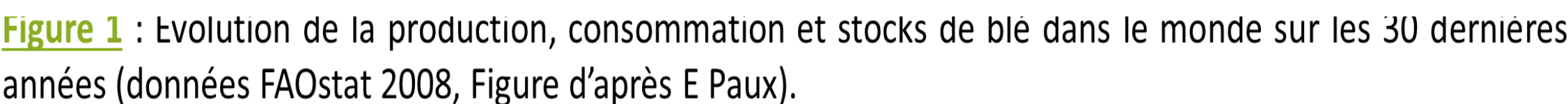


FSOV 2010 H

**Validation de marqueurs liés à des gènes d'intérêt en vue de
l'établissement d'une base de données pour la sélection assistée
chez le blé tendre**



Pour la recherche automatisée, deux corpus ont été développés : (1) les gènes, marqueurs, méthodes, variétés connus ; (2) les phénotypes et caractères en utilisant l'ontologie *WheatPhenotype* de 624 concepts et synonymes. Le moteur de recherche sémantique AlvisIR a été développé sur une base bibliographique pertinente de 3170 articles. Il utilise comme index de recherche les corpus (1) et (2). L'entraînement du moteur est fait en utilisant des exemples dérivés d'annotations manuelles des articles.

Thématiques	# de gènes	# de gènes avec un marqueur lié	# marqueurs testés	# de marqueurs testés sur panel
Qualité boulangère	16	10	31	20
Stress biotiques	91	63	158	115
Stress abiotiques	6	3	9	2
Développement	13	13	100	47
Total	126	101	298	184

Afin d'estimer la corrélation entre la présence d'allèles particuliers aux locus marqueurs et d'allèles favorables pour les gènes d'intérêt suivis, des essais en plein champ sur différentes maladies (rouilles brune et jaune, septoriose, oïdium, fusariose ; **Tableau 4**) et en conditions contrôlées (rouille jaune) ont été réalisés. Des résultats de liaison entre marqueurs et gènes ont ainsi pu être confirmés (**Figure 3A et 3B**).

Tableau 4 : Dispositif expérimental mis en place avec les partenaires pour les essais en plein champ.

Figure 1 consists of two panels, A and B, each showing box plots with overlaid individual data points for two conditions: 0 and 1.

Panel A: Gene Expression (log scale). The y-axis ranges from 2 to 8. Condition 0 has a median around 5.5, while Condition 1 has a median around 4.0.

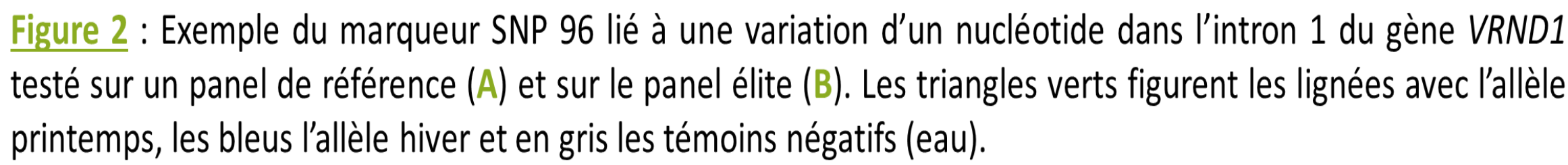
Panel B: BOLD time (sec). The y-axis ranges from 0 to 150. Condition 0 has a median around 100, while Condition 1 has a median around 75.

- 1 – INRA – UMR 1095 GDEC (Génétique, Diversité et Ecophysiologie des Céréales) 5, Chemin de Beaulieu, 63039 Clermont-Ferrand cedex
- 2 – INRA – UR 1077 MIG (Mathématique, Informatique et Génome), Domaine de Vilvert, 78352 Jouy-en-Josas
- 3 – Arvalis Institut du Végétal 3, rue Joseph et Marie Hacking, 75116 Paris
- 4 – INRA – UR BIOGER-CPP (Biologie et Gestion des Risques en Agriculture – Champignons Pathogènes des Plantes), BP01, 78850 Thiverval Grignon
- * Adresse actuelle : RAGT - Rue Émile Saigne, Site de Bourran – BP 3336, 12033 Rodéz Cedex 09
- * Coordinateur : Pierre SOURDILLE, pierre.sourdille@clermont.inra.fr, Tél. : 04 73 62 44 37

Tableau 1 : Prédiction des entités
figées mesurée en comparant la
prédiction faite par le moteur et
l'annotation manuelle. Le rappel et
la précision augmentent
significativement lorsque l'on
considère un recouvrement partiel.
Les gènes et les marqueurs sont
parfois difficile à distinguer.

Tableau 2 : Prédiction des entités non-figées mesurée en comparant la prédiction faite par le moteur et l'annotation manuelle. Le rappel et la précision augmentent significativement lorsque l'on considère un recouvrement partiel. Les gènes et les marqueurs sont parfois difficile à distinguer. AL: Analyse linguistique ; M : Méthode ; P : Paramètres (têtes ambiguës).

Des marqueurs liés ont été recherchés pour des gènes impliqués dans quatre grandes thématiques (**Tableau 3**) ce qui représente un total de 126 gènes. 298 marqueurs ont été testés et 235 ont donné un résultat satisfaisant sur un panel de 16 variétés de référence ce qui représente au moins un marqueur efficace pour 75 des 126 gènes. Des essais haut débit (SNP) ont été développés pour 125 des 235 marqueurs. 184 marqueurs validés dont 64 SNP ont été testés sur un panel élargi de 235 variétés élites et 47 lignées en cours de sélection (**Figures 2A et 2B**).



Nous avons développé un moteur de recherche pour l'extraction automatisée de l'information qui doit être amélioré en peaufinant l'entraînement à l'aide d'articles annotés manuellement. Nous avons validé 125 marqueurs utilisables à haut débit en sélection assistée. Ce chiffre peut être amélioré si des lignées témoins pertinentes possédant les gènes manquant sont mises à disposition. La base est disponible et doit être incrémentée avec les données de génotypage sur puce.

Ratkovic Z., Golik W., Warnier P. (2012) Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach". BMC Bioinformatics 2012, 13(Suppl 11):S8, 26.

