

A computational architecture designed for genome annotation: oak genome sequencing project as a use case

Nicolas Francillonne, Tina Alaeitabar, Françoise Alfama-Depauw, Loïc Couderc, Claire Guerche, Thomas Letellier, Mikaël Loaec, Isabelle Luyten, Célia Michotey, Jean-Marc Aury, et al.

▶ To cite this version:

Nicolas Francillonne, Tina Alaeitabar, Françoise Alfama-Depauw, Loïc Couderc, Claire Guerche, et al.. A computational architecture designed for genome annotation: oak genome sequencing project as a use case. JOBIM 2015 - Journées Ouvertes Biologie Informatique Mathématiques, Jul 2015, Clermont-ferrand, France. 2015. hal-01246649

HAL Id: hal-01246649 https://hal.science/hal-01246649

Submitted on 5 Jun2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



ElnterMine

Web pollo GMOD

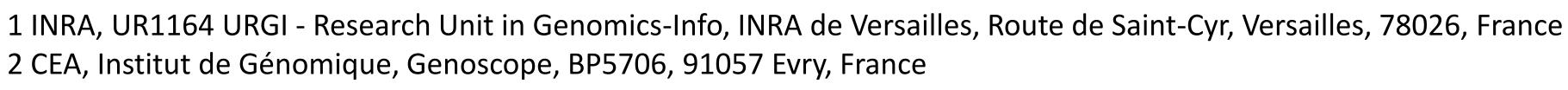
A computational architecture designed for genome annotation:

Oak genome sequencing project as a use case



N, FRANCILLONNE¹, T, ALAEITABAR¹, F, ALFAMA¹, L, COUDERC¹, C, GUERCHE¹, T, LETELLIER¹, M, LOAEC¹, I, LUYTEN¹, C, MICHOTEY¹, J, AURY², H, QUESNEVILLE¹, C. PLOMION³, J. AMSELEM¹ and the Oak genome sequencing consortium nicolas.francillonne@versailles.inra.fr

joelle.amselem@versailles.inra.fr



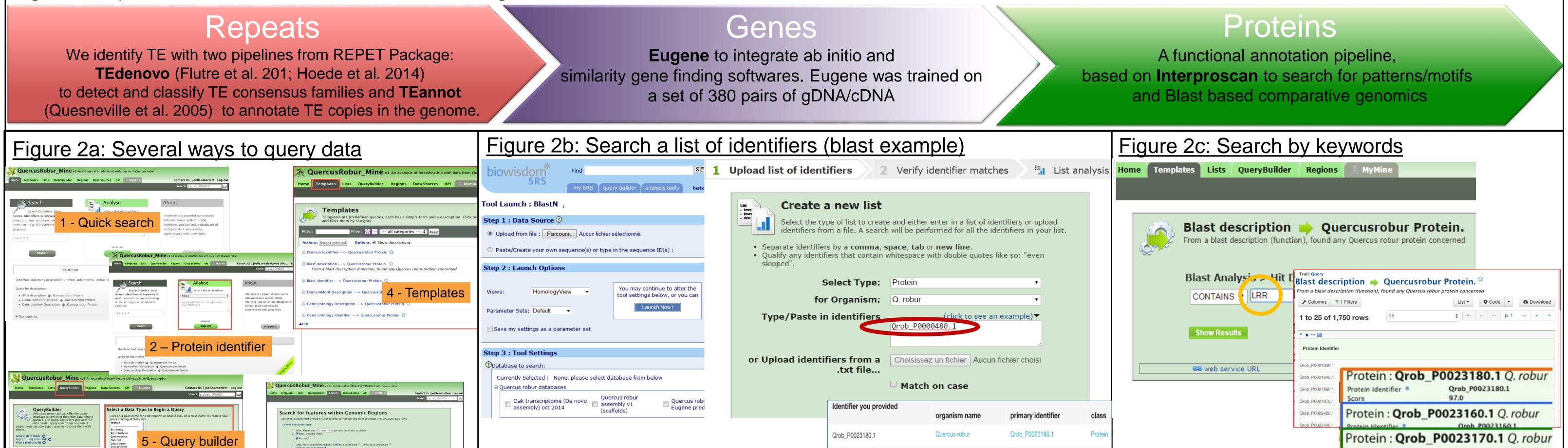
3 INRA, UMR1202 – Biodiversity, Genes and Communities - 69, route d'Arcachon, F-33610 Cestas, France

The ANR Genoak project aims to study the two key evolutionary processes that explain the remarkable diversity found within the oak genus. We performed an automated structural annotation (transposable elements (TEs) and genes) and functional annotation of predicted genes using robust pipelines i/ REPET for TEs ii/ Eugene for gene prediction iii/ FunAnnotPipe (in-house pipeline) mainly based on InterproScan for functional annotation. Further objectives were to: i/ integrate the whole genome with all the features annotated into a Genome Browser, ii/ provide an interface for gene prediction curation/validation, and iii/ provide an information system pointing towards accessibility and interoperability.

We are setting up a fast and flexible genome browser WebApollo_oak allowing the edition of genes structure. This tool based on Jbrowse is used to visualize, identify

and curate gene predictions. For functional annotation, we set up a data warehouse QuercusRoburMine based on Intermine technology. Its user friendly interface gives access to functional information which allow cross queries between different data sources. These combined tools constitute powerful resources for whole genome annotation. Next step will be to automatically update all functional information in QuercusRoburMine for gene curated through WebApollo_oak. We will present some case studies to illustrate questions raised by the users.

Figure 1: Pipelines used for structural and functional genome annotation





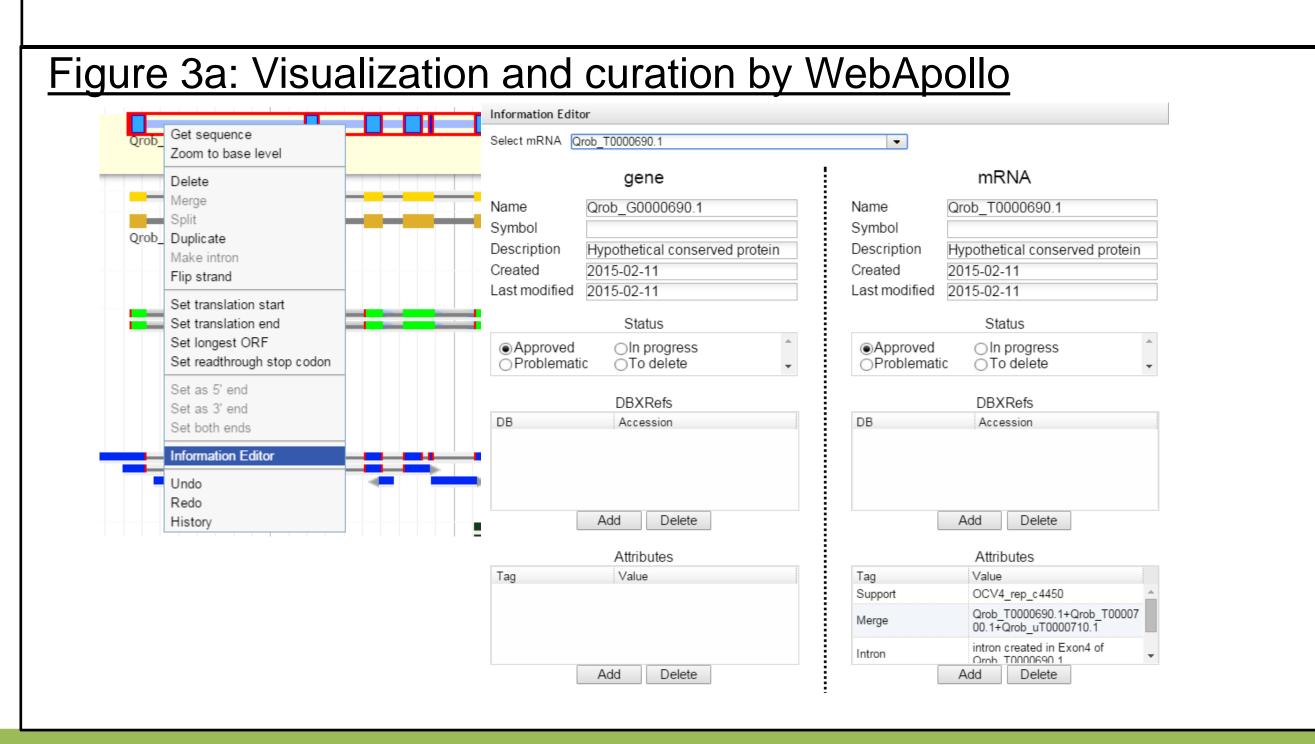
 diop_1002010011			
Qrob_P0023190.1	Quercus robur	Qrob_P0023190.1	Protein
Qrob_P0023170.1	Quercus robur	Qrob_P0023170.1	Protein
Qrob_P0023160.1	Quercus robur	Qrob_P0023160.1	Protein

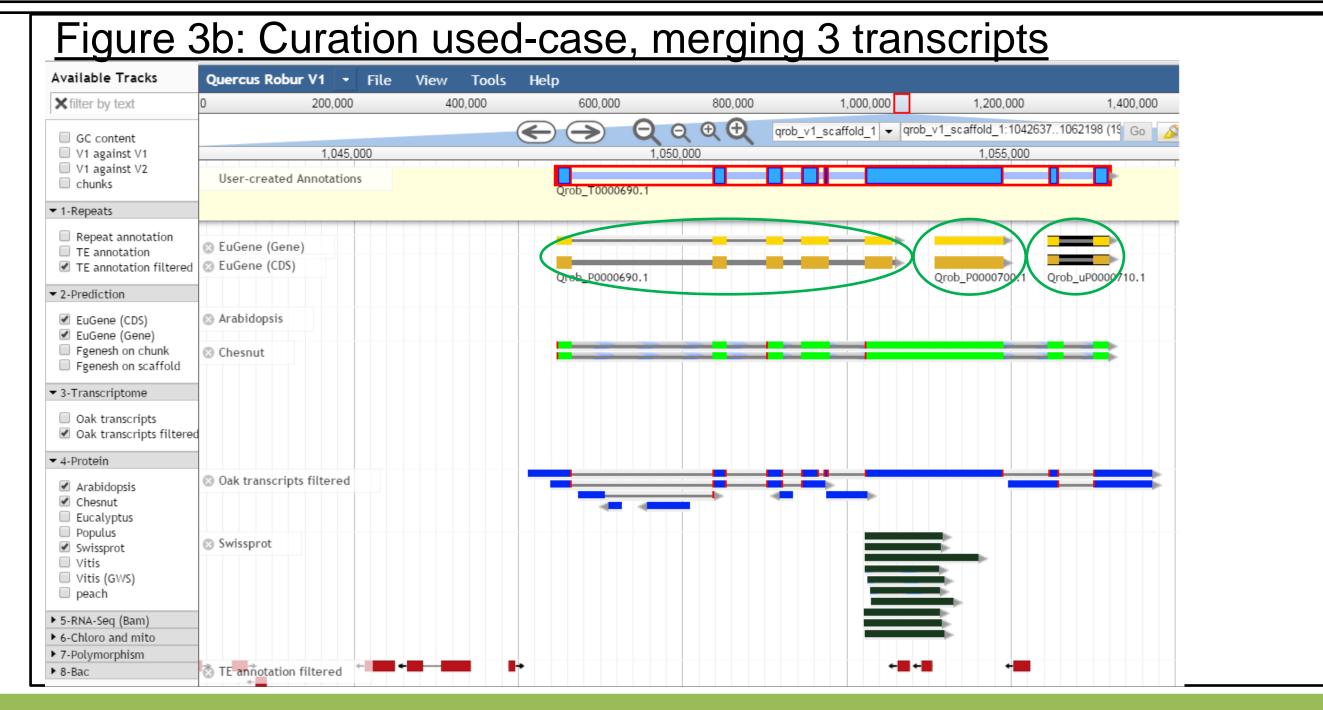
Protein Identifier	Qrob_P0023170.1	
Protein : Qrob_	P0023190.1 Q. robur	
Protein Identifier	Qrob_P0023190.1	Orga
Score	95.0	Scor

Architecture

To curate an annotation, we need to combine different sources of evidence altogether. Those information provided by structural and functional genome annotation pipeline (figure 1) are available through two powerful resources. Intermine is used to gather and give access to functional annotation using several ways of querying data (figure 2a). With a list of identifiers, obtained by blast comparison against the newly predicted set of genes (figure 2b) or using keywords based search (figure 2c), we can retrieve all functional annotation displayed through a gene card (figure 2d). Those cards linked functional annotation and structural annotation by featuring a WebApollo (figure 3). WebApollo add several tools to the jbrowse (figure 3a), which allow gene curation like merging predicted transcripts into new gene (figure 3b). This architecture evolution would be to have an automated and continuous update of all the manual annotation made through this system.

				ote													
Proteir	Qrob I	P000	0480.1		robur												
	fier ^B Qrob_P000					us robur											
Fasta	FASTA		Score		20.0												
Score Type	egn																
SHARE																	
Senome Br																	
		ove the vi	ew. Drag an	d drop tra	acks from l	eft menu into	o the main par	nel to see the data.	Clicking on individ	dual features to o	pen a report pa	ige for that fe	ature.				
Centre on Q	rob_P0000480.1																
Available	e Tracks				^	Quercus R	Robur V1	- File View	Help							Ľ	Login
X filter b	y text				0		200,000	400,000	600,0	300 000	300,000	1,000,000	0 1,200,000	0 1,400,000	1,600,00	00 1,800,00	0
	-							\leftarrow	$\rightarrow 6$	$Q \oplus G$		scaffold 1	qrob_v1_scaffold	d_1:873565877467 (3	3.1 Go 🔊		
GC cc						073 750		\bigcirc		875,000	4.00_11			876,250			8
V1 ag						873,750				875,000				070,250			8
V1 ag							(C)										
chuni	KS					🛛 EuGene (
▼ 1-Repeat	ts				3	🙁 EuGene ((CDS)	rob P0000480.1									
▼ 2-Predic	tion				4												
	r (go to WebApoll	o_oak)			•												
xpand viewe 5 GO Terms	r (go to WebApolle	o_oak)			~			Description									
xpand viewe 5 GO Terms dentifier	r (go to WebApolle s crossRef				Ŧ			Description	and non-covalently	with any protein or	protein complex (a complex of tw	vo or more proteins that	may include other nonpro	stein molecules)		
kpand viewe 5 GO Terms dentifier 50:0005515	r (go to WebApolle	1	sferring pho	sphorus-		groups		Interacting selectively						may include other nonpro	otein molecules).		
kpand viewe 5 GO Terms dentifier <u>60:0005515</u> <u>60:0016772</u> 50 Blast	r (go to WebApollo s crossRef protein binding transferase act	1	sferring pho	sphorus-(containing			Interacting selectively Catalysis of the transf	er of a phosphorus-	containing group fro	om one compound	d (donor) to an	other (acceptor).	may include other nonpro	otein molecules).		
pand viewe GO Terms dentifier <u>GO:0005515</u> <u>GO:0016772</u> 50 Blast Analysis	r (go to WebApolle s crossRef <u>protein binding</u> <u>transferase act</u> Hit	<u>a</u> iivity, trans	Start	End	<u>containing</u> Strand	Length	Note	Interacting selectively Catalysis of the transf Hit Coverage	er of a phosphorus- Hit Length	containing group fro Hit Pident	om one compound Hit Pcons	d (donor) to an E Val	other (acceptor). Hit Description				
6 GO Terms dentifier 60:0005515 60:0016772 50 Blast Analysis Mastp_kegg	r (go to WebApolle s crossRef protein binding transferase act Hit vvi:100268107	<u>a</u> iivity, trans	Start 31	End 901	containing Strand +	Length 871	Note Gaps:28	Interacting selectively Catalysis of the transf Hit Coverage 96.70	er of a phosphorus- Hit Length 878	Hit Pident 46.41	Hit Pcons	d (donor) to an E Val 0.0	Hit Description	repeat receptor-like prote	in kinase At1g35710		
kpand viewe 5 GO Terms dentifier <u>60:0005515</u> <u>60:0016772</u> 50 Blast Analysis blastp_kegg blastp_kegg	r (go to WebApollo crossRef <u>protein binding</u> <u>transferase act</u> Hit <u>vvi:100268107</u> <u>vvi:100261352</u>	a ivity, trans	Start	End 901 901	containing Strand + +	Length 871 851	Note Gaps:28 Gaps:95	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90	Hit Length 878 820	Hit Pident 46.41 49.74	Hit Pcons 16.49 17.01	E Val 0.0 0.0	Hit Description probation description	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like	
kpand viewe 5 GO Terms dentifier <u>60:0005515</u> <u>60:0016772</u> 50 Blast 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a ivity, trans	Start 31 51	End 901	containing Strand + +	Length 871	Note Gaps:28	Interacting selectively Catalysis of the transf Hit Coverage 96.70	er of a phosphorus- Hit Length 878	Hit Pident 46.41	Hit Pcons	d (donor) to an E Val 0.0	Hit Description probation description	repeat receptor-like prote	ein kinase At1g35710 rotein kinase At4g088	350-like	
kpand viewe GO Terms dentifier <u>GO:0005515</u> <u>GO:0016772</u> 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a iivity, trans	Start 31 51 51	End 901 901 901	containing Strand + + +	Length 871 851 851	Note Gaps:28 Gaps:95 Gaps:72	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07	Hit Length 878 820	Hit Pident 46.41 49.74 48.17	Hit Pcons 16.49 17.01 17.02	d (donor) to an E Val 0.0 0.0 0.0	Hit Description probationeuclourich profitible LRR recipt profitible LRR recipt	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like 350-like	
Conterms GO Terms Go:0005515 GO:0016772 GO Blast Analysis Mastp_kegg Mastp_kegg Mastp_kegg Mastp_kegg	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a ivity, trans Begin	Start 31 51 51 En	End 901 901 901 d	containing Strand + + + Length	Length 871 851 851	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cre	Hit Pcons 16.49 17.01 17.02 oss Ref	d (donor) to an E Val 0.0 0.0 0.0 Desci	hit Description probabilised on probabilised on probabilised on probabilised on protocological of the term of term of the term of	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like 350-like Inter Pro	
Conterms GO Terms Go:0005515 GO:0016772 GO Blast Analysis Mastp_kegg Mastp_kegg Mastp_kegg Mastp_kegg Mastp_kegg Mastp_kegg ProSiteProfiles	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a <u>iivity, trans</u> Begin 610	Start 31 51 51 En 885	End 901 901 901 901	Containing Strand + + + Length 280	Length 871 851 851 D	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru	Hit Pcons 16.49 17.01 17.02 oss Ref ne	d (donor) to an E Val 0.0 0.0 0.0 Descr Protei	hit Description probable december of the protoble LRR recept protoble LRR recept protoble LRR recept	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like 350-like Inter Pro <u>IPR000719</u>	
kpand viewe 5 GO Terms dentifier 50:0005515 50:0016772 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg 204 Domair Analysis ProSiteProfiles ProSiteProfiles	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a ivity, trans Begin 610 150	Start 31 51 51 51 En 889 171	End 901 901 901 901 d	Containing Strand + + + Length 280 22	Length 871 851 851 D P3 P3	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cro noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne	d (donor) to an E Val 0.0 0.0 0.0 Desci Protei Leucin	hit Description probabilised on probabilised on probabilised on probabilised on protocological of the term of term of the term of	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like 350-like Inter Pro IPR000719 IPR001611	
Conterms GO Terms Go:0005515 GO:0005515 GO:0016772 GO Blast Analysis Mastp_kegg Mastp_ke	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	a <u>iivity, trans</u> Begin 610	Start 31 51 51 En 885	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	Containing Strand + + + Length 280	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probable december of the protoble LRR recept protoble LRR recept protoble LRR recept	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like 350-like Inter Pro <u>IPR000719</u>	
kpand viewe GO Terms dentifier GO:0005515 GO:0016772 50 Blast Analysis blastp_kegg blast	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 3D SA:1.10.510	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probable converted probable converted proferble LRR recept proferble LRR recept ription n kinase domain profile.	repeat receptor-like prote or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like Inter Pro IPR000719 IPR001611 none	
kpand viewe 5 GO Terms dentifier 50:0005515 50:0016772 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg 204 Domair Analysis ProSiteProfiles ProSiteProfiles ProSiteProfiles 12 Localiza	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 3D SA:1.10.510	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru noi noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne ne	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probable access rich protoble LRR recept protoble LRR recept ription n kinase domain profile. ne-rich repeat profile.	repeat receptor-like prote or-like serine/threonine-pr or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088 rotein kinase At4g088	350-like Inter Pro IPR000719 IPR001611 none	
xpand viewe 5 GO Terms dentifier <u>30:0005515</u> <u>30:0016772</u> 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg 204 Domair Analysis ProSiteProfiles Gene3D ProSiteProfiles 12 Localiza Analysis	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 3D SA:1.10.510	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru noi noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probable converted probable converted proferble LRR recept proferble LRR recept ription n kinase domain profile.	repeat receptor-like prote or-like serine/threonine-pr or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088	350-like Inter Pro IPR000719 IPR001611 none	
xpand viewe 5 GO Terms dentifier <u>GO:0005515</u> <u>GO:0016772</u> 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg blastp_kegg 204 Domair Analysis ProSiteProfiles Gene3D ProSiteProfiles 12 Localiza Analysis SignalP_EUK	r (go to WebApollo crossRef protein binding transferase act Hit vvi:100268107 vvi:100261352 vvi:100254590	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 3D SA:1.10.510	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru noi noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne ne Start	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probabilities or probabilities of the protoble LRR recept protoble LRR recept ription n kinase domain profile. he-rich repeat profile.	repeat receptor-like prote or-like serine/threonine-pr or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088 rotein kinase At4g088	350-like Inter Pro IPR000719 IPR001611 none	
kpand viewe 5 GO Terms dentifier <u>GO:0005515</u> <u>GO:0016772</u> 50 Blast Analysis blastp_kegg blastp_kegg blastp_kegg 204 Domain Analysis ProSiteProfiles ProSiteProfiles 12 Localiza Analysis SignalP_EUK TMHMM	r (go to WebApollo crossRef <u>protein binding</u> <u>transferase act</u> Hit <u>vvi:100268107</u> <u>vvi:100261352</u> <u>vvi:100254590</u> Motifs	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 01 01 01 01 01 01 01 01 01 01 01 01 0	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 3D SA:1.10.510	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cru noi noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne ne Start	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probabilities of the probabilities of the probability of the pr	repeat receptor-like prote or-like serine/threonine-pr or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088 rotein kinase At4g088 Length	350-like Inter Pro IPR000719 IPR001611 none	
xpand viewe	r (go to WebApollo crossRef <u>protein binding</u> <u>transferase act</u> Hit <u>vvi:100268107</u> <u>vvi:100261352</u> <u>vvi:100254590</u> Motifs	Begin 610 150 678	Start 31 51 51 51 En 889 171 879	End 901 901 901 d 901	containing Strand + + + Length 280 22 202	Length 871 851 851 D P3 P3 G	Note Gaps:28 Gaps:95 Gaps:72 omain Identifi S50011 S51450 <u>3D SA:1.10.510</u> S51450	Interacting selectively Catalysis of the transf Hit Coverage 96.70 93.90 94.07 er	Hit Length 878 820	Hit Pident 46.41 49.74 48.17 Cro noi noi	Hit Pcons 16.49 17.01 17.02 oss Ref ne ne ne Start 1 549	d (donor) to an E Val 0.0 0.0 0.0 Desct Protei Leucin none	hit Description probabilitieuce, prich profitible LRR recipt profitible LRR recipt profitible LRR recipt ription n kinase domain profile. he-rich repeat profile. End 20 568	repeat receptor-like prote or-like serine/threonine-pr or-like serine/threonine-pr	ein kinase At1g35710 rotein kinase At4g088 rotein kinase At4g088 Length 19	350-like Inter Pro IPR000719 IPR001611 none	









More on Oak sequencing genome at: http://www.oakgenome.fr/



