Université du Québec à Montréal

INF4500

Examen intra

Par :
Guillaume Lahaie
LAHG04077707

Remis à :
Abdoulaye Baniré Diallo

Date de remise :
Le 9 décembre 2013

Table des matières

1 Introduction

Le but de ce travail est d'annoter des contigs du génome du blé. Nous n'avons pas d'information concernant la provenance de ces contigs, ou même l'espèce exacte de provenance. Afin de pouvoir fournir une information pertinente, j'ai tout d'abord recherché ce qui est connu concernant le génome du blé.

J'ai tout d'abord cherché à connaitre l'état d'avancement des travaux de séquençage du blé. Pour ce faire, j'ai consulté la base de données des génomes de NCBI [1]. On y apprend des informations de base sur le génome du blé. On y apprend que le génome du blé a une taille de 16000 Mb distribué en 21 chromosomes. De plus, les chromosomes ont une forme allohexaploid composée de trois sous-génomes. La nature hexaploid de son génome a ralenti les efforts de séquençage.

Une première référence de génome du blé a été créée avec l'espèce Triticum urartu [2]. Ce génome est toutefois celui d'un progéniteur du Triticum aestivum, il peut être utile pour aider à améliorer le génome du blé.

On peut obtenir une information plus complète concernant l'avancement du séquençage du Triticum aestivum sur le site du International Wheat Genome Sequencing Consortium. On y retrouve deux projets parallèles : en premier lieu, un projet de survey sequencing, afin de produire un contenu de gène potentiel et un ordre de gène virtuel [3]. Un autre projet en cours est de produire une séquence de référence pour le génome du Triticum aestivum [4]. Ce projet semble être à ses débuts, car il semble être en cours d'obtention de financement.

D'autres bases de données offrent de l'information à propos du génome du blé, par exemple CerealsDB [5], ayant un génome de travail du blé. Il y a aussi beaucoup d'autres projets, considérant la place importante occupée par le blé dans l'agriculture moderne.

Basé sur ces informations, j'ai décidé de concentrer mes recherches pour l'annotation des contigs fournis sur les données déjà connues du génome du blé. Je vais donc seulement garder les résultats de Blast provenant du Triticum aestivum. Bien sûr, il s'agit ici d'une premiére étape de recherche, il serait ensuite possible d'élargir la recherche pour identifier des zones fonctionnelles possibles des contigs, ce qui ne sera pas fait dans ce travail.

2 Produisez une analyse sommaire de ces contigs en présentant la distribution des tailles et taux de GC

Afin de compiler et de représenter la taille et le taux de GC des contigs produits par CAP3, j'ai écrit un script python (question1.py) permettant d'extraire les informations du fichier seq.data.cap.contigs. Le fichier contient 346 contigs.

Le script produit deux types de graphiques, à l'aide de gnuplot. Le premier type est un histogramme, un pour la taille des contigs, et un pour le taux de GC des contigs. On peut alors remarquer la distribution de ces valeurs. Voici les deux histogrammes :

J'ai ensuite produit deux graphiques permettant de visualiser différemment ces résultats. On peut y retrouver la moyenne de taille, la moyenne de taux de GC, ainsi que les contigs se situant en haut ou en bas ce cette moyenne. On peut aussi voir les valeurs exactes dans le tableau en annexe 1

La taille moyenne des 346 contigs est de 109 nucléotides, avec un taux de GC moyen de 42,96%. Ce taux semble indiquer une prépondérance de région non-codante dans les contigs, car généralement les séquences codantes ont un taux de GC supérieur aux séquences non-codantes [6].

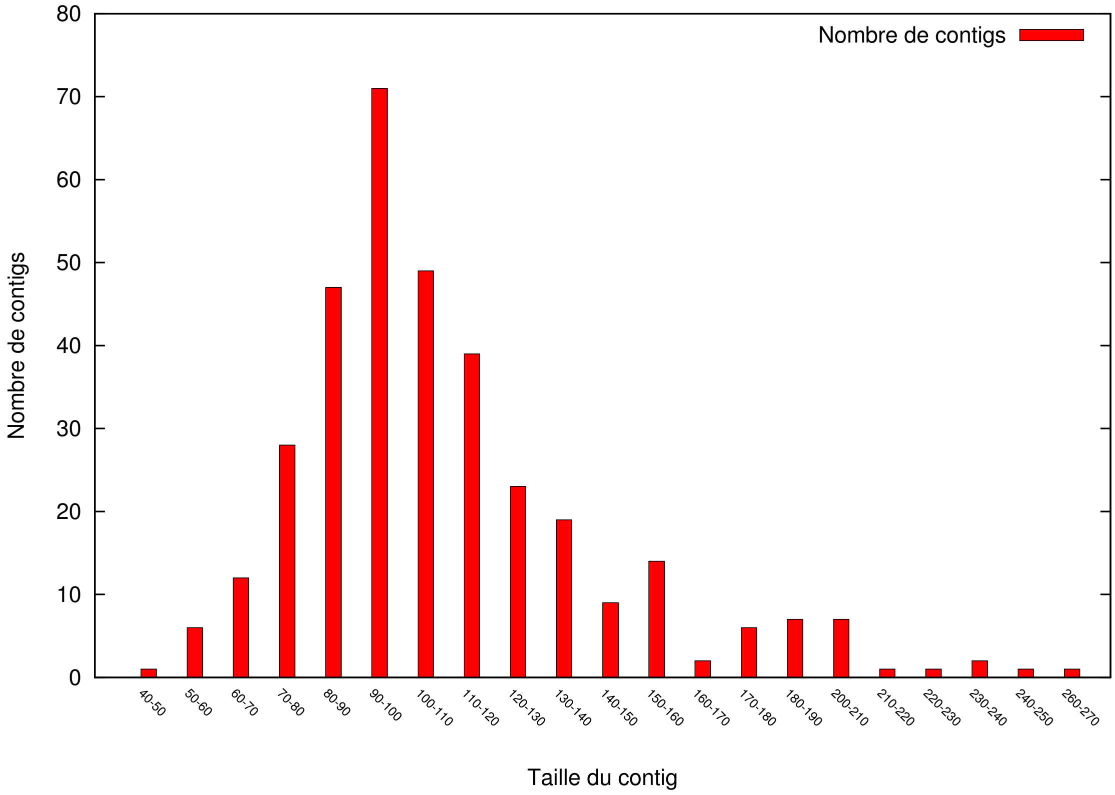

FIGURE 1 – Histogramme de la taille des contigs

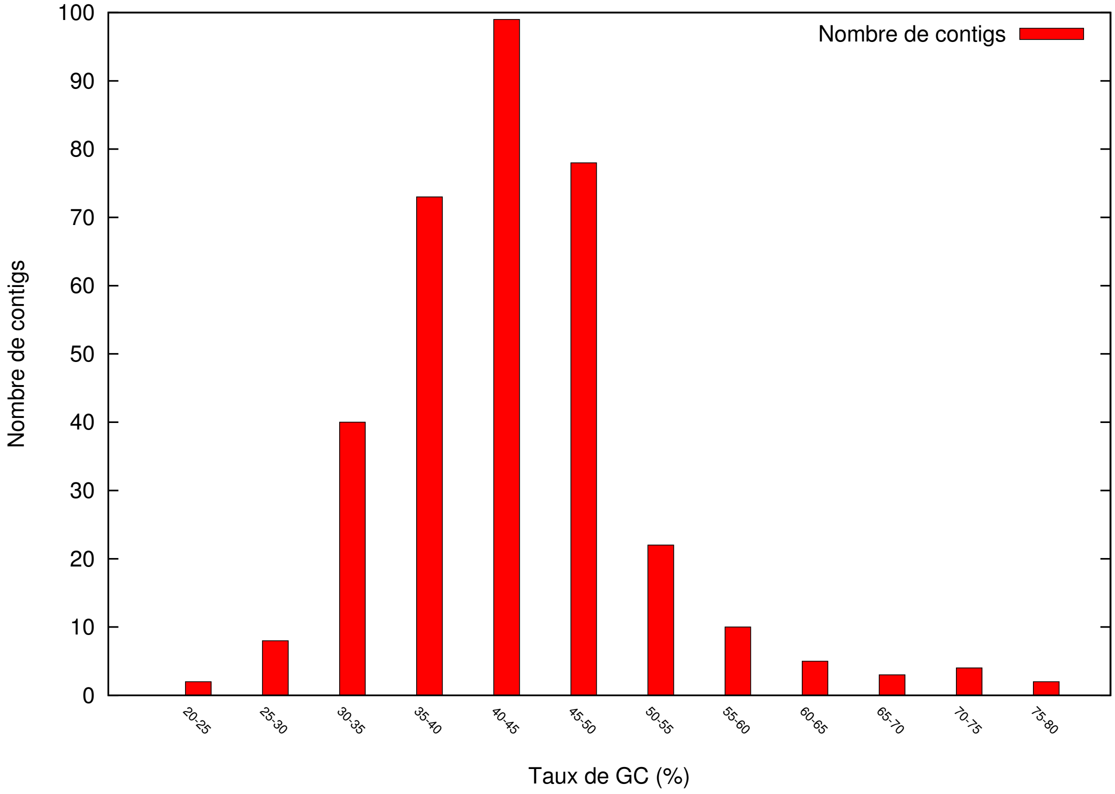

FIGURE 2 – Histogramme du taux de GC des contigs

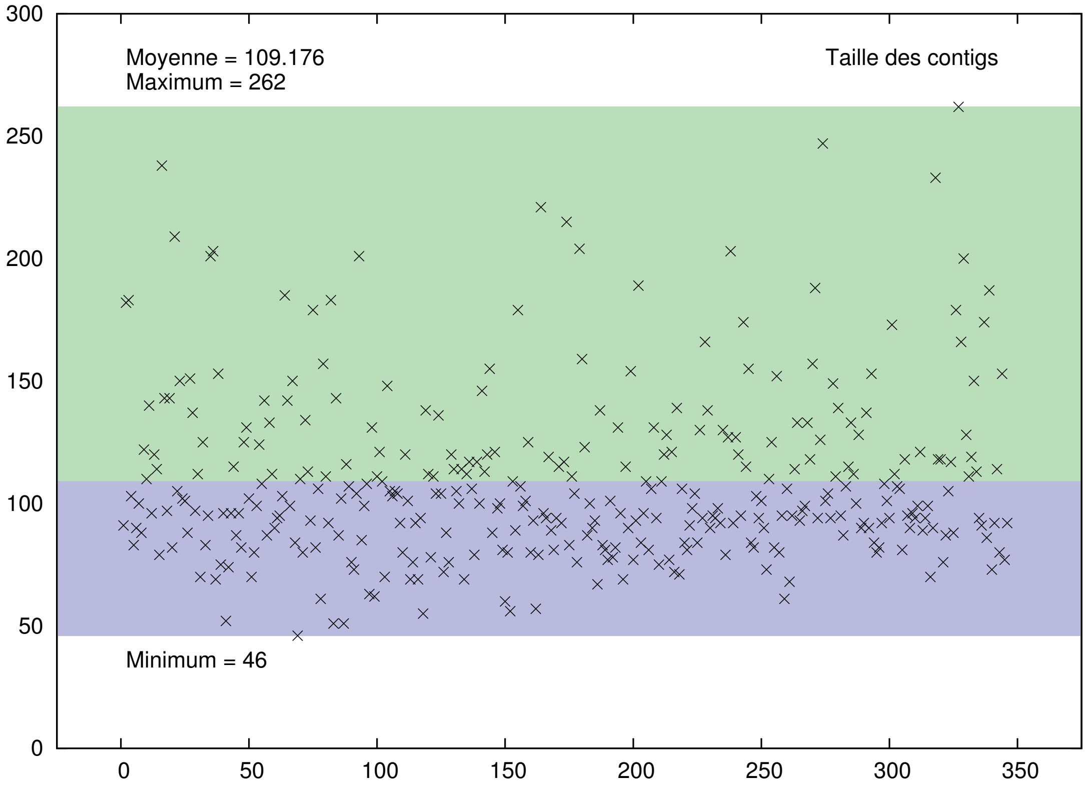

Figure 3 – Nuage de points de la taille des contigs

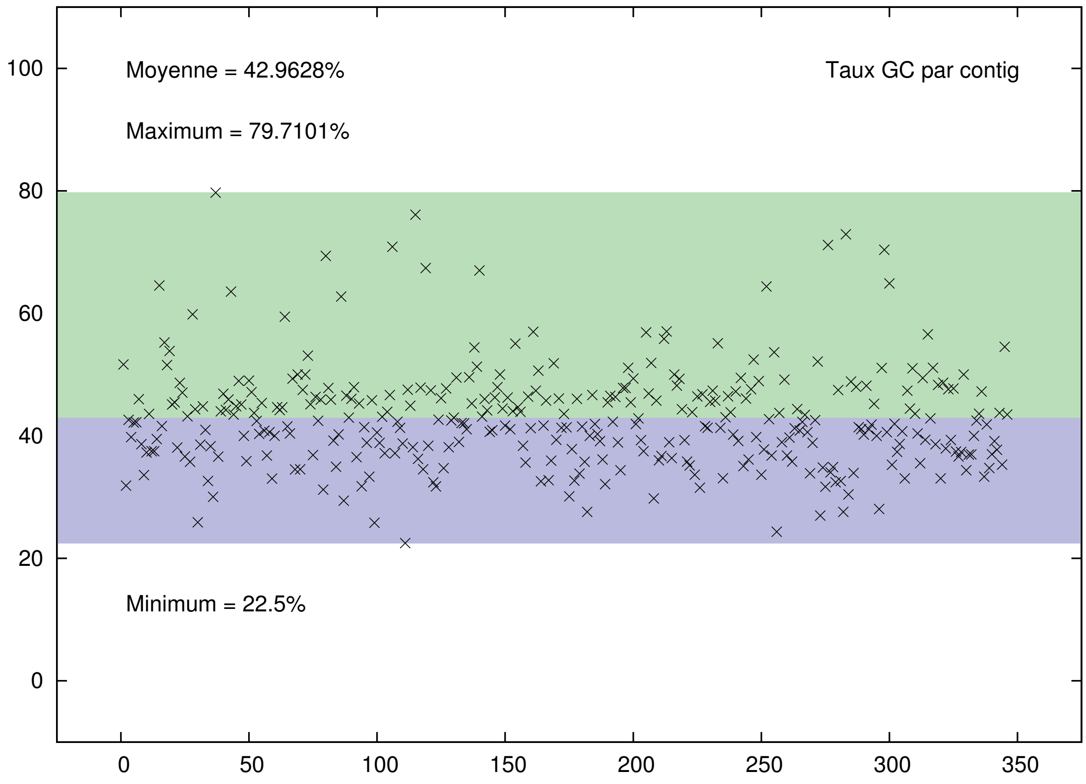

Figure 4 – Nuage de points du taux de GC des contigs

3 Identifiez les annotations Genbank de ces contigs et présentez les dans une table contenant les colonnes : contigs, numéros Accession, description, uniref id

Pour trouver les annotations Genbank des contigs, j'ai tout d'abord effectué un blast de chaque contig sur la base de données nr/nt de NCBI [8]. J'ai utilisé le script biopython question2.py pour effectuer tous les blasts, et enregistrer les résultats.

En examinant les résultats de façon sommaire, on remarque une très grande différence entre la qualité des résultats. Certains ont des E-value très haute, alors que certains ont des valeurs indiquant un résultat de haute qualité. On peut s'attendre à cela, considérant la grande variabilité des contigs.

Pour traiter les contigs selon leur taille, je calcule la valeur médiane des E-value pour les contigs plus petits que la taille moyenne. Je fais le même exercice pour les contigs plus grands que la moyenne. Pour le moment, je m'intéresse au meilleur résultat obtenu seulement pour la médiane.

Comme mentionné en introduction, comme cette analyse s'intéresse seulement au contigs ayant des résultats pour le Triticum, je ne considère pas dans mes résultats les valeurs de blast pour des espèces différentes du blé. Je prends donc, dans les résultats de blast, le premier correspondant à un match avec le blé.

J'ai enregistré les résultats dans les fichiers evalue_lower.txt et evalue_higher.txt, à l'aide du script q2_meanEvalue.py. On peut remarquer que la grande majorité des résultats obtenus ont des E-values de bonne qualité, avec un ordre de grandeur permettant d'avoir une grande confiance dans le hit. Basé sur ces données, je garderai donc tous les résultats, peu importe la taille du contig, ayant une E-value plus petite que 0.01.

Afin d'obtenir les données de numéro d'accession, j'ai modifié le script précédent pour créer un fichier associant le numéro du contig avec le hit gardé (pour le moment, je garde seulement le premier hit de blé du résultat), avec le numéro d'accession et la description du hit. Ces données sont gardées seulement si le hit correspond aux exigences de E-value et de description de hit.

Pour obtenir un Uniref pour les contigs retenus, j'ai ensuite utilisé le module bioservices de python permettant de se connecter au service idmapping de uniprot, pour trouver les identifiants uniref des contigs conservés.

Des 227 contigs restant, 113 ont obtenu des résultats de mapping. Avant de sortir les résultats, j'ai vérifié le format des données obtenues par ce mapping. Pour certains contigs, un seul résultat est obtenu, alors que pour certains, on obtient plusieurs mappings différents. Les fichiers XML ne comprennent aucune information concernant le meilleur résultat, toutefois le service REST utilisé pour le mapping demande de trier les résultats selon le meilleur score.

Afin de vérifier le résultat, j'ai tenté de blaster un des contigs directement sur la base de données Uniref100, sur le site `http://www.uniprot.org`. Le résultat a été surprenant. J'ai utilisé le contig 2 comme essai, et le blastx sur Uniref100 n'a retourné aucun hit. Afin de confirmer ce résultat, j'ai effectué le même blastx en utilisant le service d'EBI et en blastant sur toutes les bases de données de protéines de uniprot. J'ai obtenu le même résultat.

Je crois que ce résutat est dû au mécanisme de mapping. Comme nous avons pu le constater à la questions 1, la plupart des contigs donnés ont une longueur moyenne de 109 nucléotides. Toutefois, le numéro d'accession donnée pour effectuer le id mapping peut correspondre à une très longue séquence. C'est le cas du numéro d'accession pour le contig 2, il s'agit en fait d'un chromosome complet du blé, ce qui explique les nombreux résultats du mapping.

J'ai donc décidé de procéder différemment pour obtenir

J'ai produit plusieurs scripts python afin de produire les résultats pour cette question, pour éviter d'avoir à refaire certaines étapes plus longues.

J'ai tout d'abord écrit un script python pour effectuer un blast sur chaque contig. Ce blast a été fait sur la base de données nr/nt de NCBI. Je n'ai pas utilisé l'option megablast pour tenter d'obtenir des résutlats pour chaque contig, même s'il s'agit d'un résultat d'une qualité inférieure.

Le résultat de chaque blast a été enregistré dans un fichier xml dans le répertoire blastNCBI. J'ai préféré travailler de cette façon pour éviter de refaire les blasts plusieurs fois.

J'ai ensuite écrit un script permettant de choisir le résultat du blast qui serait considéré (q2_parse_ncbi.py). Plutôt que d'écrire des règles d'affaires dans le script, j'ai écrit un script qui me permet de faire un choix parmi les dix premièrs résultats du blast. Le résultat est ensuite enregistré dans un fichier texte (resultatNCBI.txt).

J'ai gardé dans ce fichier les informations utiles pour la question 2, mais aussi pour la question 3. Ce fichier contient donc le numéro d'accession pour le hit choisi, mais aussi les positions des hits dans le contig et dans la séquence choisie.

Comme la plupart des contigs sont assez courts, tel que vu à la question 1, j'ai privilégié la longueur du hit et la similarité plutôt que le e-value. J'ai aussi décidé d'inclure des résultats ayant une e-value très importante (> 1), cela pourrait être une première piste de solution pour ces contigs, mais il ne faut pas se fier au résulat.

J'ai ensuite écrit un script pour aller télécharger le fichier genbank lié à un contig (getGB.py). Ce script a télécharger tous les fichiers dans le répertoir genbank, permettant de traiter les fichiers plus tard.

J'ai effectué deux démarches différentes pour obtenir l'uniref de chaque contig. J'ai tout d'abord effectué des recherches pour voir les différentes façons d'obtenir la valeur. Une approche prometteuse était d'utilisé les services de EBI pour effectué un blast directement sur la base de données UNIREF100. Il s'agit donc d'un blastx, qui transforme la séquence nucléotidique du contig en séquence protéinique.

Afin d'utiliser ce service d'EBI, j'ai fait appel au module bioservices de python. Ce module permet un accès facile aux services REST de EBI. J'ai donc écrit un script qui a fait un blastx de chaque contig, et qui a enregistré un fichier de résultat pour chaque blast dans le dossier blastEBI.

J'ai effectué la même démarche pour examiner les résultats de ces blasts. J'ai choisi un résultat à la main pour chaque blast à l'aide du script q2_parse_ebi.py, et les résultats ont été enregistrés dans le fichier resultatEBI.txt. Pour un nombre important de contigs, blast n'a trouvé aucun résultat.

J'ai aussi utilisé le service de ID mapping de uniprot (ww.uniprot.org).

Contig	Accession	Description	Uniref - EBI	Uniref - mapping

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
1	AK354634	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1009F18.		M0YC96
2	AF343493	Secale cereale clone pla 3-phosphoglycerate kinase (Pgk-1) gene, partial cds ; nuclear gene for plastid product.		Q8LLS6
3	EF109232	Triticum aestivum strain CRB-INRA-CFD-13471 malate dehydrogenase (Mdh4B) gene, partial cds.	UniRef100_M8D509	A8QR46
4	AF277253	Australopyrum velutinum isolate H6724 disrupted meiotic cDNA 1 protein (DMC1) gene, partial cds.		Q9FQ40
5	AK331959	Triticum aestivum cDNA, clone : WT002_M17, cultivar : Chinese Spring.	UniRef100_I1I3L3	
6	AK332278	Triticum aestivum cDNA, clone : WT003_J14, cultivar : Chinese Spring.		
7	AK335464	Triticum aestivum cDNA, clone : WT012_P12, cultivar : Chinese Spring.	UniRef100_I1I3T9	
8	AK357915	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1064M07.	UniRef100_M0VVW7	M0VVW8
9	XM_005180954	PREDICTED : Musca domestica transcription factor grauzone-like (LOC101901076), mRNA.		
10	JQ240472	Triticum urartu clones BAC 70G09, BAC 169L13, and BAC 78P09, complete sequence.		M8A4Z7
11	AK374032	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3051L02.		F2ED07
12	AK376212	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3118E15.		F2DBW7
13	AK371461	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2134K02.		F2E5N7
14	AK332744	Triticum aestivum cDNA, clone : WT004_M05, cultivar : Chinese Spring.		
15	AK357832	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1062N11.	UniRef100_H8MLR6	M0VUI6
16	AK332362	Triticum aestivum cDNA, clone : WT003_M19, cultivar : Chinese Spring.	UniRef100_D5QG09	
17	XM_003617172	Medicago truncatula hypothetical protein (MTR_5g089180) mRNA, complete cds.	UniRef100_UPI0003039078	G7KCV8
18	U73217	Triticum aestivum cold acclimation protein WCOR615 (Wcor615) mRNA, complete cds.		P93614
19	GQ905535	Zea mays clone zma-miR167b precursor miRNA zma-miR167b, precursor RNA, complete sequence.	UniRef100_UPI00035C8ECC	
20	XM_003564504	PREDICTED : Brachypodium distachyon uncharacterized LOC100831523, transcript variant 2 (LOC100831523), mRNA.	UniRef100_M7Z4P7	
21	DQ286562	Triticum aestivum putative lipid transfer protein mRNA, complete cds.	UniRef100_C5XMF8	A0MAU6

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
22	KC816724	Triticum urartu cultivar G1812 clone BAC 288D18 chromosome 3AL, complete sequence.	UniRef100_M8ADN3	M7YFA9
23	AK335482	Triticum aestivum cDNA, clone : WT013_A03, cultivar : Chinese Spring.	UniRef100_E3IRR7	
24	AK330641	Triticum aestivum cDNA, clone : SET4_P05, cultivar : Chinese Spring.	UniRef100_N1QXB8	
25	AK331680	Triticum aestivum cDNA, clone : SET1_K05, cultivar : Chinese Spring.		
26	AK332086	Triticum aestivum cDNA, clone : WT003_B19, cultivar : Chinese Spring.	UniRef100_R7W7J3	
27	EU660894	Triticum turgidum subsp. durum clone BAC 1053F12+1054I5 cytosolic acetyl-CoA carboxylase (Acc-2) and putative amino acid permease genes, complete cds.	UniRef100_G8TCZ2	B2ZGK1
28	AK357333	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1051A22.	UniRef100_M8C2Q7	F2D0C6
29	BT008986	Triticum aestivum clone wdk2c.pk008.b17 :fis, full insert mRNA sequence.		
30	HQ596874	Triticum aestivum voucher AP212 trnH-psbA intergenic spacer, partial sequence ; chloroplast.		
31	AK373191	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3023F03.	UniRef100_R7W1Q2	M0WVU2
32	AK353711	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1002E02.	UniRef100_M8CYA4	M0UFE1
33	AF354298	Triticum aestivum sucrose-phosphate synthase (SPS8) mRNA, partial cds.	UniRef100_D9CJB0	Q6EZE2
34	AM932685	Triticum aestivum 3B chromosome, clone BAC TA3B95F5.		B4ERX4
35	EU159424	Triticum turgidum haplotype B DNA repair protein Rad50 gene, complete cds.	UniRef100_M8BE75	A8IE27
36	EU146234	Secale cereale ALMT1-M77.1 gene, partial cds.		B3FI77
37	AK427458	Brachypodium distachyon mRNA, clone : PL016C01-A-020_P14.	UniRef100_C6JSC2	
38	AC192066	Pan troglodytes BAC clone CH251-396E2 from chromosome 22, complete sequence.	UniRef100_E3HMV6	
39	AJ318783	Triticum sp. partial mRNA for replication factor C, large subunit (rfc-1 gene).	UniRef100_Q8L6A5	Q8L6A5
40	AK355959	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1028F07.	UniRef100_I1H723	F2CWF6
41	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
42	FP016070	Pig DNA sequence from clone CH242-325I8 on chromosome 2, complete sequence.		
43	XM_005104617	PREDICTED : Aplysia californica PTB domain-containing engulfment adapter protein 1-like (LOC101845414), transcript variant X6, mRNA.	UniRef100_M8CRT6	
44	AJ784900	Triticum aestivum mRNA for type 1 non-specific lipid transfer protein precursor (ltp9.4 gene).	UniRef100_I3JGF9	Q5NE29

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
45	AK368023	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv2066I09.	UniRef100_M7ZPU9	F2DVV5
46	AK330669	Triticum aestivum cDNA, clone : SET1_G08, cultivar : Chinese Spring.	UniRef100_I2PWX4	
47	AK331428	Triticum aestivum cDNA, clone : WT007_H14, cultivar : Chinese Spring.	UniRef100_M8AEN7	
48	AK336109	Triticum aestivum cDNA, clone : SET1_D13, cultivar : Chinese Spring.	UniRef100_J3L200	
49	AK332525	Triticum aestivum cDNA, clone : SET1_N11, cultivar : Chinese Spring.		
50	AK370651	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2113O12.		F2E3C8
51	AK331813	Triticum aestivum cDNA, clone : WT002_G19, cultivar : Chinese Spring.		
52	AK363672	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2018A04.		F2DIF4
53	AK249285	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf27g13, mRNA sequence.	UniRef100_R7W372	
54	AK362882	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2010N11.	UniRef100_M8BQI0	F2DG65
55	XM_003580986	PREDICTED : Brachypodium distachyon cysteine-rich receptor-like protein kinase 19-like (LOC100830795), mRNA.	UniRef100_M8BTN0	
56	HQ390245	Triticum turgidum clone UCDTA00696 genomic sequence.	UniRef100_UPI000359F2FB	
57	AJ862529	Hordeum vulgare subsp. vulgare transposon Islay, clone SQ001T7E5.		
58	JX295577	Aegilops tauschii chromosome 1Ds prolamin gene locus, complete sequence.	UniRef100_M8BJR6	L7VIF5
59	AK372315	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2149K01.		M0WID3
60	FN645450	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0011b.	UniRef100_M7ZZ34	D8LAL5
61	XM_003577644	PREDICTED : Brachypodium distachyon cysteine-rich receptor-like protein kinase 25-like (LOC100832903), mRNA.	UniRef100_R7WEG5	I1IMG6
62	AC159711	Mus musculus 10 BAC RP23-214N15 (Roswell Park Cancer Institute (C57BL/6J Female) Mouse BAC Library) complete sequence.		
63	FN564428	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0091b.		D8L9J2
64	AK248619	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf140g03, mRNA sequence.	UniRef100_M8BND9	
65	AK252349	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf154e03, mRNA sequence.	UniRef100_F4GUE6	
66	AC216454	Populus trichocarpa clone POP028-J04, complete sequence.		
67	AK332970	Triticum aestivum cDNA, clone : WT005_F05, cultivar : Chinese Spring.	UniRef100_M7YP29	

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
68	FJ477092	Hordeum vulgare subsp. vulgare cultivar Haruna Nijo Rym4 and MCT-1 genes, complete cds.		M0WAR2
69	JQ455953	Uncultured bacterium clone 069100_148 16S ribosomal RNA gene, partial sequence.		
70	AK332566	Triticum aestivum cDNA, clone : WT004_E21, cultivar : Chinese Spring.	UniRef100_M8BJG8	
71	AK334580	Triticum aestivum cDNA, clone : SET1_C02, cultivar : Chinese Spring.		
72	XM_004960918	PREDICTED : Setaria italica UDP-glucose 4-epimerase 1-like (LOC101782923), mRNA.	UniRef100_K5X144	K3Z7G6
73	CT009625	Aegilops tauschii.	UniRef100_M7ZVV5	Q15MP7
74	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.		D7F4N2
75	AB238931	Triticum monococcum TmABI1 gene for protein phosphatase 2C, complete cds.	UniRef100_M7YVM1	A5A6P9
76	BT009089	Triticum aestivum clone wkm2c.pk0002.a3 :fis, full insert mRNA sequence.	UniRef100_K7WDG3	
77	AK357589	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1057D01.		F2CSJ6
78	AK335897	Triticum aestivum cDNA, clone : SET2_L19, cultivar : Chinese Spring.	UniRef100_M8B5H2	
79	JQ917466	Blumeria graminis f. sp. tritici strain 08-10-3-1 heat shock protein 70 (hsp70) mRNA, complete cds.		I2DB62
80	AK330275	Triticum aestivum cDNA, clone : SET4_A24, cultivar : Chinese Spring.	UniRef100_R7W6A1	
81	HE996341	Triticum aestivum cv. Arina SNP, chromosome 3B, clone Taes_arina_ctg_16989.		
82	AK251163	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf108l03, mRNA sequence.	UniRef100_M7ZR64	
83	XM_002457340	Sorghum bicolor hypothetical protein, mRNA.		C5XQ46
84	AK249125	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf13o12, mRNA sequence.	UniRef100_M8B5C8	
85	AK252351	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf151g16, mRNA sequence.		
86	AK360584	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1121G24.	UniRef100_F2D3Z5	F2D9M2
87	XM_004287239	PREDICTED : Fragaria vesca subsp. vesca topless-related protein 4-like (LOC101312082), mRNA.		
88	FN564432	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0616b.		D8L9P6
89	JQ003179	Hordeum brevisubulatum calcineurin B-like protein 3 (CBL3) mRNA, complete cds.	UniRef100_M0V180	H9BE61
90	FP017181	Zebrafish DNA sequence from clone CH73-108E8 in linkage group 15, complete sequence.		Q4W897
91	FM242577	Aegilops speltoides, storage protein activator (spa) locus region, S genome, clone BAC sho42-9k3.	UniRef100_M8A7Y3	C1KV19
92	U76215	Triticum aestivum NBS-LRR type protein pseudogene, complete sequence.		

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
93	GQ419475	Oryza sativa Japonica Group cultivar Khao Hawm putative precursor microRNA R395n-s gene, complete sequence.		
94	HE996549	Triticum aestivum cv. Arina SNP, chromosome 3B, clone Taes_arina_ctg_58561.		
95	AY487917	Triticum aestivum Mla-like protein mRNA, partial cds.	UniRef100_Q6RW52	Q6RW52
96	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_T1MEW5	D7F4N2
97	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.		D7F4N2
98	AK333621	Triticum aestivum cDNA, clone : WT006_O21, cultivar : Chinese Spring.	UniRef100_S4RA61	
99	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.		D7F4N2
100	AK333932	Triticum aestivum cDNA, clone : WT008_N17, cultivar : Chinese Spring.	UniRef100_T1N9G3	
101	EU626553	Triticum urartu clone BAC 261N5, complete sequence.		
102	KF562709	Oryza rufipogon cultivar DongXiang chloroplast, complete genome.	UniRef100_C5WNJ6	
103	AK253124	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf62c15, mRNA sequence.		
104	AK376929	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3144F24.	UniRef100_M8BDQ9	M0UTL8
105	AK249924	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf53g17, mRNA sequence.	UniRef100_M0ZDL7	
106	AK373644	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv3038H09.	UniRef100_F2DZT8	F2DZT8
107	AK357163	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1047F19.	UniRef100_UPI00032A5479	F2CZV7
108	DQ286562	Triticum aestivum putative lipid transfer protein mRNA, complete cds.	UniRef100_E9B0Q3	A0MAU6
109	AK335062	Triticum aestivum cDNA, clone : WT011_P12, cultivar : Chinese Spring.	UniRef100_M8BY61	
110	AK333238	Triticum aestivum cDNA, clone : WT005_P18, cultivar : Chinese Spring.		
111	AE014187	Plasmodium falciparum 3D7 chromosome 14, complete sequence.	UniRef100_G3WB51	Q8ILI6
112	AK332529	Triticum aestivum cDNA, clone : WT004_D08, cultivar : Chinese Spring.	UniRef100_M7ZA56	
113	AK250397	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf73d02, mRNA sequence.	UniRef100_I1MUY1	
114	XM_003066908	Coccidioides posadasii C735 delta SOWgp hypothetical protein, mRNA.		C5PE76
115	JF489233	Secale cereale external transcribed spacer, 18S ribosomal RNA gene, internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence ; and 26S ribosomal RNA gene, partial sequence.	UniRef100_J6CJ14	

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
116	X59874	T.aestivum L. mRNA for TATA binding protein (TFIID).		P26356
117	AK334519	Triticum aestivum cDNA, clone : WT010_C18, cultivar : Chinese Spring.	UniRef100_UPI00037D8F91	
118	AK249091	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf13i21, mRNA sequence.		
119	AK333035	Triticum aestivum cDNA, clone : WT005_H19, cultivar : Chinese Spring.	UniRef100_Q9FT38	
120	CT009735	Triticum aestivum.	UniRef100_M0X4A9	P33432
121	XM_003566361	PREDICTED : Brachypodium distachyon serine carboxypeptidase II-1-like (LOC100823672), mRNA.	UniRef100_Q9FYP7	I1HB12
122	EU626553	Triticum urartu clone BAC 261N5, complete sequence.		
123	AK355723	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1024M03.	UniRef100_N1R2N4	F2CVS0
124	AK362210	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2003G07.	UniRef100_N1R1W3	F2DE93
125	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_F8RPK4	D7F4N2
126	FN667741	Xenorhabdus bovienii SS-2004 chromosome, complete genome.		D3UWL6
127	AK330423	Triticum aestivum cDNA, clone : SET4_G18, cultivar : Chinese Spring.	UniRef100_R7W9V1	
128	XM_003568915	PREDICTED : Brachypodium distachyon putative uncharacterized protein DDB_G0277003-like (LOC100834914), mRNA.	UniRef100_M8BEY4	I1HM15
129	AL161898	Human DNA sequence from clone RP11-270H22 on chromosome 13, complete sequence.	UniRef100_N1QQV8	Q9UEF7
130	AK358856	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1084G14.		M0XNP1
131	HF541871	Triticum aestivum chromosome 3B specific BAC library, BAC clone TaaCsp3BFhA_0037C18.	UniRef100_M7ZGW4	
132	AK357546	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1056E05.	UniRef100_M8BNG4	F2D0Y9
133	XM_003579270	PREDICTED : Brachypodium distachyon probable cleavage and polyadenylation specificity factor subunit 1-like (LOC100831691), mRNA.	UniRef100_M7YZ81	I1IWJ9
134	AK363003	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2012C19.	UniRef100_F2DGI6	F2DGI6
135	HQ391329	Triticum aestivum clone UCDTA01780 genomic sequence.	UniRef100_UPI00020625E8	
136	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_D5LMK7	D7F4N2
137	AK332255	Triticum aestivum cDNA, clone : WT003_I14, cultivar : Chinese Spring.	UniRef100_M0Y6M7	
138	AK428173	Brachypodium distachyon mRNA, clone : PL016C01-A-025_I24.	UniRef100_I1PCF2	

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
139	XM_001360285	Drosophila pseudoobscura pseudoobscura GA13769 (Dpse\GA13769), mRNA.	UniRef100_UPI000328E9B4	Q292G5
140	AK427458	Brachypodium distachyon mRNA, clone : PL016C01-A-020_P14.	UniRef100_T1L6P5	
141	XM_002004130	Drosophila mojavensis GI19749 (Dmoj\GI19749), mRNA.	UniRef100_M7YLM0	B4KPZ9
142	AK365545	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2034O17.	UniRef100_N1QUB1	M0V0C8
143	AK362799	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2010C22.		F2DFY2
144	AK332897	Triticum aestivum cDNA, clone : WT005_C09, cultivar : Chinese Spring.		
145	JF750561	Silene conica chromosome 74 mitochondrion, complete sequence.		
146	AF508970	Triticum aestivum translationally controlled tumor protein mRNA, complete cds.	UniRef100_M7YF70	Q8LRM8
147	AK369375	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2090A21.	UniRef100_M8AIN6	F2D0H1
148	AJ001117	Triticum aestivum mRNA for sucrose synthase type I.		O82073
149	XM_004292152	PREDICTED : Fragaria vesca subsp. vesca transcription factor bHLH155-like (LOC101296543), mRNA.		
150	AK330745	Triticum aestivum cDNA, clone : SET5_D06, cultivar : Chinese Spring.	UniRef100_M0V3G8	
151	AK358091	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1068I08.	UniRef100_M0UQD7	M0UQD6
152	AK335725	Triticum aestivum cDNA, clone : SET2_K04, cultivar : Chinese Spring.	UniRef100_M7ZVF6	
153	AK368264	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2071B24.		F2D3Q3
154	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
155	AK250053	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf63k10, mRNA sequence.	UniRef100_R7W208	
156	AK374366	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3062L03.	UniRef100_M5WKF5	M0VLL9
157	DQ537335	Triticum aestivum clones BAC 1031P08 ; BAC 754K10 ; BAC 1344C16, complete sequence.		Q41553
158	AK331581	Triticum aestivum cDNA, clone : SET1_J20, cultivar : Chinese Spring.		
159	DQ862833	Triticum monococcum S-adenosylhomocysteine hydrolase mRNA, partial cds.	UniRef100_N4UPG8	A6XMZ1
160	XM_003557202	PREDICTED : Brachypodium distachyon cation-chloride cotransporter 1-like (LOC100840956), mRNA.	UniRef100_M0XUD7	I1H1W9
161	AK330639	Triticum aestivum cDNA, clone : SET4_P03, cultivar : Chinese Spring.		

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
162	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_M0ZEQ9	D7F4N2
163	XM_003578780	PREDICTED : Brachypodium distachyon chaperone protein DnaJ-like (LOC100821453), mRNA.	UniRef100_I1ITW1	
164	DQ432014	Triticum aestivum vacuolar proton-ATPase subunit A mRNA, complete cds.	UniRef100_B7FFL1	Q1W681
165	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_M0ZEQ9	D7F4N2
166	EU835980	Triticum aestivum clone BAC 502E09, complete sequence.	UniRef100_T1MQ35	B6Z259
167	FN564426	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0005b.	UniRef100_M0ZDG8	D8L9G1
168	AK332496	Triticum aestivum cDNA, clone : WT004_B23, cultivar : Chinese Spring.		
169	AK363775	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2018M19.	UniRef100_M7YNS6	F2DIQ7
170	AK359234	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1092G20.		F2D5S4
171	AK363357	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2014M01.	UniRef100_M7ZEI4	M0VKB6
172	XM_004263706	PREDICTED : Orcinus orca spectrin repeat containing, nuclear envelope 1 (SYNE1), mRNA.		
173	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
174	AK333177	Triticum aestivum cDNA, clone : WT005_N11, cultivar : Chinese Spring.	UniRef100_M8D509	
175	AJ132439	Triticum aestivum mRNA for protein encoded by lt1.1 gene, partial.		Q9FEH6
176	AK331581	Triticum aestivum cDNA, clone : SET1_J20, cultivar : Chinese Spring.		
177	AK360900	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1127P21.	UniRef100_M0X4A9	M0Z1X6
178	AK102753	Oryza sativa Japonica Group cDNA clone :J033106N01, full insert sequence.	UniRef100_M8A918	
179	FJ436986	Aegilops tauschii Lr34 locus, partial sequence.	UniRef100_S2Y442	B8XSN7
180	FN564430	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0464b.	UniRef100_M8CQ09	D8L9N5
181	FJ427399	Triticum turgidum clone BAC 738D05 chromosome 4B, partial sequence.	UniRef100_R7W5L4	B7U385
182	AY943294	Hordeum vulgare subsp. vulgare clone BAC 673I14, complete sequence.		M0YT37

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
183	AY534123	Aegilops tauschii transposons Caspar, XJ, Angela, and XJ, complete sequence ; LRR protein WM1.7 (WM1.7) and LRR protein WM1.12 (WM1.12) genes, complete cds ; transposons Ophelia2, Angela3s, and XJ3, complete sequence ; LRR protein WM1.3 (WM1.3) gene, complete cds ; Stowaway MITE, transposons XJ1, Jody, Angela, and XJ and Stowaway MITE, complete sequence ; LRR protein WM1.2 (WM1.2) and LLR protein WM1.1 (WM1.1) genes, complete cds ; transposons XJ, XA, Angela, and Fred and WM1.11 gene, complete sequence ; RPM1-like sequence and LRR protein WM1.10 (WM1.10) genes, complete cds ; and transposon XJ, complete sequence.		Q6QM06
184	AK330153	Triticum aestivum cDNA, clone : SET3_M02, cultivar : Chinese Spring.	UniRef100_N1QZJ5	
185	AK252215	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf147e20, mRNA sequence.	UniRef100_M8BPE5	
186	XM_005568345	PREDICTED : Macaca fascicularis BTB (POZ) domain containing 3 (BTBD3), transcript variant X3, mRNA.		
187	AY951945	Triticum monococcum TmBAC 60J11 FR-Am2 locus, genomic sequence.	UniRef100_M7YVM1	Q2VQ32
188	NM_001175821	Zea mays LOC100383156 (umc1982), mRNA.	UniRef100_C0PDN9	C0PDN9
189	NM_001174628	Zea mays uncharacterized LOC100381836 (LOC100381836), mRNA.	UniRef100_M7YKC3	C0HJ82
190	AK359234	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1092G20.	UniRef100_J3LN75	F2D5S4
191	GQ419475	Oryza sativa Japonica Group cultivar Khao Hawm putative precursor microRNA R395n-s gene, complete sequence.		
192	XM_003562591	PREDICTED : Brachypodium distachyon uncharacterized LOC100836004 (LOC100836004), mRNA.		I1GS36
193	AK332664	Triticum aestivum cDNA, clone : WT004_I22, cultivar : Chinese Spring.		
194	AK249069	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf21o09, mRNA sequence.	UniRef100_M8ABV0	
195	HE774676	Triticum aestivum chromosome arm 3DS-specific BAC library, contig ctg447.		I0JTU1
196	HE601631	Schistosoma mansoni strain Puerto Rico chromosome W, complete genome.		C4PYP8
197	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
198	AK334078	Triticum aestivum cDNA, clone : WT009_E03, cultivar : Chinese Spring.		
199	AK369070	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2084N12.	UniRef100_J3MD51	F2DYV0
200	AC100740	Mus musculus chromosome 1, clone RP24-421N21, complete sequence.	UniRef100_M8CHA7	

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
201	AK331183	Triticum aestivum cDNA, clone : SET6_K07, cultivar : Chinese Spring.		
202	AK363930	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2020C20.	UniRef100_I1GZY8	F2DJ62
203	AK355756	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1025F06.		F2CVV3
204	AK251945	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf134b21, mRNA sequence.		
205	CP001848	Pirellula staleyi DSM 6068, complete genome.	UniRef100_L9JYY2	D2QW97
206	FN564430	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0464b.	UniRef100_M7ZWV8	D8L9N5
207	XM_391032	Gibberella zeae PH-1 actin-like protein 3 partial mRNA.	UniRef100_R8BWG8	I1S270
208	FJ345689	Triticum aestivum MITE Tourist-3 MITE Islay Tourist, complete sequence.		
209	DQ245666	Zea mays clone 18950 mRNA sequence.	UniRef100_UPI00030AEB8F	
210	AK250087	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf63p07, mRNA sequence.		
211	GU817319	Triticum aestivum clone BAC_2383A24 chromosome 3B, complete sequence.		F2VPV0
212	AP011170	Acetobacter pasteurianus IFO 3283-12 DNA, complete genome.		C7L2T7
213	AK362464	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2005K17.	UniRef100_M7YLY4	F2DEZ7
214	AK372026	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2145B17.	UniRef100_M0WEW6	M0WEW6
215	AB838408	Oryza sativa Indica Group Hd3a gene for complete cds, bio_material : MAFF¡JPN¿ :WRC100, cultivar : Vandaran.		
216	AP013107	Aegilops speltoides mitochondrial DNA, complete sequence.	UniRef100_R4IUU2	
217	AK364086	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2021I23.		F2DJL8
218	AK374654	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv3071M20.		F2EES8
219	CP003745	Bibersteinia trehalosi USDA-ARS-USMARC-192, complete genome.		M4R6I7
220	AK364979	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2029P19.		F2DM61
221	AK334145	Triticum aestivum cDNA, clone : WT009_O11, cultivar : Chinese Spring.	UniRef100_I1HJ55	
222	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_A6YM14	D7F4N2
223	AK372166	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2147J05.	UniRef100_M8C1S0	M0YVJ9

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
224	GU817319	Triticum aestivum clone BAC_2383A24 chromosome 3B, complete sequence.		F2VPV0
225	AK369940	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2101J05.		F2E1B9
226	AK334286	Triticum aestivum cDNA, clone : WT009_F09, cultivar : Chinese Spring.		
227	AK332238	Triticum aestivum cDNA, clone : WT003_H22, cultivar : Chinese Spring.	UniRef100_M0V5L4	
228	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.	UniRef100_E1SBT5	D8L9S5
229	AF459639	Triticum monococcum BAC clones 116F2 and 115G1 gene sequence.	UniRef100_M7Z092	Q8SAE0
230	AK355852	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1026O16.	UniRef100_M8BX29	M0Z6N7
231	GU211169	Triticum aestivum clone 09d3 gliadin/avenin-like seed protein mRNA, complete cds.	UniRef100_D2KFH0	D2KFH0
232	AK369019	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2083N17.	UniRef100_M0YFE4	M0YFE1
233	AF532601	Triticum aestivum multidrug resistance associated protein MRP2 mRNA, complete cds.	UniRef100_M7ZK96	Q71CZ3
234	AC162123	Neofelis nebulosa clone CH87-231N4, complete sequence.		
235	XM_003580986	PREDICTED : Brachypodium distachyon cysteine-rich receptor-like protein kinase 19-like (LOC100830795), mRNA.	UniRef100_T1LCX9	
236	AK333949	Triticum aestivum cDNA, clone : WT008_P23, cultivar : Chinese Spring.		
237	AK374367	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv3062L09.	UniRef100_T1M3K5	M0ZEK5
238	FN564428	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0091b.	UniRef100_M8CQ09	D8L9J2
239	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
240	JF758491	Triticum aestivum clone 515O12 genomic sequence.	UniRef100_Q9S9A8	F5CPR7
241	XM_003576169	PREDICTED : Brachypodium distachyon uncharacterized LOC100827707 (LOC100827707), mRNA.	UniRef100_M8A1Y6	
242	AK376851	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv3142C22.	UniRef100_F2E3Y2	M0Z3K0
243	AY643842	Hordeum vulgare subsp. vulgare clone BAC 519K7 hardness locus region.	UniRef100_N1R2N4	
244	AK361510	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1142F15.	UniRef100_N1QYD6	M0VBM7
245	AK333242	Triticum aestivum cDNA, clone : WT005_P22, cultivar : Chinese Spring.	UniRef100_M8BJR6	

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
246	XM_003557582	PREDICTED : Brachypodium distachyon U4/U6 small nuclear ribonucleoprotein Prp31-like (LOC100828224), mRNA.		
247	KF562709	Oryza rufipogon cultivar DongXiang chloroplast, complete genome.	UniRef100_A6N1H4	
248	AK367892	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2064E17.	UniRef100_N1QRV8	M0XT09
249	AK372363	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2150K10.		F2E889
250	AK362971	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2011O01.		F2DGF4
251	DQ537335	Triticum aestivum clones BAC 1031P08 ; BAC 754K10 ; BAC 1344C16, complete sequence.		Q41553
252	AK335757	Triticum aestivum cDNA, clone : WT013_L07, cultivar : Chinese Spring.		
253	FJ225148	Triticum aestivum ferritin 2A gene, complete cds.	UniRef100_R9M0F6	B6UZ90
254	HE996525	Triticum aestivum cv. Arina SNP, chromosome 3B, clone Taes_arina_ctg_58249.	UniRef100_C8ZBZ2	
255	AP013107	Aegilops speltoides mitochondrial DNA, complete sequence.	UniRef100_M0UC49	
256	FN645450	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0011b.	UniRef100_M8JK39	D8LAL5
257	AK332440	Triticum aestivum cDNA, clone : WT003_P20, cultivar : Chinese Spring.		
258	AK333064	Triticum aestivum cDNA, clone : WT005_I23, cultivar : Chinese Spring.		
259	AK332804	Triticum aestivum cDNA, clone : WT004_O17, cultivar : Chinese Spring.		
260	FN564430	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0464b.	UniRef100_M0W8A2	D8L9N5
261	AK335863	Triticum aestivum cDNA, clone : WT013_P13, cultivar : Chinese Spring.	UniRef100_M7YYG6	
262	AB646974	Triticum aestivum PRR gene for pseudo-response regulator, complete cds, allele : Ppd-B1a.1.	UniRef100_B6AXU4	A7J5T4
263	GU817319	Triticum aestivum clone BAC_2383A24 chromosome 3B, complete sequence.		F2VPV0
264	AK332413	Triticum aestivum cDNA, clone : WT003_O18, cultivar : Chinese Spring.	UniRef100_M8CS21	
265	FJ427399	Triticum turgidum clone BAC 738D05 chromosome 4B, partial sequence.	UniRef100_T1NSR2	B7U385
266	AF548379	Aegilops tauschii isoamylase gene, complete cds.		Q7XA16
267	FN564433	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0661b.		D8L9Q3
268	AK334924	Triticum aestivum cDNA, clone : WT011_I02, cultivar : Chinese Spring.		
269	AK334063	Triticum aestivum cDNA, clone : WT009_A23, cultivar : Chinese Spring.	UniRef100_M7YKC3	
270	EU379326	Poa palustris isolate 1-2 phosphoglucose isomerase (PgiC) gene, partial cds.		B2CBC4

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
271	DQ167201	Triticum aestivum eukaryotic translation initiation factor 5A1 gene, complete cds.		Q3S4I1
272	AK362302	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2004B14.		F2DEI5
273	AC188502	Gymnogyps californianus clone CH262-225F20, complete sequence.		
274	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
275	AF325197	Triticum aestivum LRK33 (Lrk33) and TAK33 (Tak33) genes, complete cds.		Q9ATQ4
276	AK353995	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1004E18.	UniRef100_M8C4Y8	F2CQU3
277	GQ409824	Triticum turgidum subsp. durum cultivar Langdon clone BAC 406B11, complete sequence.	UniRef100_M0X4A9	E2CZH0
278	AK360479	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1118O11.	UniRef100_M8B455	F2D9B7
279	AF343493	Secale cereale clone pla 3-phosphoglycerate kinase (Pgk-1) gene, partial cds ; nuclear gene for plastid product.		Q8LLS6
280	AK334173	Triticum aestivum cDNA, clone : WT009_C16, cultivar : Chinese Spring.		
281	JX978695	Triticum urartu clone BAC Tu-JJ1, complete sequence.	UniRef100_R7W8A4	M1FWA6
282	XM_001454549	Paramecium tetraurelia hypothetical protein (GSPATT00020842001) partial mRNA.		A0DVX2
283	FN554889	Streptomyces scabiei 87.22 complete genome.	UniRef100_K1VB68	C9Z3U4
284	AK332097	Triticum aestivum cDNA, clone : WT003_C06, cultivar : Chinese Spring.	UniRef100_H0XXW9	
285	AK367463	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv2057J24.	UniRef100_M7ZN31	M0XAJ8
286	EU626553	Triticum urartu clone BAC 261N5, complete sequence.		
287	AK355592	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1023C07.	UniRef100_M8BNQ1	F2CVE0
288	XM_004352180	Dictyostelium fasciculatum hypothetical protein (DFA_09576) mRNA, complete cds.		F4Q807
289	FN564432	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0616b.		D8L9P6
290	AK335883	Triticum aestivum cDNA, clone : SET2_L04, cultivar : Chinese Spring.		
291	AC158794	Mus musculus chromosome 1, clone RP23-306I10, complete sequence.		
292	HQ390278	Triticum aestivum clone UCDTA00729 genomic sequence.		
293	AB238931	Triticum monococcum TmABI1 gene for protein phosphatase 2C, complete cds.	UniRef100_R7W5L4	A5A6P9
294	JQ269664	Triticum aestivum cultivar WL 711 betaine aldehyde dehydrogenase-like protein mRNA, partial cds.	UniRef100_H9NAU5	H9NAU4

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
295	AK357215	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1048G09.	UniRef100_M7YMZ5	F2D008
296	AK335270	Triticum aestivum cDNA, clone : WT012_H16, cultivar : Chinese Spring.	UniRef100_R7W8A4	
297	AY968588	Triticum aestivum ice recrystallization inhibition protein 1 precursor, mRNA, complete cds.		Q56B90
298	XM_004330590	PREDICTED : Tursiops truncatus uncharacterized LOC101330279 (LOC101330279), mRNA.	UniRef100_F0VKZ7	
299	AK357801	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1062C02.	UniRef100_I1IWA3	M0VW83
300	AK332508	Triticum aestivum cDNA, clone : WT004_C11, cultivar : Chinese Spring.	UniRef100_M0VZ60	
301	DQ537335	Triticum aestivum clones BAC 1031P08 ; BAC 754K10 ; BAC 1344C16, complete sequence.	UniRef100_G7YAJ2	Q41553
302	JQ740834	Aegilops speltoides isolate SPE0661 chloroplast, complete genome.	UniRef100_Q8HUN3	D7F4N2
303	AK372309	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2149I20.	UniRef100_M0W8A2	F2E835
304	AK358630	Hordeum vulgare subsp. vulgare mRNA for predicted protein, partial cds, clone : NIASHv1080I05.	UniRef100_M0YLY1	M0YLY2
306	AK364228	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2022N09.	UniRef100_J3MQC6	F2DK10
307	AK335953	Triticum aestivum cDNA, clone : SET1_C22, cultivar : Chinese Spring.	UniRef100_M8A0S9	
308	KF602231	Poa alpina ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit gene, partial cds ; chloroplast.		
309	AK369720	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2096K06.	UniRef100_Q5BU44	F2E0P9
310	EF450765	Brachypodium sylvaticum isolate 2-6E8 microsatellite sequence.		

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
311	JF946485	Triticum aestivum retrotransposons Gypsy TREP 3245_Sabrina, Copia TREP 3161_WIS, Gypsy TREP 3208_Laura, Copia TREP 3161_WIS, and Gypsy TREP 3173_Derami and transposon TREP 3040_Harbinger, complete sequence ; pseudo-response regulator (Ppd-B1) gene, Ppd-B1a allele, complete cds ; retrotransposons Copia TREP 3161_WIS and Gypsy TREP 3173_Derami and transposon TREP 3040_Harbinger, complete sequence ; pseudo-response regulator (Ppd-B1_i1) gene, Ppd-B1_i1-Ppd-B1a allele, complete cds ; retrotransposons Gypsy TREP 3457_Danae, Copia TREP 3161_WIS, and Gypsy TREP 3173_Derami and transposon TREP 3040_Harbinger, complete sequence ; pseudo-response regulator (Ppd-B1_i2) gene, Ppd-B1_i2-Ppd-B1a allele, complete cds ; retrotransposons Gypsy TREP 3457_Danae, Copia TREP 3161_WIS, and Gypsy TREP 3173_Derami and transposon TREP 3040_Harbinger, complete sequence ; pseudo-response regulator (Ppd-B1_i3) gene, Ppd-B1_i3-Ppd-B1a allele, complete cds ; and retrotransposons Gypsy TREP 3457_Danae and Gypsy TREP 3196_Fatima, transposon CACTA TREP 3004_Boris, retrotransposons Gypsy TREP 3196_Fatima and Copia TREP 3529_Angela, complete sequence.		A7J5T2
312	AY106124	Zea mays PCO123686 mRNA sequence.	UniRef100_P42057	
313	AK336081	Triticum aestivum cDNA, clone : SET3_C24, cultivar : Chinese Spring.	UniRef100_M7YMK8	
314	AM932685	Triticum aestivum 3B chromosome, clone BAC TA3B95F5.	UniRef100_G0CWD7	B4ERX4
315	GQ905540	Zea mays clone zma-miR167d-4 precursor miRNA zma-miR167d, precursor RNA, complete sequence.		
316	AK332840	Triticum aestivum cDNA, clone : WT005_A03, cultivar : Chinese Spring.		
317	AK365987	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIA-SHv2039J08.		M0YMH3
318	FN564428	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0091b.	UniRef100_M8CQ09	D8L9J2
319	AK355497	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIA-SHv1021I10.	UniRef100_M0YYW0	F2CV45
320	AK330205	Triticum aestivum cDNA, clone : SET3_O05, cultivar : Chinese Spring.	UniRef100_D5J6W4	
321	AK367678	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIA-SHv2060E19.	UniRef100_M8A6Z7	F2DUW0
322	XM_002299628	Populus trichocarpa predicted protein, mRNA.		B9GEV5
323	AK365855	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIA-SHv2038E22.	UniRef100_M8BPB4	F2DPN7

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
324	AK371630	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2138M01.		F2E656
325	XM_001553870	Botryotinia fuckeliana B05.10 hypothetical protein (BC1G_07480) partial mRNA.	UniRef100_N1JDU0	
326	KC573058	Triticum monococcum subsp. monococcum cultivar DV92 Sr35 region, genomic sequence.	UniRef100_F2E545	S5A8C3
327	BT009452	Triticum aestivum clone wlmk8.pk0022.f7 :fis, full insert mRNA sequence.	UniRef100_UPI000347AF2A	
328	FN564429	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0382b.	UniRef100_N1R0I0	D8L9K0
329	AK358388	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1075G23.	UniRef100_K3YVC8	F2D3D1
330	AC187026	Canis Familiaris chromosome 17, clone XX-266E21, complete sequence.	UniRef100_K7GQP5	
331	AK249338	Hordeum vulgare subsp. vulgare cDNA clone : FLbaf33p14, mRNA sequence.	UniRef100_I1IAN7	
332	AK363298	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2014C09.	UniRef100_M8B4D8	F2DHD1
333	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.	UniRef100_D9R5A1	D8L9S5
334	AK333292	Triticum aestivum cDNA, clone : WT006_B19, cultivar : Chinese Spring.	UniRef100_F2D105	
335	XM_003571147	PREDICTED : Brachypodium distachyon pentatricopeptide repeat-containing protein At2g32230, mitochondrial-like (LOC100845397), mRNA.	UniRef100_M7ZJX1	
336	BT009004	Triticum aestivum clone wdk2c.pk018.c16 :fis, full insert mRNA sequence.	UniRef100_Q5AVI5	
337	AY963808	Triticum aestivum putative S-locus receptor kinase gene, partial cds ; and mitochondrial Mn-superoxide dismutase (MnSOD) gene, complete cds, nuclear gene encoding mitochondrial protein.	UniRef100_M7YVM1	Q56DH9
338	AC246851	Solanum lycopersicum strain Heinz 1706 chromosome 3 clone slm-68a14 map 3, complete sequence.		
339	AK335226	Triticum aestivum cDNA, clone : WT012_G01, cultivar : Chinese Spring.		
340	FN564434	Triticum aestivum chromosome 3B-specific BAC library, contig ctg0954b.		D8L9S5
341	HQ435325	Triticum aestivum clone BAC 1J9 Tmemb_185A domain-containing protein (1J9.1), EamA domain-containing protein (1J9.2), and Rht-D1b (Rht-D1b) genes, complete cds, complete sequence.		I3NM21
342	HQ390713	Triticum aestivum clone UCDTA01164 genomic sequence.	UniRef100_A8WZ18	
343	AK356287	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv1032K07.	UniRef100_M7YPF4	F2CXD4

Contig	Accession	Description	Uniref - EBI	Uniref - mapping
344	AK336242	Triticum aestivum cDNA, clone : SET1_E02, cultivar : Chinese Spring.	UniRef100_M8B455	
345	AK368463	Hordeum vulgare subsp. vulgare mRNA for predicted protein, complete cds, clone : NIASHv2073P17.	UniRef100_M0XRZ8	F2DX45
346	HQ391329	Triticum aestivum clone UCDTA01780 genomic sequence.		

4 Question 3

a Identificateur du vecteur pANNE.

5 Question 4

a Alignez ces CDS en utilisant ClustalW, dialign et Mavid

b Discutez de la performance en terme de facilité d'utilisation et de rapidité des différents programmes.

c Analysez et discutez de la qualité de l'alignement donné par chaque méthode.

d Quel est votre meilleur alignement ? Justifiez votre choix.

6 Annexes

1 Tableau de la taille et du taux de GC des contigs

Contig	Taille	Taux GC	Contig	Taille	Taux GC
1	91	51.65	2	182	31.87
3	183	42.62	4	103	39.81
5	83	42.17	6	90	42.22
7	100	46.00	8	88	38.64
9	122	33.61	10	110	37.27
11	140	43.57	12	96	37.50
13	120	37.50	14	114	39.47
15	79	64.56	16	238	41.60
17	143	55.24	18	97	51.55
19	143	53.85	20	82	45.12
21	209	45.45	22	105	38.10
23	150	48.67	24	102	47.06
25	101	36.63	26	88	43.18
27	151	35.76	28	137	59.85
29	97	44.33	30	112	25.89
31	70	38.57	32	125	44.80
33	83	40.96	34	95	32.63
35	201	38.31	36	203	30.05
37	69	79.71	38	153	36.60
39	75	44.00	40	96	46.88
41	52	44.23	42	74	45.95
43	96	63.54	44	115	43.48
45	87	44.83	46	96	48.96
47	82	45.12	48	125	40.00
49	131	35.88	50	102	49.02
51	70	47.14	52	80	43.75
53	99	42.42	54	124	40.32
55	108	45.37	56	142	40.85
57	87	36.78	58	133	40.60
59	112	33.04	60	90	40.00
61	94	44.68	62	95	44.21
63	103	44.66	64	185	59.46
65	142	41.55	66	99	40.40
67	150	49.33	68	84	34.52
69	46	50.00	70	110	34.55
71	80	47.50	72	134	50.00
73	113	53.10	74	93	45.16
75	179	36.87	76	82	46.34
77	106	42.45	78	61	45.90
79	157	31.21	80	111	69.37
81	92	47.83	82	183	45.90
83	51	39.22	84	143	34.97
85	87	40.23	86	102	62.75
87	51	29.41	88	116	46.55
89	107	42.99	90	76	46.05
91	73	47.95	92	104	36.54
93	201	45.27	94	85	31.76
95	99	41.41	96	108	38.89
97	63	33.33	98	131	45.80
99	62	25.81	100	111	40.54

101	121	38.84	102	109	43.12
103	70	37.14	104	148	43.92
105	105	46.67	106	103	70.87
107	105	37.14	108	104	42.31
109	92	41.30	110	80	38.75
111	120	22.50	112	101	47.52
113	69	44.93	114	76	38.16
115	92	76.09	116	69	36.23
117	94	47.87	118	55	34.55
119	138	67.39	120	112	38.39
121	78	47.44	122	111	32.43
123	104	31.73	124	136	42.65
125	104	46.15	126	72	34.72
127	88	47.73	128	76	38.16
129	120	42.50	130	114	42.98
131	105	49.52	132	100	39.00
133	114	42.11	134	69	42.03
135	112	41.07	136	117	49.57
137	106	45.28	138	79	54.43
139	117	51.28	140	100	67.00
141	146	43.15	142	113	46.02
143	120	44.17	144	155	40.65
145	88	40.91	146	121	46.28
147	98	47.96	148	100	50.00
149	81	44.44	150	60	41.67
151	80	46.25	152	56	41.07
153	109	44.04	154	89	55.06
155	179	44.69	156	107	43.93
157	99	38.38	158	101	35.64
159	125	46.40	160	80	41.25
161	93	56.99	162	57	47.37
163	79	50.63	164	221	32.58
165	96	41.67	166	94	45.74
167	119	32.77	168	89	35.96
169	81	51.85	170	94	39.36
171	115	46.09	172	92	41.30
173	117	43.59	174	215	41.40
175	83	30.12	176	111	37.84
177	104	32.69	178	76	46.05
179	204	33.82	180	159	41.51
181	123	35.77	182	87	27.59
183	100	40.00	184	90	46.67
185	93	41.94	186	67	40.30
187	138	39.13	188	83	36.14
189	81	32.10	190	77	45.45
191	101	46.53	192	78	42.31
193	82	46.34	194	131	38.93
195	96	34.38	196	69	47.83
197	115	47.83	198	90	51.11
199	154	45.45	200	77	49.35
201	93	41.94	202	189	42.86
203	84	39.29	204	96	37.50
205	109	56.88	206	81	46.91
207	106	51.89	208	131	29.77
209	94	45.74	210	75	36.00

211	109	36.70	212	120	55.83
213	128	57.03	214	77	38.96
215	121	36.36	216	72	50.00
217	139	48.20	218	71	49.30
219	106	44.34	220	84	39.29
221	81	35.80	222	91	35.16
223	98	43.88	224	104	33.65
225	84	46.43	226	130	31.54
227	94	46.81	228	166	41.57
229	138	41.30	230	90	45.56
231	95	47.37	232	94	45.74
233	98	55.10	234	92	41.30
235	130	33.08	236	79	43.04
237	127	46.46	238	203	43.84
239	92	40.22	240	127	47.24
241	120	39.17	242	95	49.47
243	174	35.06	244	115	46.09
245	155	36.13	246	84	47.62
247	82	52.44	248	103	39.81
249	94	48.94	250	101	33.66
251	90	37.78	252	73	64.38
253	110	42.73	254	125	36.80
255	82	53.66	256	152	24.34
257	80	43.75	258	95	38.95
259	61	49.18	260	106	36.79
261	68	39.71	262	95	35.79
263	114	41.23	264	133	44.36
265	93	40.86	266	97	42.27
267	99	43.43	268	133	40.60
269	118	33.90	270	157	38.85
271	188	42.55	272	94	52.13
273	126	26.98	274	247	34.82
275	101	31.68	276	104	71.15
277	94	34.04	278	149	34.90
279	111	32.43	280	139	47.48
281	95	32.63	282	87	27.59
283	107	72.90	284	115	30.43
285	133	48.87	286	112	33.93
287	100	48.00	288	128	41.41
289	90	41.11	290	92	40.22
291	137	48.18	292	90	41.11
293	153	41.83	294	84	45.24
295	80	40.00	296	82	28.05
297	92	51.09	298	108	70.37
299	101	40.59	300	94	64.89
301	173	35.26	302	112	41.96
303	107	37.38	304	106	38.68
305	81	40.74	306	118	33.05
307	95	47.37	308	90	44.44
309	96	51.04	310	94	43.62
311	99	40.40	312	121	35.54
313	89	49.44	314	94	39.36
315	99	56.57	316	70	42.86
317	90	51.11	318	233	38.63
319	118	48.31	320	118	33.05

321	76	48.68	322	87	37.93
323	105	47.62	324	117	39.32
325	88	47.73	326	179	37.43
327	262	36.64	328	166	37.35
329	200	50.00	330	128	34.38
331	111	36.94	332	119	36.97
333	150	40.00	334	113	42.48
335	94	43.62	336	91	47.25
337	174	33.33	338	86	41.86
339	187	34.76	340	73	36.99
341	92	39.13	342	114	37.72
343	80	43.75	344	153	35.29
345	77	54.55	346	92	43.48

346 contigs, taille moyenne : 109.176300578 Taux GC moyen : 42.9628288283

2 Script biopython de calcul des fréquences nucléotidiques

```
# -* coding:utf-8 *-#
from Bio import SeqIO
from Bio.SeqRecord import SeqRecord
handle = open(”NC_000002_202564986-202645895.gb”, ”r”)
seq_record = SeqIO.parse(handle, ’gb’)
for seq in seq_record:
    dist_a = seq.seq.count(”A”)
    dist_c = seq.seq.count(”C”)
    dist_g = seq.seq.count(”G”)
    dist_t = seq.seq.count(”T”)
    print ”A:␣␣count:␣” + str(dist_a) + ”␣%␣=␣” + \
        str(float(dist_a)/len(seq)*100)
    print ”C:␣␣count:␣” + str(dist_c) + ”␣%␣=␣” + \
        str(float(dist_c)/len(seq)*100)
    print ”G:␣␣count:␣” + str(dist_g) + ”␣%␣=␣” + \
        str(float(dist_g)/len(seq)*100)
    print ”T:␣␣count:␣” + str(dist_t) + ”␣%␣=␣” + \
        str(float(dist_t)/len(seq)*100)
    print ”total␣=␣” + str(dist_a+dist_c+dist_g+dist_t)
```

3 Script biopython Pour choisir 5 contigs au hasard à partir du résulat de CAP3, et d’effectuer un blast sur ces contigs

```
# *- coding:utf-8 -* #
import random
from Bio.Blast import NCBIWWW

contigs = {}
contig_no = None
contig_seq = ””
contig_size = 0

with open(”seq.data.cap.contigs”, ”r”) as f:
    for line in f:
        # on regarde d'abord si c'est un contig ou non
        if line[0] == '>':
            if contig_no == None:
                contig_no = int(line[7:])
            if contig_seq != "":
                contigs.update({contig_no:contig_seq})
                contig_seq = ""
                contig_no = int(line[7:])
                contig_size+= 1
        else :
            contig_seq = contig_seq + line.replace("\n","")
    contigs.update({contig_no:contig_seq})
    contig_size +=1

# Maintenant, on a nos contigs, on en choisit 5 au hasard
random_contig = []

#Je m'assure ici de ne pas avoir de doublon
for i in range(5):
    random_c = random.randint(1, contig_size)
    while random_c in random_contig:
        random_c = random.randint(1,contig_size)
    random_contig.append(random_c)

#On blast maintenant les contigs choisis:
for i in random_contig:
    result_handle = NCBIWWW.qblast("blastn", "nr", contigs[i])

    #on enregistre le r sultat
    nom_fichier = "blast_contig_" + str(i) + ".xml"
    save_file = open(nom_fichier, "w")
    save_file.write(result_handle.read())
    save_file.close()
    result_handle.close()

print "5␣contigs␣cherch s"
```

4 Script Biopython pour obtenir les fichiers d'accession des 10 premiers résultats du blast pour un contig donné.

```
# *- coding:utf-8 -* #

#Parser pour un fichier XML de resultat blast
#Specifique a la question 2 du devoir 1. Je sais
#ici que chaque hit a seulement un hsp, donc en
#specifiant le # d'accession et le sbjct_start et end,
#j'obtiens ce que je cherche

import sys
import os
from Bio.Blast import NCBIXML
from Bio import SeqIO
from Bio.SeqRecord import SeqRecord
from Bio import Entrez

#On choisit une E-VALUE
```

```
E_VALUE_THRESH = 0.04
Entrez.email = "glahaie@gmail.com"
path = "annexes/question_2/"

path_fichier = path + "blast_contig_"+sys.argv[1] + ".xml"

with open(path_fichier) as fichier:
    blast_record = NCBIXML.read(fichier)
    i = 0
    path_result = path + "contig_"+sys.argv[1]+"/"
    if not os.path.exists(path_result):
        os.makedirs(path_result)
    for alignment in blast_record.alignments:
        for hsp in alignment.hsps:
            if hsp.expect < E_VALUE_THRESH:
#On obtient alors le fichier genbank
                handle = Entrez.efetch(db="nucleotide", rettype="gb",
                  retmode="text", id=alignment.accession,
                  seq_start=hsp.sbjct_start, seq_stop=hsp.sbjct_end)
                seq_record= SeqIO.read(handle, "gb")
                handle.close()
                nom_fichier = path_result + alignment.accession + ".gb"
                SeqIO.write(seq_record, nom_fichier, "gb")

        i += 1
        if i > 10:
            break
```

5 Log de l'exécution de clustalw2.

```
 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: lcl|XM_518463.3_cdsid_XP_518463.2        2058 bp
Sequence 2: lcl|XM_003833312.1_cdsid_XP_003833360.1  2004 bp
Sequence 3: lcl|XM_004043991.1_cdsid_XP_004044039.1  2004 bp
Sequence 4: lcl|XM_002816867.2_cdsid_XP_002816913.1  2043 bp
Sequence 5: lcl|XM_003266293.1_cdsid_XP_003266341.1  2004 bp
Sequence 6: lcl|XM_005553053.1_cdsid_XP_005553110.1  2004 bp
Sequence 7: lcl|NM_001266091.1_cdsid_NP_001253020.1  2043 bp
Sequence 8: lcl|XM_003922988.1_cdsid_XP_003923037.1  2004 bp
Start of Pairwise alignments
Aligning...

Sequences (1:2) Aligned. Score:  98
Sequences (1:3) Aligned. Score:  97
Sequences (1:4) Aligned. Score:  97
Sequences (1:5) Aligned. Score:  97
Sequences (1:6) Aligned. Score:  96
Sequences (1:7) Aligned. Score:  96
Sequences (1:8) Aligned. Score:  96
Sequences (2:3) Aligned. Score:  99
Sequences (2:4) Aligned. Score:  98
Sequences (2:5) Aligned. Score:  98
Sequences (2:6) Aligned. Score:  97
Sequences (2:7) Aligned. Score:  97
Sequences (2:8) Aligned. Score:  97
Sequences (3:4) Aligned. Score:  98
Sequences (3:5) Aligned. Score:  98
Sequences (3:6) Aligned. Score:  98
Sequences (3:7) Aligned. Score:  97
Sequences (3:8) Aligned. Score:  96
Sequences (4:5) Aligned. Score:  98
Sequences (4:6) Aligned. Score:  98
Sequences (4:7) Aligned. Score:  98
Sequences (4:8) Aligned. Score:  97
Sequences (5:6) Aligned. Score:  98
Sequences (5:7) Aligned. Score:  98
Sequences (5:8) Aligned. Score:  96
Sequences (6:7) Aligned. Score:  99
Sequences (6:8) Aligned. Score:  96
Sequences (7:8) Aligned. Score:  96
Guide tree file created:   [foxp4_ortho.dnd]

There are 7 groups
Start of Multiple Alignment

Aligning...
Group 1: Sequences:   2      Score:37737
Group 2: Sequences:   2      Score:37091
Group 3: Sequences:   3      Score:37148
Group 4: Sequences:   4      Score:37047
Group 5: Sequences:   5      Score:37481
Group 6: Sequences:   7      Score:37220
Group 7: Sequences:   8      Score:36779
Alignment Score 440506

CLUSTAL-Alignment file created  [foxp4_ortho.aln]
```

6 Log du travail de RepeatMasker sur le fichier foxp4_ortho.fa.

```
There were no repetitive sequences detected in /usr/local/rmserver/tmp/RM2_foxp4_ortho.fa_1383262617
```

7 Log de l'exécution de Mavid sur foxp4_ortho.fa.

```
./utils/randtree/randtree foxp4_ortho.fa
./mavid ./mavid.ph foxp4_ortho.fa

*******************************************************
*                                                     *
*                  Welcome to MAVID.                  *
*                  (version 2.0, build 4)             *
*                                                     *
*******************************************************


Aligning 1 versus 1
Aligning [0,2003] to [0,2057]
Aligning 1 versus 2
Aligning [0,2003] to [0,2058]
Aligning 1 versus 1
Aligning [0,2003] to [0,2042]
Aligning 1 versus 1
Aligning [0,2003] to [0,2003]
Aligning 1 versus 2
Aligning [0,2042] to [0,2003]
Aligning 2 versus 3
Aligning [0,2042] to [0,2042]
Aligning 3 versus 5
Aligning [0,2058] to [0,2042]
MAVID worked!

clustalw2 ./mavid.mfa -tree



 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: lcl|XM_004043991.1_cdsid_XP_004044039.1  2059 bp
Sequence 2: lcl|XM_005553053.1_cdsid_XP_005553110.1  2059 bp
Sequence 3: lcl|XM_518463.3_cdsid_XP_518463.2        2059 bp
Sequence 4: lcl|XM_003266293.1_cdsid_XP_003266341.1  2059 bp
Sequence 5: lcl|NM_001266091.1_cdsid_NP_001253020.1  2059 bp
Sequence 6: lcl|XM_002816867.2_cdsid_XP_002816913.1  2059 bp
Sequence 7: lcl|XM_003833312.1_cdsid_XP_003833360.1  2059 bp
Sequence 8: lcl|XM_003922988.1_cdsid_XP_003923037.1  2059 bp

Phylogenetic tree file created:   [./mavid.ph]

../utils/root_tree/root_tree ./mavid.ph
./mavid ./mavid.ph foxp4_ortho.fa

*******************************************************
*                                                     *
*                  Welcome to MAVID.                  *
*                  (version 2.0, build 4)             *
*                                                     *
*******************************************************


Aligning 1 versus 1
Aligning [0,2003] to [0,2042]
Aligning 1 versus 1
Aligning [0,2057] to [0,2003]
Aligning 1 versus 2
Aligning [0,2003] to [0,2003]
Aligning 3 versus 1
Aligning [0,2003] to [0,2042]
Aligning 1 versus 4
Aligning [0,2003] to [0,2042]
Aligning 2 versus 5
Aligning [0,2042] to [0,2042]
Aligning 1 versus 7
Aligning [0,2003] to [0,2042]
MAVID worked!
```

```
clustalw2 ./mavid.mfa -tree



 CLUSTAL 2.1 Multiple Sequence Alignments


Sequence format is Pearson
Sequence 1: lcl|XM_003922988.1_cdsid_XP_003923037.1  2098 bp
Sequence 2: lcl|XM_005553053.1_cdsid_XP_005553110.1  2098 bp
Sequence 3: lcl|NM_001266091.1_cdsid_NP_001253020.1  2098 bp
Sequence 4: lcl|XM_003266293.1_cdsid_XP_003266341.1  2098 bp
Sequence 5: lcl|XM_004043991.1_cdsid_XP_004044039.1  2098 bp
Sequence 6: lcl|XM_518463.3_cdsid_XP_518463.2        2098 bp
Sequence 7: lcl|XM_003833312.1_cdsid_XP_003833360.1  2098 bp
Sequence 8: lcl|XM_002816867.2_cdsid_XP_002816913.1  2098 bp

Phylogenetic tree file created:   [./mavid.ph]

../utils/root_tree/root_tree ./mavid.ph
```

8 Script biopython pour identifier la composition du vecteur pANNE

```
# *- coding:utf-8 -* #

# Script pour le num ro 3 du devoir 1: Cette partie ne fait
#qu'envoyer la requ te blast au serveur du NCBI, et ensuite
#enregistre le r sultat dans un fichier.

from Bio.Blast import NCBIWWW
from Bio.Blast import NCBIXML

path_fichier = ”annexes/question_3/”
nom_resultat = ”blast_fichier”
LEN_THRESH = 100
E_THRESH = 1e-50
# Tout d'abord on ouvre le fichier

sequence = ””
with open(path_fichier+”pANNE.txt”, ’r’) as f:
    for line in f:
        sequence = sequence + line.strip()

#On enl ve les retour de chariot du fichier
i = 1
while len(sequence) > LEN_THRESH:

#Maintenant, on fait le blast
    print ”i␣=␣” + str(i)
    print ”on␣fait␣un␣blast␣sur␣la␣s quence␣de␣longeur␣”
      + str(len(sequence))
    result_handle = NCBIWWW.qblast(”blastn”, ”nr”,
      sequence, megablast=True)

#on enregistre le r sultat
    save_file = open(path_fichier+nom_resultat+str(i)+”.xml”, ”w”)
    save_file.write(result_handle.read())
    save_file.close()
    result_handle.close()

    list_start = []
    list_end = []
    sequences = []
#Maintenant on enl ve de la s quence les zones identifi es
    with open(path_fichier+nom_resultat+str(i)+".xml", "r") as result:
        blast_record = NCBIXML.read(result)
        alignment = blast_record.alignments[0]
        for alignment in blast_record.alignments:
            for hsp in alignment.hsps:
#On met   jour la s quence pour enlever ce r sultat
                if hsp.expect < E_THRESH:
                    list_start.append(hsp.query_start)
                    list_end.append(hsp.query_end)

            break
#On a les points   enlever
#sort sur les listes
        list_start.sort()
        list_end.sort()
        start= -1
        end = -1
        for s_start, s_end in zip(list_start, list_end):
            if end < 0:
                sequences.append(sequence[: s_start -1])
                end = s_start-1
            else:
                end = s_start-1
                sequences.append(sequence[start: end])
            start = s_end -1
        sequences.append(sequence[start:])
        sequence = ""
        for fragment in sequences:
            sequence +=fragment
#Pour v rifier les r sutats, j'enregistre la nouvelle
#s quence dans un fichier
        print "On␣ crit ␣le␣reste␣de␣la␣s quence␣avec␣i␣=␣" + str(i)
        with open(path_fichier+"pANNE"+str(i)+".txt", "w") as f:
            f.write(sequence)
    i +=1
```

9 Temps d’exécution des alignements multiples

Programme	Temps d’exécution
ClustalW	real 0m4.840s
dialign	real 0m18.424s
Mavid	real 0m0.296s

10 Script Biopython pour obtenir les fichiers d’accession des 10 premiers résultats du blast pour un contig donné.

```python
#- coding:utf-8 -#

from reportlab.lib import colors
from reportlab.lib.units import cm
from Bio.Graphics import GenomeDiagram
from Bio.Graphics.GenomeDiagram import CrossLink
from reportlab.lib import colors

gd_diagram = GenomeDiagram.Diagram("Composition␣du␣vecteur␣pANNE.txt")
gd_track_for_features = gd_diagram.new_track(1, name="Annotated␣Features",
  start=0, end=6627)
gd_feature_set = gd_track_for_features.new_set()
from Bio.SeqFeature import SeqFeature, FeatureLocation

colors = [colors.green,colors.lightgreen, colors.teal, colors.darkgreen,
  colors.seagreen,colors.lawngreen,colors.olivedrab]

#Essai  la main,  la lecture des fichiers XML
#blast #1: pHT2
feature = SeqFeature(FeatureLocation(1, 1056), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[0],
  sigil="ARROW", arrowhead_length=0.5,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(5342, 5798), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[1],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(3495, 5341), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[2],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(5798, 6627), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[3],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(3039, 3494), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[4],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(1742, 2057), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[5],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(2776, 3039), strand = +1)
gd_feature_set.add_feature(feature, name="pHT2", label=True, color=colors[6],
  sigil="ARROW", arrowhead_length=1,arrowshaft_height=0.1)

#blast 2:
feature = SeqFeature(FeatureLocation(1057, 1741), strand = +1)
gd_feature_set.add_feature(feature, name="PGeneClip", label=True, color="blue",
  sigil="ARROW", arrowhead_length=1)

#blast 3
feature = SeqFeature(FeatureLocation(2058, 2770), strand = +1)
gd_feature_set.add_feature(feature, name="Cloning␣vector␣EN.Cherry", label=True,
  color="red", sigil="ARROW",arrowhead_length=1)

#On essai d'ajouter d'autres track pour repr senter la position des blasts
gd_track_for_features = gd_diagram.new_track(1, name="pHT2", start=0, end=4924)
gd_feature_set = gd_track_for_features.new_set()
feature = SeqFeature(FeatureLocation(2246, 4092), strand = None)
gd_feature_set.add_feature(feature, name="pHT2", label=False, color=colors[2],
  sigil="ARROW", arrowhead_length=0.2,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(1, 1058), strand = None)
```

```
gd_feature_set.add_feature(feature, name=”pHT2”, label=False, color=colors[0],
  sigil=”ARROW”, arrowhead_length=0.2,arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(4093, 4922), strand = +1)
gd_feature_set.add_feature(feature, name=”pHT2”, label=False, color=colors[3],
  sigil=”ARROW”, arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(2795, 2339), strand = -1)
gd_feature_set.add_feature(feature, name=”pHT2”, label=False, color=colors[1],
  sigil=”ARROW”, arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(2340, 2795), strand = +1)
gd_feature_set.add_feature(feature, name=”pHT2”, label=False, color=colors[4],
  sigil=”ARROW”, arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(743, 1058), strand = +1)
gd_feature_set.add_feature(feature, name=”pHT2”, label=False, color=colors[5],
  sigil=”ARROW”, arrowshaft_height=0.1)
feature = SeqFeature(FeatureLocation(1984, 2246), strand = +1)
gd_feature_set.add_feature(feature, name=”pHT2”, label=False,
  color=colors[6],sigil=”ARROW”,arrowshaft_height=0.1)

#On essai d'ajouter d'autres track pour repr senter la position des blasts
gd_track_for_features = gd_diagram.new_track(1, name=”PGeneClip”, start=0,
  end=5267)
gd_feature_set = gd_track_for_features.new_set()
feature = SeqFeature(FeatureLocation(1879, 2563), strand = +1)
gd_feature_set.add_feature(feature, name=”PGeneClip”, label=False, color=”blue”,
  sigil=”ARROW”, arrowhead_length=1,arrowshaft_height=0.1)

gd_track_for_features = gd_diagram.new_track(1,
  name=”Cloning_vector_EN.Cherry”, start=0, end=10649)
gd_feature_set = gd_track_for_features.new_set()
feature = SeqFeature(FeatureLocation(7102, 7813), strand = +1)
gd_feature_set.add_feature(feature, name=”Cloning_Vector_EN.Cherry”, label=False,
  color=”red”, sigil=”ARROW”, arrowhead_length=1,arrowshaft_height=0.1)

gd_diagram.draw(format='linear', pagesize=”LETTER”, orientation=”portrait”,
  fragments=1, start=0, end=10649)
gd_diagram.write(”GD_labels_default.eps”, ”eps”)
```

11 Fichiers genbank utilisés pour ce rapport

Nom	Numéro d'accession	Nom	Numéro d'accession
Cloning Vector EN.Cherry, complete sequence	HM771696.1	PGeneClip hMGFP Vector, complete sequence	AY744386.1
Expression vector pHT2, complete sequence	AY773970.1	Homo sapiens chromosome 6, GRCh37.p13 Primary Assembly	NC_000006.11
SARS coronavirus MA15 ExoN1 isolate d3om5, complete genome	JF292906.1	SARS coronavirus MA15 isolate d2ym4, complete genome	JF292909.1
SARS coronavirus MA15 isolate d4ym5, complete genome	JF292915.1	SARS coronavirus HKU-39849 isolate recSARS-CoV HKU-39849, complete genome	JN854286 .1
SARS coronavirus HKU-39849 isolate UOB, complete genome	JQ316196.1	SARS coronavirus isolate Tor2/FP1-10912, complete genome	JX163923.1
SARS coronavirus isolate Tor2/FP1-10851, complete genome	JX163924.1	SARS coronavirus isolate Tor2/FP1-10895, complete genome	JX163925.1

Nom	Numéro d'accession	Nom	Numéro d'accession
SARS coronavirus isolate Tor2/FP1-10912, complete genome	JX163926.1	SARS coronavirus isolate Tor2/FP1-10851, complete genome	JX163927.1
SARS coronavirus isolate Tor2/FP1-10895, complete genome	JX163928.1	SARS coronavirus SinP3, complete genome	AY559090.1
SARS coronavirus HKU-39849 isolate TCVSP-HARROD-00001, complete genome	GU553363.1	SARS coronavirus HKU-39849 isolate recSARS-CoV HKU-39849, complete genome	JN854286.1
SARS coronavirus HKU-39849 isolate TCVSP-HARROD-00002, complete genome	GU553364.1	SARS coronavirus HKU-39849 isolate TCVSP-HARROD-00003, complete genome	GU553365.1
SARS coronavirus Sin850, complete genome	AY559096.1	SARS coronavirus MA15 isolate P3pp3, complete genome	FJ882948.1
SARS coronavirus MA15 ExoN1 isolate P3pp3, complete genome	FJ882951.1	SARS coronavirus MA15 isolate P3pp4, complete genome	FJ882952.1
SARS coronavirus MA15, complete genome	FJ882957.1	SARS coronavirus MA15 isolate P3pp7, complete genome	FJ882958.1
SARS coronavirus MA15 ExoN1 isolate P3pp6, complete genome	FJ882959.1	SARS coronavirus MA15 isolate P3pp5, complete genome	FJ882961.1
SARS coronavirus ExoN1 isolate c5P1, complete genome	JF292922.1	SARS coronavirus ExoN1 isolate c5P10, complete genome	JX162087.1
SARS coronavirus ExoN1 strain	KF514407.1	PREDICTED : Pan troglodytes forkhead box P4, transcript variant 2 (FOXP4), mRNA	XM_518463.3
PREDICTED : Pan paniscus forkhead box P4, transcript variant 2 (FOXP4), mRNA.	XM_003833312.1	PPREDICTED : Gorilla gorilla gorilla forkhead box P4, transcript variant 2 (FOXP4), mRNA.	XM_004043991.1
PREDICTED : Pongo abelii forkhead box P4, transcript variant 1 (FOXP4), mRNA.	XM_002816867.2	PREDICTED : Nomascus leucogenys forkhead box P4, transcript variant 2 (FOXP4), mRNA.	XM_003266293.1
PREDICTED : Macaca fascicularis forkhead box P4 (FOXP4), transcript variant X3, mRNA.	XM_005553053.1	Macaca mulatta forkhead box P4 (FOXP4), mRNA.	NM_001266091.1
PREDICTED : Saimiri boliviensis boliviensis forkhead box P4, transcript variant 2 (FOXP4), mRNA	XM_003922988.1	Homo sapiens chromosome 2, GRCh37.p13 Primary Assembly	NC_000002.11
Homo sapiens amyotrophic lateral sclerosis 2 (juvenile) (ALS2), transcript variant 1, mRNA	NM_020919.3	Homo sapiens amyotrophic lateral sclerosis 2 (juvenile) (ALS2), transcript variant 2, mRNA	NM_001135745.1
Pan troglodytes chromosome 2B, Pan_troglodytes-2.1.4	NC_006470.3	Macaca mulatta chromosome 12, Mmul_051212, whole genome shotgun sequence	NC_007869.1
Canis lupus familiaris breed boxer chromosome 37, CanFam3.1, whole genome shotgun sequence	NC_006619.3	Bos taurus breed Hereford chromosome 2, Bos_taurus_UMD_3.1, whole genome shotgun sequence	AC_000159.1
Mus musculus strain C57BL/6J chromosome 1, GRCm38.p1 C57BL/6J	NC_000067.6	Rattus norvegicus strain BN/SsNHsdMCW chromosome 9, Rnor_5.0	NC_005108.3
Gallus gallus isolate #256 breed Red Jungle fowl, inbred line UCD001 chromosome 7, Gallus_gallus-4.0, whole genome shotgun sequence	NC_006094.3	Danio rerio strain Tuebingen chromosome 6, Zv9	NC_007117.5

Nom	Numéro d'accession	Nom	Numéro d'accession
Homo sapiens chromosome 2 genomic contig, GRCh37.p13 Primary Assembly	NT_005403.17	alsin isoform 1 [Homo sapiens]	NP_065970.2
alsin [Pan troglodytes]	NP_001073389.1	forkhead box protein P4 isoform 1 [Homo sapiens]	NP_001012426.1

12 Fichiers de gène de NCBI utilisé pour ce rapport

Nom	Gene ID	Nom	Gene ID
ALS2 amyotrophic lateral sclerosis 2 (juvenile) [Homo sapiens (human)]	57679	ALS2 amyotrophic lateral sclerosis 2 (juvenile) [Pan troglodytes (chimpanzee)]	470613
ALS2 amyotrophic lateral sclerosis 2 (juvenile) [Macaca mulatta (Rhesus monkey)]	703263	ALS2 amyotrophic lateral sclerosis 2 (juvenile) [Canis lupus familiaris (dog)]	100856109
ALS2 amyotrophic lateral sclerosis 2 (juvenile) [Bos taurus (cattle)]	535750	Als2 amyotrophic lateral sclerosis 2 (juvenile) [Mus musculus (house mouse)]	74018
Als2 amyotrophic lateral sclerosis 2 (juvenile) [Rattus norvegicus (Norway rat)]	363235	FOXP4 forkhead box P4 [Homo sapiens (human)]	116113

Références

[1] Triticum aestivum (ID 11) - Genome - NCBI (2013). Retrieved December 17, 2013 from http ://www.ncbi.nlm.nih.gov/genome/11.

[2] Ling HQ, Zhao S, Liu D, Wang J, Sun H, Zhang C, Fan H, Li D, Dong L, Tao Y, et al. Draft genome of the wheat A-genome progenitor Triticum urartu. Nature. 2013 Apr 4 ;496(7443) :87-90. doi : 10.1038/nature11997. Epub 2013 Mar 24. PubMed PMID : 23535596.

[3] Whole Chromosome Survey Sequencing (2013). Retrieved December 17, 2013 from http ://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat- Projects/Sequencing/Whole-Chromosome-Survey-Sequencing

[4] Sequencing Projects (2013). Retrieved December 17, 2013 from http ://www.wheatgenome.org/Projects/IWGSC-Bread-Wheat-Projects/Sequencing

[5] Wilkinson, P.A., Winfield, M.O., Barker, G.L.A., Allen, A.M., Burridge, A, Coghill, J.A., Burridge, A. and Edwards, K.J. 2012. CerealsDB 2.0 : an integrated resource for plant breeders and scientists. BMC Bioinformatics 13 : 219.

[6] GC content. In Wikipedia. Retrieved December 17, 2013, from `http://en.wikipedia.org/wiki/GC-content`

[7] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res.* 2002 Jun ;12(6) :996-1006.

[8] Basic Local Alignment Search Tool (Altschul et al., J Mol Biol 215 :403-410 ; 1990).