



Projet AGENAE

Lettre d'information de SIGENAE



N°7

du 1^{er} juillet 2004 – Rédacteur : CK

Cette lettre d'information est destinée aux membres des équipes de recherche participant aux projets AGENAE (Analyse des GENômes des Animaux d'Elevage). Elle a pour but de vous informer sur les évolutions de l'équipe SIGENAE, les nouveaux outils et services mis en place.

L'équipe SIGENAE : Composition :

<i>Prénom Nom</i>	<i>Localisation</i>	<i>Tel</i>	<i>Email</i>
Christelle Dantec	Toulouse	05 61 28 51 19	christelle.dantec@toulouse.inra.fr
Elodie Retout	Rennes	02 23 48 54 61	elodie.retout@irisa.fr
Patrice Dehais	Toulouse	05 61 28 54 44	patrice.dehais@toulouse.inra.fr
Eddie Iannucelli	Toulouse	05 61 28 54 44	eddie.iannucelli@toulouse.inra.fr
Christophe Klopp	Toulouse	05 61 28 50 36	christophe.klopp@toulouse.inra.fr
Cédric Cabau	Jouy-en-Josas	01 34 65 28 83	Cedric.Cabau@jouy.inra.fr

1/ Contamination des EST AGENAE et mauvaise annotation des contigs :

• Contamination des EST AGENAE

Suite à l'annotation manuelle des contigs truie (assemblage V3, mars 2003) et au choix des clones pour la puce à ADN générique, des biologistes du SCRIBE nous ont alarmé sur la présence de contigs de séquences d'E. Coli parmi les contigs Sigenae. Après analyse, nous nous sommes aperçu que la chaîne de nettoyage des séquences n'extrayait pas les annotations trouvées, par le logiciel CROSSMATCH, sur le brin moins.

Certaines des séquences contaminées appartiennent aux banques AGENAE et ont été publiées.

Tableau du nombre de séquences contaminées par espèce :

<i>Espèce</i>	Bovin	Porc	Poule	Truite
Nombre de séquences des banques AGENAE	474	123	396	312

Les travaux faits ou à faire sont :

1. Modifier le programme fautif de la chaîne = fait,
2. Retraiter les séquences pour rechercher les séquences contaminées = fait,
3. Modifier les publication = à faire (prévu suite à une discussion avec les personnes ressource),
4. Renommer les contigs du dernier assemblage pour visualiser les contigs contaminés (à faire),
5. Refaire des assemblages en supprimant les séquences contaminées (à faire).

Dans la phase de choix des clones pour les puces à ADN génériques, les clones ayant des séquences contaminées ont été systématiquement supprimées.

Les nouveaux assemblages seront réalisés une fois que le nouveau système de nettoyage de séquences (SURF, cf. point 2) sera mis en production.

• Mauvaise annotation des contigs

Suite au choix des clones pour la puce à ADN, des biologistes du LGC nous ont indiqué ne pas trouver certaines protéines (hits SWISSPROT) dans les listes fournies. Après analyse, ceci provient d'une erreur dans l'extraction de la liste des hits pertinents (e-value < 1e-5) pour un contig. Le logiciel extrayait les 10 derniers hits pertinents au lieu des 10 premiers. Ce problème n'est apparu que pour les versions d'assemblages suivantes : truie - version 4, porc - version 5, bovin - version 2.

La mise à jour des bases est faite pour les porcins et en cours pour les bovins et les truites.

2/ Présentation du nouvel environnement SIGENAE :

L'équipe SIGENAE travaille à la mise en place des nouveaux outils de traitement, de visualisation de séquences et de gestion du site web. Ces outils seront sous peu assez avancés pour que nous vous les présentions et que vous puissiez les tester.

Ces outils sont :

- **SURF** (SeqUence Repository and Feature detection) : nouveau système de nettoyage des séquences. Les évolutions majeures viennent de l'intégration de parties faites précédemment dans différents logiciels (qualité, GPI et publication). Toutes les séquences présentes dans les assemblages seront traitées de la même manière et pourront être visualisées par les mêmes interfaces.
- **Ensembl** : nouveau système d'interrogation des assemblages, de visualisation des contigs et de leurs annotations. Les nouveautés viennent ici de l'augmentation du nombre et du type d'interrogations possibles ainsi que de l'interface de visualisation graphique des assemblages et des annotations.
- **SMAT** (Sigenae Manual Annotation Tool) : nouveau système d'annotation manuelle. Adossé à Ensembl il permettra aux biologistes d'ajouter leurs annotations à celles obtenues par similarité.
- **Typo3** : nouveau logiciel de gestion de contenu. Typo3 permet de créer des sites web propres à des projets et de déléguer l'administration de ces sites aux responsables des projets.

Trois réunions, d'une matinée chacune (10h-13h), sont prévues pour vous présenter ces outils et discuter des évolutions possibles. Pensez à vous inscrire en écrivant à sigenasupport@jouy.inra.fr.

Les lieux et dates sont :

<i>Lieu</i>	<i>Date</i>
BIA Toulouse	Jeudi 16 septembre 2004
SCRIBE Rennes	Jeudi 7 octobre 2004
SGQA Jouy-en-Josas	Jeudi 30 septembre 2004

3/ Listes des clones pour les puces génériques truie, bovine et porcine :

- La liste des clones truies a été transmise au CRB le 26 avril 2004.
- La liste des clones bovins a été transmise au CRB le 4 juin 2004.
- La sélection des clones porcins est en cours. Les outils de choix des clones sont disponibles sur le site web SIGENAE (accès via la page d'accueil).

4/ Évolutions de l'infrastructure matérielle SIGENAE :

De manière à assurer de bonnes performances d'interrogation et de traitement sur BASE, 3 processeurs de topaze sur 8 ont été affectés au moteur du système de gestion de base de données Postgresql (pgagena). Merci de nous informer de toute lenteur de traitement constatée : sigenasupport@jouy.inra.fr.

5/ Enquête en cours: quel identifiant pour les séquences AGENAE publiées?

Suite aux demandes de modification d'identifiants des séquences faites par certains biologistes, nous avons mis sur le site web un questionnaire à ce propos. Les résultats actuels sont :

<i>Identifiant à utiliser</i>	<i>Nombre de réponses</i>
AGENAE ID : tcay0034b.d.11_5.1	4
GenBank Acc: BX317614	5

Merci à ceux d'entre vous qui n'ont pas encore répondu de le faire. Le lien vers le questionnaire se trouve dans la partie « les nouvelles » de la page d'accueil du site web.

6/ Liste d'information :

L'équipe SIGENAE a mis en place un alias (sigenae_user@jouy.inra.fr). Il est utilisé pour diffuser des informations relatives aux ruptures d'accès au site web dues soit à des travaux de maintenance planifiée soit à des problèmes sur le serveur ou le réseau. Envoyez un mail à sigenasupport@jouy.inra.fr pour vous inscrire.

7/ Assemblages faits dans la période écoulée ou prévus :

- Poule : un nouvel assemblage a été réalisé (473 846 séquences, 107 103 contigs) en juin 2004.
- Truite : un nouvel assemblage est prévu dans les semaines à venir, une fois SURF en production.

Les résultats de ces assemblages vous seront mis à disposition en septembre 2004 via Ensembl.