Recherche de gènes *de novo* dans le génome d'eucaryotes unicellulaires : les levures

Contact : ingrid . lafontaine @ sorbonne – universite . fr

Depuis l'obtention de génomes complètement séquencés, dès 1996, des phases codant pour des protéines, ne présentant aucune similarité avec des gènes déjà connus sont annotés dans les génomes. Ces gènes peuvent avoir des homologues qui n'ont pas été détectés soit parce qu'ils appartiennent à des génomes qui n'ont pas encore été séquencés, soit parce que leurs séquences ont trop divergé pour que la similarité soit encore détectable. Il est aussi possible qu'ils correspondent à des erreurs d'annotation, *i.e.* que ces phases ouvertes de lectures ne correspondent pas à des gènes codant pour des protéines. Enfin, si toutes ces hypothèses sont écartées, certains de ces gènes pourraient correspondre à des gènes nouvellement créés à partir de séquences non-codantes (gènes *de novo*). Depuis une dizaine d'années, plusieurs cas ont ainsi été identifiés chez les eucaryotes, de la levure à l'homme.

On définit comme gènes « taxonomiquement restreints » (Taxonomically Restricted Gene, TRG), les gènes présents dans un ou plusieurs génomes d'espèces proches. Une sous-classe de ces TRGs est appelée ORFan (proposé par Bernard Dujon en 1996) et correspond à ceux présents dans un seul génome.

Vous travaillerez sur les gènes de 5 levures : *Saccharomyces cerevisiae*, *S. mikatae* (IFO 1815 ^{T}), *S. bayanus* var. *uvarum* (CBS 7001), *S. kudriavzevii (ZP 591)* et *S. paradoxus*. Un événement de duplication totale du génome a eu lieu dans un ancêtre commun de toutes ces levures (cf. arbre http://www.saccharomycessensustricto.org/cgi-bin/s3.cgi).

1. Dresser la liste des TRGs présents dans les génomes des 5 levures étudiés.
   Pour sélectionner les TRGs candidats, établir des critères de « non-similarité ». Les outils de recherche de similarité comme BLAST, PSI-BLAST, RPS-BLAST, CD-Search, PFAM, etc. pourront être utilisés. Vous pouvez également (mais ce n'st pas obligatoire) utiliser des données transcriptomiques et protéomiques disponibles pour affiner vos critères de sélection.
2. Essayer de caractériser les TRGs, à l'aide de méthodes ne reposant pas sur la similarité des séquences. Par exemple, vous pouvez analyser la composition nucléotidique, des biais d'usage, utiliser des outils de prédiction de structure 2D/3D, etc.

Les fichiers contenant les gènes et les protéines correspondantes contenus dans ces génomes vont sont fournis (avec la bibliographie indiquée ci-dessous)

Vous pourrez les retrouver à cette adresse :
http://www.saccharomycessensustricto.org/cgi-bin/s3.cgi?data=Annotations&version=current.
Vous trouverez également dans la colonne *« Ohnolog Pairs »* (extension « .ohnologs ») des fichiers recensant, sur chaque ligne, les gènes homologues issus de la duplication totale du génome qui a eu lieu dans l'ancêtre commun de toutes les levures étudiées dans ce projet.

Quelques références bibliographiques (contactez-moi si vous en voulez plus):

1. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I. 2018. A Molecular Portrait of De Novo Genes in Yeasts. *Mol Biol Evol* **35**: 631–645.
2. McLysaght A, Hurst LD. 2016. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet 17: 567–578.
3. Bornberg-Bauer E, Schmitz J, Heberlein M. 2015. Emergence of de novo proteins from "dark genomic matter" by "grow slow and moult." *Biochemical Society Transactions* **43**: 867–873.
4. Schlötterer C. 2015. Genes from scratch – the evolutionary fate of de novo genes. *Trends in Genetics* **31**: 215–219.
5. Light S, Basile W, Elofsson A. 2014. Orphans and new gene origination, a structural and evolutionary perspective. *Current Opinion in Structural Biology* **26**: 73–83.