Who has the largest (genome)?

Contacts

Annabelle Haudry, maître de conférence Université Lyon 1
annabelle.haudry@univ-lyon1.fr
Tel. +33 (0) 472 432 918

Laurent Duret, directeur de recherche CNRS
laurent.duret@univ-lyon1.fr
Tel. +33 (0) 472 431 388

Simon Penel, ingénieur d'étude CNRS
simon.penel@univ-lyon1.fr

Lab

Laboratoire de Biométrie et Biologie Evolutive -Université Lyon 1 - UMR CNRS 5558
43 bd du 11 novembre 1918
69622 Villeurbanne cedex FRANCE

Project

Context

Genome size can vary up to 200,000 fold among eukaryotes genomes. Such variations can not be fully explained by changes in the level of "complexity" of species (estimated via the number of genes), as it was originally suggested[1]. Sequencing projects of the past two decades have revealed that the differences between genomes were mainly due to differences in their content in repeated DNA. Transposable elements (TEs) are mobile segments of DNA capable of being cut or copied in the genome, that make them potentially highly repeated. TEs represent a very large proportion of Eukaryotic genomes in general (~ 50% of the human genome) and their content in the genome is positively correlated with the genome size[2].

The number of available complete genomes has dramatically increased in the past few years. It is now possible to quantify more precisely the TE content (based on sequencing data) and to calculate some genomic features for a large number of species. A large database aiming to centralize available molecular information on sequenced prokaryotic and eukaryotic species (number of genes, the number of proteins, the content of ET, GC%) and estimates the genome size (through the assembly and C-value) is being developed in the laboratory (LBBE, Lyon1).

Objectives

In this context, we propose to use a *de novo* approach to identify repeats in genomes sequenced with high-throughput methods sequencing (Illumina or 454). This approach, independent of the existing description of TEs for a given species, will complement the current estimate made from a database of TEs. The ultimate goal of the database is to compile a dynamic list of all species sequenced respectively with the size of their genome, the genomics fraction corresponding to TEs, or to genes and other genomic

features or life history traits in order to better understand the factors influencing the variation of genome size. During the internship, the student will analyze genomes to identify *de novo* Tes and feed the database with ecological characteristics of species found in the literature. With all the collected the data the variation of sizes of the genomes could be analyzed in regards with molecular and ecological traits of species.

Skills

This project implies using bioinformatics tools to analyze Next Generation Sequencing whole-genome data. The knowledge of some programming languages will be required for file manipulation (perl, python or C), database updating (MySQL/PostGreSQL), and statistics analysis (R).

1. Mirsky, A. E. & Ris, H. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J. Gen. Physiol.* 451–462 (1951).
2. Biémont, C. & Vieira, C. Junk DNA as an evolutionary force. *Nature* **443,** 521–524 (2006).