Offre de stage de M2 recherche

Evolution de l'architecture des génomes à l'échelle de l'arbre de la vie

Annabelle Haudry (département GINSENG),

Damien M. de Vienne et **Laurent Duret** (département PEGASE),

Simon Penel (pôle informatique)

Contexte. La taille des génomes peut varier jusqu'à 200 000 fois chez les Eucaryotes, de 2,9Mb pour une microsporidie parasite intracellulaire des Vertébrés jusqu'à 670Gb pour le gigantesque génome d'une amibe qui est, elle, microscopique. De telles variations ne peuvent pas être expliquées par des variations du degré de « complexité » des espèces (leur nombre de gènes) comme suggéré initialement. Les projets de séquençage ont en fait révélé que les différences de taille observées entre les génomes étaient principalement dues à des différences dans leur teneur en ADN répété. Parmi les éléments répétés des génomes, les éléments transposables (ET) sont des séquences ayant la capacité de se reproduire et se déplacer dans le génome ; les ET représentent une très grande proportion des génomes Eucaryotes en général, et environ 50% du génome humain.

Parallèlement, de nombreux traits liés à l'architecture des génomes varient entre espèces (tels que les taux de GC, le nombre d'exons par gène, la taille des introns, etc) sans que leur dynamique évolutive ne soit encore très bien comprise.

Objectifs. Le but de ce projet est de mieux comprendre les facteurs influençant la variation des tailles de génomes, et ce dans une approche évolutive. Pour aborder cette question, nous proposons de créer une base de données exhaustive de toutes les espèces séquencées contenant des informations génomiques et écologiques les concernant:

- taille du génome
- nombre de chromosomes
- ploïdie
- fraction génomique correspondant à des ET
- fraction génomique correspondant à des gènes
- teneur en GC
- nombre et taille des introns par gène
- nombre et taille des exons par gène

Différentes bases de données, spécifiques de certains groupes d'espèces, existent et contiennent des estimations la quantité d'ADN dans les cellules, basées sur des méthodes biochimiques, indépendantes du séquencage du génome (Gregory et al. 2007). Il est désormais possible d'allier ces estimations de taille de génome avec les caractéristiques génomiques grâce à l'accumulation du nombre de génomes entièrement séquencés, et dont les données sont disponibles publiquement (ensemblgenomes.org). Nous pouvons donc envisager de quantifier plus précisément la teneur en ET et d'autres caractéristiques génomiques pour un grand nombre d'espèces afin d'identifier les facteurs associés aux changements de taille au cours de l'évolution. Les objectifs de ce projet sont les suivants:

1. Développer une pipeline d'analyse qui permette de récolter l'ensemble de ces données automatiquement (soit directement d'après des données publiques soit suite à l'analyse du contenu des génomes ; par ex : nombre de gènes, nombre de protéines, teneur en ET, GC%).
2. Créer une large base de données destinée à centraliser des informations moléculaires disponibles sur les espèces séquencées et à y accéder et les visualiser directement depuis l'arbre de la vie présenté dans **Lifemap** (de Vienne, 2016; http://lifemap.univ-lyon1.fr/).
3. Mener une méta-analyse de l'évolution de l'architecture des génomes au sein de plusieurs groupes taxonomiques en utilisant la phylogénie des espèces (voir Sessegolo et al. 2016; et qui pourra directement être extraite via LifeMap).

Méthodes et moyens

Aujourd'hui, il y a ~20,000 génomes dans la base de données des génomes séquences du LBBE entretenue par Simon Penel. Pour ces génomes, on a les fichiers EMBL d'annotation et Simon a des scripts qui renseignent sur un grand nombre des traits génomiques qui nous intéressent. Il faut (i) rajouter les estimations de la teneur en ET des génomes et (ii) intégrer ces informations à LifeMap. Nous avons envisagé trois approches pour estimer la teneur en éléments répétés dans les génomes :
- annotation à partir de bibliothèque d'éléments connus sur le génome assemblé eu en utilisant RepeatMasker.
- estimation indépendante de la qualité du séquençage, en utilisant l'outil DnaPipeTE (Goubert et al. 2015): une méthode d'identification de novo de séquences répétées pour les génomes séquencés avec des méthodes de séquençage haut-débit (Illumina ou 454).
- DnaPipeTE à partir de reads simulés à partir de l'assemblage, afin de rechercher des éléments répétés non répertoriés dans les bibliothèques d'ET.

Le stagiaire M2 prendra en charge les tâches 1, 2 et 3 en utilisant plusieurs estimateurs complémentaires d'ET sur les génomes, en intégrant à Lifemap (de Vienne 2016) les outils requis et en exploitant la base pour réaliser une méta-analyse. Ce stage implique l'utilisation d'outils bioinformatique d'analyse comparative de génomes, de la programmation (bases de données, langage de script, bash, environnement linux) ainsi que des analyses statistiques.

Références

- Sessegolo C, Burlet N, Haudry A (2016) Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol Lett 12:20160407. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5014035/

- de Vienne DM (2016) Lifemap: Exploring the Entire Tree of Life. PLOS Biology 14(12): e2001624. https://doi.org/10.1371/journal.pbio.2001624

- Goubert C et al. (2015) De novo assembly and annotation of the Asian tiger mosquito (*Aedes albopictus*) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (*Aedes aegypti*). Genome Biol Evol. 2015 Mar 11;7(4):1192-205. https://doi.org/10.1093/gbe/evv050

- Gregory TR et al. (2007) Eukaryotic genome size databases. Nucleic Acids Research, 35(Database issue), D332–D338. http://doi.org/10.1093/nar/gkl828

- RepeatMasker:

http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasetsAlt.html