Epistack : Visualisation de profils épigénétiques

Safia Saci[11*] , Laura Morel[1], Guillaume Devailly[1†]

Résumé

Epistack est un package R (prochainement soumis à Bioconductor) permettant la génération d'une visualisation informative très utilisée en bioinformatique : les piles de profils épigénétiques centrées sur un type de région donnée (promoteurs de gènes, sommets de piques, etc.). Il se veut (un peu) plus simple d'utilisation que les méthodes alternatives existantes, tout en restant suffisamment flexible pour s'adapter à un grand nombre de situations. Une attention particulière a été portée au problème d'*overplotting* lorsqu'un grand nombre de régions (> 10 000) est visualisée en même temps.

Mots-clefs : bioinformatique – génomique – épigénétique – Bioconductor – visualisation

Développement

Les analyses épigénomiques aboutissent souvent à la génération de pistes génomiques sous la forme de scores le long du génome (données de ChIP-seq, ATAC-seq, analyses de la méthylation de l'ADN, etc.). Souvent, les profils moyens sont visualisés autour de points d'intérêts (sommet de piques, promoteurs de gènes, etc.), mais ils masquent la variabilité des données parmi ces régions. En complément de ces profils moyens, des piles de profiles épigénomiques sont représentées sous forme d'heatmaps. Ces Heatmaps peuvent être générés avec divers outils, tels que seqMINER (Ye *et al*, 2011) ou deepTools (Ramírez *et al*, 2016). En R, les packages bioconductor Repitools (Statham *et al*, 2010), seqplots (Stempor & Ahringer, 2016), et surtout EnrichedHeatmap (Gu *et al*, 2018) permettent la réalisation de ce type de plots, qu'il est aussi possible de générer avec des fonctions plus bas niveaux en y mettant plus d'effort.

Nous développons epistack, un package R prochainement soumis à Bioconductor. Ces originalités sont :

- Une méthode originale de limitation du problème d'overplotting (quand il y a plus de régions à empiler que de pixels sur l'image de sortie, par exemple lorsqu'on veut plotter les 40 000 promoteurs annotés du génome humain) en réduisant la taille des matrices au dernier moment avant le plot.
- La réalisation de profils moyens affichant l'erreur standard autour de ces profils moyens pour refléter la variabilité intrinsèque à chaque groupe.
- Une relative simplicité d'utilisation toute en conservant une grande flexibilité d'utilisation.

Plusieurs pistes épigénétiques peuvent être visualisées côte à côte. Les régions peuvent être triées et/ou groupées, par exemple selon le niveau d'expression des gènes, la hauteur des piques ou par clustering. Epistack est disponible sur GitHub : github.com/GenEpi-GenPhySE/epistack

Nous sommes en train d'utiliser epistack pour générer automatiquement un grand nombre de figures pour chacun des ChIP-seq du jeu de données FAANG (data.faang.org, regroupant des données pour l'annotation des génomes d'animaux de rentes).

[1] GenPhySE, Université de Toulouse, INRAE, ENVT, 31326, Castanet Tolosan, France

* safiasaci1995@gmail.com

† guillaume.devailly@inrae.fr

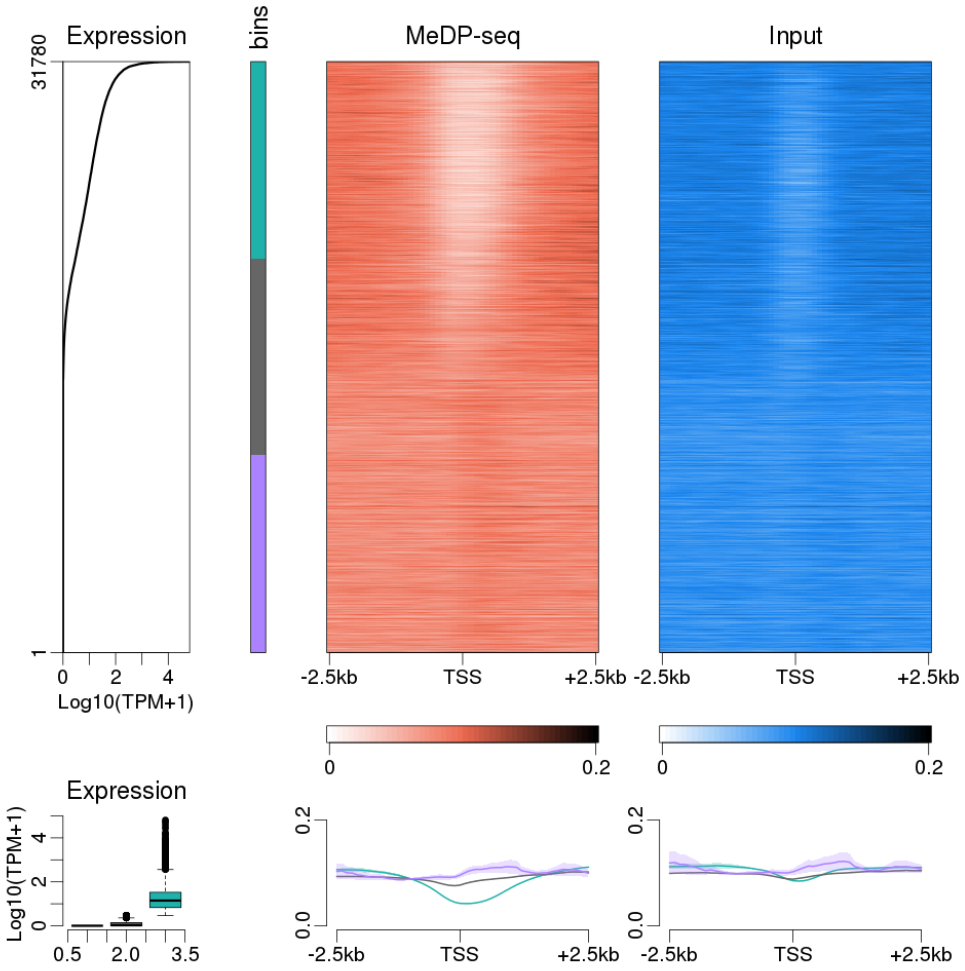

Figure 1 : exemple de rendu avec epistack. Les profils de méthylation de l'ADN des promoteurs de 31780 gènes sont représentés en fonction du niveau d'expression des gènes. Les gènes les plus exprimés ont le promoteur le moins méthylé.

Références

Gu Z, Eils R, Schlesner M & Ishaque N (2018) EnrichedHeatmap: an R/Bioconductor package for comprehensive visualization of genomic signal associations. *BMC Genomics* 19: 234

Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dündar F & Manke T (2016) deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research* 44: W160–W165

Statham AL, Strbenac D, Coolen MW, Stirzaker C, Clark SJ & Robinson MD (2010) Repitools: an R package for the analysis of enrichment-based epigenomic data. *Bioinformatics* 26: 1662–1663

Stempor P & Ahringer J (2016) SeqPlots - Interactive software for exploratory data analyses, pattern discovery and visualization in genomics. *Wellcome Open Res* 1: 14

Ye T, Krebs AR, Choukrallah M-A, Keime C, Plewniak F, Davidson I & Tora L (2011) seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res* 39: e35