GeneNeighborhood: an R package to explore the direct neighbors of your favorite gene set

Pascal GP MARTIN[1]

[1] INRA ToxAlim, UMR1331 INRA, 180 ch de Tournefeuille, 31027, Toulouse, Cedex 3, France

Pascal.Martin@inra.fr

1 Introduction

It is increasingly recognized that the organization of the genomes in terms of gene orientations [1] and intergenic distances [2] is not random but rather reflects complex interactions between genes, especially adjacent ones [3,4]. Studies of correlated gene expression and regulation [5], together with the analysis of phylogenetic conservations by comparative genomics [6] have defined large expression "neighborhoods" in multiple genomes. These are likely under the influence of chromatin domains and 3D genome organization [7], as defined by high throughput methods such as ChIP-seq and Hi-C.

In addition, advances in DNA sequencing have brought a redefinition of what fraction of a genome is transcriptionally "active". A large part of what was once called the "transcriptional dark matter" of the human genome is now considered as actively transcribed [8]. Sequencing deep enough and in the right experimental conditions seems now the limiting factor to reveal what a region of DNA produces or binds, rather than if it produces or binds anything.

Pervasive transcription of the genome and the existence of large scale positional patterns of gene co-expression raise the question of how adjacent genes might interact, interfere or be insulated from each other. However, apart from a few examples such as tandem gene duplications or bidirectional promoters, our understanding of the molecular mechanisms influencing how adjacent genes interact is still limited.

While studying a novel family of negative transcription elongation factors in Arabidopsis thaliana, we uncovered their unexpected role in protecting the downstream genes of tandem gene pairs from transcriptional interferences [9]. Building on this study, I developed a set of R functions designed to explore the direct neighbors of any set of pre-defined genes.

High-throughput sequencing experiments (and others!) frequently yield lists of genes ("gene sets"). In some cases these sets of genes might be enriched for neighbors with a specific orientation, or that tend to be closer or farther than expected by chance. The statistical and graphical summaries provided by the *GeneNeighborhood* R package allow to explore such situations, potentially opening avenues to study how adjacent genes interact.

2 Analyzing gene neighborhood

Analyzing the neighborhood of one or several sets of genes with the *GeneNeighborhood* package involves the following steps:

1) *Defining all direct neighborhoods*. Starting from objects obtained from or created with Bioconductor packages (http://bioconductor.org/), a single function gathers informations on the orientation and intergenic distances of the direct upstream and downstream neighbors for all genes in a genome (Fig. 1). This task is straightforward in the absence of annotation overlaps (Fig. 1) but quickly becomes complicated in the presence of overlapping annotations. So far, the *GeneNeighborhood* package handles cases of a single overlapping annotation, thus typically covering >75-80% of the situations, even in well annotated genomes. In certain cases, the annotations may be pre-filtered by the user to answer specific questions (e.g. focus on transcribed gene only), further improving the coverage of the analysis.

2) *Analyzing neighbors orientations*. Enrichment of a specific orientation (same strand / opposite strand, with or without overlap) for the upstream and downstream gene neighbors is evaluated by a Fisher exact test. The percentages of each orientation are displayed as stacked bars.

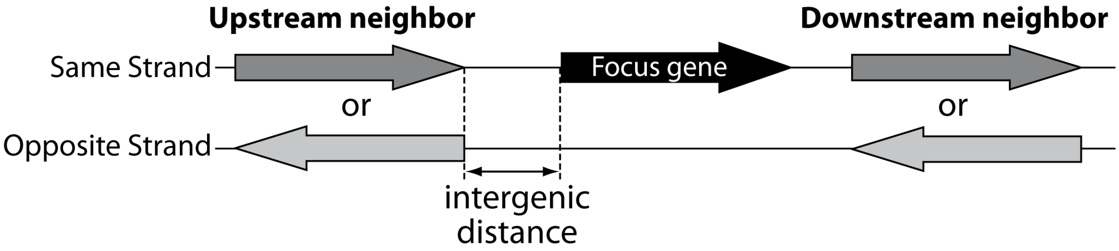

Figure 1. Definition of direct gene neighborhood.

3) *Analyzing the proximity of gene neighbors*. Descriptive statistics on intergenic distances are gathered for the gene sets and for the gene universe (e.g. all genes). The distributions of intergenic distances for the gene sets and for the gene universe are plotted and compared using different statistical tests in order to evaluate if the upstream or downstream neighbors of the genes of interest are located at distances shorter or larger than expected by chance.

4) *Additional graphical representations* are proposed. Metagene profiles of annotation coverage allow to integrate the information of distance and orientation. Cumulative frequency of genes with a neighboring gene border (TSS or TES) as a function of increasing upstream or downstream intergenic distance allows to focus on single point features (i.e. annotation borders) that typically play specific roles (TSS vs TES).

3 Conclusion and future development

Overall, the statistical summaries, tests and graphical representations allow to evaluate if the upstream and downstream neighbors of a given gene set present specific orientations or proximity/distance. The *GeneNeighborhood* R package is available on github (github.com/pgpmartin/GeneNeighborhood). Future developments will focus on integrating genes with more than one overlap, on analyzing sorted lists of genes (e.g. by differential expression) and on further improving graphical and statistical outputs.

Acknowledgements

Pascal GP Martin has received the support of the EU in the framework of the Marie-Curie FP7 COFUND People Programme, through the award of an AgreenSkills+ fellowship (under grant agreement n°609398).

References

[1] XQ. Li, D. Du. Gene direction in living organisms. *Sci Rep.* Vol. 2:982, 2012.

[2] A. Gherman, R. Wang, D. Avramopoulos. Orientation, distance, regulation and function of neighbouring genes. *Hum Genomics.* Vol. 3(2), pp. :143-56, 2009.

[3] Y. Chen, AA. Pai, J. Herudek *et al.* Principles for RNA metabolism and alternative transcription within closely spaced promoters. *Nat Genet.* Vol. 48(9), pp. :984-94, 2016.

[4] YH. Woo, WH. Li Gene clustering pattern, promoter architecture, and gene expression stability in eukaryotic genomes. *Proc Nat Acad Sci.* Vol. 108(8), pp. :3306-11, 2011

[5] P. Michalak. Coexpression, coregulation, and functionality of neighboring genes in eukaryotic genomes. *Genomics.* Vol. 91, pp. :243-48, 2008

[6] AT Ghanbarian, LD Hurst. Neighboring genes show correlated evolution in gene expression. *Mol Biol Evol.* Vol. 32(7), pp. :1748-66, 2015.

[7] S. Rennie, M. Dalby, L. van Duin, R. Andersson. Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions. *Nat Commun.* Vol. 9(1), pp. :487, 2018.

[8] P. Kapranov, AT. Willingham, TR Gingeras. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet.* Vol. 8, pp. :413-23, 2007.

[9] KE. Shearwin, BP. Callen, JB. Egan. Transcriptional interference – a crash course. *Trends Genet.* Vol. 21(6), pp. : 339-45, 2005.