Scientific Project

Master GENIOMHE

2023–2024

Samuel Ortion

Further development on FTAG Finder, a pipeline to identify Gene Families and Tandemly Arrayed Genes

Advisors:
Carène Rizzon
Franck Samson
Laboratoire de Mathématiques et Modélisation d'Évry
carene.rizzon@univ-evry.fr
franck.samson@inrae.fr
+33 (0) 1 64 85 35 40
IBGBI
23 Bd. de France
91037 Évry Cedex

Abstract: *D uplicate genes is an important component of genomes. They have a particular role in genome evolution, allowing species to explore new gene functionality offering a pool of usable genes to build on. TODO:*

keywords: duplicate genes, tandemly arrayed genes, pipeline

Contents

List of Figures

Acronyms

FTAG Families and Tandemly Arrayed Gene 6

1 Context

It is estimated that between 46% and 65.5% of human genes could be considered as duplicate genes (Correa et al., 2021). Duplicate genes offers a pool of genetic material available for further experimentation during species evolution.

1.1 Duplication mechanisms

Multiple mechanisms may lead to gene duplication. We review them in this section.

1.1.1 Segment duplication

1.1.2 Retroduplication

Retrotransposons, or RNA transposons are one type of transposable elements. Retrotransposons share similar structure and mechanism with retroviruses. They may replicate in the genome through a mechanism known as "copy-and-paste". These transposons are typically composed of a reverse transcriptase gene. This enzyme may proceed in the reverse transcription of an mRNA transcript into DNA sequence which can then be inserted elsewhere in the genome. More generally, retroduplication refers to the duplication of a region of a chromosome through reverse transcription or a RNA transcript.

1.1.3 Transduplication

DNA transposons are another type of transposable element whose transposition mechanism can also lead to gene duplication. This type of transposable element moves in the genome through a mechanisms known as "cut-and-paste". A typical DNA transposon contains a transposase gene. This enzyme recognize two sites surrounding the donnor transposon sequence in the chromosome resulting in a DNA cleavage and excision of the transposon. The transposase can then insert the transposon in a new place of the genome. Similarly to retrotransposon, if a gene was present between the two cleavage sites of the donnor transposon, it may move with the transposed sequence.

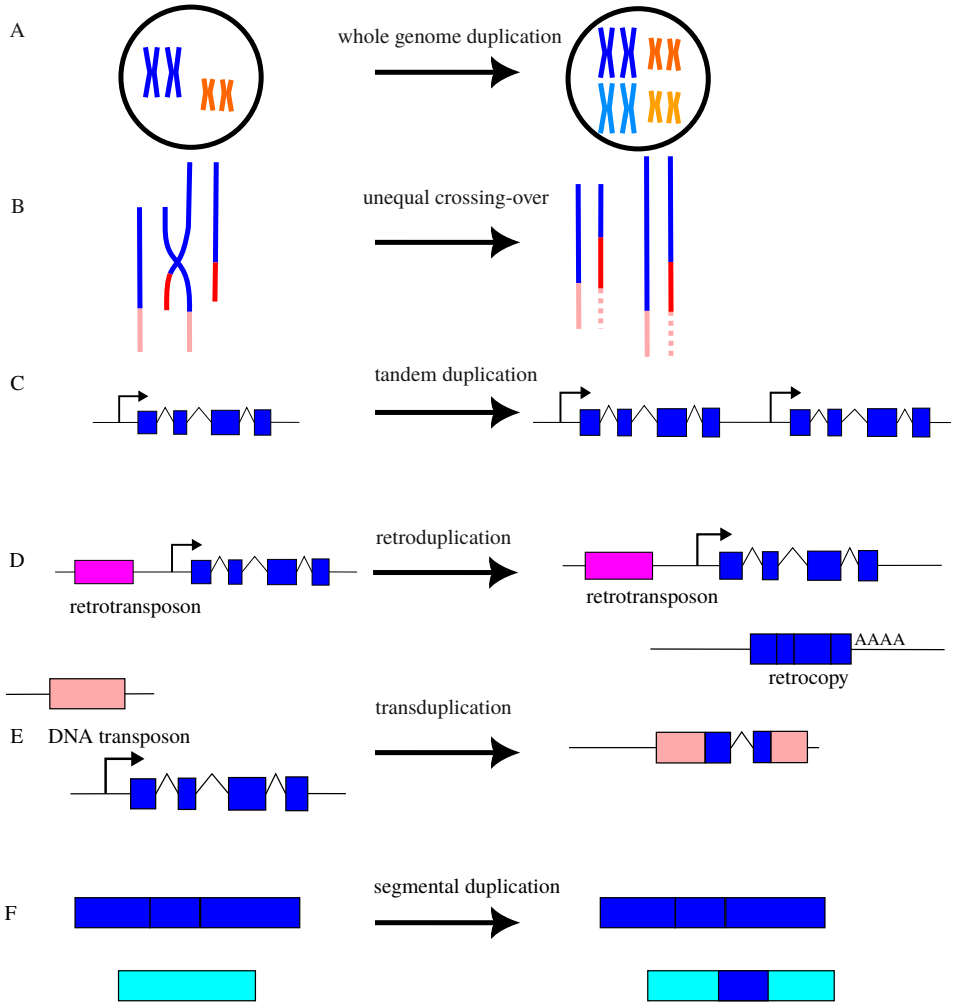

Figure 1.1: Different types of duplications. (A) Whole genome duplication. (B) An unequal crossing-over leads to a duplication of a fragment of a chromosome. (C) In tandem duplication, two (set of) genes are duplicated one after the other. (D) Retrotransposon enables retroduplication: a RNA transcript is reverse transcribed and inserted back without introns and with a polyA tail in the genome. (E) A DNA transposon can acquire a fragment of a gene. (F) Segmental duplication corresponds to long stretches of duplicated sequences with high identity. **Source** Adapted from (LALLEMAND et al., 2020).

1.1.4 Tandem Duplication

1.1.5 Polyploidisation and Whole Genome Duplication

In an event of whole genome duplication, the entire set of genes present on the chromosomes is duplicated. Whole genome duplication is more frequent in plants. A striking example is probably the *Triticum* genus (wheat) in which some species (such as *T. aestivum*) are hexaploid, due to hybridisation events (Golovnina et al., 2007).

We distinguish two kinds of polyploidisation, based on the origin of the duplicate genome:

- Allopolyploidisation occurs when the supplementary chromosomes comes from an other species. This is the case for *Triticum aestivum* hybridisation.
- Autopolyploidisation consist in the hybridisation of the genome within the same species.

Whole genome duplication can occur thanks to polyspermy or in case of a non-reduced gamete, for instance.

1.1.6 Unequal crossing-over

A crossing-over may occur during cell division. A fragment of chromosome is exchanged between two chromatids of a pair of chromosome. If the cleavage of the two chromatids occured at different positions on both chromosomes, the shared fragments may have different lengths. When the repair of missing fragment is performed, the resulting chromosome will incorporate a duplicate region of the chromosome, leading to a potential duplication for genes present in this region, as represented in figure **??**. This mechanism leads to the duplication of the whole set of genes present in the inserted fragment. An array of genes is duplicated after the original array and are thus called Tandemly Arrayed Genes.

1.2 Role of duplicate genes in genome evolution

In his book *Evolution by Gene Duplication*, Susumu Ohno proposed that gene duplication plays a major role in species evolution (Ohno, 1970).

1.3 Methods to identify duplicate genes

Lallemand et al. review the different methods used to detect duplicate genes. These methods are dependant on the type of duplicate genes they target (Lallemand et al., 2020).

1.3.1 FTAG Finder

Developped in the LaMME laboratory, this pipeline targets the detection of gene families and tandemly arrayed genes from a given species' proteome (Bouillon et al., n.d.).

Estimation of homology links between genes

This steps consists in establishing a relation between each genes in a genome. In this step, the typical tool involved is `BLAST` (Basic Local Alignment Search Tool) (Altschul et al., 1990) run on the whole proteome.

Several `BLAST` metrics can be used as an homology measure, such as bitscore, identity percentage, E-value or variations on those. The metrics choice may have an impact on the results of graph clustering in the following step (Gibbons et al., 2015).

Identification of gene families

Based on the homology links between each pair of genes, we construct a weighted undirected graph whose vertices corresponds to genes and edges to homology links. Then, a graph clustering algorithm is applied on this graph in order to infer the gene families.

The team chosed to propose three clustering algorithms: Single linkage, Markov Clustering or Walktrap.

2 Objectives

2.1 Extend the existing Galaxy pipeline

Galaxy is a web-based platform for performing accessible data analysis pipeline, first designed for use in genomic data analysis (Goecks et al., 2010).

Last year, Séanna Charles, worked on the Galaxy's version of the FTAG Finder pipeline during her M1 internship (Charles, 2023). I will continue this work.

2.2 Port FTAG Finder pipeline on a workflow manager

Another objective of my internship will be to port FTAG Finder on a workflow manager better suited to larger and more reproducible analysis.

We will have to make a choice for the tool we will use. The two main options are Snakemake and Nextflow. Snakemake is a python powered workflow manager based on rules *à la* GNU Make (Köster and Rahmann, 2012). Nextflow, is a groovy powered workflow manager, which rely on data flows (Di Tommaso et al., 2017). Both are widely used in the bioinformatics community, and their use have been on the rise since they came out in 2012 and 2016 respectively (Djaffardjy et al., 2023).

These tools ease the deployment of large scale data analysis workflow with reproducible output.

Bibliography

Altschul, Stephen F. et al. (Oct. 5, 1990). "Basic Local Alignment Search Tool". In: *Journal of Molecular Biology* 215.3, pp. 403–410. issn: 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. url: https://www.sciencedirect.com/science/article/pii/S0022283605803602 (visited on 04/30/2023).

Bouillon, Bérengère et al. (n.d.). *FTAG Finder: Un Outil Simple Pour Déterminer Les Familles de Gènes et Les Gènes Dupliqués En Tandem Sous Galaxy*.

Charles, Séanna (2023). *Finalisation du pipeline FTAG (Families and TAG) Finder, un outil de détection des gènes dupliqués sous Galaxy*. Internship Report. Laboratoire de Mathématiques et Modélisation d'Évry.

Correa, Margot et al. (May 1, 2021). "The Transposable Element Environment of Human Genes Differs According to Their Duplication Status and Essentiality". In: *Genome Biology and Evolution* 13.5, evab062. issn: 1759-6653. doi: 10.1093/gbe/evab062. url: https://doi.org/10.1093/gbe/evab062 (visited on 09/15/2023).

Di Tommaso, Paolo et al. (Apr. 2017). "Nextflow Enables Reproducible Computational Workflows". In: *Nature Biotechnology* 35.4, pp. 316–319. issn: 1546-1696. doi: 10.1038/nbt.3820. url: https://www.nature.com/articles/nbt.3820 (visited on 03/26/2024).

Djaffardjy, Marine et al. (2023). "Developing and Reusing Bioinformatics Data Analysis Pipelines Using Scientific Workflow Systems". In: *Computational and Structural Biotechnology Journal* 21, p. 2075. doi: 10.1016/j.csbj.2023.03.003. pmid: 36968012. url: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10030817/ (visited on 03/26/2024).

Gibbons, Theodore R. et al. (Dec. 2015). "Evaluation of BLAST-based Edge-Weighting Metrics Used for Homology Inference with the Markov Clustering Algorithm". In: *BMC Bioinformatics* 16.1, p. 218. issn: 1471-2105. doi: 10.1186/s12859-015-0625-x. url: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0625-x (visited on 03/19/2024).

Goecks, Jeremy et al. (2010). "Galaxy: A Comprehensive Approach for Supporting Accessible, Reproducible, and Transparent Computational Research in the Life Sciences". In: *Genome Biology* 11.8, R86. issn: 1474-760X. doi: 10.1186/gb-2010-11-8-r86. pmid: 20738864.

Golovnina, K. A. et al. (Apr. 1, 2007). "Molecular Phylogeny of the Genus Triticum L". In: *Plant Systematics and Evolution* 264.3, pp. 195–216. issn: 1615-6110. doi: 10.1007/s00606-006-0478-x. url: https://doi.org/10.1007/s00606-006-0478-x (visited on 03/27/2024).

Köster, Johannes and Sven Rahmann (Oct. 1, 2012). "Snakemake–a Scalable Bioinformatics Workflow Engine". In: *Bioinformatics (Oxford, England)* 28.19, pp. 2520–2522. issn: 1367-4811. doi: 10.1093/bioinformatics/bts480. pmid: 22908215.

Lallemand, Tanguy et al. (Sept. 4, 2020). "An Overview of Duplicated Gene Detection Methods: Why the Duplication Mechanism Has to Be Accounted for in Their Choice". In: *Genes* 11.9, p. 1046. issn: 2073-4425. doi: 10.3390/genes11091046. url: https://www.mdpi.com/2073-4425/11/9/1046 (visited on 03/19/2024).

Ohno, Susumu (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. isbn: 978-3-642-86661-6 978-3-642-86659-3. doi: 10.1007/978-3-642-86659-3. url: http://link.springer.com/10.1007/978-3-642-86659-3 (visited on 03/21/2024).

Summary