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

keywords: duplicate genes, tandemly arrayed genes, pipeline

Contents

List of Figures

Acronyms

FTAG Finder Families and Tandemly Arrayed Genes Finder 10

TAG Tandemly Arrayed Genes 6, 10

WGD Whole Genome Duplication 6

1 Scientific 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 Gene duplication mechanisms

Multiple mechanisms may lead to gene duplication. The following sections review them.

1.1.1 Polyploidisation and whole genome duplication

In an event of Whole Genome Duplication (WGD), the entire set of genes present on the chromosomes is duplicated (figure 1.1 (A)). WGD 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 duplication of the genome within the same species.

WGD can occur thanks to polyspermy or in case of a non-reduced gamete.

1.1.2 Unequal crossing-over

A crossing-over may occur during cell division. Two chromatids may exchange a fragment of chromosome. If the cleavage of the two chromatids occurs at different positions, the shared fragments may have different lengths. Homologous recombination of such uneven crossover results in the incorporation of a duplicate region, as represented in figure 1.1 (B, C). This mechanism leads to the duplication of the whole set of genes present in the inserted fragment. These duplicate genes locate one set after the other, and are thus called Tandemly Arrayed Genes (TAG).

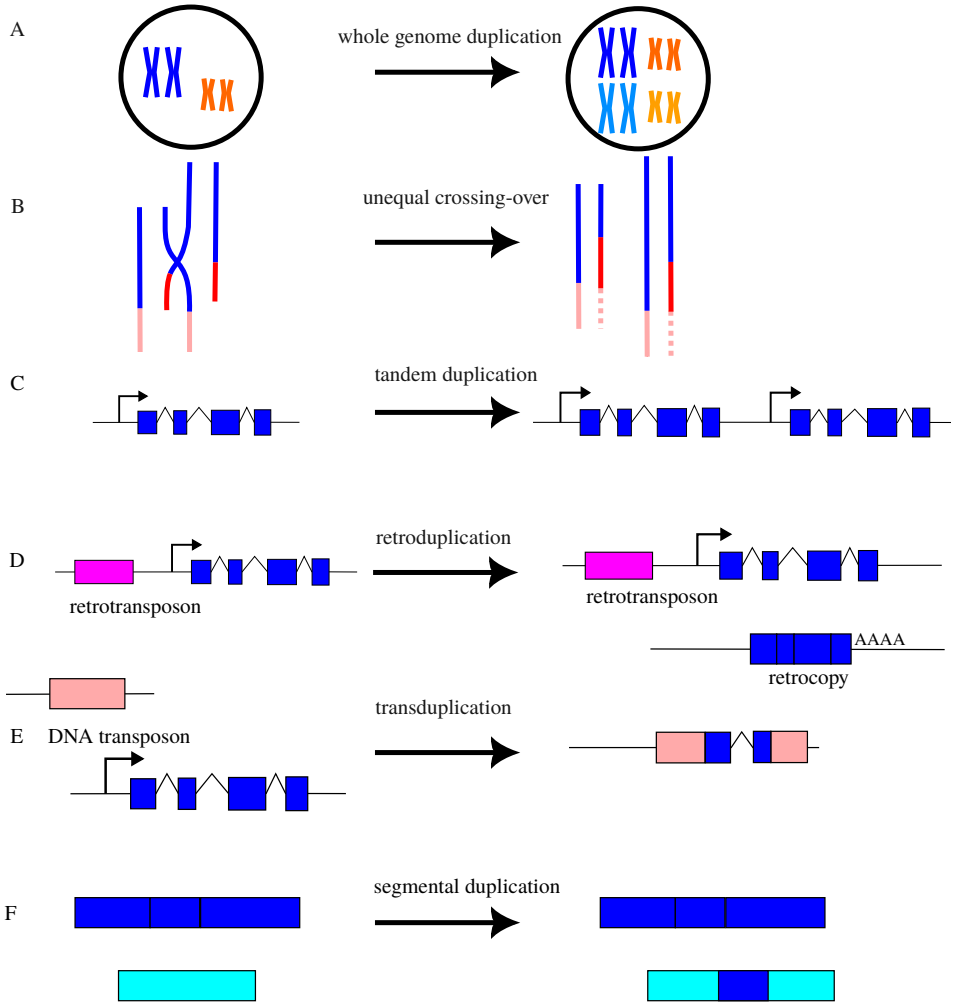

Figure 1.1: Different types of duplication. (A) Whole genome duplication. (B) An unequal crossing-over leads to a duplication of a fragment of a chromosome. (C) In tandem duplication, two (set of) genes are duplicated one after the other. (D) Retrotransposon enables retroduplication: a RNA transcript 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. Adapted from (Lallemand et al., 2020) (fig. 1).

1.1.3 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 typically contain 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 of a RNA transcript. In this case the duplicate gene lost its intronic sequences and brings a polyA tail with it (figure 1.1 (D)).

1.1.4 Transduplication

DNA transposons are another type of transposable elements 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 genome locus. A transposon can bring a fragment of a gene during its transposition in the other locus (figure 1.1 (E)).

1.1.5 Segment duplication

Segment duplications, also called low copy repeats are long stretches of DNA with high identity score (figure 1.1 (E)). Their exact duplication mechanisms remains unclear (Lallemand et al., 2020), they may results from an accidental replication, distinct from an uneven cross-over or a double stranded breakage. Nevertheless, transposable elements may well be involved as a high enrichment of transposable elements has been found at segment extremities, in *Drosophila* (Lallemand et al., 2020).

1.2 Fate 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), as it provides a new genetic material to build on new phenotypes while keeping a backup gene for the previous function.

Duplicate genes may be inactivated becoming pseudogenes, be deleted or conserved.

1.2.1 Pseudogenisation

Duplicate genes may be inactivated and become pseudogenes. These pseudogenes keep a gene-like structure, which degrades as and when further genome modifications occur, but are no longer expressed.

1.2.2 Neofunctionalisation

Duplicate genes may be conserved and gain a new function. For instance, in *Drosophila*, the set of olfactory receptor genes result from several duplication and deletion events (Nozawa and Nei, 2007), after which the duplicate may specialize in the detection of a particular chemical compound.

1.2.3 Subfunctionalisation

Two duplicate genes with the same original function may encounter a subfunctionalisation during which each gene conserves only one part of the function.

1.2.4 Functional redundancy

Two copies may keep the ancestral function: in this case the organism may increase the quantity of gene product.

1.3 Methods to identify duplicate genes

Lallemand et al. review the different methods used to detect duplicate genes. These methods depend on the type of duplicate genes they target, and vary on computation burden (Lallemand et al., 2020).

1.3.1 Paralog detection

Paralogs are homologous genes derived from a duplication event. They can be identified as homologous genes located in the same genome, or as homologous genes between different species once we filtered out orthologous genes (homologous genes derived from a speciation event).

Two gene characteristics can be used to assess to assess homology between two genes: gene structure of sequence similarity. The sequence similarity can be tested with a sequence alignment tool, such as `BLAST` (Altschul et al., 1990), `Psi-BLAST`, and `HMMER3` (Johnson et al., 2010), or `diamond` (Buchfink et al., 2021), which are heuristic algorithm, which means they may not provide the best results, but do so way faster than exact algorithms, such as the classical Smith

and Waterman algorithm (Smith and Waterman, 1981) or its optimized versions `PARALIGN` or `SWIMM`.

1.3.2 FTAG Finder

Developed in the LaMME laboratory, the Families and Tandemly Arrayed Genes Finder (FTAG Finder) pipeline targets the detection of gene Families and Tandemly Arrayed Genes from a given species' proteome (Bouillon et al., 2016).

The pipeline proceeds in three steps. First, it estimates the homology links between each pair of genes; then, it deduce the gene families and finally, it detects TAG.

Estimation of homology links between genes

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

Several `BLAST` metrics can be used as homology measures, such as bitscore, identity percentage, E-value or variations of these. The choice of metrics can affect the results of graph clustering in the following step, and should therefore be chosen carefully (Gibbons et al., 2015).

Identification of gene families

Based on the homology links between each pair of genes, we construct a undirected weighted graph whose vertices correspond to genes and edges to homology links between them. We apply a graph clustering algorithm on the graph in order to infer the gene families.

FTAG Finder proposes three clustering algorithm alternatives: single linkage, Markov Clustering (van Dongen, 1998) or Walktrap (Pons and Latapy, 2005).

Detection of TAGs

The final step of FTAG Finder consists in the determination of TAG from the gene families and the chromosome sequence. For a given chromosome, the tool seeks genes belonging to the same family and located close to each other. The tool allows a maximal number of genes between the homologous genes, with a parameter set by the user.

2 Objectives for the internship

2.1 Scientific questions

The underlying question of FTAG Finder is the study of the evolutionary fate of duplicate genes in Eukaryotes.

2.2 Extend the existing FTAG Finder Galaxy pipeline

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

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

2.3 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 2013 respectively (Djaffardjy et al., 2023).

Bibliography

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost (Apr. 2021). "Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND". In: *Nature Methods* 18.4, pp. 366–368. issn: 1548-7105. doi: 10.1038/s41592-021-01101-x. url: https://www.nature.com/articles/s41592-021-01101-x (visited on 03/28/2024).

Bouillon, Bérengère et al. (2016). *FTAG Finder: Un Outil Simple Pour Déterminer Les Familles de Gènes et Les Gènes Dupliqués En Tandem Sous Galaxy*.

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.

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.

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).

Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly (Aug. 18, 2010). "Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure". In: *BMC Bioinformatics* 11.1, p. 431. issn: 1471-2105. doi: 10.1186/1471-2105-11-431. url: https://doi.org/10.1186/1471-2105-11-431 (visited on 04/09/2024).

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).

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).

Nozawa, Masafumi and Masatoshi Nei (Apr. 24, 2007). "Evolutionary Dynamics of Olfactory Receptor Genes in Drosophila Species". In: *Proceedings of the National Academy of Sciences* 104.17, pp. 7122–7127. DOI: 10.1073/pnas.0702133104. URL: https://www.pnas.org/doi/full/10.1073/pnas.0702133104 (visited on 04/02/2024).

Di Tommaso, Paolo et al. (Apr. 2017). "Nextflow Enables Reproducible Computational Workflows". In: *Nature Biotechnology* 35.4, pp. 316–319. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.3820. URL: https://www.nature.com/articles/nbt.3820 (visited on 03/27/2024).

Pons, Pascal and Matthieu Latapy (Dec. 12, 2005). *Computing Communities in Large Networks Using Random Walks (Long Version)*. DOI: 10.48550/arXiv.physics/0512106. arXiv: physics/0512106. URL: http://arxiv.org/abs/physics/0512106 (visited on 03/30/2024). preprint.

Van Dongen, S. (Jan. 1, 1998). "A New Cluster Algorithm for Graphs". In: R 9814. URL: https://ir.cwi.nl/pub/4604 (visited on 03/22/2024).

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.

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).

Ohno, Susumu (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-642-86661-6. 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).

Smith, T. F. and M. S. Waterman (Mar. 25, 1981). "Identification of Common Molecular Subsequences". In: *Journal of Molecular Biology* 147.1, pp. 195–197. ISSN: 0022-2836. DOI: 10.1016/0022-2836(81)90087-5. URL: https://www.sciencedirect.com/science/article/pii/0022283681900875 (visited on 04/29/2023).

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).

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://b

mcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0625-x (visited on 03/19/2024).

Summary