SCIENCES

BIOLOGY

Genetics, Epigenetics

Function and Evolution of Repeated DNA Sequences

Coordinated by
Guy-Franck Richard

ISTE

WILEY

Table of Contents

List of Tables

List of Illustrations

Chapter 8

SCIENCES

Biology, Field Director – Marie-Christine Maurel

Genetics, Epigenetics, Subject Head – Bernard Dujon

Function and Evolution of Repeated DNA Sequences

Coordinated by

Guy-Franck Richard

First published 2023 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2022949377

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78945-119-1

ERC code:
LS2 Genetics, 'Omics', Bioinformatics and Systems Biology

LS2_5 Epigenetics and gene regulation

Foreword

Our modern societies are too preoccupied with immediate performances to conceive of a world where the costs and efforts to achieve a result are not rationally minimized. And yet, life offers this image as soon as we take time to study it closely. The remarkable adaptation of different organisms to their living conditions revolves around genomes which are far from the products of what we may consider to be rational engineering. There is no such thing as a minimal genome: all of them are too large in comparison to the number of genes considered necessary to produce the organism hosting them. Often far too large, ours appears 50 times too large. They all contain identically (or almost identically) repeated sequences, sometimes they are repeated numerous times in the same genome; whereas random combinations of the four nucleotides make this phenomenon extremely unlikely, if not practically impossible.

This situation was observed as far back as the mid-20th century, long before the emergence of genomics, through the study of the renaturation kinetics of DNA molecules. The excessive amount of DNA and the abundance of repeated sequences remained a puzzle that some tended to quickly dismiss by referring to junk DNA, continuing to focus their studies only on what they already knew! Genome sequencing would come to solve this enigma by demonstrating just how incomplete our prior knowledge was. A new vision of genome organization and function is now provided, in which temporal dynamics combine with the present, because all genomes are simply imperfect copies of the genomes that preceded them and not new constructs. From now on, traces of the past mix with

present events and together they lay the foundations of the future.

As the present work illustrates so remarkably, repeats found in genomes can result from major evolutionary accidents such as whole-genome duplications, which, in a singular phenomenon, tend to coincide with the transitions between major geological eras. But they may also come from repeated interactions with infectious elements – of the viral type – that eventually integrate into chromosomes and are transmitted to the offspring. Thus, there are both endogenous and exogenous causes for the existence of repeated sequences in genomes. In turn, repeats can form the basis of the formation of chromosome functional elements such as centromeres, telomeres and guanine quadruplexes. Copy number variation of long repeated sequences can play a critical role in phenotypes and in organism adaptation. Similarly, the instability of short-sequence repeats allows us to easily differentiate between individuals from the same population. However, this can sometimes lead to very serious syndromes. Finally, the different mobile elements, kinds of specialized molecular machines, present in various numbers in the different genomes cannot be ignored. By the mid-20th century, they had already been identified by their genetic effects – they are mutagenic – but we now have a much broader view of their diversity and of the consequences, sometimes considerable, of their activity.

If our knowledge of repeated genome sequences has only progressed belatedly, this is partly due to technical difficulties encountered in their sequencing. Until the recent advent of new technologies allowing longer reads, it was very difficult to correctly assemble repeated sequences and many so-called whole-genome sequences were in fact incomplete. For example, about 8% of the human genome, made up of highly repetitive sequences, remained unknown

for two decades, until the application of special technologies this year. Similarly, the study of copy number variations due to segmental duplications, which have long been underestimated, is only just beginning. And let us not forget that our exploration of the living world is not only far from complete but is very much biased in favor of the already well-known groups of organisms. We can therefore expect new discoveries, or even surprises, in the study of this part of genomes, overlooked for too long, which demonstrates just how real long-term success differs from the illusion of immediate performance.

Gif-sur-Yvette
Bernard DUJON
Professor Emeritus at *Sorbonne Université*
and the *Institut Pasteur*
Member of the *Institut de France*
(Académie des Sciences)
May 2023

Introduction
About Repeated Genomes

Guy-Franck RICHARD
Instabilités naturelles & synthétiques des génomes, Institut Pasteur, CNRS UMR3525, Paris, France

Genome ['dʒiːnəʊm] *nm Biol.*: Set of hereditary characteristics of a living being, of which a small part is composed of genes providing a function to the organism, and the majority is composed of repeated sequences for which it is unknown whether they have a function.

Taking matters a little further, this could be a modern definition of the word “genome”, in the light of the knowledge garnered across three decades from sequencing the DNA content of living beings, in particular eukaryotic organisms, more complex than those of their bacterial and archaebacterial ancestors. Biologists were already aware back in the 1960s, long before the invention of the first DNA sequencing methods, that the content of genomes was difficult to comprehend. Denaturation-renaturation experiments highlighted that the speed of renaturation of the double-helix was proportional to its concentration. The C_0t parameter was the value at which renaturation of half the genomic DNA was complete, under controlled conditions. Each organism could then be defined by the C_0t value of its genome. In trying to establish the C_0t values of genomes of the simplest organisms – phages or bacteria – or of more complex organisms, such as vertebrates, it transpired that the latter contained three types of sequences presenting very different C_0t values (see Figure I.1).

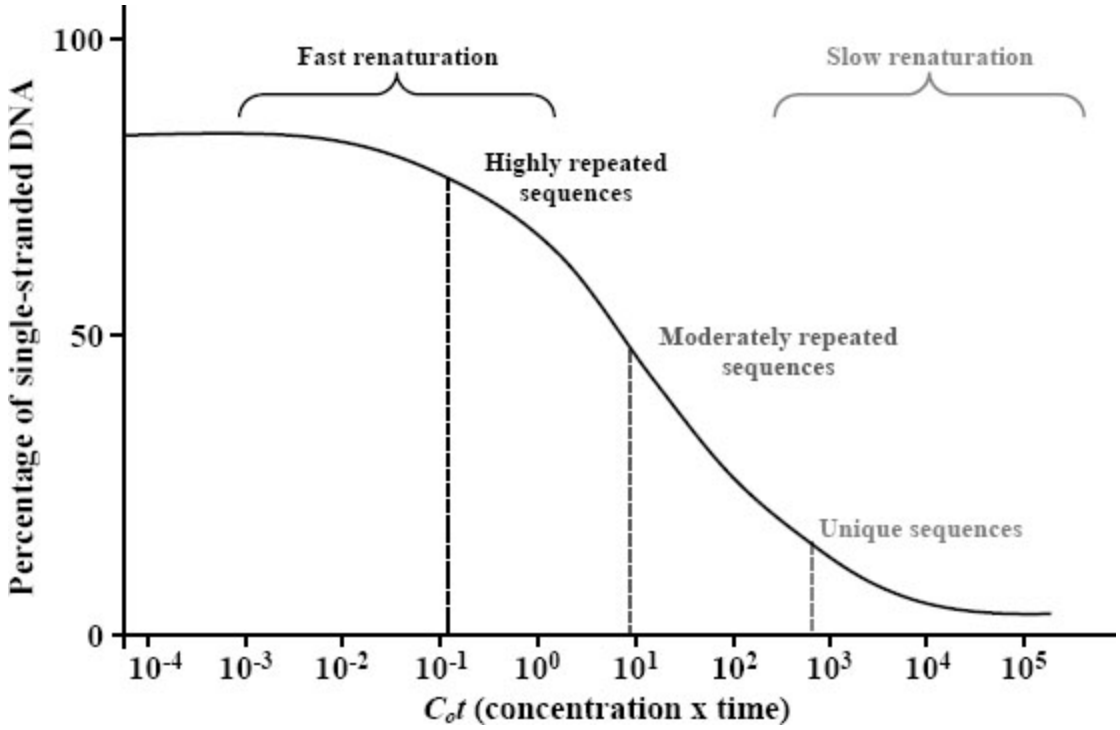

Figure I.1. *Example of* C_0t *curve*

It is thus possible to show that the mouse genome, for example, is composed of 70% unique sequences with slow renaturation, 20% moderately repeated sequences present in 1,000 to 100,000 copies per genome and 10% highly repeated sequences representing at least 1 million copies per genome and showing rapid renaturation (Britten and Kohne 1968). This approach, based on the physicochemical properties of DNA, slightly underestimated the quantity of repeated sequences because their renaturation rate depends on the identity between these sequences, divergent sequences (such as long terminal repeats (LTRs)) renaturing more slowly than identical sequences. Nowadays, C_0t curves are still sometimes used to separate the highly repetitive fraction of a genome from its unique fraction in order to sequence specific DNA of either fraction (Peterson et al. 2008).

I.1. The “C-value” paradox

From the moment it was proven that DNA was the support of heredity, and theoretically contained all the genes necessary for the development of a living being, it seemed logical that the most sophisticated organisms had to contain more genes and therefore more DNA in their genome (the “C value”) to encode these genes. This idea was to be questioned in the 1950s with the discovery that the nuclei of certain amphibians and fish contained 20 times more DNA than the nuclei of mammals. Given that the latter presented a greater developmental complexity, this appeared very much paradoxical, and was even used as an argument by the opponents of DNA being the sole support of heredity (Thomas 1971). This “C-value paradox” could finally be explained only decades later, when the first genomes were sequenced. It is now known that the number of genes in an organism has little to do with its size or level of complexity. The baker’s yeast genome contains about 6,000 genes, that of fruit flies about 14,000 and the human genome (or those of its very close cousins, great apes) contents itself with 20,000 genes, with which it manages a very sophisticated level of developmental and behavioral complexity. But what about the paramecium with its 40,000 genes, twice as many as the human genome? Or *Trichomonas vaginalis*, a parasite of the genital tract, with its 60,000 genes? Or indeed wheat and its 124,000 genes, more than six times as many as our genes? Clearly, this so-called complexity could not be measured by the number of genes in an organism. Studies of comparative genomics[1] have shown that this high number of genes in certain organisms does in fact conceal ancestral events of partial or total genome duplication, followed by variable amounts of gene losses (Wolfe and Shields 1997; Jaillon et al. 2004). These events actively participate in the genetic redundancy and their identification as well as their underlying mechanisms will be addressed in Chapter 1.

If the complexity of an organism has nothing to do with the number of genes contained, the same is true of the amount of DNA. The human genome, with just over three times as many genes as brewer's yeast, contains 200 times more DNA. The genome of a rotifer – a small animal measuring just a few millimeters that lives in freshwaters – contains three times more genes than the human genome in 12 times less DNA! (see Figure I.2).

The genomic sequence of all these organisms showed that some of them had evolved a very compact genome, with high gene density, while others contained a multitude of repeated DNA sequences whose function did not appear obvious at first glance, and that some authors did not hesitate to call them "junk DNA" (Ohno 1972).

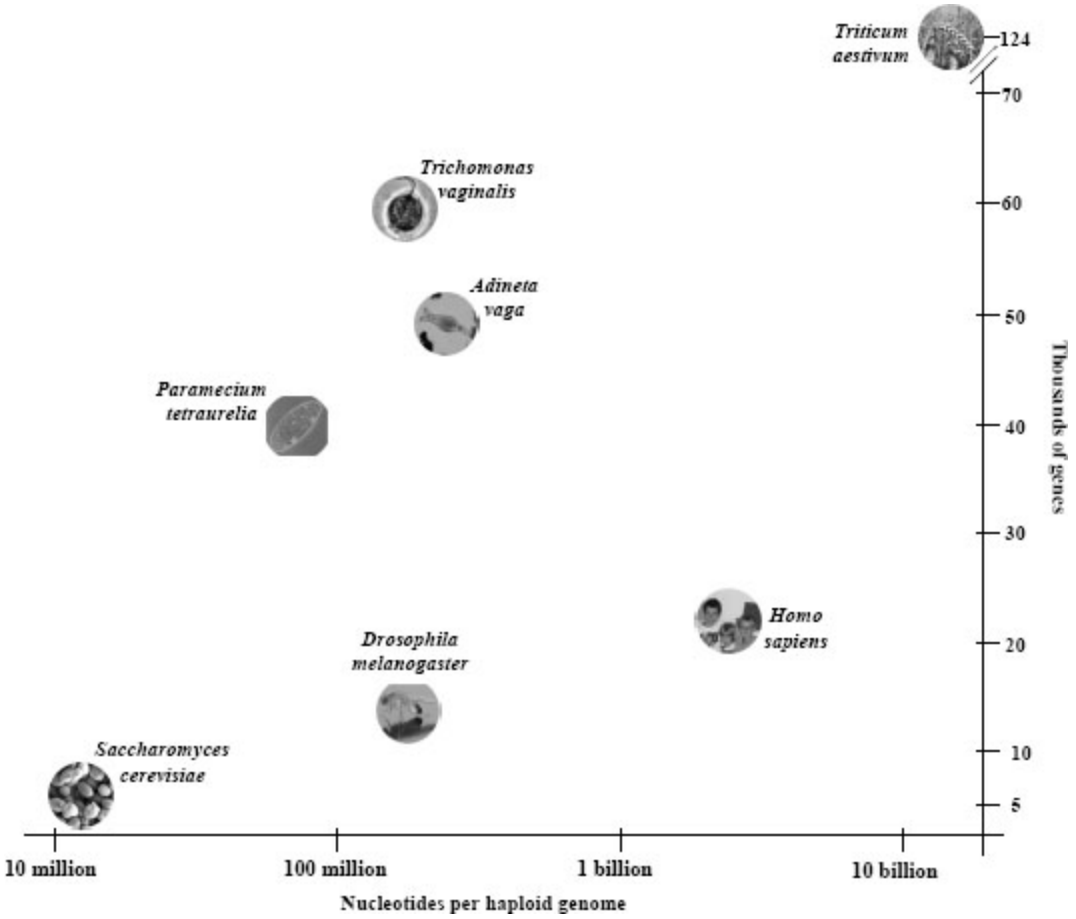

Figure I.2. *Comparison of genome sizes and gene numbers*

I.2. Recycling junk DNA

About 2% of the human genome is translated into proteins. Even by adding the untranslated genes (rRNA, tRNA, siRNA, snRNA, etc.), the percentage of “useful” DNA barely increases. So, what is the purpose of the 98% of DNA in our genome that has, apparently, no function? One conceivable answer is that it has none. The consortium led by Jeff Boeke, professor of genetics at Johns Hopkins University in Baltimore, set out to create the first synthetic yeast genome, using synthetic oligonucleotides. The brewer’s yeast *Saccharomyces cerevisiae* is a eukaryotic organism whose genome contains 12.5 million nucleotides distributed across 16 chromosomes. The synthetic chromosomes were reconstructed one by one from 70 nucleotide-long sequences assembled in blocks of 750 base pairs, themselves assembled in mega-blocks of 2–4 kb, reintroduced one after the other in a hierarchical manner into the yeast genome in replacement of the natural sequences (Muller and Koszul 2015). When designing synthetic chromosomes, it was decided that all repeated sequences would be removed from the genome. All tRNA-encoding DNAs were grouped on a single circular chromosome, specifically built to carry them. Retrotransposons, microsatellites, minisatellites and other repeated elements inessential to life were removed from the new sequence. These synthetic chromosomes, with their junk DNA removed, are perfectly able to sustain life in yeast cells containing them, without any apparent phenotypic defect, at least under laboratory growth conditions (Dymond et al. 2011; Annaluru et al. 2014). One may conclude from the results of this project that junk DNA is useless. However that would be a mistake.

The human reference genome contains about 443,000 residual elements of past retroviral invasions, covering

8.3% of the total sequence (International Human Genome Sequencing Consortium 2001). These retroviral scars are the remains of successive invasions, occurring over the past hundred million years, of our mammalian ancestors by exogenous elements, which left the trace of their passage in the form of LTR[2]. These retroviral remains are therefore part of our junk DNA. Nevertheless, as we will see, their presence in our genome testifies to their distant but indispensable role in the existence of our lineage. Therian mammals, that is, those possessing a uterus within which the fertilized egg develops, are classified into two groups. Eutherians (or placentals) like humans and mice have a very elaborate placenta connecting the wall of the uterus to the embryo and allowing it to develop in complete safety throughout the entire gestation period. Marsupials (kangaroos and koalas) do not have placentas and the development of their young takes place mainly outside the uterus. Genome sequencing showed that the two human genes specifically expressed in the placenta, *syncytin-1* and *syncytin-2*, were derived from a gene encoding an ancestral viral protein, which infected the primate lineage 25–40 million years ago. Remarkably, the genome of the mouse, another placental mammal, also contains two viral genes having the same function as human genes but deriving from a slightly more recent viral infection than that of the human lineage. Thus, the placenta was invented twice, independently, in two lineages of mammals, by capture of genes of retroviral origin (Dupressoir et al. 2009). Another example is even more striking. Sexual reproduction was invented at the origin of the eukaryotic world. From the first primitive eukaryotic cells, a syngamy[3] system was developed that allowed the nuclei of two haploid cells to fuse to give birth to a diploid cell. The protein responsible for the fusion of male and female gametes is the same in plants and animals; it is the product of the *HAP2* gene. This

protein is of viral origin and allows the envelope of a virus to fuse with the plasma membrane of its host's cells (Fédry et al. 2017). Thus, a gene essential to sexual reproduction was captured from a virus by the genome of the very first eukaryotic cells about 1.5 billion years ago.

Other examples of the capture of a piece of transposable element exist, thus creating a new gene, a new function. Junk DNA is therefore regularly recycled during the course of evolution to bring diversity and novelty. As François Jacob (1977) said more than 40 years ago, evolution "tinkers", it makes new from old, reusing bits of genes, cutting them, splicing them and fusing them with others in order to create novelty. What appears today to the geneticists of the 21st century as junk DNA perhaps served in the past – or will serve in the future – to create diversity. The tremendous success of the eukaryotic world in invading all ecological niches under all climates and latitudes stems in part from the extraordinary flexibility of its genome and its ability to accumulate genetic elements that are seemingly useless but will be recycled in the long run to create novelty and enable the appearance of new living species.

I.3. The different repeat types

There are often several ways to classify genetic elements. Some authors have chosen to distinguish between dispersed repeated elements in contrast to tandem repeats, the latter being repeated at least twice in a row at the same genetic locus, unlike the former, which are repeated at different loci (Richard et al. 2008). But some dispersed repeats are so numerous in the genomes that they appear to be tandemly repeated. This is the case for *Alu* sequences in humans, which are frequently found grouped in introns or intergenic sequences. Repeated sequences of exogenous

origin, that is to say originating from an organism other than the cell in which they are observed, could also be distinguished from repeated sequences of endogenous origin, manufactured by the cell in which they are observed. Transposable elements would belong to the first category, having invaded the genomes of eukaryotic (or prokaryotic) lineages, while the different satellite DNAs would belong to the second, being manufactured by molecular processes specific to the genomes that contain them. But other problems then arise. It is known, for example, that *Alu* elements, inactive retrotransposons that can be mobilized in *trans* by the machinery of other retroelements, are of endogenous origin. They result in fact from the duplication of the non-coding 7SL RNA, which is involved in the synthesis of excreted proteins. This duplication, prior to mammalian radiation, resulted in the fusion of two monomers of 130 nucleotides derived from 7SL RNA, separated by a short adenine-rich region (Ullu and Tschudi 1984). Achieving a coherent classification of the repeated sequences therefore proves a complicated task, particularly in the genomes of evolved plants and animals within which they are plethoric, both in structure and number.

We have therefore tried in the rest of this work to present the repeat elements in relation to their role (proven or assumed) in genomes, rather than according to their structure or their assumed origin (see Figure I.3).

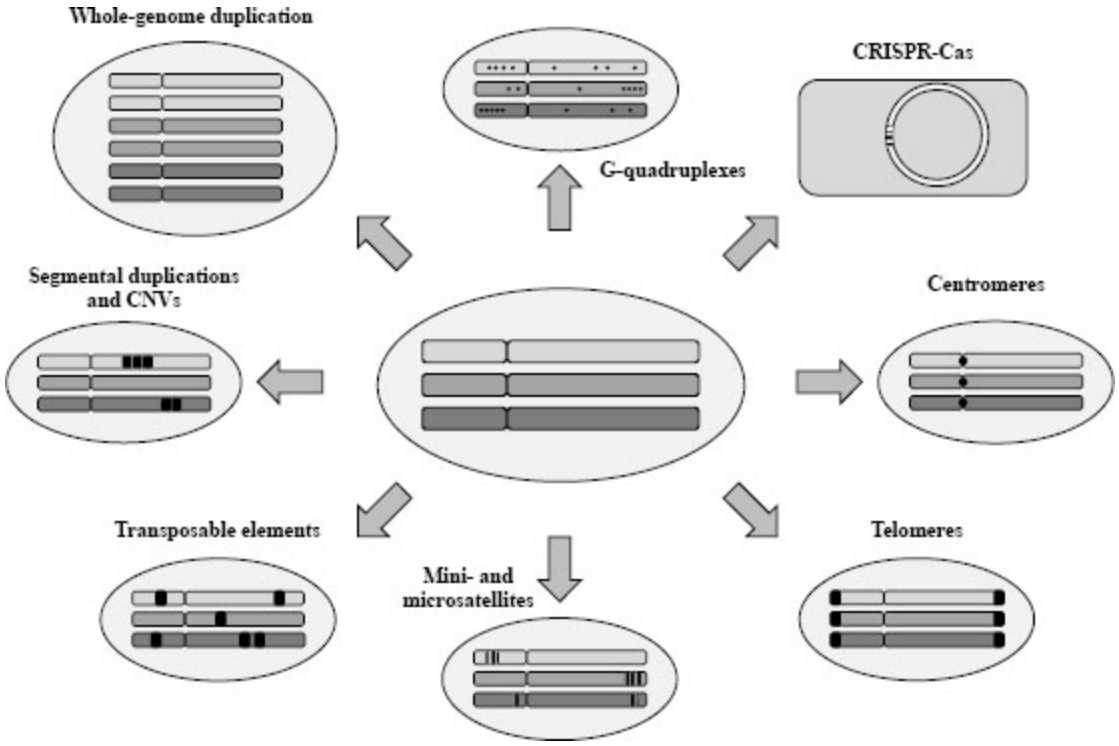

Figure I.3. *The different types of repeated DNA sequences*

After exploring total or partial genome duplications in Chapter 1, the duplications of large DNA segments, sometimes in multiple copies in tandem or dispersed within genomes, will be described in Chapter 2. These contribute significantly to the level of genetic redundancy and gene duplication and their study, although essential to understand the dynamics of complex genomes and the inheritability of certain traits is still in its infancy. Transposons and retrotransposons will be presented in Chapter 3, and their role in the generation of genetic novelties will be detailed. In most species, centromeres are present at a rate of one per chromosome. These very particular repeated elements are essential for the proper segregation of sister chromatids during cell divisions. They will be studied in Chapter 4 and as we will see, holocentric organisms depart from this rule by exhibiting several tens of centromeres per chromosome. Telomeres are highly repeated sequences found at the ends of chromosomes to

G, H

I, L

M, N

P

R, S

T, V