Head Office: Università degli studi di Padova

Department: Department of Molecular Medicine

Ph.D Course in Molecular Medicine

DEVELOPMENT AND APPLICATION OF INFORMATICS TOOLS FOR THE DETECTION AND ANALYSIS OF NON-CANONICAL DNA STRUCTURES.

Coordinator: Prof. Stefano Piccolo

Supervisor: Prof. Stefano Toppo

Co-supervisor: Dott. Enrico Lavezzo

Ph.D Student: Michele Berselli

SUMMARY

ABSTRACT

The DNA is a flexible and heterogeneous molecule that can adopt different local conformations alternative to the classical double-helix. These non-canonical structures are known as non-B DNAs. These conformers appear to play an important role in different physiological and pathological cellular conditions and influence many biochemical properties of the genome. The formation of these structures is dependent upon specific features of the DNA sequence and different patterns may lead to the formation of different non-B DNAs. Due to lack of updated and flexible computational methods, during these years I focused my work on the development of new tools for the detection of some of these patterns at a genome-wide scale. Particularly, I focused on the detection of patterns that are degenerate. For this task, I developed NeSSie and QPARSE. NeSSie efficiently and exhaustively detects sequences with symmetrical properties, such as mirrors and palindromes that are associated to the formation of hairpins, cruciforms, and triple-stranded DNA. QPARSE detects consecutive exact or degenerate runs of Gs (G-islands) that are involved in the formation of G-quadruplex (G4) and paired G-quadruplex structures, i.e. two quadruplex structures that are close to each other along the sequence and that can fold cooperatively interacting into a higher-order structure. Eventually, I started using these tools to perform analyses on *Mycobacterium spp.* and human genomes. In the genomes of *Mycobacterium spp.* that are capable of developing tuberculosis-like diseases, NeSSie revealed the enrichment of a pattern with perfect mirror properties. Experimental analyses confirmed that the pattern can fold into a previously unknown but very stable hairpin structure. In the human genome, I focused on the detection of paired G-quadruplex systems. A genome-wide analysis revealed a striking enrichment of sequences potentially involved in the formation of paired G4 systems in correspondence of the TSS (Transcription Starting Site) of thousands of

human genes. Among the predicted systems, one has been detected in correspondence of BCL2 TSS and ongoing experimental validations suggest a cooperative folding of the two G-quadruplex structures. These results contribute to the idea that non-B DNAs can play important functional and potentially structural roles. They also suggest that the folding landscape of the DNA molecule is much more complex than previously assumed, and we have a huge lack of knowledge towards the alternative structures that can form in DNA. Following these evidences, the DNA sequence needs to be widely re-evaluated considering also its structural properties addressing efforts both at computational and experimental validation levels.

ABSTRACT (ITALIAN)

La doppia elica del DNA è una molecola molto flessibile ed eterogenea, che può adottare una vasta gamma di conformazioni locali alternative. Queste conformazioni vengono collettivamente chiamate non-B DNA. Questi conformeri sembrano svolgere un ruolo importante in diverse condizioni cellulari sia fisiologiche che patologiche, ed influenzano molte proprietà biochimiche del genoma. La formazione di queste strutture dipende da caratteristiche specifiche della sequenza del DNA, e diversi motivi di sequenza possono portare alla formazione di diverse strutture non-B DNA. Durante questi anni, ho concentrato il mio lavoro sullo sviluppo di nuovi strumenti computazionali per la rilevazione di alcuni di questi motivi su scala genomica. Questo investimento di tempo è stato necessario, poiché attualmente mancano strumenti sufficientemente flessibili in grado di eseguire tali analisi. In particolare, mi sono concentrato sul rilevamento di motivi degenerati. A tale scopo, ho sviluppato NeSSie e QPARSE. NeSSie è in grado di rilevare in modo efficiente ed esauriente sequenze con proprietà simmetriche, come motivi speculari e palindromici associati alla formazione di forcine, strutture cruciformi e regioni di DNA a triplo filamento. QPARSE può rilevare ripetizioni consecutive di isole di G esatte o degenerate, che sono coinvolte nella formazione di G-quadruplex (G4) e strutture G-quadruplex appaiate (cioè due strutture quadruplex che si trovano vicine lungo la sequenza e che possono interagire formando una struttura di ordine superiore ed influenzandosi reciprocamente nel ripiegamento). Ho quindi iniziato a utilizzare questi strumenti per eseguire analisi su genomi appartenenti a specie di micobatterio e sul genoma umano. Nei genomi delle specie di micobatteri che sono in grado di sviluppare malattie simili alla tubercolosi, NeSSie ha rivelato l'arricchimento di un motivo con una perfetta simmetria a specchio. Analisi sperimentali hanno quindi confermato che questo motivo può piegarsi in una struttura a forcina precedentemente sconosciuta ma molto stabile. Nel genoma

umano, mi sono concentrato sul rilevamento di sistemi G-quadruplex accoppiati. Una analisi su tutto il genoma ha rivelato un sorprendente arricchimento di sequenze potenzialmente coinvolte nella formazione di questi sistemi in corrispondenza del TSS (Sito di inizio della trascrizione) di migliaia di geni umani. Tra i sistemi predetti, uno identificato in corrispondenza del TSS di BCL2 è in corso di validazione sperimentale e i risultati preliminari sono promettenti. Questi risultati contribuiscono all'idea che i non-B DNA possano svolgere importanti ruoli funzionali e potenzialmente strutturali. Suggeriscono anche che il panorama di strutture che possono formarsi nella molecola di DNA sia molto più complesso di quanto ipotizzato, e che abbiamo ancora un'enorme mancanza di conoscenza verso queste strutture alternative. Seguendo queste evidenze, la sequenza del DNA deve essere ampiamente rivalutata non solo dal punto di vista della codifica, ma considerando anche le sue proprietà strutturali e funzionali. È quindi necessario indirizzare gli sforzi verso nuovi campi di indagine, studiando e caratterizzando queste strutture a livello genomico.

INTRODUCTION

OVERVIEW OF GENOME STRUCTURE

In the last years, a growing number of publications and evidences revealed a non-random and tightly regulated structure for the human genome. According to the data, the chromosomes three-dimensional structure and positioning are actively regulated and concur to the regulation of many different downstream processes such as replication, transcription and chromosome segregation. It is becoming clear that the genome is not only the storage location for the genetic information, but it has a fundamental role in governing the accessibility to this information. Interestingly, a common architectural pattern is emerging and it appears to be conserved throughout the evolution (Badrinarayanan et al., 2015; Fraser et al., 2015). The first insight on chromosomes positioning came from microscopy studies that revealed a spatial organization for the interphase nucleus with individual chromosomes residing in distinct chromosomal territories (Figure 1). This organization is reproducible and it is maintained in different cells with a limited intermingling, leading to a separation between active and inactive regions of chromatin inside the nucleus. Inactive regions are often found in the proximity of the nuclear envelope, whereas active chromatin is generally located in a more internal position within the nucleus (Cremer et al., 2015). More recently, capture-based studies revealed more on the chromatin structure. These techniques allow to quantify the interactions between genomic loci in the three-dimensional space (Dekker et al., 2013). Based on interaction frequencies and epigenetic marks, several layers of chromosome organization have been identified over different length scales. Chromosomes can be roughly partitioned into megabase-sized compartments of active and inactive chromatin (Lieberman-Aiden et al., 2009) (Figure 1), that can further be divided into sub-megabase domains

called topological associated domains or TADs (Dixon et al., 2012; Nora et al., 2012; Ramírez et al., 2018; Sexton et al., 2012) (Figure 1). These domains appear to partition the genomes into distinct structural units, that can remain spatially distant even if they are next to each other along the chromosomes. The borders of these domains are enriched with actively transcribed genes and architectural proteins, and longer range interactions are described as well, even between different chromosomes. Additionally, a sub-TADs level of organization is also emerging given by loop regions that can form inside TADs or at the TADs boundaries (Rao et al., 2014) (Figure 1). How this complex structure is achieved is currently far from being known, however, different mechanisms and actors appear to be responsible for the formation of the different layers of compartments and TADs-loops, respectively (Schwarzer et al., 2017). This can possibly lead to the different behavior observed for these organization patterns, with compartments being more stable and conserved across cell types with respect to TADs, that are more variable and heterogeneous in size from one cell to another (from few Kilobases up to few Megabases) (Stevens et al., 2017).

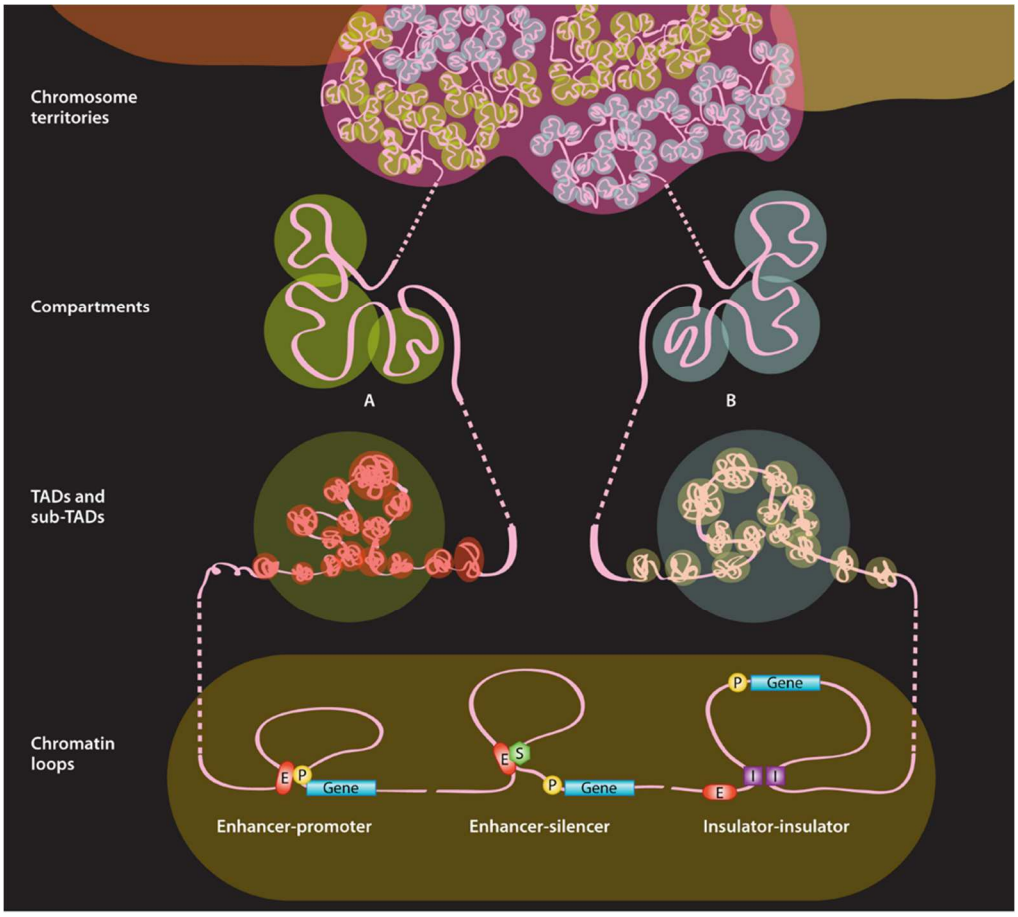


Figure 1. Chromatin organization at different genomic scales. Conformations are presented from lower (top) to higher (bottom) resolution. The chromatin fiber is shown in pink. At the very top level different chromosome territories are shown. Each chromosome can be further divided into two different compartments (A and B) with a distinct nature. At megabase scale each compartment can be partitioned into self-interacting regions (Topological associated domains or TADs). At the bottom a sub-TADs level of organization is shown as different looping interactions. Adapted from Fraser et al., 2015 (Copyright © 2015, American Society for Microbiology).

COMPARTMENTS

Compartments are defined as genomic regions of the same chromatin type characterized by a strong enrichment in reciprocal interactions compared to interactions with the other compartments. Currently, two broad compartments have been identified that correspond to active euchromatic regions (A compartment) and repressed heterochromatic regions (B compartment), respectively (Lieberman-Aiden et al., 2009). Interestingly, compartments appear to reflect and arise because of a precise nuclear organization. Chromatin distribution is polarized inside the

nucleus and the heterochromatic regions preferentially associate with the nuclear lamina (LADs) (Meuleman et al., 2013; Peric-Hupkes et al., 2010) or the nucleolar region (NADs) (van Koningsbruggen et al., 2010). This leads to a very precise nuclear distribution of chromatin that is characterized by an outer B compartment ring close to the nuclear periphery, an internal B compartment region around the nucleoli and an inner A compartment ring in between (Stevens et al., 2017). Compartments represent a stable and conserved organization but the mechanisms through which they are generated and maintained are currently unknown. Although initially considered a manifestation of TADs on a higher scale, recent evidences suggest that these structural entities are very different from each other and independent mechanisms are involved in their formation. Schwarzer et al. (Schwarzer et al., 2017) recently demonstrated that compartments formation is independent of cohesin, while cohesin is necessary for TADs formation. Interestingly, not only compartments and TADs originate from different mechanisms but probably represent orthogonal and possibly antagonistic types of chromatin organization. The loss of TADs, due to impairment in cohesin loading on the DNA, not only does not affect compartments organization but enhances and strengthens intra-compartment interactions. The loss of TADs also reveals the emergence of a finer organization for compartments, where shorter compartmental intervals of B-like regions appear inside the A regions. Interestingly, this finer compartmentalization better reflects the local transcriptional activity and the chromatin state, compared to the map obtained in the presence of TADs. Together, these observations suggest that the TADs can span regions intrinsically belonging to opposing compartments and antagonize the intrinsic tendency of chromatin to segregate into compartments by bringing together regions with opposite chromatin states. They also showed that TADs are not directly correlated with the epigenetic state as compartments are. Recently, Nothjunge et al. (Nothjunge et al., 2017) demonstrated that the establishment of compartments precedes and

determines DNA methylation during the maturation of cardiac myocytes. Compartments are established early in embryonic stem cells-progenitors and the impairment or reduction of DNA methylation in these cells appear not to impact the overall compartments organization, with methylation being established in preformed compartments as a continuous process until maturation to adult cardiac myocytes.

TOPOLOGICAL ASSOCIATED DOMAINS (TADS)

Topological associated domains were initially revealed by Hi-C analyses and are identified as genomic regions which preferentially self-interact and are insulated from the surrounding regions. Hi-C is a technique that allows to probe the three-dimensional architecture of whole genomes by coupling proximity-based ligation with massively parallel sequencing (Lieberman-Aiden et al., 2009). TADs are not as stable and conserved as the compartments and can be very heterogeneous in size, ranging from few Kilobases up to few Megabases in length (Dixon et al., 2012; Nora et al., 2012; Ramírez et al., 2018; Sexton et al., 2012). Except this definition, little is known about TADs and several hypotheses are on hold to explain their role and the mechanisms underlying their formation. According to the most recent model, a TAD originates through the formation of a long-range closed loop in DNA that acts as a physical barrier insulating the enclosed region from the surroundings. This prevents the formation of cross-interactions with other regions outside the loop or regions enclosed in different loops. The loop formation also contributes to increasing the spatial vicinity between different regions of the enclosed domain thus favoring self-interactions within the loop. Interestingly, this mechanism of compartmentalization through looping is not only associated with TADs but is emerging as a more general mechanism that is widely used by insulator elements. Through the formation of closed loops, insulators can modulate the interactions between regulatory elements by separating them into different loops or holding them together in the same loop. Interactions will be favored between elements residing within the same loop and inhibited

between elements residing in different loops (Doyle et al., 2014; Geyer and Corces, 1992; Muravyova et al., 2001; Ong and Corces, 2014; Udvardy et al., 1985). Despite the apparent generality of the mechanism, the actors involved are heterogeneous and very dependent on the organism, and it is not clear yet how relevant their role is in the formation of the topological associated domains (Phillips-Cremins and Corces, 2013). Recently, lot of attention and credit has been attributed to the CCCTC-binding factor (or CTCF), and particularly, this insulator protein has been widely described as the most important factor involved in the formation of TADs in eukaryotic genomes (Dixon et al., 2012; Ong and Corces, 2014; Phillips-Cremins and Corces, 2013; Zuin et al., 2014). Indeed, in mammalian chromosomes CTCF binding sites are often found at the boundaries of the topological associated domains (Dixon et al., 2012; Rao et al., 2014), and Zuin et al. (Zuin et al., 2014) demonstrated that CTCF is necessary for the formation of TADs. In their model, the depletion of CTCF leads to the disruption of the organization of TADs with an increase in interactions between domains and a decrease in interactions within domains. However, the situation is proving to be far more complex than initially pictured and conflicting evidences are emerging. For example, Schwarzer et al. (Schwarzer et al., 2017) couldn't find any correlation between CTCF and TADs and described a model where the depletion of cohesin leads instead to the disruption of the organization of TADs, with no loss of or changes in CTCF occupancy. Although apparently in contradiction, if considered together these results can probably provide a more realistic picture of the situation. Since the two studies use very different models and approaches, it is reasonable to hypothesize that they capture different moments in the life cycle of cells and different cellular conditions. This suggests that while CTCF may have a relevant involvement in TADs formation at some point, it can probably be replaced or compensated for by other actors in different moments of the cell life or under different conditions, being not essential anymore. Indeed, in *Drosophila melanogaster* other proteins different than CTCF appear to

be far more relevant for the formation of the topological associated domains (Cattoni et al., 2017; Phillips-Cremins and Corces, 2013; Ramírez et al., 2018). In a large study on TADs in *Drosophila*, Ramirez et al. (Ramírez et al., 2018) demonstrated that eight motifs (6 of which associated to known proteins) are enriched at the TADs boundaries, and could correlate different combinations of motifs with the strength and the role of the boundaries. Interestingly, the motifs associated with Beaf-32 and M1BP are the most abundant while the motif associated with CTCF is the less enriched among the detected motifs. This suggests that while CTCF may have a role in the formation of the topological associated domains, it is not working alone and we need to widen our search and consider multiple actors, some of which are probably currently unknown, to fully understand how TADs are created, regulated, and maintained. On top of that, TADs-like structures have been also demonstrated in plants and bacteria that lack CTCF homologs. This emphasizes the idea that despite the generality of looping as a regulatory and structural mechanism that leads to the formation of insulated and self-interacting regions, different organisms may use very different strategies and actors to reach the same goal. From a functional perspective, the biological role of the TADs is currently elusive. The main hypothesis is that these insulated regions are important for the regulation of the expression of genes. By partitioning the genome into self-interacting domains, TADs can define and guide the proper interactions between regulatory elements and their target promoters. It has also been suggested that TADs formation can drive and coordinate the co-expression of the genes residing within the same domains. A structural role is also possible and the TADs can be important for the three-dimensional folding of the chromatin inside the nucleus. However, there is a lack of strong evidences and the available data are often contradictory, leaving the functional and structural implications of TADs obscure.

LOOPS

Hi-C data also revealed that a subset of shorter domains can form local loops at a sub-TADs scale. These loops can form inside the TADs or in the correspondence of their boundaries (Rao et al., 2014). As for the TADs, their function is currently not fully understood but the mechanisms involved in their formation appear to be similar. The hypothesis is that, like the TADs, they represent a mechanism of compartmentalization just on a finer scale. Likewise, the actors involved in their formation are heterogeneous and dependent on the organisms. Recently, Cattoni et al. (Cattoni et al., 2017) proposed an interesting hypothesis that shifts the attention toward these shorter looping domains as the basic units of chromosome folding in *Drosophila* and possibly in other metazoans. Their study suggests that the TADs in *Drosophila* are not generated as stable long-range loops, as it appears to be in mammals, but are the result of the combination of a multitude of local intra-TADs contacts. According to these evidences, it is tempting to hypothesize that these shorter and more dynamic loops are indeed these contacts and represent the sub-modules that progressively interact and sum up to define the final TADs organization.

FROM 3D GENOME STRUCTURE TO DNA SEQUENCE

Despite the emerging evidences on the three-dimensional structure of genomes, little is currently known about the mechanisms truly responsible for the development, maintenance, and regulation of such a complex scheme. Its biological implications are equally obscure. However, while proteins have been the focus of the investigation, the DNA and the RNAs have been mostly overlooked even though they represent a relevant piece of the puzzle. These molecules can represent a great source of information and can provide new insights toward the comprehension of the three-dimensional folding of the genome.

3D GENOME FOLDING

The emergence of a common architectural pattern that is conserved throughout the evolution strongly suggests that a general and global mechanism is underneath the three-dimensional folding of genomes. DNA binding proteins and, to a lesser degree, non-coding RNAs have been investigated as the main determinants for the process. However, none of these elements alone holds as the '*primum movens*' and is sufficient to guide and regulate the whole process. Both these molecules are organism-specific and lack universality and conservation. Moreover, the way these molecules recognize and interact with the DNA hardly reconciles with such a complex and tightly regulated process. DNA-binding proteins often recognize small sequence patterns that are widespread in the genome but bind selectively only to a subset of these patterns lacking a general recognition rule. Non-coding RNAs bind instead to several different sites through unknown mechanisms other than complementary pairing, since the sequence similarity between the targeted regions is often missing. A mechanism based on the collaboration of different actors seems therefore more reasonable. Such a mechanism should also consider the involvement of the DNA that, although being completely overlooked so far, represents the most relevant piece of the puzzle. Given that neither DNA nor proteins or RNAs are capable of reasoning over a strategy to build the final genome structure, a self-assembly process is more plausible. In this scenario, a disordered system evolves through subsequent steps until an organized structure is reached in a process that is driven by energetic advantages that lead to specific local interactions among the different players. To some extent, the folding of the genome could mirror the folding of proteins (Anfinsen, 1973). This process follows a "divide and conquer" approach towards the energetic stabilization. Short regions (micro-domains) spontaneously assume local structures that are only marginally stable (e.g. alpha-helices, beta-strands). These structures, that represent the 'seeds' of the folding, diffuse under the action of internal and external forces and

collide one against the other to form higher aggregates up to the final protein (Karplus and Weaver, 1976, 1994). The whole process is initially governed by the properties of the local structures and by their interactions rather than by the single amino acids. Eventually, long-range interactions between residues which are far apart in the linear sequence intervene to stabilize both the micro-domains and the whole molecules. The process is assisted by other elements like proteins and RNAs. Since a sub-diffusive motion was reported also for prokaryotic and eukaryotic chromosomes (Albert et al., 2012; Weber et al., 2010), a model like the one developed for protein folding can reasonably hold for genomes as well. Analogously to protein folding, that is driven by transient structures that form locally, the structural organization of genomes could be propelled by the local formation of secondary structures in DNA. These structures represent the 'seeds' of the folding in the proposed model and are the driving force for the formation of higher-order structures through direct DNA-DNA interaction or by recruiting other factors such as DNA-binding proteins or non-coding RNAs (Figure 2). Following this idea, the initial force leading to the three-dimensional folding of the genome is encoded in the primary DNA sequence as structural motifs that form alternative structures in DNA.

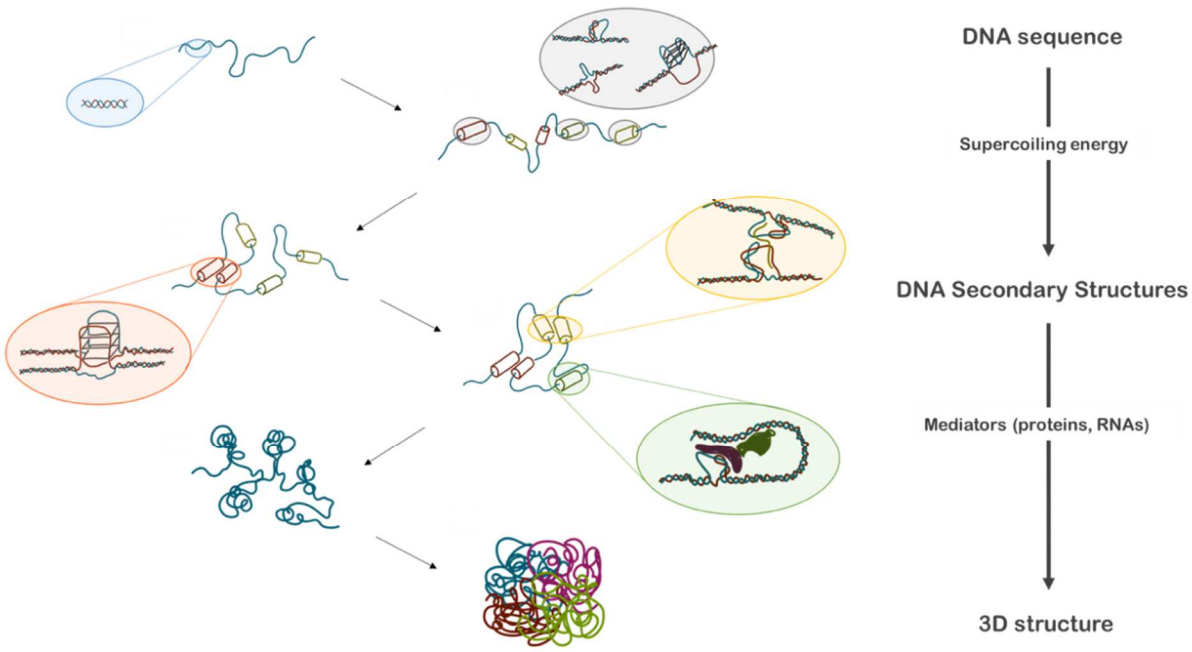


Figure 2. Self-assembly model proposed for the three-dimensional folding of the genome.

THE DNA ALTERNATIVE CONFORMATIONS

The DNA double-helix is a very flexible and heterogeneous molecule that can adopt a wide range of alternative local conformations that are collectively named non-B DNAs (Figure 3). These conformers usually form in regions that contain distinct features at the primary sequence level. For example, regions with symmetrical properties (e.g. palindromic and mirror symmetries) or different degrees of repetition (e.g. direct repeats), as well as regions with a bias in base composition, are usually associated with the formation of such structures (Kouzine et al., 2017; Ohyama, 2005; Sinden, 1994) (Figure 4). Some of these non-B DNAs have been also widely described in RNA. Interestingly, the non-B DNAs appear to play an important role in different physiological and pathological cellular conditions and influence many biochemical properties of the genome.

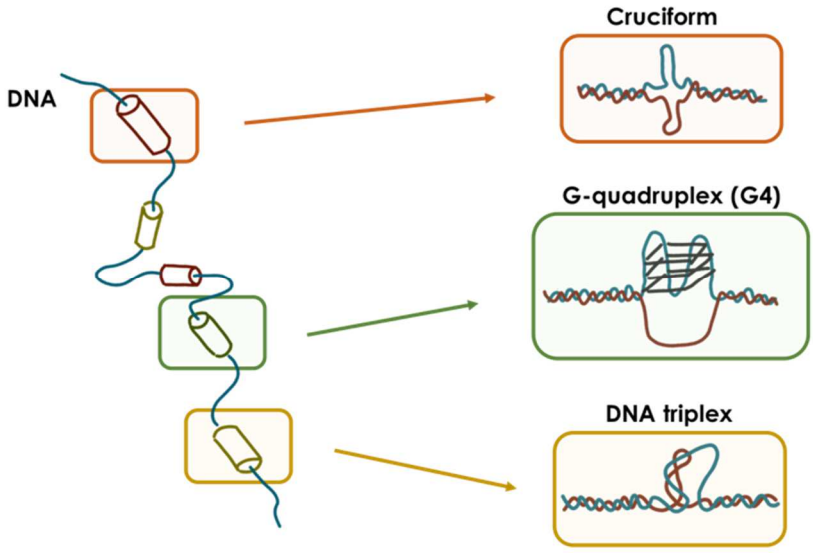


Figure 3. Non-B DNA structures.

HAIRPINS AND CRUCIFORM STRUCTURES

A hairpin is a stem-loop structure that forms in a single-stranded DNA (or RNA) when two regions within the same strand base-pair to form a double-helix that ends in an unpaired loop. In double-stranded DNA this leads to the formation of a cruciform structure with two hairpins that form specularly in the opposite strands (Figure 4). As far as it is known, a cruciform can form in a region with a palindromic symmetry (i.e. inverted repeat, Figure 4) and the opening of the 8-10 bp corresponding to the center of symmetry triggers the subsequent event of nucleation of intra-strand base pairing. The process is driven by the presence of a critical level of DNA supercoiling that is necessary to open the double-helix at the center of symmetry. Regions enriched in A + T melt more easily when compared to G + C rich regions and have a faster rate of cruciform structures formation. This is due to lower stacking energies of the A + T pairs and the formation of fewer hydrogen bonds. The relevance of hairpins and cruciform structures is mostly elusive, but the prevalence of inverted repeats in bacterial, eukaryotic, and viral DNAs suggest a biological role. Palindromes are

frequently associated with the origins of DNA replication together with other structures known as DNA unwinding elements. To this regard, a glimpse of a possible function for cruciform structures is provided by the plasmid pT181. This plasmid contains a small palindromic sequence that can form a cruciform *in* vivo acting as the origin of replication (Noirot et al., 1990). In supercoiled DNA, the initiation protein RepC recognizes and binds (possibly stabilizing it) the cruciform and introduces a specific nick in the loop region to create a starting point for replication (a replication primer with a free 3'-OH end). Another example is provided by the bacteriophage N4. For the early transcription, the phage requires a phage-encoded RNA polymerase (that is packaged within the virion) and a supercoiled DNA template on which the bacterial single-strand DNA-binding protein (SSB protein) acts as a transcriptional activator (Markiewicz et al., 1992). Interestingly, the early gene promoters contain palindromic sequences that are required for transcription, and the symmetry is more important than the specific base sequence (Glucksmann et al., 1992). According to the model, the DNA is supercoiled by the DNA gyrase that produces a single-stranded region triggering the formation of a cruciform structure at the inverted repeat (or stem-loop structures in unwound DNA). The SSB protein then binds and stabilizes the cruciform and facilitate the binding of the N4 RNA polymerase that starts the transcription. As further examples, cruciform structures are also involved in Holliday recombination (Duckett et al., 1988) and seem partially responsible for the genetic instability of long inverted repeats (>150 bp, slippage during replication due to cruciform formation) (Albertini et al., 1982; Warren and Green, 1985). It is noteworthy, that also in the linear form inverted repeats may play an important biological role as recognition sites. Indeed, many palindromes (ranging in size from 4 to 20 bp) are recognized as binding sites by specific dimeric proteins (e.g. restriction enzymes and methylases). Interestingly, also quasi-inverted repeats (i.e. degenerate palindromes) are recognized as possible binding sites (e.g. binding of dimeric repressors). Therefore, it is mandatory to detect

and study these regions genome-wide to better understand their involvement in possible regulatory and structural mechanisms.

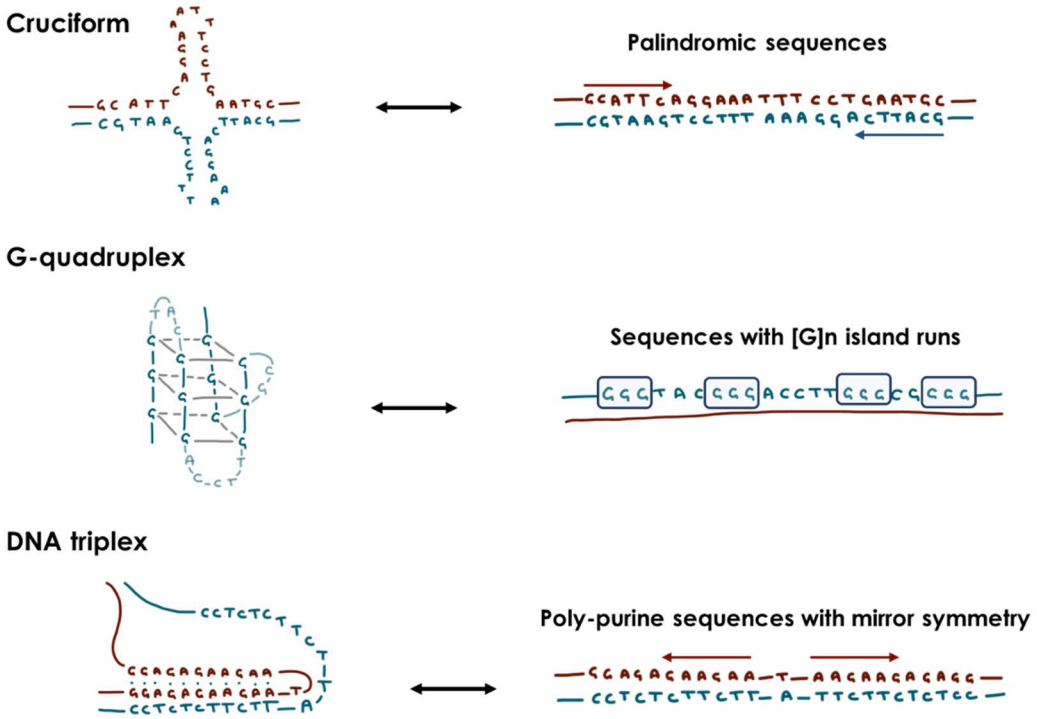


Figure 4. Formation of non-B DNA structures is dependent on distinct features and patterns at the primary sequence level.

DOUBLE-STRANDED DNA ALTERNATIVES CONFORMATIONS

B-DNA: The B-form (Figure 5.a) is the most common and stable structure of the DNA. It is a right-handed helix with a fixed step of approximately 10.5 bp per each helical turn. The main feature of this form is the presence of two distinct grooves, a major and a minor groove. This provides very distinct surfaces of interaction for proteins as different functional groups on the bases are accessible from the different grooves. Indeed, different DNA binding proteins (as well as certain chemicals and drugs) have domains that interact specifically with either the major or the minor groove. Another feature is that in this structure the Watson-Crick hydrogen bonding surfaces

are not available to solvent or proteins since they interact with each other in the complementary base-pairing. This makes the center of the double-helix a relatively safe and chemically inert place to store genetic information.

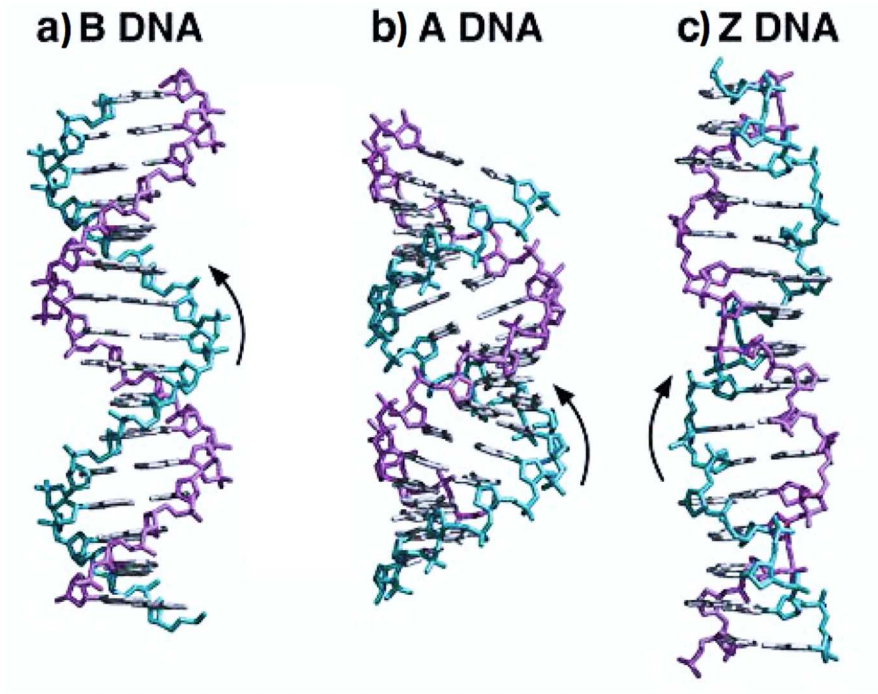


Figure 5. Double stranded DNA alternatives conformations. a) B-form, b) A-form, c) Z-form.

A-DNA: The A-form (Figure 5.b) is an alternative conformation of the right-handed DNA. Compared to the B-form the turns are longer (approximately 11 bp for each turn), the bases are much more tilted (approximately 20°), and the grooves are not as deep. Another significant difference is that the sugar pucker is C3' endo compared with the C2' endo in B-DNA. In biological systems, this alternative conformation appears to form in correspondence of sequences characterized by homopurine • homopyrimidine runs [e.g. poly(dG) • poly(dC)]. Therefore, within a generally B-like DNA molecule, specific regions may exist in an A-DNA form depending on the bases composition of the sequence. RNA also frequently

exists in a double-helical conformation (e.g. tRNA, rRNA, and parts of mRNA) that can form A-like regions as well, with the distinguishing ribose configuration (C3' endo).

C-, D- and T-DNA: Numerous other subtle variations in the shape of the DNA double-helix have been described and, in specialized situations, may have biological relevance. C-DNA has 9.3 bp per turn. D-DNA is a structure with a helix repeat of 8.5 bp per turn and is believed to originate from runs of poly(dA) • poly(dT). T-DNA has a helix repeat of 8 bp per turn and is described in bacteriophages T2, T4, and T6. T-DNA is quite different from most DNA and the cytosine residues contain a hydroxymethyl on the 5' carbon. In addition to this modification, a glucose residue is added to the hydroxymethyl, making glucosylated DNA. These changes reflect in the different shape of the helix of bacteriophage T4 DNA.

Z-DNA: The Z-form of the DNA (Figure 5.c) is a left-handed helix that differs from the other variants that have been described so far. It can form in regions of alternating purine-pyrimidine bases [e.g. poly(dG-dC) • poly(dC-dG)] when specific conditions are met. Its formation generally requires a high salt-concentration, the presence of certain divalent cations, and a sufficient level of DNA supercoiling. As for the sequence requirements, the $(GC)_n$ sequence is the one most likely to form the Z-DNA. The $(GT)_n$ sequence can also form the Z-DNA but it requires a higher stabilization energy (higher levels of negative supercoiling). The $(AT)_n$ region generally does not form the Z-DNA since the formation of cruciform structures is more likely and generally prevails. The $(AT)_n$ can form Z-DNA only if a sequence longer than 10 bases is embedded within a $(GC)_n$ or $(GT)_n$ region and the conditions are appropriate (high negative supercoiling, high salt concentration, and $NiCl_2$ is present). As a rule, the longer is the region that forms the Z-DNA, the easiest is the transition from the B- to the Z-form. Except for the antiparallel nature of the two strands and the Watson-Crick hydrogen bonding, there are major structural differences if compared to the B-form. The phosphate backbone has a "zig-zag" structure and there is

no distinction between a major and minor groove. The major groove is a nearly flat surface and the only visible groove, that is structurally analogous to the minor groove, is deep and narrow. The helix is narrower in diameter and has wider turns with approximately 12 bp each. The relative position of the bases is also different. In the B-DNA the bases form a cylinder within the double helix, with the hydrogen bonds at the center of the helix. In the Z-DNA, the bases are instead positioned toward the outside of the helix so that neither the bases nor the hydrogen bonds overlap with the center of the helix. This leads to a structure that is more reactive since the bases are partially exposed to solvent and not protected as in the B-DNA conformation. The formation of a Z-DNA region within a larger B-form DNA molecule requires the formation of junctions between the different types of helices. The best estimation is that these junctions consist of a region of probably 3-4 unpaired or weakly paired bases, with no sharp transition from the left- to the right-handed helices but a region of a few base pairs that is partially unwound. The Z-DNA can exist readily in bacterial cells and, although sequences that can form Z-DNA are widely found in the human DNA, the biological role of the Z-DNA in eukaryotic cells is still elusive. From a regulatory and functional perspective, it is interesting that certain chemical modifications of DNA (e.g. the methylation of the N7 position of guanine) drive the equilibrium in favor of the formation of the Z-DNA. It seems likely that these modifications can be used as a regulation system to control the Z-DNA formation. Interestingly, proteins (e.g. Topoisomerase II) that bind specifically to the Z-DNA have been identified in many organisms including bacteria, fruit-flies and higher eukaryotes, and support for a biological role of this alternative DNA form (Sinden, 1994).

DNA TRIPLEXES

A DNA triplex (triple-stranded DNA) is an unwound non-B DNA conformation that forms when a single-stranded DNA (ssDNA) or RNA interacts within the major groove of a double-stranded DNA in the B-form (dsDNA). The result is a triple-helix structure (Figure 4, Figure 6). In this

conformation, the Watson-Crick base-pairing surfaces are locked in the complementary binding at the center of the dsDNA and are not available to further interactions. Therefore, the ssDNA forms hydrogen bonds with the other surfaces that are accessible through the major groove using a Hoogsteen base-pairing scheme (Figure 6). This creates a sequence constraint for the central strand of the triplex that must be composed of purine (Pu) bases. Pyrimidines (Py) do not have two hydrogen bonding surfaces to form more than one hydrogen bond at the time to sustain the formation of the triplex. According to the base-pairing scheme, a central G can pair with a C^+ (protonated cytosine) or a G, while a central A can pair with a T or an A (Figure 6). Depending on the origin of the ssDNA, the triple-stranded DNA can form as an intra-molecular structure (Figure 6) or as an inter-molecular one. Moreover, when the ssDNA comes from a dsDNA the formation of the triple-stranded structure leaves an unpaired strand that forms a single-stranded loop. While a continuous purine strand is the only requirement for the formation of inter-molecular triplexes, the formation of the intra-molecular ones additionally requires that the poly-Pu • poly-Py region contains a mirror symmetry (Figure 4, Figure 6). However, there can be exceptions to these rules and regions that contain some mismatches in the mirror symmetry or a mixed Pu-Py composition can still sustain triplex formation. Another exception is a quasi-mirror sequence containing G residues interspersed by As on one half and Ts on the other half (e.g. GGAGGGAGGGGA-TGGGGTGGGTGG). This sequence can form an intra-molecular triplex using the G•G•C and T•A•T base-pairing schemes. The formation of an intra-molecular triplex starts with the opening of the dsDNA at the center of symmetry that is followed by the nucleation of the triple-stranded structure and the formation of the Hoogsteen bonds. The process is driven by the presence of a critical level of DNA supercoiling that is necessary for the initial opening of the double-helix. A + T regions are less stable and the super-helical density required is proportional to the A + T content. As the A + T content increases, the negative super-helical density

required for the formation of the triplex decreases. Once the triplex is formed, the thermal stability is dependent on the G + C content. The higher is the G + C content the more thermally stable is the structure. As for the possible functional implications, it is noteworthy that homopurine • homopyrimidine regions are widespread in the eukaryotic DNA with a much higher frequency than expected by chance. Many of these sequences with the potential to form triplex structures are also commonly associated with regulatory regions suggesting a potential biological function (Kouzine et al., 2017). Indeed, growing evidences indicate that the formation of DNA triplexes may have a role in the regulation of transcription (Buske et al., 2011; Ohyama, 2005). The two principal mechanisms involved are the promoter occlusion and the elongation arrest (Lee et al., 1984; Praseuth et al., 1999). In the promoter occlusion, the formation of a DNA triplex interferes with the binding of a specific protein to the promoter region. If the occluded protein is a transcription activator, the transcription is inhibited. Vice versa, if the occluded protein is a transcriptional repressor, the transcription is enhanced. In the elongation arrest, the formation of a triple-stranded structure in the transcribed region impedes the movement of the RNA polymerase across that region. Despite the triplex alone cannot effectively impede the elongation, its formation can direct a subsequent covalent modification of the template (e.g. a cross-link or strand break) that renders it unsuitable for the elongation. Although other roles for the DNA triplexes remain more elusive, it is possible to hypothesize that these structures might take part in several other regulatory and structural mechanisms: (1) The formation of an intra-molecular triplex can influence the local and global structure of the DNA by regulating the level of DNA supercoiling in the surrounding topological domain. (2) The formation of a triplex can affect the nucleosome organization and positioning (nucleosome phasing) by creating a region that is not accessible for the nucleosome formation (Goobes et al., 2002; Ruan and Wang, 2008; Westin et al., 1995). (3) The transition between duplex and triplex conformations

can provide a molecular switch that determines whether a regulatory or structural protein can bind or not to the DNA (Lee et al., 1984). (4) The single-stranded loop generated by the formation of an intra-molecular triplex can provide a recognition site or an entry point for specific proteins (e.g. RNA polymerase) or RNAs (Zheng et al., 2010). (5) Potential intramolecular-triplex-forming sequences may act as replication terminators. It has been observed that replication in the Z-DNA-to-triplex direction occurs more easily than replication in the triplex-to-Z-DNA direction, suggesting that the Z-triplex DNA region can act as an "orientation-specific replication fork gate" (Brinton et al., 1991). (6) Intra-molecular triplexes can also play a role in the initiation of DNA replication by providing a stable unpaired strand that might represent an entry site for the proteins involved in the replication process. (7) Triplexes formation can contribute to the stability of the linear eukaryotic chromosomes. The repetitive telomere sequences can form a variety of triplex structures that can protect their ends providing stability (Veselkov et al., 1993). (8) Triplex structures can also represent anchorage points between different DNA molecules or mediate long-range interactions within the same molecule (Hampel et al., 1993). Recently, Brázdová et al. (Brázdová et al., 2016) demonstrated that the p53 protein can bind with a very high affinity to a T•A•T triple-stranded DNA structure. Using an *in-silico* approach, they further searched the human promoters for the p53 consensus sequence and the motifs associated with the formation of T•A•T triplexes. They identified a set of candidate genes that represent potential targets for p53 suggesting that a triplex-dependent mechanism can contribute to regulate their transcription. Cooney et al. (Cooney et al., 1988) reported another example of the modulation of gene expression through the formation of an inter-molecular triplex. Upstream of the human c-myc oncogene there is a poly-Pu • poly-Py region that is sensitive to the S1 nuclease and can form an intra-molecular triple-stranded structure *in vitro* (Boles and Hogan, 1987; Kinniburgh, 1989). This region interacts with

transcription factors including one ribonucleoprotein complex that can bind through the formation of an inter-molecular triplex or through the formation of an RNA-DNA hybrid to the unpaired strand generated by the formation of the intra-molecular triplex (Davis et al., 1989; Postel et al., 1989). The addition of an oligonucleotide designed to form an inter-molecular triplex structure with the purine region results in the inhibition of the transcription *in vitro*. Considered together, all these evidences reveal that indeed triplex-dependent mechanisms can contribute to the regulation of the transcription process. Possibly, triple-stranded structures can be determinant for the epigenetics and the three-dimensional structure of chromatin. It is therefore necessary to better detect and study these structures and their surrounding regions genome-wide to fully comprehend their biological role.

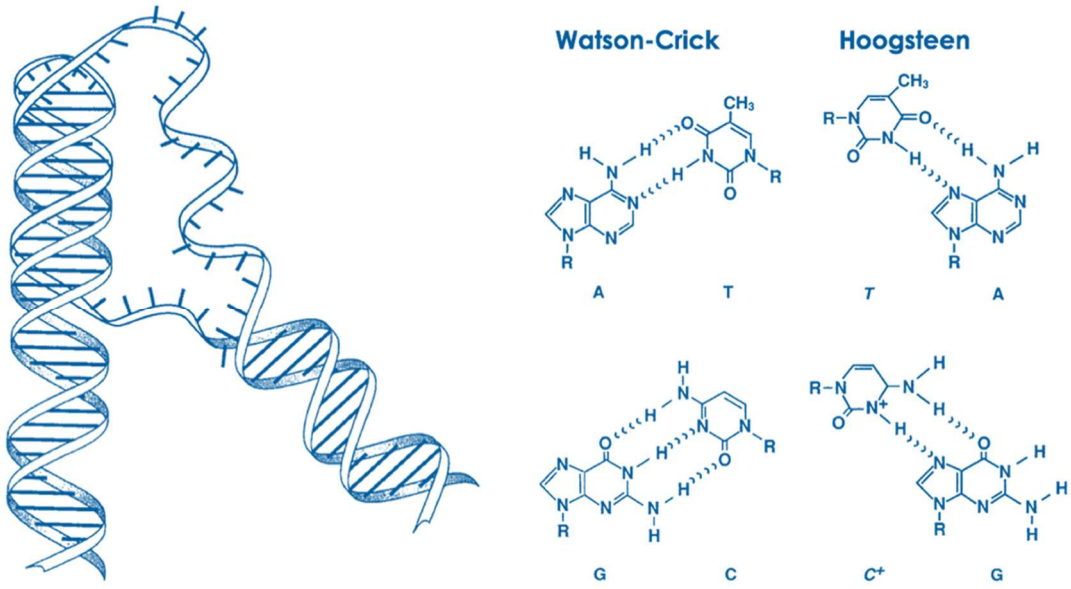


Figure 6. Intra-molecular DNA triplex and base pairings schemes (Watson-Crick on the left and Hoogsteen on the right).

G-QUADRUPLEX STRUCTURES

A G-quadruplex (G4) is a four-stranded non-canonical DNA structure that forms in single-stranded guanine-rich DNAs (or RNAs). Four guanine bases can arrange within a planar quartet (G-quartet, Figure 7) via Hoogsteen-type base-pairings and the stacking of at least two G-quartets leads to the formation of the G4 structure (Figure 4, Figure 7). The intervening sequences are extruded as single-strand loops and the structure is further stabilized by the presence of monovalent cations (e.g. K^+). G4s can form as intra-molecular or inter-molecular structures (e.g. two G-hairpins can interact to form a G-quadruplex, Figure 7) depending on whether the interacting strands come from the same or from different molecules. The formation of a G4 in a double-stranded DNA requires the local opening of the double-helix that is followed by the nucleation of the four-stranded structure in the G-rich strand. A complementary structure called i-motif may form in the opposite C-rich strand (Day et al., 2014; Zeraati et al., 2018). The process is driven by the presence of a critical level of DNA supercoiling that is necessary to initially melt the double-helix. Intra-molecular G4 structures can form in G-rich regions that share a common pattern for the distribution of the Gs ($G_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}N_xG_{\geq 2}$) (Figure 4). Basically, at least four continuous runs of guanines (G-islands) are necessary and each G-island should contain at least two Gs. The intervening sequences between the islands can be instead of variable length (generally maximum 7-12 bp depending on the length of the islands). However, there can be exceptions to these rules and G4 structures can still form in regions that diverge from the optimal pattern. For example, degenerate G-islands that contain gaps can still sustain the formation of G4 structures (Hon et al., 2017). G4s can form *in vivo* under physiological conditions (Lipps and Rhodes, 2009; Rhodes and Lipps, 2015) and more than 10,000 G4 structures have been recently experimentally validated in the human genome (Hänsel-Hertsch et al., 2016). These structures associate with definite genomic regions such as gene promoters, DNA replication origins, telomeres regions,

immunoglobulin switch regions and recombination sites (Maizels and Gray, 2013) and represent key functional and structural regulators (Lipps and Rhodes, 2009; Rhodes and Lipps, 2015). The importance of G4 structures *in vivo* is consolidated by the discovery of cellular proteins that specifically recognize and interact with the four-stranded DNAs (Qiu et al., 2015; Tosoni et al., 2015) by stabilizing or destabilizing their structure. G4s can regulate the transcription process. As a demonstration, targeting G4 structures with selective ligands alter the transcription of the genes that contain these structures in their promoters such as the oncogenes *KRAS* (Cogoi and Xodo, 2006) and *MYC* (Siddiqui-Jain et al., 2002). A recent study also demonstrates in zebrafish embryos that small molecules that specifically target the G4s in the promoters of developmental genes reduce their transcription and cause the degenerate associated phenotype (David et al., 2016). G4 DNA structures may also be involved in the initiation of the replication process. G4s have been predicted at the human replication origins (Besnard et al., 2012), and the human origin recognition complex (ORC) preferentially binds to the G4 structures that form in DNA (and RNA) sequences *in vitro* (Hoshina et al., 2013). In the absence of the DNA helicases, the stable G4 structures can interfere with the progression of the DNA polymerases. This can lead to replication stalling, DNA damages, and ultimately genomic instability (Mendoza et al., 2016). Indeed, Paeschke et al. (Paeschke et al., 2013) demonstrated that in *Saccharomyces cerevisiae* the helicases Pif1 and Rrm3 are essential for the suppression and the prevention of the damages induced in the DNA by the formation of G4 structures. Ligands that stabilize G4 structures can induce DNA breakages in human cells (Rodriguez et al., 2012) and are responsible for the formation of indels (insertions – deletions) at predicted G4s positions in yeast (Paeschke et al., 2013; Ribeyre et al., 2009). In mouse cells, the regulator of telomere elongation helicase 1 (RTEL1) can resolve G4 structures that form at the telomeres and it is important for the maintenance of their integrity (Vannier et al., 2012). G4s can form also in RNAs and, indeed, four-stranded

structures have been detected in different classes of RNAs such as messenger RNAs (mRNAs), non-coding RNAs (ncRNAs) (e.g. long non-coding RNAs or lncRNAs) (Jayaraj et al., 2012) and precursor microRNAs (pre-miRNAs) (Mirihana Arachchilage et al., 2015). This supports the potential role of the G4s in regulating the expression of genes both at the pre- and the post-transcriptional levels (Agarwala et al., 2015; Cammas and Millevoi, 2017). Finally, G4 structures that form in RNA are suggested to affect also DNA processes (Hirashima and Seimiya, 2015; Takahama et al., 2013; Zheng et al., 2010). Given the wide distribution and the broad panel of possible roles for the G4s, it is surprising how little we know and how scattered this knowledge is. It is therefore necessary to widen the study of these structures and extend the analysis to their surroundings (both in linear and three-dimensional space) to gain a deeper and more consistent knowledge of their biology. In doing so, it is important to consider the heterogeneity of the four-stranded DNAs and improve the detection methods to identify also those structures that form from degenerate patterns.

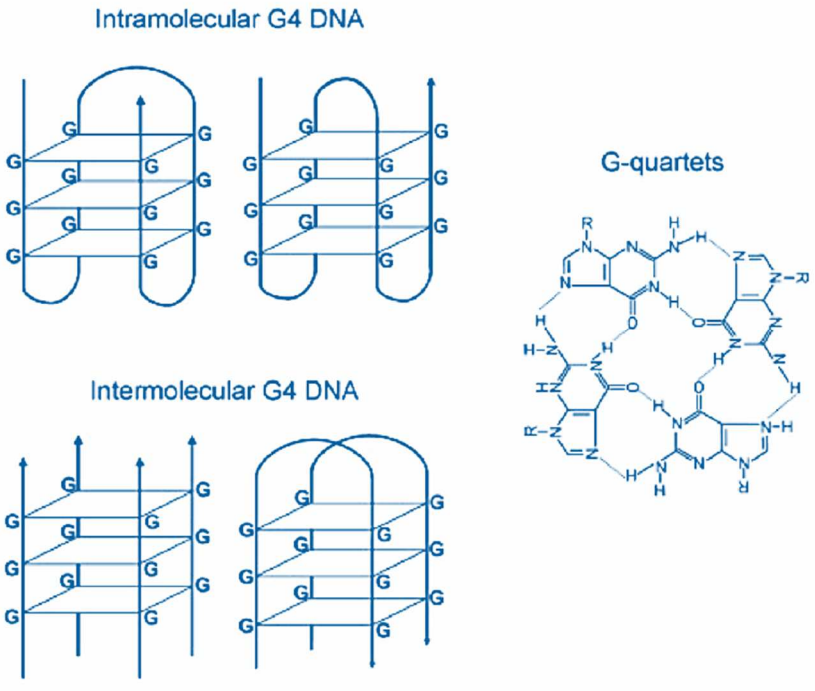


Figure 7. Examples of G-quadruplex structures (left) and G-quartets structure (right).

OTHER NON-B DNA STRUCTURES

DNA unwinding elements: The DNA unwinding elements (or DUEs) are A + T rich regions in DNA that acquire a single-stranded structure under supercoil tension. DUEs are 30-100 bp in length and have been identified both in prokaryotes and eukaryotes. DUEs are commonly associated with replication origins and are required for the initiation of the replication process at these sites. Indeed, the shortening and deletion of these regions result in the loss of replication origins (Umek and Kowalski, 1988).

Nodule DNA: The nodule DNA is a structure that forms between an intra-molecular DNA triplex and an inter-molecular one that is adjacent in the linear or in the three-dimensional space. Basically, the single-stranded loop that originates from the formation of the intra-molecular triplex acts as the third strand (ssDNA) in the formation of the adjacent inter-molecular triplex (Sinden, 1994). The biological significance of the nodule DNA is currently unknown but it is tempting to hypothesize a possible structural role. The formation of the nodule could represent an anchoring point between different DNA molecules or between distant regions within the same molecule.

Mirror motifs and G-hairpin: Despite the large number of motifs with a mirror symmetry scattered across the genome, their biological meaning is still mostly elusive. Those occurring in poly(Pu) regions have been related to the formation of triple-stranded structures in DNA, but no real meaning has been attributed to other mirror sequences. Only recently, Gajarský et al. (Gajarský et al., 2017) demonstrated that the short mirror sequence d(GTGTGGGTGTG) can fold into a particular structure defined as a G-hairpin. This suggests that this class of motifs may have a primary role as a determinant for the formation of alternative non-B DNA conformations, most of which are probably currently unknown. This also demonstrates that the folding landscape of short DNA single strands is much more complex

than previously assumed and needs to be further investigated (Satange et al., 2018).

AIM

Given the complexity of the three-dimensional structure of genomes, and considering that the biological molecules are not capable of reasoning over a strategy to build such a complex architecture, a self-assembly process for the chromatin seems more reasonable. Like the folding of proteins, the folding of chromatin can be imagined as a multi-step process driven by energetic advantages. The process is driven by the local formation of transient structures that diffuse and collide to progressively originate bigger and more stable aggregates up to the final structure. The hypothesis is that the structures that form locally and drive this self-assembly process are the alternative conformers that can form in the DNA (non-B DNAs). Indeed, the DNA is a very flexible molecule and emerging evidences show its ability to form a wide-spectrum of alternative conformations. These structures have been demonstrated to interact directly and/or to recruit proteins and RNAs. Evidences also suggest that they act as potential functional and structural regulators. Following this idea, it is reasonable to hypothesize that the initial force leading to the three-dimensional folding of the genome is encoded in the primary DNA sequence as structural motifs that can form non-B DNAs. A first parallel between proteins and DNA goes back to Kajava (Kajava, 2001), that suggested a parallel between the repeats in proteins and DNA. The idea was lately refined (Kajava, 2012) to consider for the appearance of new three-dimensional structures. In proteins, different types of repeats are associated to the formation of different types of structures that are used to classify the proteins into different classes. The author suggests a direct correspondence between these repeats and those that appear in the DNA. The class I crystalline structures correspond to microsatellites repeats,

class II and III structures are formed by minisatellites, and class IV and V repeats correspond to satellites (Figure 8). This idea of a structural role for the repeats in DNA is supported by the observation that different types of repeats are indeed associated with the formation of alternative non-B DNAs. Considering these evidences, the DNA sequence needs to be widely re-evaluated not only for the encoded genetic information, but considering also its structural properties. Formation of the non-B DNAs is dependent on the primary DNA sequence and specific patterns are associated with the formation of specific structures. It is therefore justified to direct efforts towards new fields of investigation by studying and characterizing these patterns genome-wide. Moreover, other characteristics of the DNA sequence can be of interest and hide important information on the role of DNA. The DNA sequence is often characterized by unbalances in base composition and k-mers (words of size k) distribution. Since the DNA sequence is non-random and tightly shaped by the evolution, these unbalances from the random expectation are likely to reflect underlying functional and structural roles for the DNA molecule that are dependent on these sequence features. Given the size of genomes, performing such analyses is not trivial and a great amount of computational automatization is required. In the past, algorithms were developed for sequence/pattern search with limited resources and lack of both biological data and knowledge. Nowadays the scenario has completely changed and previously developed tools are inadequate or do not support genome-wide searches. Indeed, those few programs that are still maintained and currently available are mostly limited by the size of the datasets they can analyze or lack the functionality and flexibility that are required. In this scenario, I focused my work on the development of computational tools for the detection at a genomic scale of the exact and degenerate patterns that can potentially form alternative structures in DNA. This is a first step towards a more comprehensive characterization of

those structures that can form in DNA and that may play a role (structural or functional) in the context of the three-dimensional genome.

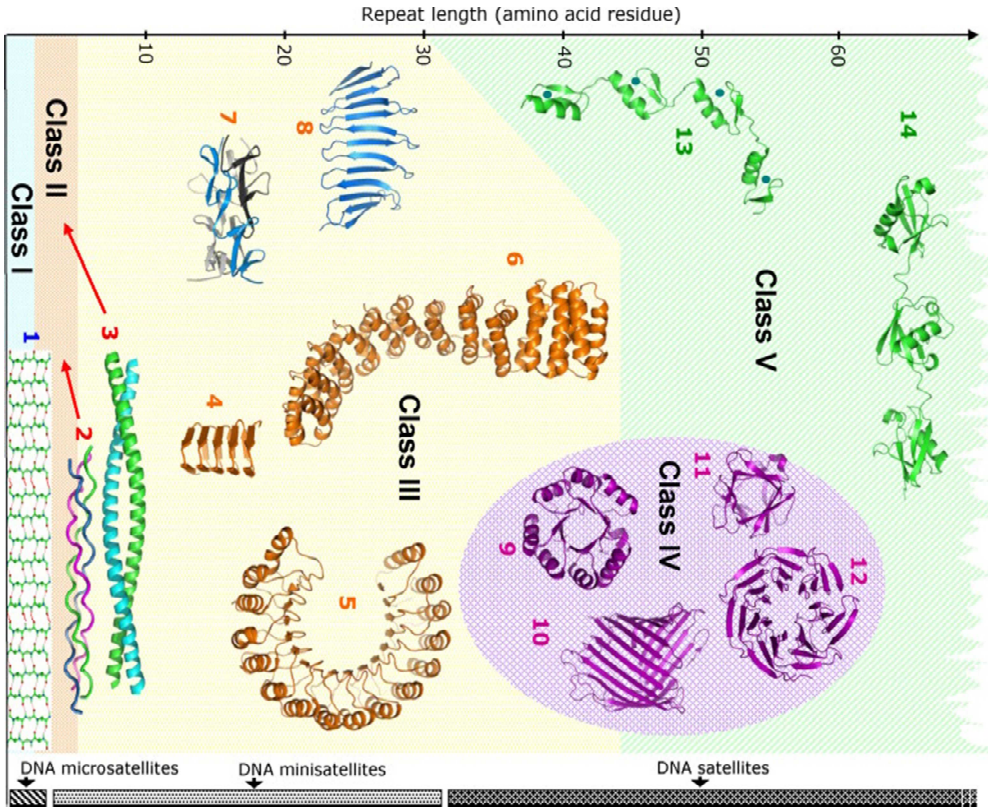


Figure 8. Structural classification of repeated proteins based on the length of their repeats and a parallel with repeats in DNA. Adapted from Kajava, 2012 (Copyright © 2011 Elsevier Inc. All rights reserved).

TOOLS AND DATABASES

NESSIE (NUCLEIC-ACIDS ELEMENTS OF SEQUENCE SYMMETRY IDENTIFICATION)

Formation of non-B DNA structures is mostly dependent on the primary sequence and the presence of specific patterns in DNA (see Introduction section). Different patterns may have different features that lead to the formation of different non-B DNAs. Interestingly, many non-B DNA structures show symmetrical properties such as palindrome and mirror sequences that can form hairpins, cruciform structures, and triplexes. By developing programs that recognize local symmetries in the DNA sequences, it is possible to automatically detect these structures. However, while searching for perfect symmetries is simple, searching for symmetries that are degenerate to a certain level (e.g. containing gaps or mismatches that impair the symmetry) is computationally expensive and challenging from an algorithmic perspective. To date, there is a lack of updated tools that can perform an exhaustive search of such motifs in a reasonable amount of time. The efficiency is important when the targets of the analysis are sequences representing whole genomes (millions to billions of bp). Trying to address the problem and compensate for the lack of such tools, I developed NeSSie (Berselli et al., 2018). NeSSie is based on dynamic programming and is implemented as a C/C++ 64-bit library and tool easily customizable. The tool can quickly scan for perfect and degenerate DNA palindromes, mirrors and potential triplex forming patterns. In addition, the tool can compute linguistic complexity and Shannon entropy measures to verify the repetitive nature of the DNA.

ALGORITHM

To detect symmetrical patterns, I implemented a strategy based on dynamic programming applied to sliding windows:

I. A sliding window of a fixed length, that is defined by the user, is used to progressively scan the input sequence shifting by one base at a time.

II. For each sliding window a global alignment approach is applied to search for potential symmetries. However, imperfection-tolerant palindromes and mirrors may contain mismatches, insertions, or deletions that impair the symmetry. This poses the problem of finding the best symmetry point that divides the sequence in two self-complementary halves when evaluating the best alignment. To solve the problem, I implemented a modified version of the Needleman-Wunsch algorithm for global alignment (Needleman and Wunsch, 1970), which can identify the best symmetry point and generate the optimal alignment between the two halves of the symmetric motif. The algorithm steps are the following:

   - The sequence and its inverted copy are placed in the horizontal and vertical axes of an alignment matrix respectively.
   - The alignment matrix is constructed according to Needleman-Wunsch algorithm.
   - Two different scoring matrices are used to test for the mirror and palindromic symmetry (Figure 11). The gap opening score is -1.
   - The backtracking to retrieve the optimal alignment starts from the highest scoring cell along the diagonal that connects top-right cell and bottom-left cell. This boundary imposes a constraint for the entire sequence to be aligned and identifies the best symmetry point for the sequence.

The example in Figure 9 shows the result of a mirror search in the sequence 'ATCAAGTTGCCA'. The scoring matrix used to build the alignment matrix is shown in Figure 11 (gap open -1). The best possible alignment (Figure 10) is obtained following the optimal path as shown by arrows in Figure 9, where the green cell with the highest score along the diagonal is the starting point of backtracking.

		A	T	C	A	A	G	T	T	G	C	C	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	
C	-2	0	0	1	0	-1	-2	-3	-4	-5	-6		
C	-3	-1	-1	1	0	-1	-2	-3	-4	-5			
G	-4	-2	-2	0	0	-1	0	-1	-2				
T	-5	-3	-1	-1	-1	-1	-1	1					
T	-6	-4	-2	-2	-2	-2	-2						
G	-7	-5	-3	-3	-3	-3							
A	-8	-6	-4	-4	-2								
A	-9	-7	-5	-5									
C	-10	-8	-6										
T	-11	-9											
A	-12												

Figure 9. Alignment matrix built for the test sequence 'ATCAAGTTGCCA' searching for mirror symmetry where parameter -t 25 (NeSSie parameter considering percentage of degeneracy including both gaps and mismatches) is used. Being the sequence 12 bp long and allowing up to 25% of degeneracy, at most three imperfections are permitted in the self-complementary alignment positions of the mirror motif (highlighted in red in the sequence alignment of Figure 10 and in the backtrack of the matrix).

```
ATCAAGT
|  |   | |
ACC—GT
```

Figure 10. Optimal alignment calculated from the alignment matrix Figure 9.

a - mirror	A	T	C	G
A	1	-1	-1	-1
T	-1	1	-1	-1
C	-1	-1	1	-1
G	-1	-1	-1	1

b - palindrome	A	T	C	G
A	-1	1	-1	-1
T	1	-1	-1	-1
C	-1	-1	-1	1
G	-1	-1	1	-1

Figure 11. Scoring matrices for mirror and palindromic symmetries.

SEARCH STRATEGIES

While using an overall approach based on sliding windows, the tool implements slightly different search modalities that are as follows:

- *Detection of motifs of length equal to sliding window size and satisfying the selected scoring cutoffs (-k N parameter of NeSSie, see Figure 12):* the entire DNA sequence captured by the sliding window (size defined by -k) is tested with the global alignment algorithm and reported if it satisfies the selected scoring cutoffs (degeneracy parameters).

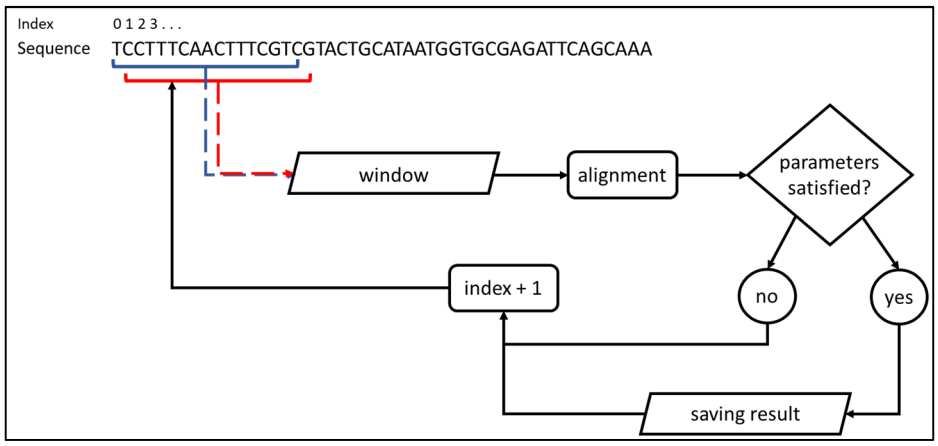


Figure 12. Motifs of length equal to the sliding window size and satisfying the selected scoring cutoffs (-k n parameter of NeSSie).

- *Detection of all motifs falling in the range of [min…max] lengths and satisfying the selected scoring cutoffs (-k N, -K N parameters of NeSSie, see Figure 13):* all the sequences from max to min length (window size defined by –K, max length) are tested with the global alignment algorithm and only those that satisfy the selected scoring thresholds (degeneracy parameters) are reported at each position.

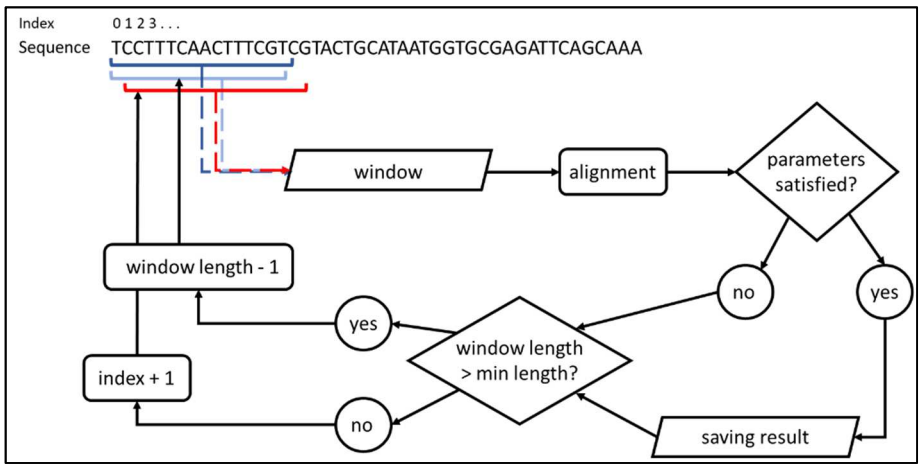


Figure 13. All motifs falling in the range of [min…max] lengths and satisfying the selected scoring cutoffs (-k n, -k n parameters of NeSSie).

- *Detection of the longest highest scoring motif falling in the range of [min…max] lengths and satisfying the selected scoring cutoffs (-MAX, -k N, -K N parameters of NeSSie see Figure 14):* all the sequences from max to min length (window size defined by –K, max length) are tested with the global alignment algorithm and only the longest sequence that satisfies also the selected scoring thresholds (degeneracy parameters) is reported at each position.

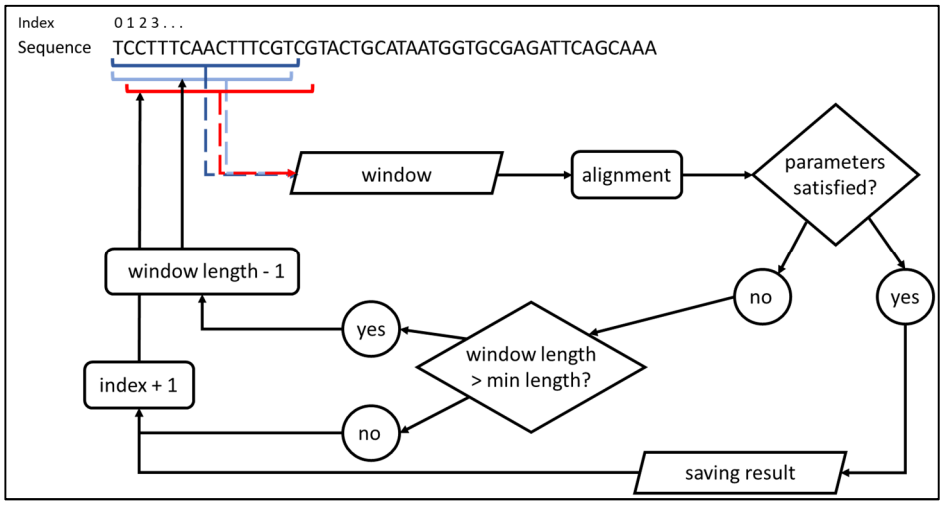


Figure 14. Longest max scoring motif falling in the range of [min…max] lengths and satisfying the selected scoring cutoffs (-max, -k n, -k n parameters of NeSSie).

ENTROPY AND COMPLEXITY MEASURES

Repetitive regions are associated with the formation of non-canonical structures in DNA and different types of repeats drive the folding of the different non-B DNAs. Linguistic complexity and Shannon entropy are measures that when applied to DNA can evaluate its repetitive nature. These measures, in combination with the detection of specific patterns, can extend the information about the structural and functional potential of a DNA molecule. For example, a low complexity is associated with an enrichment for specific patterns that reflects an unbalance in k-mers (words of length k) distribution across the genome, suggesting for a distinct regulatory or structural functionality for that region dependent on that specific k-mer.

SHANNON ENTROPY

The Shannon entropy (Shannon, 1948) is an index derived from the information theory that measures the amount of non-compressible information contained in a message. It can be applied to any symbolic sequence and allows to measure the order state (or entropy) of the sequence by the analysis of the symbols distribution. Practically, when applied to a DNA sequence (Machado and A, 2012), it allows to detect unbalances in base composition. The entropy is calculated according to the formula:

$$H = -\sum_{i=1}^{n} p(x_i)\, log_n p(x_i)$$

where $p(x_i)$ represents the probability for the symbol x_i to occur at any position of the sequence and n is the size of the alphabet x (n = 4 for DNA).

LINGUISTIC SEQUENCE COMPLEXITY

The Linguistic Sequence Complexity (Trifonov, 1990) is an index that measures the vocabulary richness of a sequence. It can be applied to any symbolic sequence and is calculated as the level of repetition of the k-mers (words of length k) in the sequence. The more complex the sequence is, the richer is the vocabulary it contains; vice versa, the lower the complexity is, the more repetitive the sequence is. The complexity is calculated according to the formula (Orlov and Potapov, 2004):

$$C = \sum_{k=1}^{N} V_k \Big/ \sum_{k=1}^{N} V_{\max k}$$

where V_k is the actual number of different words of length k in the sequence, $V_{\max k}$ is the maximum number of possible words of length k in the sequence, and N is the maximum length of the words considered in the score calculation. $V_{\max k}$ is calculated as the $\min(A^k, L - k + 1)$, where A is the alphabet size and L is the length of the sequence.

PERFORMANCE EVALUATION

The functionality of the tool has been evaluated and compared with others software currently available that represent the state-of-the-art. The datasets, benchmark procedures and results are presented in the supplementary materials for Berselli et al., 2018 (see attachment NESSIE SUPPLEMENTARY MATERIALS).

QPARSE (QUADRUPLEX AND PAIRED QUADRUPLEX SEARCH)

Mounting evidences are supporting a main role for G-Quadruplex (G4) structures as functional and structural regulators (see G-Quadruplex section in the Introduction). Although of extreme interest, these evidences are mostly scattered and there is a need to obtain a more comprehensive knowledge of G4s to fully understand their biological implications. To answer this need, several software tools have been developed to perform an automatic detection of G4s forming sequences in DNA (Huppert and Balasubramanian, 2005; Kikin et al., 2006; Perrone et al., 2017; Scaria et al., 2006). Despite providing different implementations, all these tools are mostly based on the same algorithm described by Huppert et al. and perform an exact pattern matching of the sequence d($G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$). While exhaustive in the detection of such a pattern, they are mostly limited in their functionality and cannot detect slightly degenerate variations of the pattern (i.e. G-islands containing gaps). This is a major limitation since new evidences are showing that the G-island pattern is not so strict as previously thought but G4s can also originate from slightly degenerate sequences (Mukundan and Phan, 2013; Varizhuk et al., 2017). G4Hunter (Bedrat et al., 2016) and pqsfinder (Hon et al., 2017) have been developed trying to answer this problem and implement different algorithms that allow to detect also degenerate patterns that can form G4s. However, both these tools present some limitations. G4Hunter detect only regions enriched in G runs, without providing the exact sequence of the potential quadruplex. pqsfinder search is instead restricted to the detection of patterns with only four subsequent islands. While this works perfectly for the strict search of G4s, Rigo and Sissi (Rigo and Sissi, 2017) demonstrated that longer patterns with more than four G-islands can be of extreme interest and can form paired G4s that interact. Basically, two G4s that are next to each other along the sequence can cross-talk and

physically interact, forming a higher order DNA structure. To date there is no tool capable of searching for such a pattern. To answer this need, I developed QPARSE (paper in preparation). QPARSE can search for exact and degenerate patterns in DNA that are potentially involved in the formation of G4 or paired G4 structures. QPARSE is fast and allows to easily perform analyses at a genome wide scale. Compared to other available tools, QPARSE is more flexible and allows to search not only for degenerate patterns, but also for longer patterns with more than four G-islands (with no potential upper limit to that). To perform the patterns detection, the tool exploits a graphs-based algorithm that is the more interesting and innovative feature introduced by the tool. Using direct acyclic graphs, the algorithm can model all the possible G4 patterns considering all the combinations of islands that are allowed by the parameters (see Algorithm section below).

PARAMETERS AND COMMAND LINE

The tool is very flexible and allows the user to select different parameters that can be defined to refine and customize the analysis. The parameters that can be used are:

- **-b** [**--base]**. The tool does not search automatically for G but allows the user to define which base to use in the search. This allows flexibility and extends the analysis to other letters, if desired. Since the tool does not automatically performs the search in the complementary strand, C and G can be used to search for G4s in both strands.
- **-m [--minlen]**. This parameter defines the length for the island. It is also possible to select a range of lengths by adding the parameter **–M [--maxlen]**. This allows to detect all the islands that are in the range of length [m…M].
- **-p [--perfect]**. This parameter defines the minimum number of islands that are required to be "perfect" within each pattern. Since the tool allows to detect also islands that are degenerate, this parameter

imposes a constraint for the minimum number of islands that cannot contain any degeneration within each pattern.

- **-g [--gapnum]**. This parameter defines the maximum number of gaps that can be opened per island. Alternatively, or in combination, the parameter **–l [--gaplen]** defines the maximum cumulative length of the gaps that is permitted per island. Finally, the parameter **–nocore** defines whether at least two consecutive bases are required to define an island. This Boolean parameter allows to detect also islands that do not have a "core" of at least two consecutive bases (e.g. G, GTGAG).
- **-n [--islandnum**]. This parameter defines the number of islands that need to be consecutive in the same pattern.
- **-L [--maxloop]**. This parameter defines the maximum distance (loop distance) that is allowed between two consecutive islands within the same pattern.

ALGORITHM

The detection of patterns potentially involved in the formation of G4 or paired G4 structures that are degenerate is not trivial and imposes a lot of computational and decisional challenges. When considering islands that are degenerate to a certain level, the number of possible islands that need to be considered increases exponentially. To be exhaustive in the search, it is therefore necessary to evaluate all the possible combinations considering all the possible islands that can be defined along the sequence. However, when allowing for degeneration, it is not straightforward to define when an interruption between Gs must be considered as a potential gap for an island or a loop between different islands. All the possibilities need to be evaluated and this further increases the combinatorial complexity. The picture is further complicated by increasing the number of G runs that are considered that expand the amount of possible combinations. To model the problem, I developed an exhaustive algorithm based on direct acyclic graphs (DAG) that allows to

map and evaluate all the possible combinations of islands that respect the parameters defined by the user. This is an improvement over the algorithms based on exact pattern-matching techniques that are not able to model the whole arrangement of G-islands forming a G4, i.e. a G-island can be contained in a long loop rather than participating in the tetrads stacking of the G4 as observed in hTERT quadruplex (Palumbo et al., 2009).

Algorithmic steps:

I. *Detection of the putative islands* (1st step): the analysis starts with a scan of the input sequence to identify all the possible islands that respect the parameters defined for the analysis. The Figure 15.I shows the results obtained by searching for islands of length three, with a maximum of one gap and a maximum gap-length of two [*-m 3 –g 1 –l 2*]. The islands are progressively detected using a finite-state machine that scans the sequence starting from each G and retrieves all the possible islands that start from that position.

Pseudocode for the finite-state machine that detects all the possible islands starting at the *i*-th G:

Input parameters:

max gaps num =[-g], max gaps len = [-l], min island len = [-m], max island len = [-M] if ([-M] and [-M] > [-m]) else [-m]

Function:

gaps num = 0, gaps len = 0,
island len = min island len,
base num = 1, last base = G,
start index = i, index = 1,
island list = []

```
while (gaps num <= max gaps num) and (gaps len <= max gaps len):
        next base = sequence[start index + index]
        if next base == G:
                base num += 1
        else:
                gap len += 1
                if last base == G:
                        gaps num += 1
                #END IF
        #END IF

        if base num == max island len:
                island list.append(ISLAND)
                return island list
        #END IF

        if base num == island len:
                island list.append(ISLAND)
                island len += 1
        #END IF

        last base = next base
        index += 1
#END WHILE

return island list
```

II. *Construction of the DAG* (2nd step): the identified islands are modeled as nodes of a direct acyclic graph. To build the graph, the non-overlapping islands that reside within the maximum length

allowed for the loops are connected through edges. Figure 15.II shows the graph that is obtained using the islands identified in Figure 15.I and loops with a maximum length of 20 bases [*-L 20*]. Such a graph intrinsically contains and models all the possible relations between all the islands that respect the parameters defined for the analysis. As a result, the potential G4 forming patterns are represented in the graph by continuous sequences of nodes that are connected by edges (i.e. continuous islands that reside within the loop distance).

III. *Navigation of the DAG* (3rd step): the paths modeled in the graph and starting from one island represent all the potential G4 forming patterns that begin with that island. The full navigation of the graph starting from each of the islands (each of the nodes) returns every potential G4 forming pattern in the sequence that respects the parameters defined for the analysis. The Figure 15.III shows the full output obtained while searching for patterns with four subsequent islands using the graph in Figure 15.II.

IV. *Refining the output* (4th step): since the graphs can model all the possible patterns in the sequence while considering all the possible islands and their combinations, the output that represents all these possible solutions is huge and difficult to handle "by eye". As the last step, the results are refined using a scoring system, and only the highest scoring and more promising solutions are returned. This step can be skipped and the full output can be obtained using the parameter **-all [–allresults]**. Each pattern has an associate score that is calculated as the sum of the scores of the single islands involved in the pattern. For each island the score is calculated using a finite-state machine as follows:

#the first base of the island is always a G
score = 0, increment = 0, last base = G

```
for base in island:
        if base == G:
                if last base == G:
                        increment += 1
                else:
                        increment = 1
                #END IF
                score += increment
        else:
                if last base == G:
                        increment = 1
                else:
                        increment += 1
                #END IF
                score -= increment
        #END IF
        last base = base
#END FOR
```

Such a score is conceived to reward the consecutive runs of Gs and to penalize the longer gaps following a scoring approach similar to the one described in Bedrat et al. (Bedrat et al., 2016). The Figure 15.IV shows the refined output obtained after the filtering step applied to the output shown in Figure 15.III.

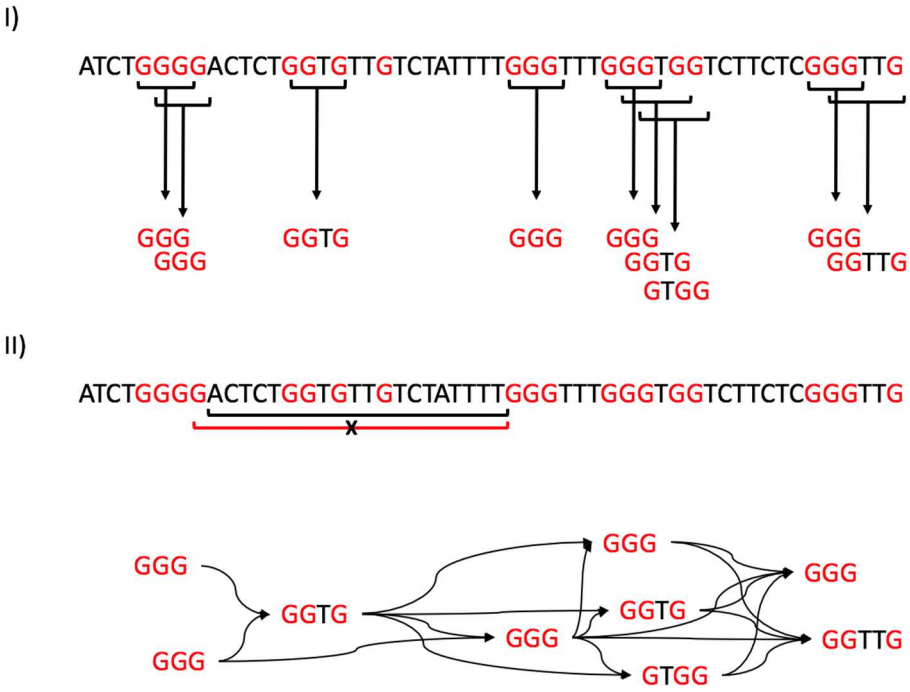


Figure 15. QPARSE algorithm: I) Detection of the putative islands, II) Construction of the direct acyclic graph (DAG), III) Navigation of the DAG, IV) Refinement of the output.

PERFORMANCE EVALUATION

Despite being an exhaustive algorithm, the tool is fast and capable of scaling on larger sequences. When evaluating the human chr1 (250 Mb) the times are as follows:

Pattern to search	**Time to search for both strands**
Perfect patterns (e.g. -m 3 -M 5 -L 7)	2 minutes
Degenerate patterns (e.g. -m 3 -M 5 -L 12 -l 5)	1 hours
Longer and degenerate patterns (e.g. -m 3 -M 5 -L 5 -l 5 –n 8 –p 5)	2 hours

G4 VIRUS

G4 structures are becoming more and more relevant. However, especially in viruses, the data on these structures are scattered and not homogeneous. In collaboration with a group inside our department, we aimed at creating a more comprehensive and complete source of data. We performed a statistical evaluation of the G4 forming sequences in the genome of all the human viruses. Particularly, we focused on the conservation of G4s within the local context of viral genomes. For each virus, we generated the multiple alignments of all its available strains and calculated the overall degree of conservation of the G-islands that are necessary and sufficient to form a G4. We then collected all the results of the analyses in a database accessible from web (www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus) to allow an easy and interactive navigation (Lavezzo et al., 2018) (currently under review). This work shows that the presence, distribution, and location of G4s are features characteristic of each virus class and family. Moreover, the statistical analyses show that their presence within the viral genome is orderly arranged, as indicated by the possibility to correctly assign up to two-thirds of viruses to their exact class based on the G4 classification. For each virus, the website provides: I) the list of all G4s formed by GG-, GGG- and GGGG-islands present in the genome (positive and negative strands), II) their position in the viral genome along with the known function of that region, III) the degree of conservation among strains of each G4 in its genomic context, IV) the statistical significance of G4 formation. The availability of these data will represent a useful resource that can expedite the research on G4s in viruses identifying the most conserved and thus potentially relevant quadruplex structures for each virus (see attached publication for more details).

ANALYSES AND PRELIMINARY DATA

OVERVIEW

The design and assessment of the new developed tools (see Tools and Databases section) have allowed the progression of the work towards both the detection of sequence patterns in genomes and their initial experimental validation. I started to apply the new tools that I developed to the analysis of genomic sequences. I focused my work on *Mycobacterium spp.* and *Homo sapiens*: I) In the *Mycobacterium spp.* genomes, I identified an enriched pattern with a perfect mirror symmetry that can fold into a previously unknown hairpin structure through the formation of mixed Watson-Crick and Hoogsteen bonds. This pattern is unique of the species capable of developing tuberculosis-like diseases and is completely absent in the human genome. The distribution of the pattern is also peculiar, being enriched only in specific regions of the genome (up to thousands of bp in length) characterized by a strong modularity, II) In the human genome, I focused my analyses on the paired quadruplex structures. Rigo and Sissi (Rigo and Sissi, 2017) recently described the formation of a paired G4s structure in the promoter of the c-KIT gene. While no exact rules are currently known, apparently, the symmetry is important for the formation of such a structure. By combining the search for paired G4s and mirror symmetries I identified a potential similar system in the promoter of the BCL2 gene that is being experimentally confirmed. Surprisingly, extending the analysis genome-wide I detected an enrichment for sequences with the potential to form such a paired quadruplex system just upstream the TSS (Transcription Starting Site) of thousands of human genes.

MYCOBACTERIUM SPP.

BACKGROUND

Pulmonary tuberculosis caused by Mycobacterium tuberculosis complex (MTC) is one of the major death-causes worldwide. Nevertheless, pathogenic mycobacteria remain very mysterious organisms and the mechanisms behind their unique pathogenic properties remain elusive. Their ability to survive inside the host cells is somehow unique and understanding the mechanisms behind these survival capabilities would be a major advance in defeating these pathogens. I started working on the genome of *Mycobacterium bovis* while collaborating with a research group in the department of Molecular Medicine interested in the study of this organism. Particularly, I focused my attention on the analyses of the complexity of the genome and the detection of sequence features potentially relevant for the formation of high-order DNA structures (e.g. non-B DNAs).

METHODS AND ANALYSES

SEQUENCE ANALYSIS

Using NeSSie, I started to perform complexity and entropy analyses while searching for patterns with symmetrical properties. Complexity (linguistic complexity) and entropy (Shannon entropy) measures allow to obtain information on the repetitive nature of the DNA. Motifs with symmetrical properties are instead associated with the formation of non-B DNA structures (see NeSSie in Tools and Databases section).

Complexity and entropy: complexity and entropy analyses revealed that the genome of *Mycobacterium bovis* AF2122/97 (Table 1) is not homogeneously complex. There are many long regions (up to thousands of bp) that are characterized by lower complexity in comparison to the surrounding regions. These regions are dispersed throughout the genome.

To obtain meaningful and balanced profiles of entropy and complexity, I had to tune the tool parameters and try different calculations because the results are strongly influenced by the choice of the window size that is used for the analysis. A narrow window is more sensible in capturing local biases but loses accuracy in the detection of meaningful trends on a wider scale. A large window is more accurate in detecting meaningful trends but loses in the sensibility to detect biases in the sequence composition. To identify the parameters to use, I explored different combinations of windows sizes and shifting intervals. Eventually, I selected a combination of parameters that provided a good compromise between the sensibility to detect biases in the base composition and the accuracy in the detection of meaningful trends. To calculate the linguistic complexity, I used a sliding window 2 Kb long with a shifting interval of 500 bp. To calculate the Shannon entropy, I followed a slightly different approach. I decided to use a shorter window (length 50 bp, shift 25 bp) to achieve a more sensible detection of the local biases. Using the shorter window, I calculated the entropy measures inside each of the main windows (2 Kb) and, from these scores, I calculated an average score for each of the larger windows.

Patterns with symmetrical properties: the search for patterns with symmetrical properties revealed that these motifs are abundant in the genome of *Mycobacterium bovis* AF2122/97. However, they are not randomly scattered across the genome, but are instead highly enriched in the regions characterized by a lower complexity (see Berselli et al., 2018). Particularly, patterns with perfect mirror properties are mostly exclusive of these regions.

ANALYSIS OF THE PERFECT MIRROR PATTERNS

The analysis of the perfect mirror patterns revealed that most of these motifs share a common consensus: 'GGCGGCAACGGCGGCAACGGCGG' (Figure 16). The core sequence 'AACGGCGGCAA' is almost perfectly conserved (Figure 16). In this analysis, I considered the motifs with a perfect

mirror symmetry that were more abundant and conserved. I calculated the reverse complement of the sequences when necessary and I performed an alignment of the motifs using muscle (Edgar, 2004). To visualize the alignment I used Jalview 2 (Waterhouse et al., 2009) (Figure 16).

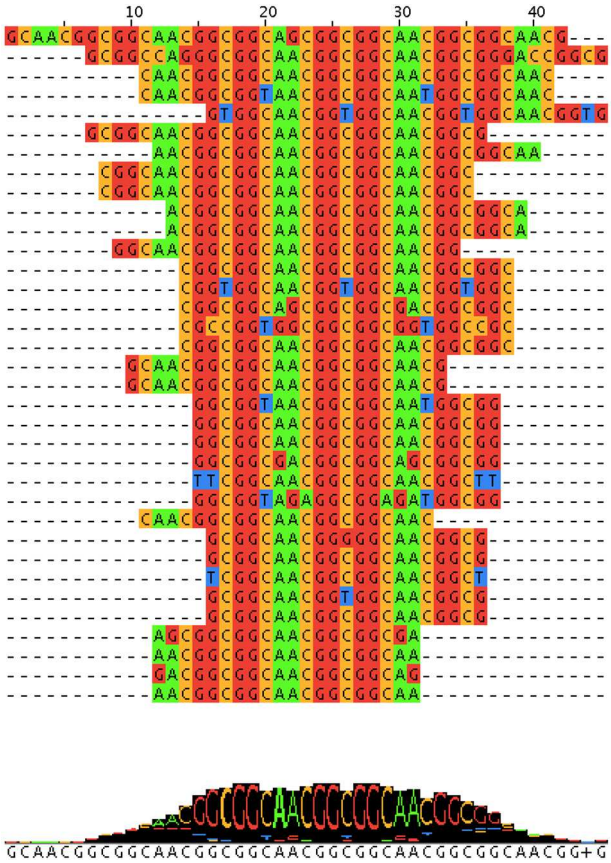


Figure 16. Multiple alignment of the perfect mirror patterns detected in the genome of *Mycobacterium bovis*. On the bottom the consensus calculated for the alignment.

COMPARATIVE ANALYSIS AMONG *MYCOBACTERIUM SPP.*

To understand if these observations are restricted to the *Mycobacterium bovis* or can be extended to other mycobacteria, I performed a comparative analysis for the same motifs detected in *M. bovis* (Figure 16) in different strains and species (see Table 1). I counted all the occurrences of the patterns detected in *M. bovis* genome in the genomes of the other bacteria. This analysis revealed almost a total conservation of the perfect

mirror patterns in the group of the Mycobacterium tuberculosis complex (MTC, marked in red in Table 1), as well as in *M. kansasi* and *M. marinum*. All the other species lack these motifs. The heatmap in Figure 17 shows the results for a subset of motifs with perfect mirror symmetry that are more enriched in the genome of *M. bovis*. For the analysis, I used a custom Python script.

Table 1. Information and accession numbers for the species and strains of *Mycobacterium* used for the analyses (in red the bacteria belonging to the Mycobacterium tuberculosis complex, in black the others).

Species	Strain	GenBank	ENA
Mycobacterium tuberculosis	H37Rv	GCA_000195955.2	NC_000962.3/AL123456.3
Mycobacterium tuberculosis	KZN 1435	GCA_000023625.1	NC_012943.1/CP001658.1
Mycobacterium tuberculosis	CCDC5180	GCA_000270365.1	NC_017522.1/CP001642.1
Mycobacterium tuberculosis	HKBS1	GCA_000572125.1	NZ_CP002871.1/CP002871.1
Mycobacterium tuberculosis	BT1	GCA_000572175.1	NZ_CP002883.1/CP002883.1
Mycobacterium tuberculosis	K	GCA_000698475.1	NZ_CP007803.1/CP007803.1
Mycobacterium abscessus	ATCC 19977	GCA_000069185.1	NC_010397.1/CU458896.1
Mycobacterium abscessus	UC22	GCA_001050395.1	NZ_CP012044.1/CP012044.1
Mycobacterium abscessus	NOV0213	GCA_001430775.1	NZ_CP013049.1/CP013049.1
Mycobacterium abscessus	FLAC045	GCA_001610615.1	NZ_CP014958.1/CP014958.1
Mycobacterium abscessus subsp. bolletii	MM1513	GCA_000758225.1	NZ_CP009447.1/CP009447.1
Mycobacterium abscessus subsp. bolletii	MC1518	GCA_000770125.1	NZ_CP009613.1/CP009613.1
Mycobacterium bovis	AF 2122/97	GCA_000195835.1	NC_002945.3/BX248333.1
Mycobacterium bovis BCG str.	ATCC 35743	GCA_000194075.3	NZ_CP003494.1/CP003494.1
Mycobacterium bovis BCG str.	Korea 1168P	GCA_000338715.2	NC_020245.2/CP003900.2
Mycobacterium bovis BCG str.	Moreau RDJ	GCA_000967285.1	NZ_AM412059.1/AM412059.2
Mycobacterium bovis BCG	3281	GCA_001043255.1	NZ_CP008744.1/CP008744.1
Mycobacterium bovis BCG str.	Tokyo 172 substr. TRCS	GCA_001580385.1	NZ_CP014566.1/CP014566.1
Mycobacterium avium subsp. paratuberculosis	MAP4	GCA_000390085.1	NC_021200.1/CP005928.1
Mycobacterium avium subsp. avium	2285 (R)	GCA_000758285.1	NZ_CP009493.1/CP009493.1
Mycobacterium smegmatis str.	MC2 155	GCA_000283295.1	NC_018289.1/CP001663.1
Mycobacterium smegmatis	INHR1	GCA_000767665.1	NZ_CP009495.1/CP009495.1
Mycobacterium africanum	GM041182	GCA_000253355.1	NC_015758.1/FR878060.1
Mycobacterium intracellulare	ATCC 13950	GCA_000277125.1	NC_016946.1/CP003322.1
Mycobacterium intracellulare	MOTT-64	GCA_000276825.1	NC_016948.1/CP003324.1
Mycobacterium kansasii	ATCC 12478	GCA_000157895.2	NC_022663.1/CP006835.1

Mycobacterium marinum	M	GCA_000018345.1	NC_010612.1/CP000854.1
Mycobacterium leprae	TN	GCA_000195855.1	NC_002677.1/AL450380.1
Mycobacterium leprae	Br4923	GCA_000026685.1	NC_011896.1/FM211192.1
Mycobacterium haemophilum	DSM 44634	GCA_000340435.3	NZ_CP011883.2/CP011883.2
Mycobacterium immunogenum	CCUG 47286	GCA_001605725.1	NZ_CP011530.1/CP011530.1
Mycobacterium chelonae	CCUG 47445	GCA_001632805.1	NZ_CP007220.1/CP007220.1
Mycobacterium neoaurum	VKM Ac-1815D	GCA_000317305.3	NC_023036.2/CP006936.2
Mycobacterium canettii	CIPT 140010059	GCA_000253375.1	NC_015848.1/HE572590.1
Mycobacterium sinense	JDM601	GCA_000214155.1	NC_015576.1/CP002329.1
Mycobacterium chubuense	NBB4	GCA_000266905.1	NC_018027.1/CP003053.1
Mycobacterium yongonense	05-1390	GCA_000418535.2	NC_021715.1/CP003347.1
Mycobacterium rhodesiae	NBB3	GCA_000230895.3	NC_016604.1/CP003169.1
Mycobacterium phlei	CCUG 21000	GCA_001583415.1	NZ_CP014475.1/CP014475.1
Mycobacterium fortuitum	CT6	GCA_001307545.1	NZ_CP011269.1/CP011269.1
Mycobacterium vaccae	95051	GCA_001655245.1	NZ_CP011491.1/CP011491.1

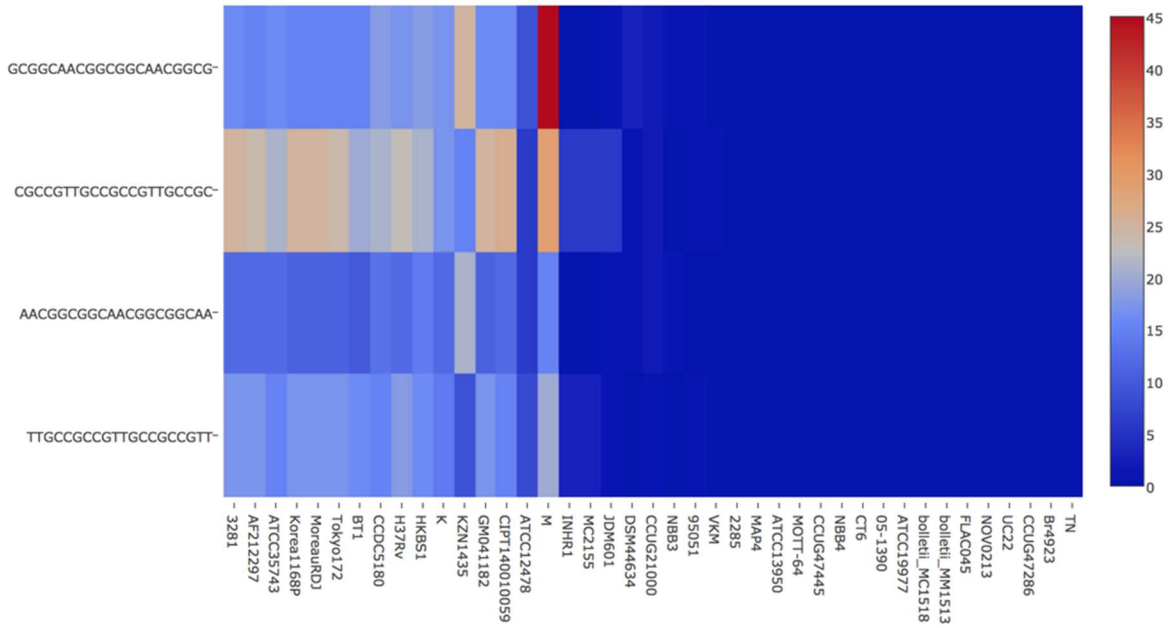


Figure 17. Conservation of motifs with a perfect mirror symmetry that are enriched in *M. Bovis* genome across the genomes of different mycobacteria (Table 1). The map shows the counts for each motif in each of the compared genomes.

K-MERS ANALYSIS

To confirm the enrichment of the consensus 'GGCAACGGCGGCAACGG' in the genomes of *Mycobacterium bovis* AF2122/97 and *tuberculosis* H37Rv, I performed genome wide analyses of all the existent k-mers (words of size k) up to a length of 20 bp. The results confirmed that the most abundant 17-mers in the genomes correspond to the consensus. Interestingly, there is no enrichment of a different sequence pattern of comparable length. To further confirm that such an enrichment in sequences with a high similarity is not commonly expected in bacteria, I repeated the analyses of 17-mers also in *Mycobacterium leprae* (a mycobacterium not included in the MTC) and *Escherichia coli* (a bacterium with a highly-repeated genome, NC_000962.3). The results confirmed that the observed enrichment is typical and not likely to happen by chance in bacteria. To perform the analyses, I used the NeSSie tool to retrieve all the k-mers, and custom Python scripts to confirm the enrichment. To evaluate the enrichment, I clustered the 17-mers based on the similarity of their sequences (calculated with GLSEARCH 36.3.5b) and allowing a maximum of two mismatches between the sequences in the same cluster. I considered only the 17-mers that are more abundant (10 or more occurrences in the genome). In *M. tuberculosis*, this led to the formation of few large clusters that contain most of the sequences (Figure 18.I). Vice versa, in *M. leprae* and *E. coli* only a small number of 17-mers group together. This led to the formation of a larger number of smaller clusters (Figure 18.II, Figure 18.III).

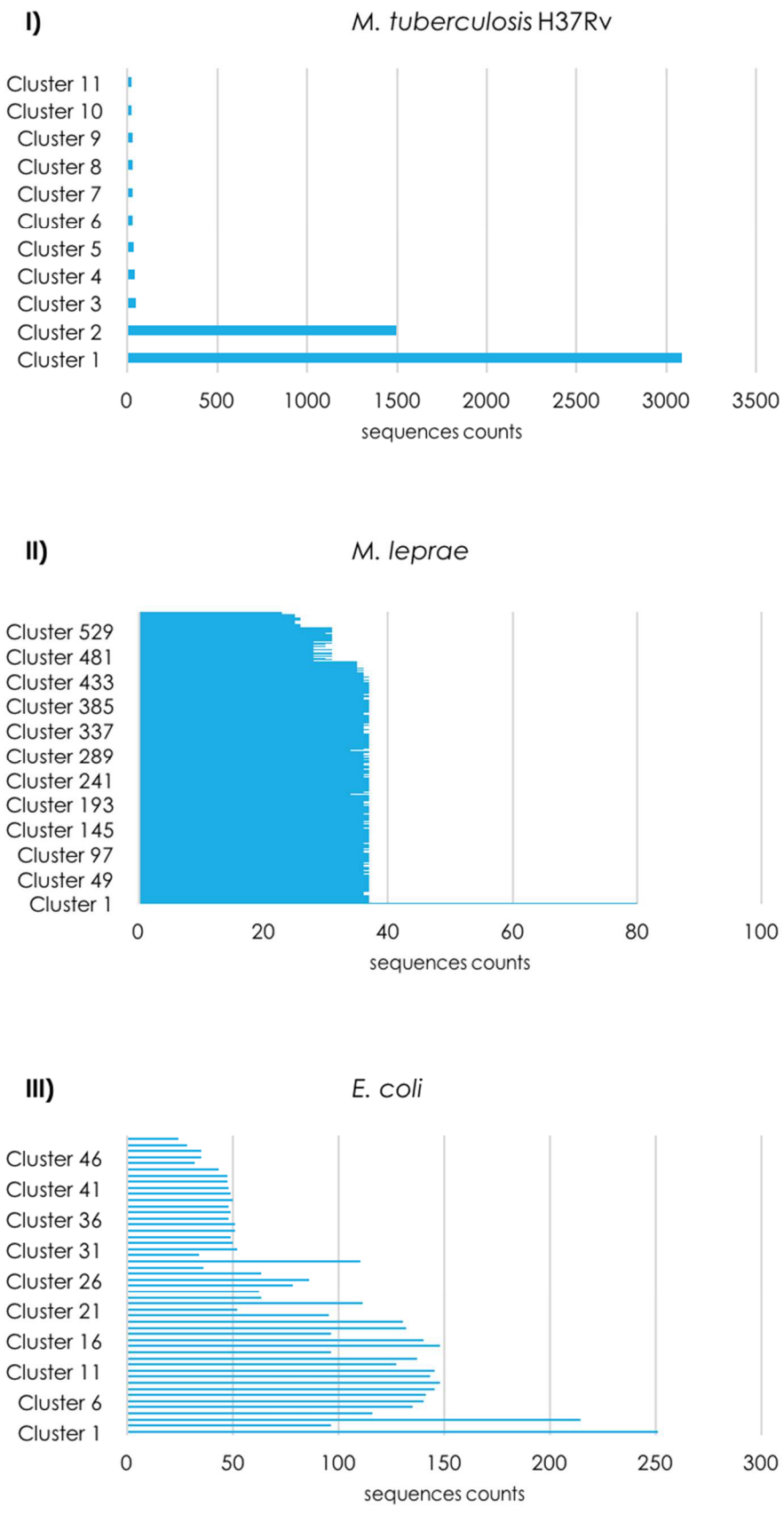


Figure 18. Analysis of the enrichment and similarity of the k-mers of 17 bp in: I) *Mycobacterium tuberculosis* H37Rv, II) *Mycobacterium leprae*, III) *Escherichia coli*.

K-MERS DISTRIBUTION IN *M. TUBERCULOSIS*

I re-evaluated the distribution of the most enriched 17-mers across the genome of *M. tuberculosis* H37Rv. I considered the 17-mers belonging to the two most enriched clusters (Cluster 1 and Cluster 2, see Figure 18.I). This confirmed what was previously observed in *M. bovis*. Also in *M. tuberculosis*, there is a strong enrichment of these motifs across specific genomic regions characterized by a lower complexity. Interestingly, these regions perfectly overlap with the putative genes belonging to the PE-PGRS family (Figure 19). Indeed, out of the 4587 17-mers that were mapped, 4403 reside within a PE-PGRS region (Figure 20). To evaluate the distribution of the 17-mers, I partitioned the genome in continuous bins of 1000 bp and counted the number of 17-mers falling into each bin (Figure 19). To evaluate the overlap with the PGRS, I counted all the 17-mers that fall into a PE-PGRS region (Figure 20). I used custom Python scripts to perform the analyses.

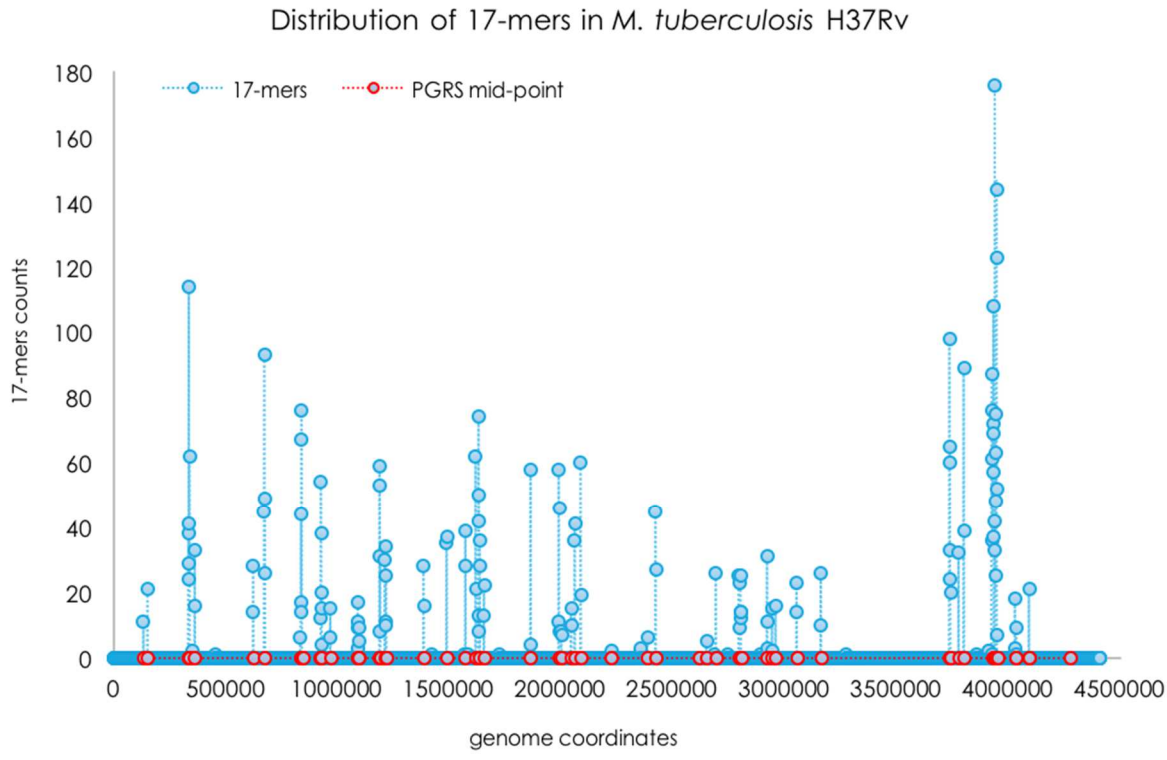


Figure 19. Distribution of the k-mers of 17 bp from Cluster 1 and Cluster 2 in Figure 18.I across *M. Tuberculosis* H37Rv genome. The genome is partitioned into continuous bins of 1000 bp. The blue points represent the

number of 17-mers falling into each of the bins. The red points represent the mid-point of the regions annotated as PE-PGRS.

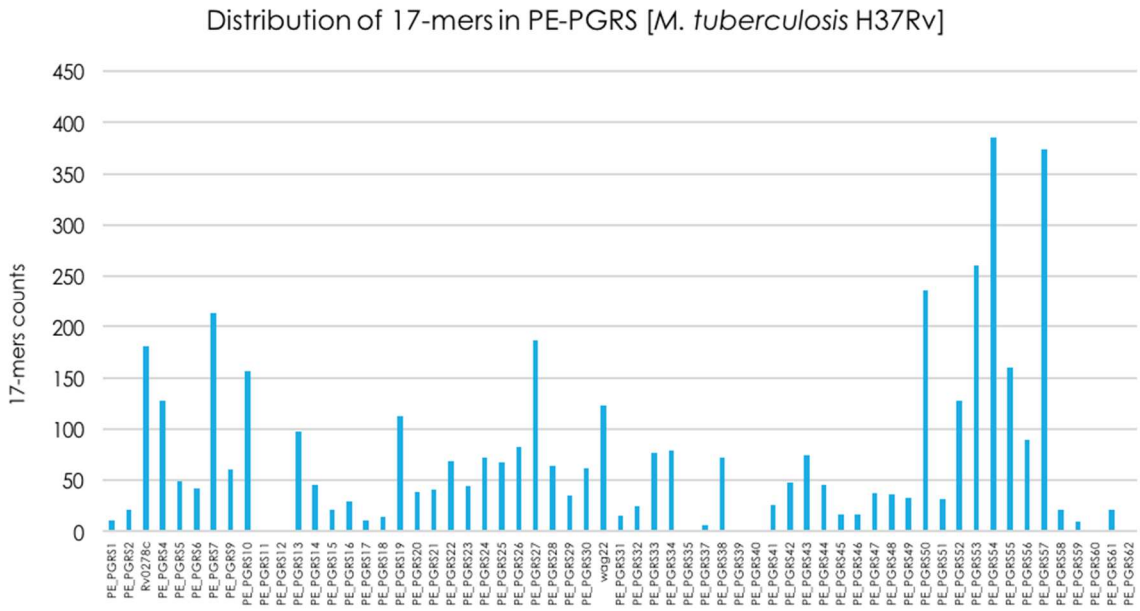
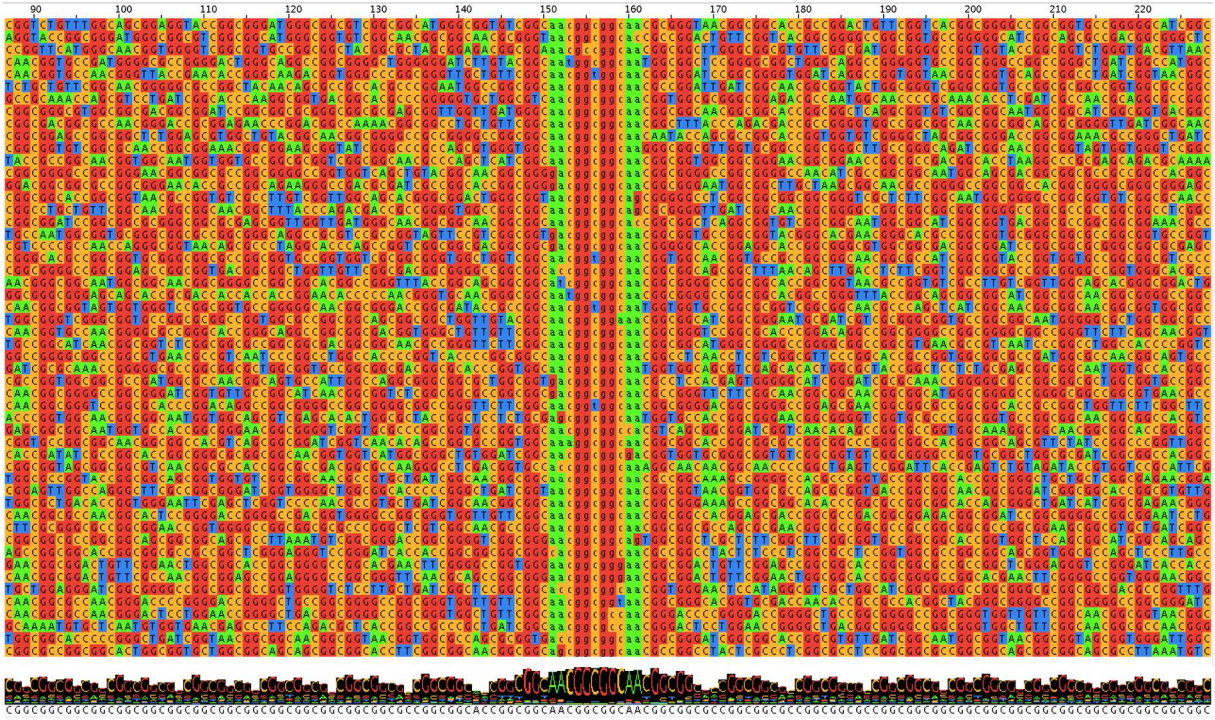


Figure 20. Distribution of the k-mers of 17 bp from Cluster 1 and Cluster 2 in Figure 18.I with respect to *M. Tuberculosis* H37Rv PE-PGRS genes.

EXTENDING THE CONSENSUS

To evaluate a wider context around the consensus, I considered and extended the analysis also to the regions surrounding the core sequence 'AACGGCGGCAA'. This motif is central to the consensus and perfectly conserved among the perfect mirror patterns. I performed the analysis in *M. bovis*, *M. tuberculosis* and *M. marinum*. To retrieve all the exact and slightly degenerate occurrences of the core sequence, I used glsearch (GLSEARCH 36.3.5b). I further extended these sequences upstream and downstream up to 150 bp. In *M. bovis* and *M. tuberculosis*, I divided all the retrieved sequences into two groups based on the relative position with respect to PE-PGRS genes. The first group contains the sequences that overlap with a PE-PGRS region. The second group contains the sequences that do not overlap with PE-PGRS regions. In *M. marinum*, since there is no

information on putative PE-PGRS, I divided the retrieved sequences based on the alignment results. The first group contains the sequences with exact occurrences of the core motif, the second group contains the sequences with degenerate occurrences of the same motif. Eventually, I aligned all the sequences for each of the groups using the core motifs as the central point of the alignment (Jalview 2). In *M. bovis* and *M. tuberculosis*, the analyses revealed the existence of a wider consensus that can be represented by the sequence d(CGGCGGCNN)$_n$. The consensus extends for thousands of bp and is unique of the sequences that are associated with the PE-PGRS (Figure 21, Figure 22). The analyses also revealed that the distribution of the core motifs is non-casual, with all the exact occurrences residing in frame within the repeated regions. This was further confirmed in *M. marinum*, where the very same consensus emerged in the group of sequences that extend the exact core motifs (Figure 23). Interestingly, while the core motifs within the repeated regions extend into the larger 'GGCGGCAACGGCGGCAACGGCGG' consensus, the core motifs outside the repeated regions are not part of any longer consensus (Figure 21, Figure 22, Figure 23). I used custom Python scripts to perform the analyses.



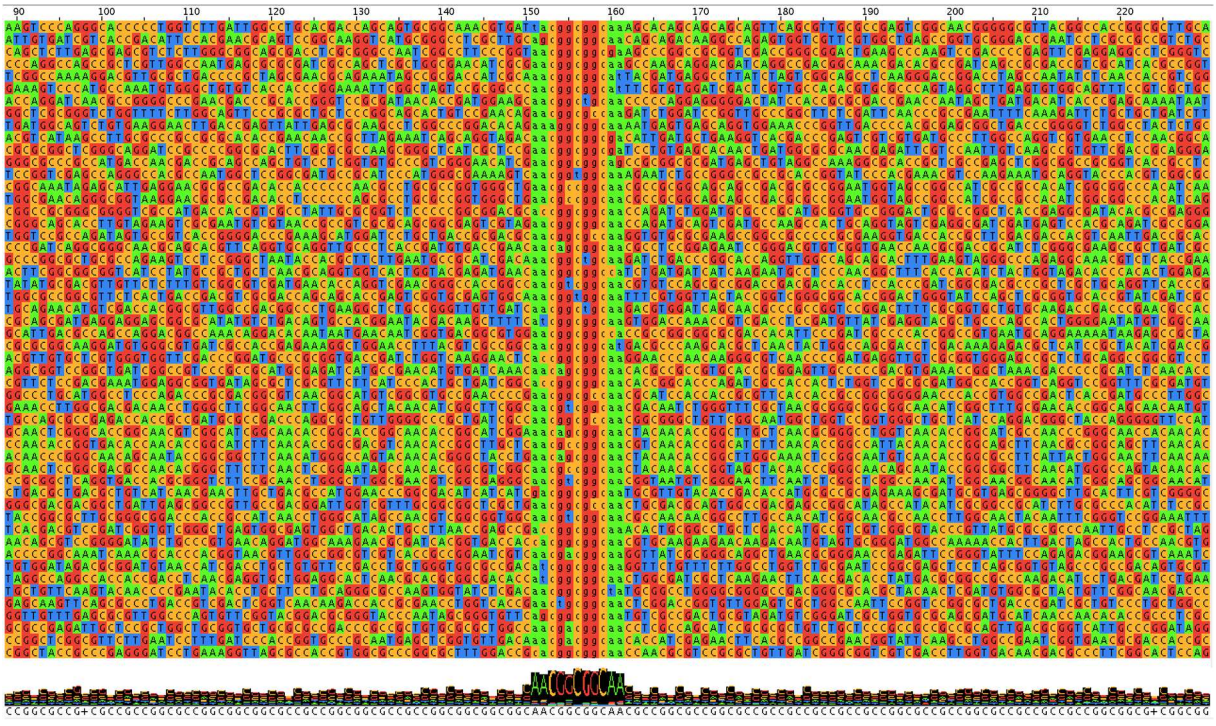
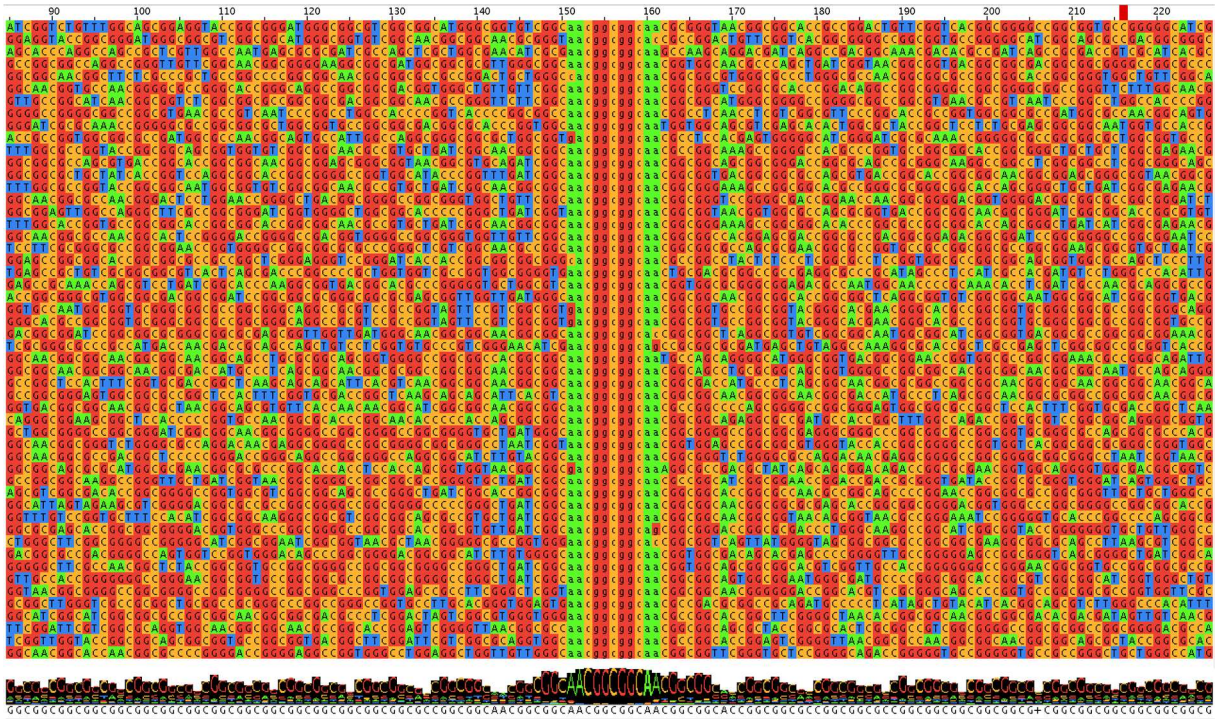


Figure 21. Alignment of *M. tuberculosis* sequences extended around the consensus 'AACGGCGGCAA'. In PE-PGRS (top), outside PE-PGRS (bottom).



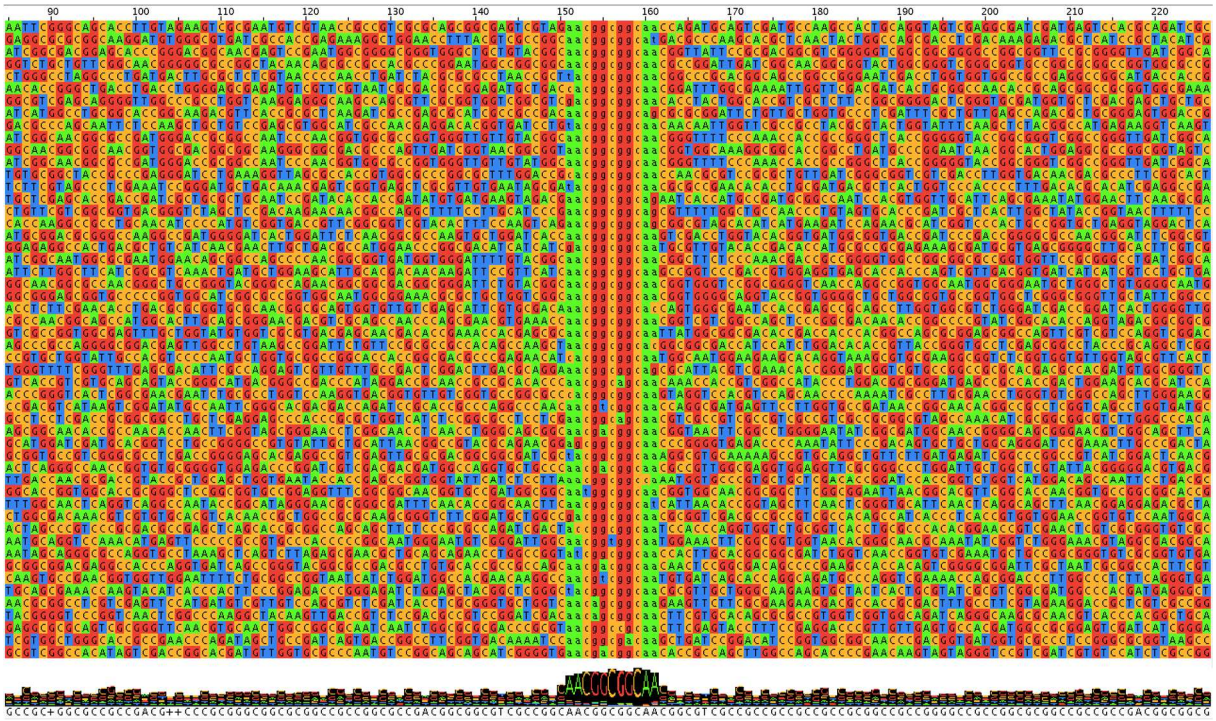
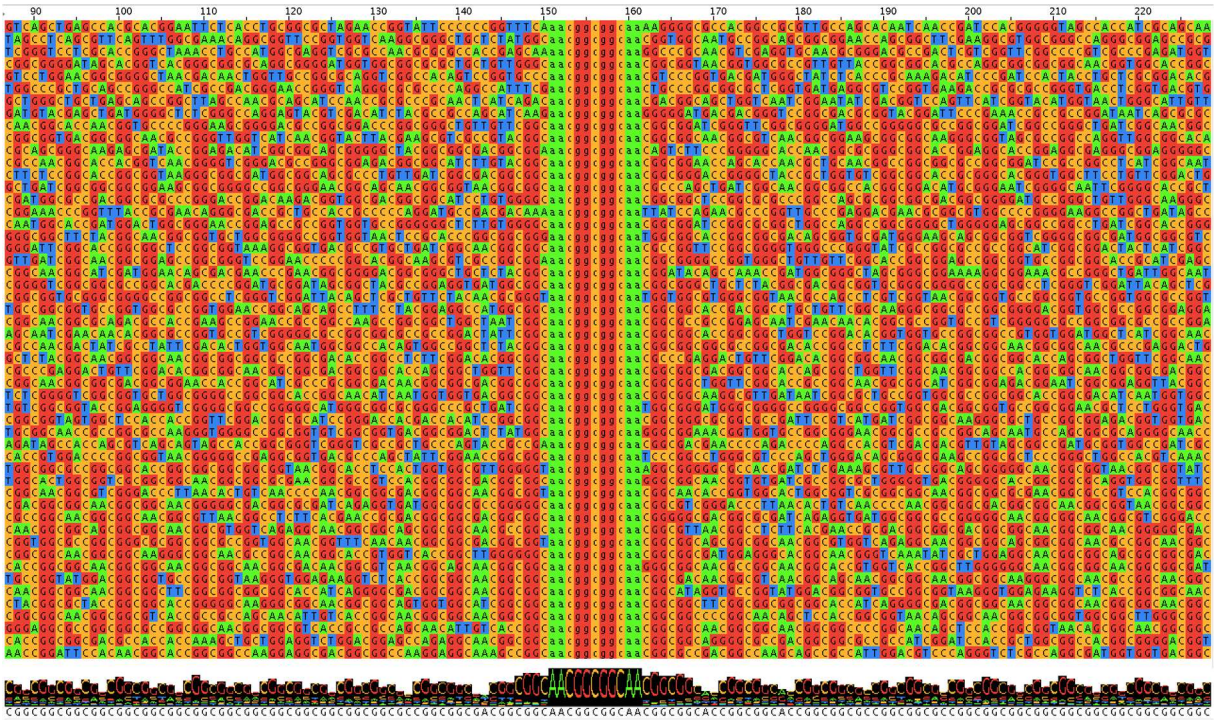


Figure 22. Alignment of *M. bovis* sequences extended around the consensus 'AACGGCGGCAA'. In PE-PGRS (top), outside PE-PGRS (bottom).



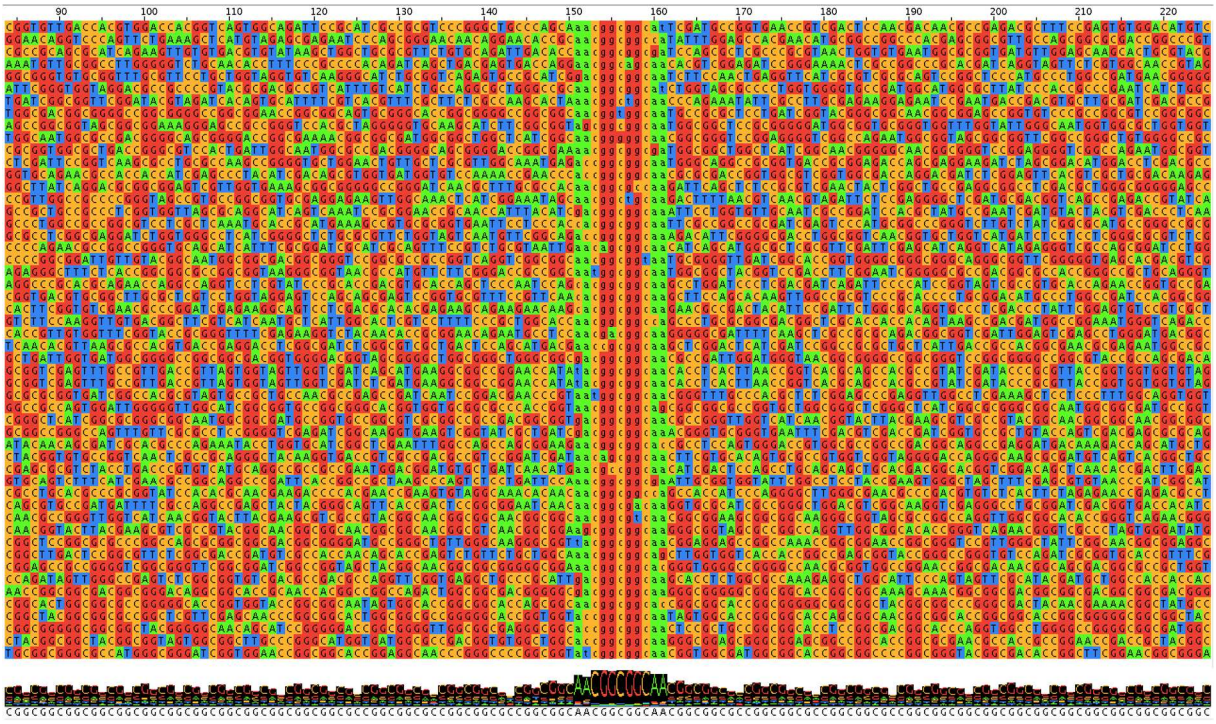


Figure 23. Alignment of *M. marinum* sequences extended around the consensus 'AACGGCGGCAA'. Sequences containing the exact consensus (top), sequences containing the degenerate consensus (bottom).

PE-PGRS

The PE-PGRS belong to a large family of glycine-rich proteins that are exclusive of several mycobacterial species. Over the past few decades, these genes have been widely investigated in relation to the pathogenic capabilities of these species (Meena, 2015). However, their function remains elusive. In *M. bovis* and *M. tuberculosis*, PE-PGRS regions are highly modular and are characterized by the unique consensus d(CGGCGGCNN)$_n$. These regions are also highly enriched for the consensus 'GGCGGCAACGGCGGCAACGGCGG', that is perfectly framed within the repeated modularity (see above). Given the strong conservation of these features, I tried to evaluate whether the overall sequence of PE-PGRS is conserved between these bacteria. I also extended the analysis to others *Mycobacterium spp.* to evaluate the conservation of PE-PGRS regions across their genomes. *Mycobacterium tuberculosis* H37Rv has 61 annotated PE-PGRS. Using glsearch (GLSEARCH 36.3.5b) to perform global alignments, I searched in the genome of each of the other mycobacteria (Table 1) for the most similar occurrence to each of these genes (Figure 24). The results revealed that PGRS regions are mostly conserved between the species belonging to the MTC. Interestingly, also *M. marinum* and *M. kansasi*, where no information of PGRS are available, have regions in their genomes that are alike to *M. tuberculosis* PE-PGRS. These are the same species that share the almost total conservation of the consensus and the perfect mirror patterns initially detected in *M. bovis* (Figure 17). However, the degree of conservation of PE-PGRS regions between these bacteria is highly variable despite their relationship. For example, different strains that belong to the same species show completely different degrees of conservation for the same PE-PGRS. This is intriguing since these bacteria are known for the extreme conservation of their genome also across different species. I used custom Python scripts to perform the analyses.

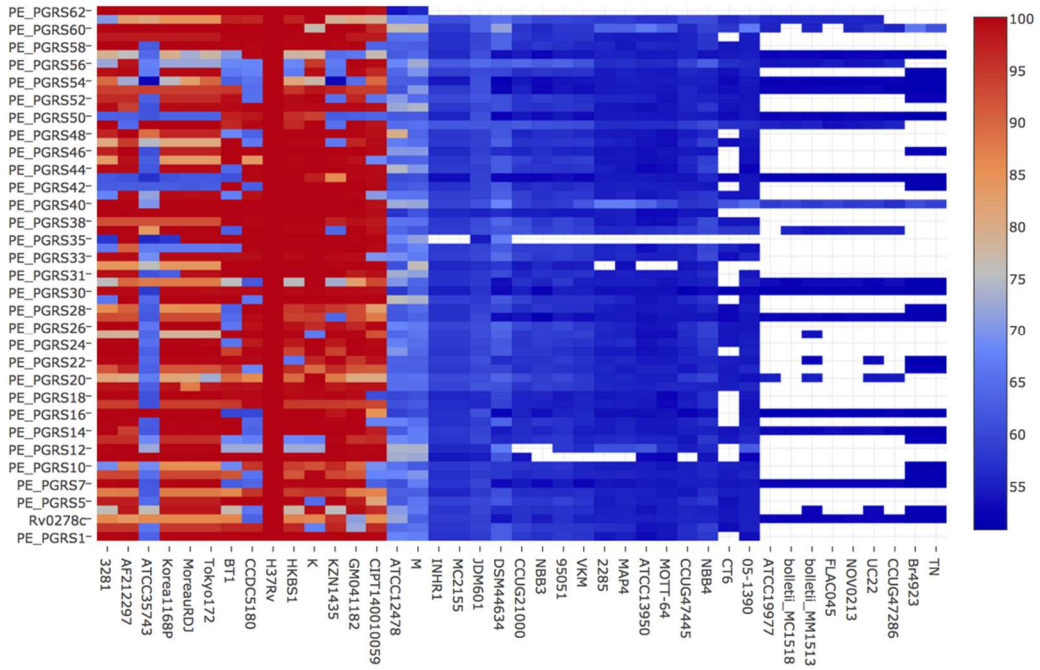


Figure 24. Conservation of *M. tuberculosis* H37Rv PE-PGRS across the different mycobacteria (Table 1). The score represents the best match as obtained by global alignment. 100 correspond to perfect identity of the sequences.

STRUCTURAL ANALYSIS OF THE CONSENSUS

The consensus 'GGCGGCAACGGCGGCAACGGCGG' is highly conserved across the repeated regions that correspond to PE-PGRS. Given the unexpected enrichment for such a sequence, collaborators from the Department of Pharmaceutical and Pharmacological Sciences performed additional experimental studies on this sequence. Using the Nuclear Magnetic Resonance (NMR), they identified the sub-sequence 'GGCGGCAACGGCGG' as the responsible for the formation of a non-B DNA structure that is extremely stable in solution. Eventually, they characterized the structure as a hairpin with mixed Watson-Crick (**-**) and Hoogsteen (•) base pairings (Figure 25). Such a structure was not previously described.

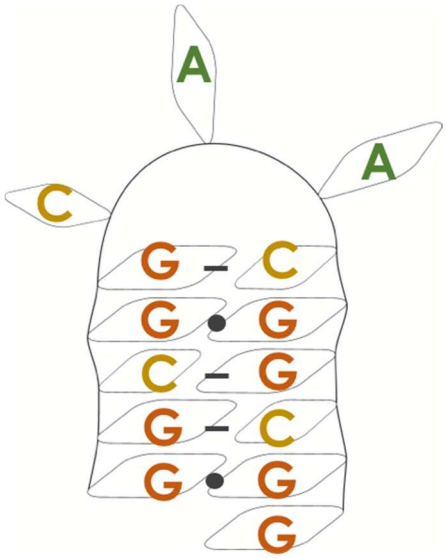


Figure 25. Secondary structure characterized for the sequence 'GGCGGCAACGGCGG'. The sequence can form a stable hairpin structure with mixed Watson-Crick (-) and Hoogsteen (•) base pairings.

DISCUSSION

The genome of *Mycobacterium bovis* AF2122/97 is not homogeneously complex and contains many regions with a lower complexity. These regions can be long up to thousands of bp and are dispersed throughout the length of the genome. Interestingly, these regions are highly enriched in motifs with symmetrical properties and patterns with a perfect mirror symmetry. The latter are exclusive of these regions and share the common consensus 'GGCGGCAACGGCGGCAACGGCGG', with the core sequence 'AACGGCGGCAA' perfectly conserved between all the motifs (Figure 16). Considering other *Mycobacterium spp.* (Table 1), it is possible to observe similar regions in the group of the Mycobacterium tuberculosis complex (MTC), as well as in *M. kansasi* and *M. marinum*. The consensus and the perfect mirror patterns are also conserved between these species (Figure 17). Interestingly, these are the species that are capable to develop tuberculosis-like diseases and are more relevant for the human health. In the genomes of *M. tuberculosis* H37Rv and *M. bovis* AF2122/97, the words of 17 bp (17-mers) that are more abundant correspond to the consensus

(Figure 18). As previously observed in *M. bovis*, also in *M. tuberculosis* these motifs are highly enriched within specific genomic regions characterized by a lower complexity. Unexpectedly, there is a perfect overlap between these regions and the putative genes belonging to the PE-PGRS family (Figure 19). Indeed, 96% of the 17-mers with more than 10 occurrences in the genome of *M. tuberculosis* fall within a PE-PGRS region (Figure 20). The PE-PGRS represent a large family of genes that is exclusive to several *Mycobacterium species*. While their role remains elusive, they have been widely investigated in relation to the pathogenic capabilities of these species. The PGRS are conserved in the MTC, in *M. marinum* and in *M. kansasi* (Figure 24). However, their degree of conservation between these bacteria is highly variable and different strains inside the same species can show completely different degrees of conservation for the same PE-PGRS (Figure 24). Mycobacteria are known for the extreme conservation of their genome even across different species. If the complete sequence of PE-PGRS is important for their functionality, such a high variability is unexpected and hard to explain. However, while the sequence identity is variable, all the PE-PGRS maintain common features that are highly conserved in all the bacteria. The PE-PGRS are modular regions that share a wider consensus that can be represented by the sequence $d(CGGCGGCNN)_n$. The consensus is unique to these regions and extends for thousands of bp (Figure 21, Figure 22). The PGRS are also enriched for the unique consensus 'GGCGGCAACGGCGGCAACGGCGG', that is perfectly framed within the modular pattern. These observations open to the possibility that these features are relevant for the PE-PGRS functionality, rather than the complete identity of the sequence. Supporting this idea, experimental data confirmed that the sequence 'GGCGGCAACGGCGG', that is highly enriched in these regions and part of the consensus, can fold into a hairpin conformation not previously described. The hairpin is extremely stable in solution and fold through the formation of mixed Watson-Crick and Hoogsteen base pairings (Figure 25).

The folding of such structures is not yet demonstrated *in vivo*; however, their extreme conservation strongly suggests a relevant role in PE-PGRS biology. Likewise, the modularity, that is equally conserved, is probably relevant for the overall picture. Given the suggested importance for PE-PGRS genes in relation to the pathogenesis of mycobacteria, unraveling the role of these stable and unique non-B DNA structures could open new possible therapeutic approaches addressed to these new potential structural targets.

PAIRED G4 STRUCTURES

BACKGROUND

G-Quadruplex (G4) structures can act as important functional and structural regulators (see G-Quadruplex section in the Introduction). However, recent evidences suggest that these structures can have higher layers of complexity than expected. Emerging data reveal that multiple G4s, that fold continuously in specific regions, can cross-talk into higher order DNA structures and potentially influence each other folding (Chaires et al., 2014; Palumbo et al., 2009; Rigo and Sissi, 2017). In collaboration with a group from the Department of Pharmaceutical and Pharmacological Sciences, I recently started to focus my work on these peculiar structures. Particularly, I tried to develop a method to detect these potential paired systems. Currently, no rules are known that allow to predict whether two G4s that are spatially-related can form such a system. Indeed, the spatial proximity is not sufficient to define the development of a paired system.

METHODS AND ANALYSES

ANALYSES OF THE C-KIT AND HTERT SYSTEMS

I started by analyzing the promoter regions of c-KIT and hTERT. These regions contain sequences that can form systems of paired G4s that have been

experimentally validated (Palumbo et al., 2009; Rigo and Sissi, 2017). Using QPARSE and NeSSie, I combined the search for multiple runs of G-islands (allowing also for several slightly degenerate islands) with information on the symmetrical properties of the sequence. Interestingly, the analyses revealed that both sequences share common features: I) They both contain 8 consecutive runs of exact or slightly degenerate G-islands spaced by short linkers (less than 6 bp); II) They both reside within the 100 bp that are upstream the transcription start site (TSS); III) Unexpectedly, they both share a very strong overall mirror symmetry (Figure 26).

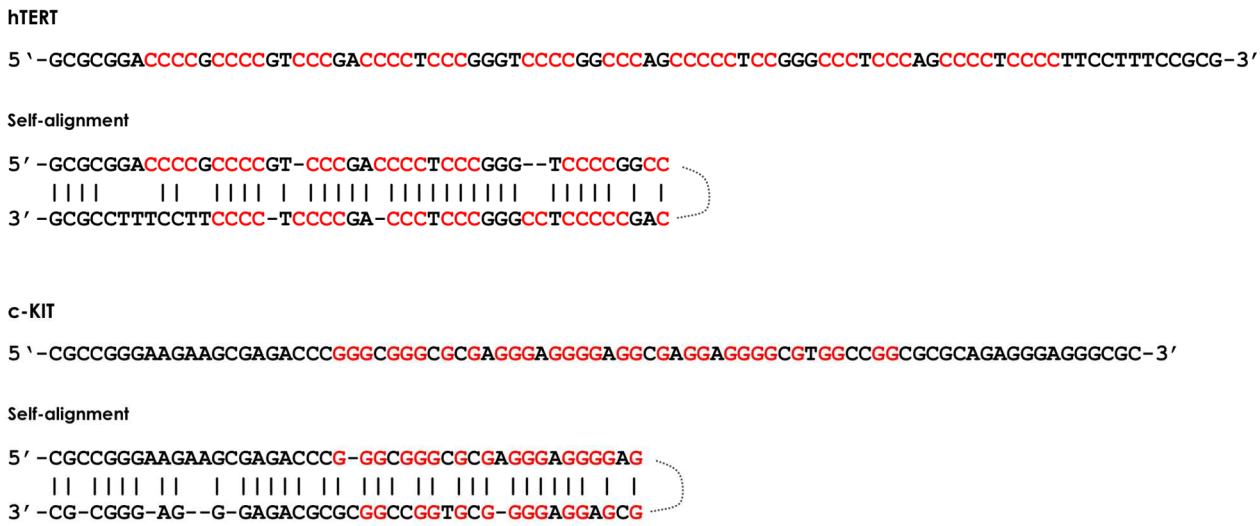


Figure 26. Mirror regions containing the sequences responsible for the formation of the paired G4s (red) in hTERT (top) and c-KIT (bottom). The self-alignment is shown for each sequence to display the mirror-symmetry.

BCL2, A NEW SYSTEM?

Using the same approach applied to hTERT and c-KIT, I extended the analysis to include also the promoter region of BCL2. It was thought that this region could have a similar system not identified yet. Indeed, as the result of the analysis, I identified a sequence that shares the very same features observed for the sequences of hTERT and c-KIT: it contains 8 consecutive runs of G-islands (with linkers that are shorter than 6 bp), and it is

characterized by an overall very strong mirror symmetry (Figure 27). Likewise, it is located within the 100 bp upstream of the TSS.

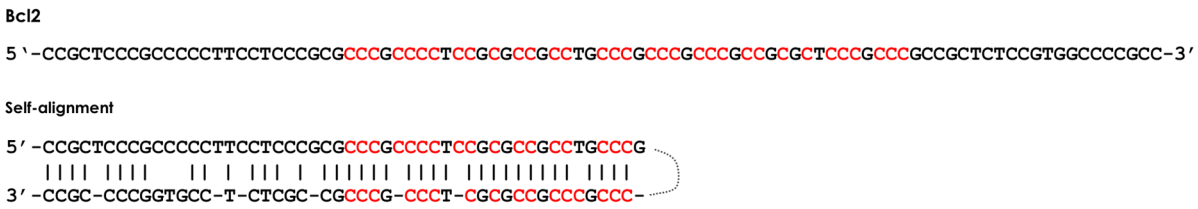


Figure 27. Mirror region containing the putative sequence responsible for the formation of the paired G4s (red) in BCL2. The self-alignment is shown to display the mirror-symmetry.

PRELIMINARY EXPERIMENTAL DATA

Given the similarity between the sequence detected in BCL2 and the sequences associated with the paired systems in c-KIT and hTERT, collaborators from the Department of Pharmaceutical and Pharmacological Sciences performed additional experimental studies on BCL2 sequence. Using the circular dichroism together with electrophoretic techniques, they demonstrated that two separate G4 structures can form simultaneously in the sequence and that these two structures are proximal enough to potentially cross-talk and interact. However, further analyses are ongoing to confirm whether, and how, these structures interact into a higher order structure and influence each other folding.

GENOME-WIDE ANALYSIS

Although the experimental data for BCL2 are preliminary, and there is a need of further analyses to confirm the cross-talk between the two G4s, they are also very promising. Indeed, the experimental evidences suggest that the two G4s can form simultaneously while being proximal enough to potentially cross-talk and interact. To evaluate the potential relevance of these structures on a wider scale, I further extended the analysis to include all the regions surrounding the TSS of all the genes that are annotated in

human. To generate the dataset, I retrieved the IDs for all the human genes that are annotated in gencode (v28) (Harrow et al., 2012). For each of the IDs, I downloaded the corresponding genomic sequence from ensembl (Zerbino et al., 2018). I considered the sequences from 15000 bp upstream up to 15000 bp downstream the TSS. To automate the downloading, I used a custom python script and the REST api provided by ensembl. For the analysis, I applied the same criteria previously used for the detection of c-KIT, hTERT, and BCL2. Basically, I searched for regions with 8 runs of G-islands and short linkers that are fully contained in wider regions with an overall strong mirror symmetry. Surprisingly, this led to the identification of a great number of putative paired systems in the TSS regions of thousands of genes (~30% of the analyzed sequences). To explore the distribution of the predicted systems, I divided the sequence range into bins of 100 bp, and for each bin calculated the number of unique genes that have a predicted system within that range. The mid-point of the mirror sequence was used to univocally associate each of the predicted systems to a unique bin. The results are shown in Figure 28. Interestingly, the predicted paired G4 systems are not randomly distributed and show a strong enrichment in the region surrounding the TSS. The enrichment is striking in correspondence of the bin that covers the 100 bp just upstream the TSS. Curiously, this is the very same location where the hTERT, c-KIT, and the putative BCL2 systems are in relation to their corresponding TSS. To perform these analyses, I used custom python scripts together with NeSSie and QPARSE tools.

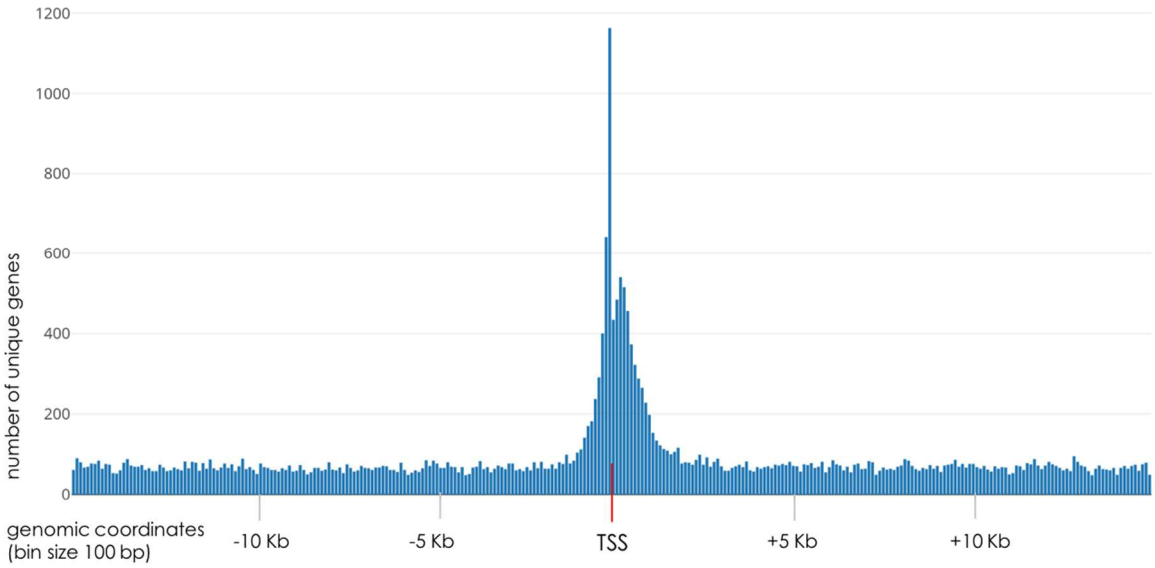


Figure 28. Distribution of putative paired G4s systems embedded in a mirror sequence around the TSS of annotated human genes. Each bar represents a bin of 100 bp. The height of each bar is the number of unique genes that have a predicted system in the correspondence of that position. The mid-point of the predicted sequences is used to univocally associate each of these putative systems to a bin.

DISCUSSION

G-Quadruplex structures are important structural and functional regulators. However, they can be far more versatile than expected. Recent evidences show that two G4s that are proximal along the sequence can cross-talk and interact. This leads to the formation of a more complex higher-order structure with the two G4s influencing each other folding. Given two G4 structures that are proximal to each other, no rules are currently known to predict whether such a complex system can form or not. Indeed, the spatial proximity of the G4s is not sufficient to predict their interaction into a paired system. Two of these systems have been experimentally demonstrated in the promoter regions of hTERT and c-KIT. The comparative analysis of these regions revealed that both the sequences share few common features. They both contain 8 consecutive G-islands (exact or slightly degenerate) that are spaced by linkers shorter than 6 bp. They are both located within the 100 bp upstream of the TSS. Finally, they are both

characterized by a very strong overall mirror symmetry (Figure 26). The latter is an unexpected feature that nobody explored so far. However, NeSSie data indeed suggest that the symmetrical properties of the sequence may be relevant for the formation of these systems. It was thought that a similar system could be located within the BCL2 promoter, and yet to be identified. Indeed, applying the same rules observed for hTERT and c-KIT, and searching for double G4s in an extended mirror region, led to the identification of a sequence that is located within the 100 bp upstream of the TSS that shares these very same features. Preliminary experimental data confirmed that two separate G4s can form simultaneously in the sequence and the two structures are proximal enough to potentially cross-talk and interact. Further analyses are scheduled to confirm this hypothesis. Although these data are still preliminary, they are strengthening the idea that the symmetry may be important for the formation of putative paired systems. Extending the same analysis to all the annotated human genes, led to the identification of a great number of putative paired systems in the promoter regions of thousands of genes. Interestingly, these systems are not distributed randomly along the sequence but are highly enriched in the immediate proximity of the TSS (Figure 28). Particularly, there is a striking enrichment for these systems immediately upstream the TSS. This is the same location where the systems in hTERT, c-KIT, and the potential one in BCL2 are located. The data are still very preliminary and need further experimental validation. Nevertheless, observing such a striking enrichment in such a small region is very unlikely to happen by chance. Given the very peculiar position with respect to the TSS, and the known role of G4 structures, it is tempting to hypothesize a potential role also for these paired systems. Supporting this idea, the genes that contain a putative system within 1000 bp around the TSS are potentially related in their biological activity. A preliminary functional enrichment analysis performed using David (Huang et al., 2009) revealed that most of the genes that are mapped in KEGG pathways (Kanehisa and Goto, 2000) are related in their

function. The most enriched pathways are associated to proliferation, development, response to chemical and mechanical signals, and cancer. Indeed, most of the genes that have been described in relation to cell proliferation and cancer have a predicted paired system in the proximity of their TSS. These observations suggest that the paired G4s systems may have a potential impact far larger than expected, involving thousands of genes. Moreover, the symmetrical properties of the sequence may potentially be relevant for their formation. To support these observations, we are now scheduling further computational and experimental analyses with our collaborators to better understand and characterize these systems. Unraveling the biological implications for these structures could improve our comprehension of a new potential regulative system that can be important for thousands of genes. Since most of these genes are involved in important diseases such as cancer, these systems could represent new potential targets for therapeutic approaches. These evidences also contribute to the idea that local structures (i.e. two G4s) can interact into higher order structures and this can represent a global mechanism that can be common to other non-B DNAs. While in this specific situation this probably applies to a regulative system, in other scenarios these higher levels of complexity can be relevant from a structural perspective.

CONCLUSIONS

During these years, I focused my research activity on the study of the alternative local structures that can form in the DNA molecule (non-B DNAs). These structures are of great interest and emerging evidences are supporting their role as important functional and structural regulators for the genome (Kouzine et al., 2017; Ohyama, 2005; Sinden, 1994). Particularly, the hypothesis is that these structures can be part of a self-assembly process that is responsible for the folding of genomes. In proteins, the transient formation of local structures (e.g. alpha-helices, beta-strands) drives a progressive folding towards the final structure by direct interactions or interactions with accessory proteins. Likewise, non-B DNAs could drive the three-dimensional folding of genomes through direct interactions or the recruitment of proteins and RNAs (see 3D Genome folding section in Introduction). This leads to a possible hierarchical model where the organization at a higher scale is defined by the sum of the interactions that progressively develop between the underlying regions, starting from the local formation of non-B DNAs. To explore such a complex model, I started by studying the basic elements of the model and I focused my attention on the alternative local structures that can form in DNA. Despite their important role, the knowledge on these structures remains limited and fragmented mostly due to the difficulties in the detection and characterization of these conformers. The formation of non-B DNAs, is dependent upon specific features of the DNA sequence and different patterns may lead to the formation of different non-B DNAs. These patterns can be detected computationally but there is currently a lack of tools that can detect these structures genome-wide. Those that are presently available are mostly limited in their search or computational efficiency and are not suitable to work with large datasets. Moreover, most of these tools are not flexible enough to detect also slightly degenerate patterns that are important for the formation of the non-B DNAs. To fill the gap, I focused my

work on the development of new computational tools for the detection of some of these patterns at a genome-wide scale. As a first step, I developed NeSSie (Berselli et al., 2018) as a tool for the detection of motifs in DNA characterized by symmetrical properties such as mirrors and palindromes. Some of these motifs are known to be associated with the formation of hairpins, cruciform structures or DNA triplexes. However, despite it is known that the mirror motifs are abundant in genomes, our knowledge of their functional and structural implications has not improved in more than 10 years. Having such a tool can hopefully contribute to extend this knowledge, allowing for a more comprehensive and complete detection and characterization of these patterns. I also worked on the detection of G-Quadruplex structures (G4s). These structures are actively studied and the knowledge is growing. However, computational tools are not keeping the pace. For example, new evidences show that also slightly degenerate patterns can sustain the formation of G4s and that multiple G4s can interact as paired units (i.e. two quadruplex structures that are close to each other along the sequence and that can fold simultaneously and interact into a higher-order structure influencing each other folding). Currently only few tools extend the search to degenerate patterns, although with some limitations, but no tools can currently detect the paired G4 systems. I tried to address this need by developing QPARSE. QPARSE can detect consecutive runs of Gs (exact or degenerate G-islands) that are involved in the formation of G4 and paired G4 structures. Another problem related to G4 data is that all the available data are very fragmented and the knowledge is scattered. With collaborators working on G4s in viruses, we tried to better organize this knowledge by performing a comprehensive evaluation of the G4s forming sequences in all the known human viruses. We collected all the data in a database accessible from web (Lavezzo et al., 2018) to provide a comprehensive resource easy to access and that can help researchers to target their work and expedite research. This revealed that G4s in viruses are highly conserved. This suggests a relevant

role for these structures that can be important functional and structural regulators. The structural role is probably extremely relevant given the single-stranded nature of the genome of many DNA and RNA viruses. Eventually, I started using these tools to perform analyses on *Mycobacterium spp.* and human genome. Preliminary data are revealing the presence of sequence patterns able to fold into previously unknown structures in both *Mycobacterium* and human genomes. In *Mycobacterium*, I identified an enriched motif with a perfect mirror symmetry that can fold into a previously unknown structure that has been characterized as a hairpin with mixed Watson-Crick and Hoogsteen base-pairings. The motif is unique of the bacteria capable of developing tuberculosis-like diseases (MTC, *M. kansasi* and *M. marinum*) and is completely absent in the human genome. The distribution of the motif is also peculiar, being enriched only in specific regions characterized by a strong modularity that correspond to the PE-PGRS regions. These regions are typical of the bacteria belonging to the MTC and are believed to be important for their unique pathogenic capabilities. While the sequence of these regions is highly variable even between very related bacteria (i.e. different strains of the same species), the enrichment in the motifs involved in the hairpins formation and the modularity are highly conserved. This suggests a prominent role of these structures and conserved patterns regardless of the overall sequence identity among PE-PGRS. The relevance of these findings and the formation *in vivo* of the hairpin structures need further validation. Their abundant and widespread distribution across the genome suggests a possible structural role both locally, for PE-PGRS, or more globally, as potential seeds involved in the formation of higher order architectures. Further analyses are ongoing to confirm these hypotheses. Given the suggested importance for PE-PGRS, understanding the role of such structures, if demonstrated, could possibly provide new insights on the biology of those *Mycobacterium* species that impact and are more relevant for the human health. These observations also support the idea

that non-B DNA structures can be involved in structural mechanisms and higher levels of organization across genomes. In the human genome, I focused my analyses on the paired-quadruplex structures. While no exact rules are currently known, apparently, the symmetry may be important for the formation of such a system. Indeed, by combining the search for coupled-quadruplex and mirror symmetries I identified a potential paired system in the promoter of BCL2 that is undergoing experimental validation. Surprisingly, by extending the analysis genome-wide I detected an enrichment for sequences with the potential to form this paired system just upstream the TSS (Transcription Starting Site) of thousands of human genes. Preliminary functional analyses indicate that these genes are mostly involved in signaling pathways that control cell growth, proliferation and interaction with the surrounding environment, being active during development. Although these observations are still preliminary, it is fascinating how these data suggest that such a complex system can have an important role in the biology of so many important genes that are associated to fearsome diseases as cancer. The data support the idea that non-B DNAs can interact into higher order structures and that the formation of these systems may represent a global and general mechanism rather than exceptions. This contributes to the idea of a hierarchical model where local structures that interact with each other lead to higher order and more complex systems that still act as functional and structural regulators. While the paired G4 systems potentially represent regulative systems of gene expression, other non-B DNAs can more likely interact into higher order structures that can be more relevant from a structural perspective. Together, these results contribute to the idea that non-B DNAs can play important functional and potentially structural roles. They also suggest that the folding landscape of the DNA molecule is much more complex than previously assumed, and we have a huge lack of knowledge towards these alternative structures. Particularly, we have a huge gap in the understanding of the higher order level of interactions that can involve the

non-B DNAs. Considering these evidences, the DNA sequence needs to be widely re-evaluated not only for the encoded genetic information, but considering also its structural properties. It is necessary to direct efforts towards new fields of investigation by studying and characterizing these structures genome-wide. Unraveling the properties of the non-B DNAs local structural conformations is the first step to progressively understand more complex and higher order structural systems towards the comprehension of the three-dimensional folding of the genome.

BIBLIOGRAPHY

Agarwala, P., Pandey, S., and Maiti, S. (2015). The tale of RNA G-quadruplex. Org. Biomol. Chem. *13*, 5570–5585.

Albert, B., Léger-Silvestre, I., Normand, C., and Gadal, O. (2012). Nuclear organization and chromatin dynamics in yeast: biophysical models or biologically driven interactions? Biochim. Biophys. Acta *1819*, 468–481.

Albertini, A.M., Hofer, M., Calos, M.P., and Miller, J.H. (1982). On the formation of spontaneous deletions: the importance of short sequence homologies in the generation of large deletions. Cell *29*, 319–328.

Anfinsen, C.B. (1973). Principles that govern the folding of protein chains. Science *181*, 223–230.

Badrinarayanan, A., Le, T.B.K., and Laub, M.T. (2015). Bacterial Chromosome Organization and Segregation. Annu. Rev. Cell Dev. Biol. *31*, 171–199.

Bedrat, A., Lacroix, L., and Mergny, J.-L. (2016). Re-evaluation of G-quadruplex propensity with G4Hunter. Nucleic Acids Res. *44*, 1746–1759.

Berselli, M., Lavezzo, E., and Toppo, S. (2018). NeSSie: a tool for the identification of approximate DNA sequence symmetries. Bioinforma. Oxf. Engl. *34*, 2503–2505.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.-M., and Lemaitre, J.-M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. *19*, 837–844.

Boles, T.C., and Hogan, M.E. (1987). DNA structure equilibria in the human c-myc gene. Biochemistry *26*, 367–376.

Brázdová, M., Tichý, V., Helma, R., Bažantová, P., Polášková, A., Krejčí, A., Petr, M., Navrátilová, L., Tichá, O., Nejedlý, K., et al. (2016). p53 Specifically Binds Triplex DNA In Vitro and in Cells. PLoS ONE *11*.

Brinton, B.T., Caddle, M.S., and Heintz, N.H. (1991). Position and orientation-dependent effects of a eukaryotic Z-triplex DNA motif on episomal DNA replication in COS-7 cells. J. Biol. Chem. *266*, 5153–5161.

Buske, F.A., Mattick, J.S., and Bailey, T.L. (2011). Potential in vivo roles of nucleic acid triple-helices. RNA Biol. *8*, 427–439.

Cammas, A., and Millevoi, S. (2017). RNA G-quadruplexes: emerging mechanisms in disease. Nucleic Acids Res. *45*, 1584–1595.

Cattoni, D.I., Cardozo Gizzi, A.M., Georgieva, M., Di Stefano, M., Valeri, A., Chamousset, D., Houbron, C., Déjardin, S., Fiche, J.-B., González, I., et al. (2017). Single-cell absolute contact probability detection reveals chromosomes are organized by multiple low-frequency yet specific interactions. Nat. Commun. *8*, 1753.

Chaires, J.B., Trent, J.O., Gray, R.D., Dean, W.L., Buscaglia, R., Thomas, S.D., and Miller, D.M. (2014). An improved model for the hTERT promoter quadruplex. PloS One *9*, e115580.

Cogoi, S., and Xodo, L.E. (2006). G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. Nucleic Acids Res. *34*, 2536–2549.

Cooney, M., Czernuszewicz, G., Postel, E.H., Flint, S.J., and Hogan, M.E. (1988). Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro. Science *241*, 456–459.

Cremer, T., Cremer, M., Hübner, B., Strickfaden, H., Smeets, D., Popken, J., Sterr, M., Markaki, Y., Rippe, K., and Cremer, C. (2015). The 4D nucleome: Evidence for a dynamic nuclear landscape based on co-aligned active and inactive nuclear compartments. FEBS Lett. *589*, 2931–2943.

David, A.P., Margarit, E., Domizi, P., Banchio, C., Armas, P., and Calcaterra, N.B. (2016). G-quadruplexes as novel cis-elements controlling transcription during embryonic development. Nucleic Acids Res. *44*, 4163–4173.

Davis, T.L., Firulli, A.B., and Kinniburgh, A.J. (1989). Ribonucleoprotein and protein factors bind to an H-DNA-forming c-myc DNA element: possible regulators of the c-myc gene. Proc. Natl. Acad. Sci. *86*, 9682–9686.

Day, H.A., Pavlou, P., and Waller, Z.A.E. (2014). i-Motif DNA: structure, stability and targeting with ligands. Bioorg. Med. Chem. *22*, 4407–4418.

Dekker, J., Marti-Renom, M.A., and Mirny, L.A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat. Rev. Genet. *14*, 390–403.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Doyle, B., Fudenberg, G., Imakaev, M., and Mirny, L.A. (2014). Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Comput. Biol. *10*, e1003867.

Duckett, D.R., Murchie, A.I., Diekmann, S., von Kitzing, E., Kemper, B., and Lilley, D.M. (1988). The structure of the Holliday junction, and its resolution. Cell *55*, 79–89.

Edgar, R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. *32*, 1792–1797.

Fraser, J., Williamson, I., Bickmore, W.A., and Dostie, J. (2015). An Overview of Genome Organization and How We Got There: from FISH to Hi-C. Microbiol. Mol. Biol. Rev. *79*, 347–372.

Gajarský, M., Živković, M.L., Stadlbauer, P., Pagano, B., Fiala, R., Amato, J., Tomáška, L., Šponer, J., Plavec, J., and Trantírek, L. (2017). Structure of a Stable G-Hairpin. J. Am. Chem. Soc. *139*, 3591–3594.

Geyer, P.K., and Corces, V.G. (1992). DNA position-specific repression of transcription by a Drosophila zinc finger protein. Genes Dev. *6*, 1865–1873.

Glucksmann, M.A., Markiewicz, P., Malone, C., and Rothman-Denes, L.B. (1992). Specific sequences and a hairpin structure in the template strand are required for N4 virion RNA polymerase promoter recognition. Cell *70*, 491–500.

Goobes, R., Cohen, O., and Minsky, A. (2002). Unique condensation patterns of triplex DNA: physical aspects and physiological implications. Nucleic Acids Res. *30*, 2154–2161.

Hampel, K.J., Burkholder, G.D., and Lee, J.S. (1993). Plasmid dimerization mediated by triplex formation between polypyrimidine-polypurine repeats. Biochemistry *32*, 1072–1077.

Hänsel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M., et al. (2016). G-quadruplex structures mark human regulatory chromatin. Nat. Genet. *48*, 1267–1272.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

Hirashima, K., and Seimiya, H. (2015). Telomeric repeat-containing RNA/G-quadruplex-forming sequences cause genome-wide alteration of gene expression in human cancer cells in vivo. Nucleic Acids Res. *43*, 2022–2032.

Hon, J., Martínek, T., Zendulka, J., and Lexa, M. (2017). pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. Bioinformatics *33*, 3373–3379.

Hoshina, S., Yura, K., Teranishi, H., Kiyasu, N., Tominaga, A., Kadoma, H., Nakatsuka, A., Kunichika, T., Obuse, C., and Waga, S. (2013). Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. J. Biol. Chem. *288*, 30161–30171.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. *37*, 1–13.

Huppert, J.L., and Balasubramanian, S. (2005). Prevalence of quadruplexes in the human genome. Nucleic Acids Res. *33*, 2908–2916.

Jayaraj, G.G., Pandey, S., Scaria, V., and Maiti, S. (2012). Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biol. 9, 81–86.

Kajava, A.V. (2001). Review: proteins with repeated sequence--structural prediction and modeling. J. Struct. Biol. *134*, 132–144.

Kajava, A.V. (2012). Tandem repeats in proteins: from sequence to structure. J. Struct. Biol. *179*, 279–288.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Karplus, M., and Weaver, D.L. (1976). Protein-folding dynamics. Nature *260*, 404–406.

Karplus, M., and Weaver, D.L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. Protein Sci. Publ. Protein Soc. *3*, 650–668.

Kikin, O., D'Antonio, L., and Bagga, P.S. (2006). QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. Nucleic Acids Res. *34*, W676-682.

Kinniburgh, A.J. (1989). A cis-acting transcription element of the c-myc gene can assume an H-DNA conformation. Nucleic Acids Res. *17*, 7771–7778.

van Koningsbruggen, S., Gierlinski, M., Schofield, P., Martin, D., Barton, G.J., Ariyurek, Y., den Dunnen, J.T., and Lamond, A.I. (2010). High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. Mol. Biol. Cell *21*, 3735–3748.

Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K.-R., Benham, C.J., Casellas, R., Przytycka, T.M., et al. (2017). Permanganate/S1 Nuclease Footprinting Reveals Non-B DNA Structures with Regulatory Potential across a Mammalian Genome. Cell Syst. *4*, 344-356.e7.

Lavezzo, E., Berselli, M., Frasson, I., Perrone, R., Palù, G., Brazzale, A., Richter, S., and Toppo, S. (2018). G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide. BioRxiv 344127.

Lee, J.S., Woodsworth, M.L., Latimer, L.J., and Morgan, A.R. (1984). Poly(pyrimidine) . poly(purine) synthetic DNAs containing 5-methylcytosine form stable triplexes at neutral pH. Nucleic Acids Res. *12*, 6603–6614.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009).

Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Lipps, H.J., and Rhodes, D. (2009). G-quadruplex structures: in vivo evidence and function. Trends Cell Biol. *19*, 414–422.

Machado, T., and A, J. (2012). Shannon Entropy Analysis of the Genome Code.

Maizels, N., and Gray, L.T. (2013). The G4 genome. PLoS Genet. *9*, e1003468.

Markiewicz, P., Malone, C., Chase, J.W., and Rothman-Denes, L.B. (1992). Escherichia coli single-stranded DNA-binding protein is a supercoiled template-dependent transcriptional activator of N4 virion RNA polymerase. Genes Dev. *6*, 2010–2019.

Meena, L.S. (2015). An overview to understand the role of PE_PGRS family proteins in Mycobacterium tuberculosis H37 Rv and their potential as new drug targets. Biotechnol. Appl. Biochem. *62*, 145–153.

Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, R.M., and Mergny, J.-L. (2016). G-quadruplexes and helicases. Nucleic Acids Res. *44*, 1989–2006.

Meuleman, W., Peric-Hupkes, D., Kind, J., Beaudry, J.-B., Pagie, L., Kellis, M., Reinders, M., Wessels, L., and van Steensel, B. (2013). Constitutive nuclear lamina-genome interactions are highly conserved and associated with A/T-rich sequence. Genome Res. *23*, 270–280.

Mirihana Arachchilage, G., Dassanayake, A.C., and Basu, S. (2015). A potassium ion-dependent RNA structural switch regulates human pre-miRNA 92b maturation. Chem. Biol. *22*, 262–272.

Mukundan, V.T., and Phan, A.T. (2013). Bulges in G-quadruplexes: broadening the definition of G-quadruplex-forming sequences. J. Am. Chem. Soc. *135*, 5017–5028.

Muravyova, E., Golovnin, A., Gracheva, E., Parshikov, A., Belenkaya, T., Pirrotta, V., and Georgiev, P. (2001). Loss of insulator activity by paired Su(Hw) chromatin insulators. Science *291*, 495–498.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Noirot, P., Bargonetti, J., and Novick, R.P. (1990). Initiation of rolling-circle replication in pT181 plasmid: initiator protein enhances cruciform extrusion at the origin. Proc. Natl. Acad. Sci. U. S. A. *87*, 8560–8564.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature *485*, 381–385.

Nothjunge, S., Nührenberg, T.G., Grüning, B.A., Doppler, S.A., Preissl, S., Schwaderer, M., Rommel, C., Krane, M., Hein, L., and Gilsbach, R. (2017). DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes. Nat. Commun. *8*, 1667.

Ohyama, T. (2005). DNA Conformation and Transcription (Boston, MA: Springer US).

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Orlov, Y.L., and Potapov, V.N. (2004). Complexity: an internet resource for analysis of DNA sequence complexity. Nucleic Acids Res. *32*, W628-633.

Paeschke, K., Bochman, M.L., Garcia, P.D., Cejka, P., Friedman, K.L., Kowalczykowski, S.C., and Zakian, V.A. (2013). Pif1 family helicases suppress genome instability at G-quadruplex motifs. Nature *497*, 458–462.

Palumbo, S.L., Ebbinghaus, S.W., and Hurley, L.H. (2009). Formation of a unique end-to-end stacked pair of G-quadruplexes in the hTERT core promoter with implications for inhibition of telomerase by G-quadruplex-interactive ligands. J. Am. Chem. Soc. *131*, 10878–10891.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular maps of the reorganization of genome-nuclear lamina interactions during differentiation. Mol. Cell *38*, 603–613.

Perrone, R., Lavezzo, E., Riello, E., Manganelli, R., Palù, G., Toppo, S., Provvedi, R., and Richter, S.N. (2017). Mapping and characterization of G-quadruplexes in Mycobacterium tuberculosis gene promoter regions. Sci. Rep. *7*.

Phillips-Cremins, J.E., and Corces, V.G. (2013). Chromatin insulators: linking genome organization to cellular function. Mol. Cell *50*, 461–474.

Postel, E.H., Mango, S.E., and Flint, S.J. (1989). A nuclease-hypersensitive element of the human c-myc promoter interacts with a transcription initiation factor. Mol. Cell. Biol. 9, 5123–5133.

Praseuth, D., Guieysse, A.L., and Hélène, C. (1999). Triple helix formation and the antigene strategy for sequence-specific control of gene expression. Biochim. Biophys. Acta *1489*, 181–206.

Qiu, J., Wang, M., Zhang, Y., Zeng, P., Ou, T.-M., Tan, J.-H., Huang, S.-L., An, L.-K., Wang, H., Gu, L.-Q., et al. (2015). Biological Function and Medicinal Research Significance of G-Quadruplex Interactive Proteins. Curr. Top. Med. Chem. *15*, 1971–1987.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K.C., Grüning, B.A., Villaveces, J., Habermann, B., Akhtar, A., and Manke, T. (2018). High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nat. Commun. 9, 189.

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell *159*, 1665–1680.

Rhodes, D., and Lipps, H.J. (2015). G-quadruplexes and their regulatory roles in biology. Nucleic Acids Res. *43*, 8627–8637.

Ribeyre, C., Lopes, J., Boulé, J.-B., Piazza, A., Guédin, A., Zakian, V.A., Mergny, J.-L., and Nicolas, A. (2009). The Yeast Pif1 Helicase Prevents Genomic Instability Caused by G-Quadruplex-Forming CEB1 Sequences In Vivo. PLOS Genet. *5*, e1000475.

Rigo, R., and Sissi, C. (2017). Characterization of G4-G4 Crosstalk in the c-KIT Promoter Region. Biochemistry *56*, 4309–4312.

Rodriguez, R., Miller, K.M., Forment, J.V., Bradshaw, C.R., Nikan, M., Britton, S., Oelschlaegel, T., Xhemalce, B., Balasubramanian, S., and Jackson, S.P. (2012). Small-molecule-induced DNA damage identifies alternative DNA structures in human genes. Nat. Chem. Biol. *8*, 301–310.

Ruan, H., and Wang, Y.-H. (2008). Friedreich's ataxia GAA.TTC duplex and GAA.GAA.TTC triplex structures exclude nucleosome assembly. J. Mol. Biol. *383*, 292–300.

Satange, R., Chang, C.-K., and Hou, M.-H. (2018). A survey of recent unusual high-resolution DNA structures provoked by mismatches, repeats and ligand binding. Nucleic Acids Res. *46*, 6416–6434.

Scaria, V., Hariharan, M., Arora, A., and Maiti, S. (2006). Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. Nucleic Acids Res. *34*, W683-685.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N.A., Huber, W., H Haering, C., Mirny, L., et al.

(2017). Two independent modes of chromatin organization revealed by cohesin removal. Nature *551*, 51–56.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shannon, C.E. (1948). A mathematical theory of communication. Bell Syst. Tech. J. *27*, 379–423.

Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H. (2002). Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription. Proc. Natl. Acad. Sci. U. S. A. *99*, 11593–11598.

Sinden, R.R. (1994). DNA Structure and Function (Gulf Professional Publishing).

Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. Nature *544*, 59–64.

Takahama, K., Takada, A., Tada, S., Shimizu, M., Sayama, K., Kurokawa, R., and Oyoshi, T. (2013). Regulation of telomere length by G-quadruplex telomere DNA- and TERRA-binding protein TLS/FUS. Chem. Biol. *20*, 341–350.

Tosoni, E., Frasson, I., Scalabrin, M., Perrone, R., Butovskaya, E., Nadai, M., Palù, G., Fabris, D., and Richter, S.N. (2015). Nucleolin stabilizes G-quadruplex structures folded by the LTR promoter and silences HIV-1 viral transcription. Nucleic Acids Res. *43*, 8884–8897.

Trifonov, E.N. (1990). Making sense of the human genome. Struct. Methods *1*, *Human Genome Initiative and DNA Recombination*, 69–77.

Udvardy, A., Maine, E., and Schedl, P. (1985). The 87A7 chromomere. Identification of novel chromatin structures flanking the heat shock locus that may define the boundaries of higher order domains. J. Mol. Biol. *185*, 341–358.

Umek, R.M., and Kowalski, D. (1988). The ease of DNA unwinding as a determinant of initiation at yeast replication origins. Cell *52*, 559–567.

Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M.I.R., Ding, H., and Boulton, S.J. (2012). RTEL1 dismantles T loops and counteracts telomeric G4-DNA to maintain telomere integrity. Cell *149*, 795–806.

Varizhuk, A., Ischenko, D., Tsvetkov, V., Novikov, R., Kulemin, N., Kaluzhny, D., Vlasenok, M., Naumov, V., Smirnov, I., and Pozmogova, G. (2017). The expanding repertoire of G4 DNA structures. Biochimie *135*, 54–62.

Veselkov, A.G., Malkov, V.A., Frank-Kamenetskll, M.D., and Dobrynin, V.N. (1993). Triplex model of chromosome ends. Nature *364*, 496.

Warren, G.J., and Green, R.L. (1985). Comparison of physical and genetic properties of palindromic DNA sequences. J. Bacteriol. *161*, 1103–1111.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinforma. Oxf. Engl. *25*, 1189–1191.

Weber, S.C., Spakowitz, A.J., and Theriot, J.A. (2010). Bacterial chromosomal loci move subdiffusively through a viscoelastic cytoplasm. Phys. Rev. Lett. *104*, 238102.

Westin, L., Blomquist, P., Milligan, J.F., and Wrange, O. (1995). Triple helix DNA alters nucleosomal histone-DNA interactions and acts as a nucleosome barrier. Nucleic Acids Res. *23*, 2184–2191.

Zeraati, M., Langley, D.B., Schofield, P., Moye, A.L., Rouet, R., Hughes, W.E., Bryan, T.M., Dinger, M.E., and Christ, D. (2018). I-motif DNA structures are formed in the nuclei of human cells. Nat. Chem. *10*, 631.

Zerbino, D.R., Achuthan, P., Akanni, W., Amode, M.R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C.G., et al. (2018). Ensembl 2018. Nucleic Acids Res. *46*, D754–D761.

Zheng, R., Shen, Z., Tripathi, V., Xuan, Z., Freier, S.M., Bennett, C.F., Prasanth, S.G., and Prasanth, K.V. (2010). Polypurine-repeat-containing RNAs: a novel class of long non-coding RNA in mammalian cells. J. Cell Sci. *123*, 3734–3744.

Zuin, J., Dixon, J.R., van der Reijden, M.I.J.A., Ye, Z., Kolovos, P., Brouwer, R.W.W., van de Corput, M.P.C., van de Werken, H.J.G., Knoch, T.A., van IJcken, W.F.J., et al. (2014). Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proc. Natl. Acad. Sci. U. S. A. *111*, 996–1001.

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Application Note

NeSSie: a tool for the identification of approximate DNA sequence symmetries

Michele Berselli[1], Enrico Lavezzo[1] and Stefano Toppo[1,*]

[1]Department of Molecular Medicine, University of Padova, viale G. Colombo,3, I-35131 Padova (ITALY)

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Non-B DNA conformations play an important role in genomic rearrangements, structural three-dimensional organization and gene regulation. Many non-B DNA structures show symmetrical properties as palindromes and mirrors that can form hairpins, cruciform structures or triplexes. A comprehensive tool, capable to perform a fast genome wide search for exact and degenerate symmetrical patterns, is needed for further investigating nucleotide tracts potentially forming non-B DNA structures.
Results: We developed NeSSie, an easy customizable C/C++ 64-bit library and tool, based on dynamic programming, to quickly scan for perfect and degenerate DNA palindromes, mirrors, and potential triplex forming patterns. In addition, the tool computes linguistic complexity and Shannon entropy measures to verify the repetitive nature of the DNA regions enriched in these motifs. As a case study, the analysis of the *Mycobacterium bovis* genome is presented.
Availability: http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=nessie
https://github.com/B3rse/nessie
Contact: stefano.toppo@unipd.it
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Canonical Watson-Crick double helix of DNA may adopt more than 12 alternative conformations collectively named non-B DNA structures (Wells, 1988). These conformers play an important role in different physiological and pathological cellular conditions, and influence many biochemical properties of the genome ranging from DNA rearrangements to transcription regulation (Sinden, 1994). The regions potentially forming non-B DNA structures usually contain distinct features at primary sequence level, which may be easily recognized with classical exact pattern matching techniques. Nonetheless, experimental evidences revealed that sequence patterns forming non-B DNA structures may be highly polymorphic and degenerate (Kaushik and Kukreti, 2006). This has been reported for many non-B DNA conformers like G-quadruplexes (G4), triplexes, cruciforms, hairpins, etc. This led to the development of adapted algorithmic strategies for imperfection-tolerant searches as, for example, the recent pqsfinder tool for the detection of degenerate G4s (Hon, et al., 2017). To this respect, there are still few tools covering exhaustively and efficiently other potentially forming non-B DNA structures carrying non-perfect sequence patterns. We focused our work on symmetrical DNA sequence patterns, mirrors and palindromes. Mirrors are inverted repeats occurring within each individual strand. Recently, some mirror repeats with a mixed composition of purines and pyrimidines have been found to form a stable G-hairpin conformation (Gajarsky, et al., 2017) and it is known that polypyrimidine/polypurine tracts may form intramolecular interactions and generate triplexes or H-DNA. Palindromes are also inverted repeats but occurring over two strands of DNA; this creates self-complementary tracts within each strand with the potential to generate hairpins and cruciform structures. Addressing these non-B DNA sequence motifs, we developed a customizable imperfection-tolerant algorithm based on dynamic programming for the optimal search of motifs with such symmetrical properties. NeSSie (**N**ucleic-acids **e**lements of **S**equence **S**ymmetry **id**entification) is a C/C++ 64-bit library and a comprehensive tool capable to perform a genome wide exhaustive search, together with the calculation of Shannon entropy and simple linguistic complexity measures to correlate the potential enrichment of identified patterns in genomic regions with a peculiar nucleotide composition.

2 Methods

NeSSie library and tool implement different measures and searches (details are reported in Suppl. Mat. S1).

2.1 Search for degenerate mirrors and palindromes

We apply the optimal global alignment based on Needleman-Wunsch algorithm of dynamic programming to identify degenerate motifs. Perfect mirrors and palindromes are repeats with a point symmetry, but allowing for gaps and mismatches poses the problem of finding where the best central inversion is. To work around this issue, the alignment matrix is generated using the sequence itself and its inverted copy. Two different scoring matrices are used to test for mirror and palindromic symmetries respectively. Backtracking of the self-alignment starts from the highest score along the diagonal of the scoring matrix up to the first cell of the matrix itself. In this way, the optimal solution of the alignment is reported, where the central point of symmetry can be offset from the exact center of the sequence. By applying a sliding window, the tool also searches for all the possible combinations within a specified length range and reports the highest scoring solution satisfying the chosen threshold. The tool uses pattern matching and speeds up if exact mirrors and palindromes are requested.

2.2 Shannon entropy of DNA sequence

Shannon entropy measures the amount of non-compressible information contained in a message. It estimates the order state (or entropy) of the sequence and allows to detect unbalances in base composition of the DNA.

2.3 Simple linguistic complexity of DNA sequence

The Linguistic Sequence Complexity is an index that measures the vocabulary richness of a sequence. It is calculated as the level of repetition of the k-mers (words of length k) in the sequence. The more complex the sequence is, the richer is the vocabulary it contains; vice versa, the lower the complexity is, the more repetitive is the sequence.

2.4 Output wrappers

Python tools (v 2.7) are provided to transform NeSSie output in a more user friendly and readable format. The tools can also generate gff/wig files that can be imported in compatible genome browsers.

3 Results

We have benchmarked and compared NeSSie with some available tools (details are reported in Suppl. Mat. S1).

3.1 Benchmark design and datasets

We randomly generated a set of 1200 mirrors and 1200 palindromes from 15bp to 30bp in length and two third of these motifs were degenerated to affect their starting perfect symmetry. Motifs were inserted in: i) a repeated 'ACTG' sequence 5 Mb long originally not containing mirrors and palindromes (from herein ACGT), ii) *E. Coli* genome (Accid: U00096.3), possibly containing other mirrors and palindromes (from herein COLI).
For searching potential triplex forming patterns, the dataset containing 'true' triplexes retrieved by Lexa et al. (Lexa, et al., 2011) was employed. These triplexes were also used as seeds to generate five decoy sets.

3.2 Benchmark results of mirrors and palindromes

We checked the ability of NeSSie to retrieve all the inserted motifs in ACGT and COLI. All true positives were detected by NeSSie both in ACGT and COLI and no false positives were found in ACGT, as expected. In COLI, preexisting potentially forming mirrors and palindromes found by NeSSie were not considered.

3.3 Benchmark results of triplexes

NeSSie performance was compared with specialized Triplex tool designed to search for imperfection-tolerant triplexes (Lexa, et al., 2011). We assessed the ability of the tools to discriminate between true and decoy triplexes in the differently designed datasets. NeSSie performed slightly better in the datasets from 25% to 40% of decoy degeneracy, proving to be competitive.

3.4 Genome browser visualization: *M. bovis* example

We scanned the whole genome of *M. bovis* with NeSSie. We observed an uneven distribution of symmetric motifs in the genome and particularly

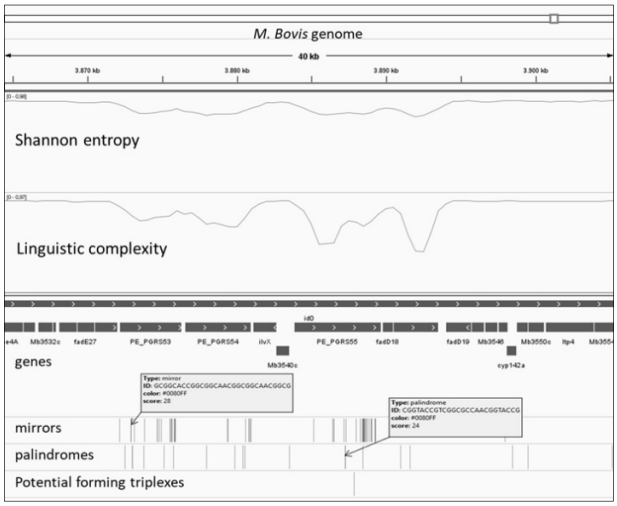


Figure 1 Screenshot of a zoomed region of *M. bovis* visualized in IGV. Motifs are reported as vertical bars where gray shades reflect scores increasing from light to dark gray. Shannon entropy and linguistic complexity profiles range from 0 to 1.

clustered in correspondence of PE-PGRS glycine-rich protein genes, whose function is still elusive (Meena, 2015). In addition, a parallel drop in both linguistic complexity and Shannon entropy measures was detected in these regions revealing that these sequence tracts display a peculiar base composition worth investigating in detail. An example of this trend is shown in Figure 1, where a zoomed region of *M. bovis* extracted from IGV genome browser (Robinson, et al., 2011) is reported. A detailed analysis and complete visualization data are reported in Suppl. Mat. S1 and NeSSie web site respectively.

4 Conclusions

We developed NeSSie as a tool and C/C++ 64-bit library to accomplish different tasks, especially exhaustive and fast searches of imperfection-tolerant symmetric motifs at genome wide scale. To our knowledge, this is the first approach that can perform searches of mirrors, potential triplex forming patterns, and palindromes effectively using slight different modifications of the implemented dynamic programming algorithm.

Funding

This work has been supported by University of Padova grant: CPDA138081/13
Conflict of Interest: none declared.

References

Gajarsky, M., *et al.* Structure of a Stable G-Hairpin. *Journal of the American Chemical Society* 2017;139(10):3591-3594.

Hon, J., *et al.* pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* 2017;33(21):3373-3379.

Kaushik, M. and Kukreti, S. Structural polymorphism exhibited by a quasipalindrome present in the locus control region (LCR) of the human beta-globin gene cluster. *Nucleic acids research* 2006;34(12):3511-3522.

Lexa, M., *et al.* A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics* 2011;27(18):2510-2517.

Meena, L.S. An overview to understand the role of PE_PGRS family proteins in Mycobacterium tuberculosis H37 Rv and their potential as new drug targets. *Biotechnology and applied biochemistry* 2015;62(2):145-153.

Robinson, J.T., *et al.* Integrative genomics viewer. *Nature biotechnology* 2011;29(1):24-26.

Sinden, R.R. DNA structure and function. San Diego CA: Academic Press; 1994.

Wells, R.D. Unusual DNA structures. *The Journal of biological chemistry* 1988;263(3):1095-1098.

NESSIE SUPPLEMENTARY MATERIALS

TABLE OF CONTENTS

TOOL FEATURES

ALGORITHM TO DETECT DEGENERATE MOTIFS

Imperfection-tolerant palindromes and mirrors may contain mismatches but also insertions/deletions that impair the symmetry of the motifs. This poses the problem of finding the best symmetry point that divides the sequence in two self-complementary arms. To accomplish this task, we implemented a strategy based on dynamic programming applied to sliding windows as follows:

1) The input sequence is scanned with a sliding window, whose length is defined by the user, shifting along the sequence one base at a time.
2) For each sliding window a global alignment approach based on Needleman-Wunsch algorithm [1] is applied to search for potential symmetries with the following different modalities:
   a. **Motifs of length equal to sliding window size and satisfying the selected scoring cutoffs (-k N parameter of NeSSie, see Figure 1).**
   The entire DNA sequence captured by the sliding window (size defined by -k) is tested with the global alignment algorithm and reported if satisfies the selected scoring cutoffs (degeneracy parameters).

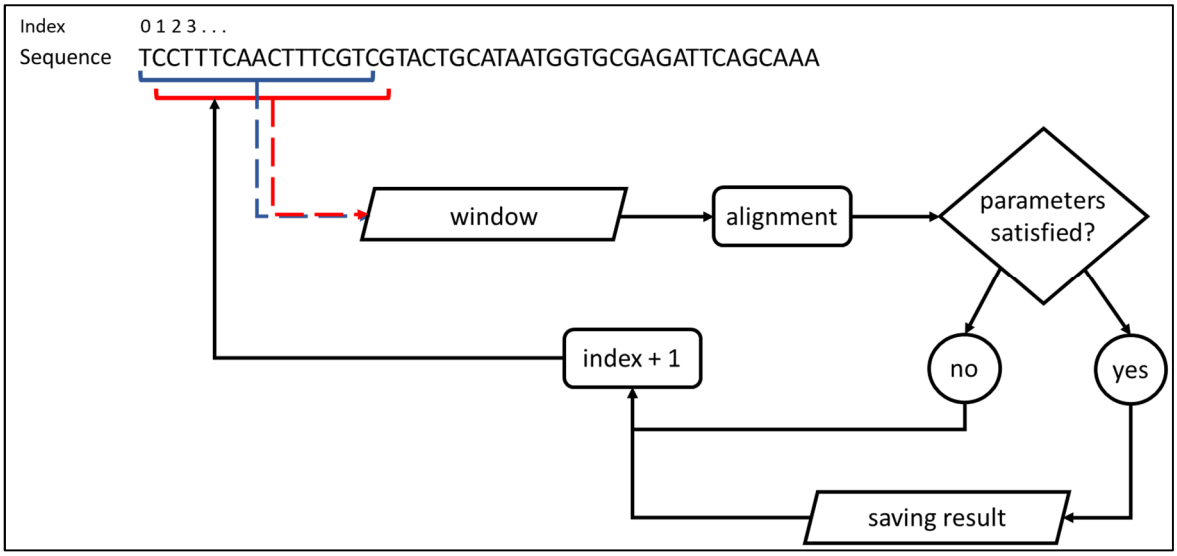


Figure 1 - Motifs of length equal to the sliding window size and satisfying the selected scoring cutoffs (-k N parameter of NeSSie).

   b. **All motifs falling in the range of min/max lengths and satisfying the selected scoring cutoffs (-k N, -K N parameters of NeSSie, see Figure 2)**
   All the sequences from max to min length (window size defined by –K, max length) are tested with the global alignment algorithm and only those that satisfy the selected

scoring thresholds (degeneracy parameters) are reported at each position.

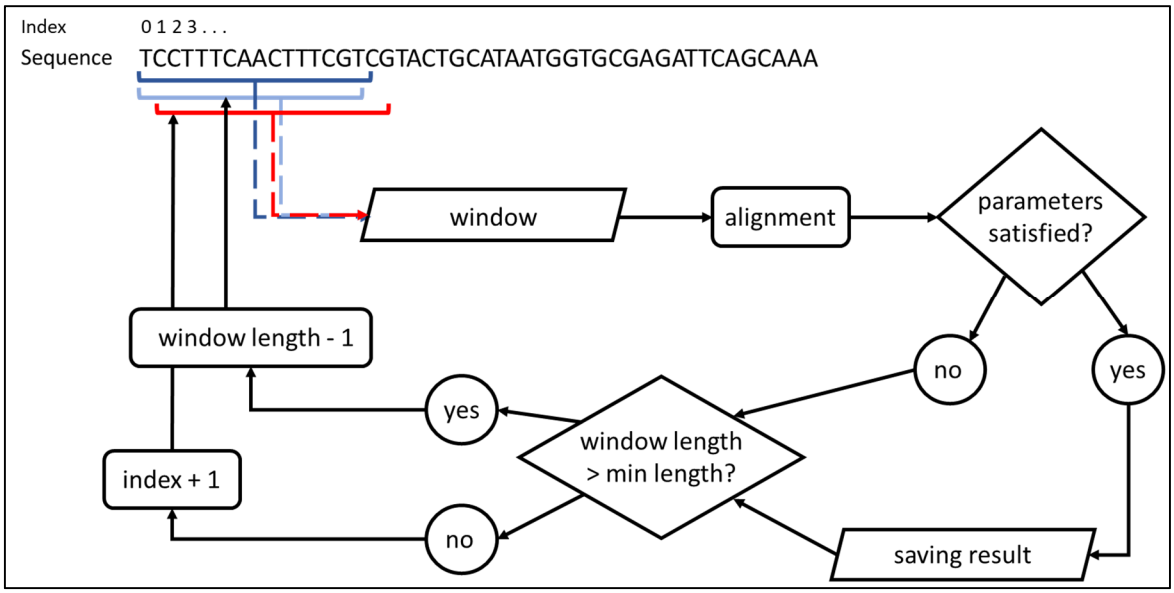


Figure 2 - All motifs falling in the range of min/max lengths and satisfying the selected scoring cutoffs (-k N, -K N parameters of NeSSie).

c. Longest max scoring motif falling in the range of min/max lengths and satisfying the selected scoring cutoffs (-MAX, -k N, -K N parameters of NeSSie see Figure 3)

All the sequences from max to min length (window size defined by –K, max length) are tested with the global alignment algorithm and only the longest sequence that satisfies also the selected scoring thresholds (degeneracy parameters) is reported at each position.

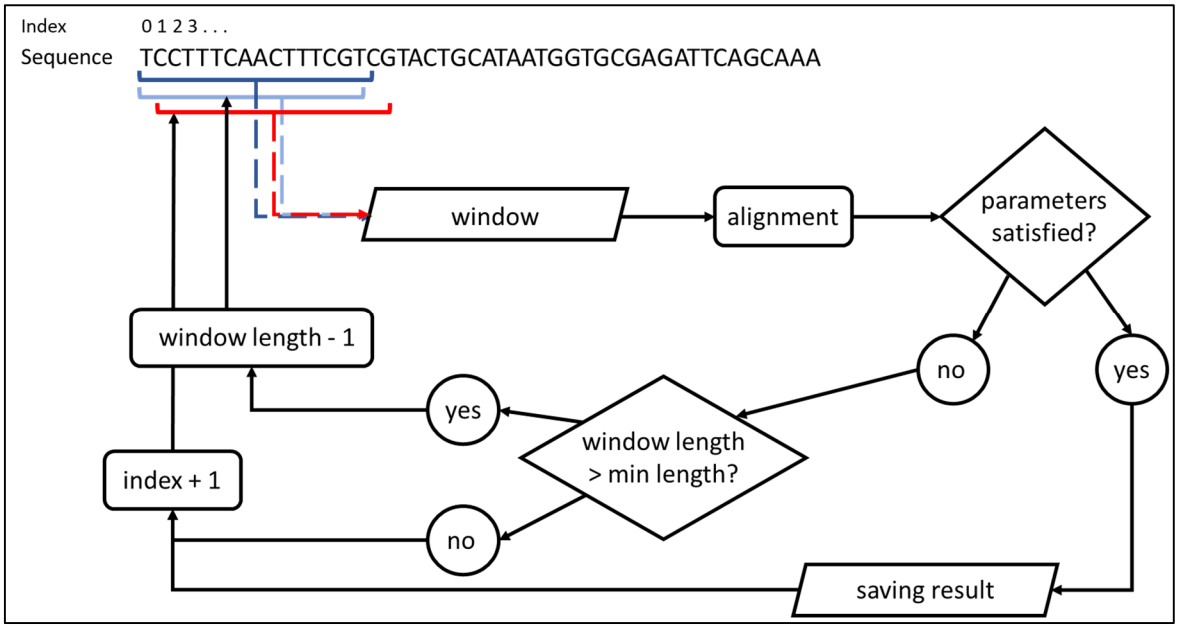


Figure 3 - Longest max scoring motif falling in the range of min/max lengths and satisfying the selected scoring cutoffs (-MAX, -k N, -K N parameters of NeSSie).

3) To manage insertions and deletions that impair the symmetry, we implemented a modified version of the Needleman-Wunsch algorithm for global alignment, which can identify the best

symmetry point and generate the optimal alignment between the two arms of the symmetric motif. The algorithm steps are the following:

a. The sequence and its inverted copy are placed in the horizontal and vertical axes of the matrix respectively (see Figure 4).
b. The alignment matrix is filled according to Needleman-Wunsch algorithm.
c. Two different scoring matrices are used to test for the mirror and palindromic symmetry (see Figure 6); gap opening score is -1.
d. The backtracking to retrieve the optimal alignment starts from the highest scoring cell found along the diagonal connecting top-right cell and bottom-left cell. This boundary defines a constraint for the entire sequence to be aligned.

The example in Figure 4 shows the result of a mirror search in the sequence "ATCAAGTTGCCA". The scoring matrix used for filling the matrix is shown in Figure 6a (gap open -1). The best possible alignment (Figure 5) is finally obtained following the optimal path as shown by arrows in Figure 4 where the green cell containing the highest score along the diagonal is the starting point of backtracking.

		A	T	C	A	A	G	T	T	G	C	C	A
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10	-11	-12
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	
C	-2	0	0	1	0	-1	-2	-3	-4	-5	-6		
C	-3	-1	-1	1	0	-1	-2	-3	-4	-5			
G	-4	-2	-2	0	0	-1	0	-1	-2				
T	-5	-3	-1	-1	-1	-1	-1	1					
T	-6	-4	-2	-2	-2	-2	-2						
G	-7	-5	-3	-3	-3	-3							
A	-8	-6	-4	-4	-2								
A	-9	-7	-5	-5									
C	-10	-8	-6										
T	-11	-9											
A	-12												

Figure 4 – Alignment matrix built for the test sequence "ATCAAGTTGCCA" searching for mirror symmetry where parameter -t 25 (NeSSie parameter considering percentage of degeneracy including both gaps and mismatches) is used. Being the sequence 12bp long with allowed up to 25% of degeneracy, at most 3 imperfections are permitted in the self-complementary alignment positions of the mirror motif (highlighted in red in the sequence alignment of Figure 5 and in the backtrack of the matrix).

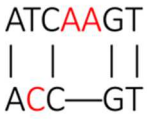


Figure 5 – Optimal alignment calculated from the filled scoring matrix of Figure 4.

a - mirror	A	T	C	G
A	1	-1	-1	-1
T	-1	1	-1	-1
C	-1	-1	1	-1
G	-1	-1	-1	1

b - palindrome	A	T	C	G
A	-1	1	-1	-1
T	1	-1	-1	-1
C	-1	-1	-1	1
G	-1	-1	1	-1

Figure 6 – Scoring matrices for mirror and palindrome symmetries.

DETECTION OF CRUCIFORMS: THE LOOP REGION PROBLEM

NeSSie evaluates the overall symmetry of a motif without considering the potential presence of a long loop between the two paired arms. Loops are not involved in the self-complementary pairing of the stem and it is consequently expected that an increase of insertions/deletions/mismatches will occur if this sequence tract is forced to be aligned. The present work-around is to use tolerant parameters. For example, the motif forming the cruciform shown in Figure 7 can be detected by allowing a 20% of degeneration in the symmetry.

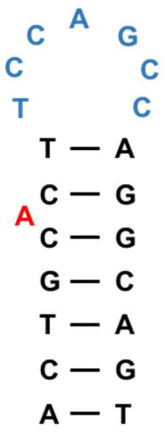


Figure 7 – "ACTGCACTTCCAGCCAGGCAGT" alignment as cruciform.

SHANNON ENTROPY

The Shannon entropy [2] is an index derived from the information theory that measures the amount of non-compressible information contained in a message. It can be applied to any symbolic sequence and allows to measure the order state (or entropy) of the sequence by the analysis of the symbols distribution. Practically, when applied to a DNA sequence [3], it allows to detect unbalances in base composition. The entropy is calculated according to the formula:

$$H = -\sum_{i=1}^{n} p(x_i)\ log_n p(x_i)$$

where $p(x_i)$ represents the probability for the symbol x_i to occur at any position of the sequence and n is the size of the alphabet x (n = 4 for DNA).

LINGUISTIC COMPLEXITY

The Linguistic Sequence Complexity [4] is an index that measures the vocabulary richness of a sequence. It can be applied to any symbolic sequence and is calculated as the level of repetition of the k-mers (words of length k) in the sequence. The more complex the sequence is, the richer is the vocabulary it contains; vice versa, the lower the complexity is, the more repetitive is the sequence. The complexity is calculated according to the formula [5]:

$$C = \sum_{k=1}^{N} V_k \Big/ \sum_{k=1}^{N} V_{\max k}$$

where V_k is the actual number of different words of length k in the sequence, $V_{\max k}$ is the maximum number of possible words of length k in the sequence, and N is the maximum length of the words considered in the score calculation. $V_{\max k}$ is calculated as the $\min(A^k, L - k + 1)$, where A is the alphabet size and L is the length of the sequence.

OUTPUT WRAPPERS AND GENOME BROWSER VISUALIZATION

Two python scripts are provided with NeSSie to transform the raw output in a more user-friendly and human readable format.

NESSIEOUTPARSER.PY

This parser works with the results obtained for the search of *mirror* and *palindromic* motifs, as well as the motifs with a *DNA-triplex* forming potential. The whole sequence is usually partitioned in overlapping sliding windows and each block is analyzed separately. If the same motif is detected in the overlap region of two different blocks, the hit will be reported for every block with the associated indexes at which it is found in that block. This can lead to a redundancy of some hits in the results. This parser allows to join these redundant motifs together under one hit while ordering the results. The results can be ordered by indexes (lowest to highest) as a default, by counts (highest to lowest) or by score (highest to lowest). The parser allows also to generate an output where the retrieved best alignment is made explicit in a human readable format. Finally, the parser can generate a GFF format file to simplify the visualization of the results using a genome browser such as the Integrative Genomics Viewer (IGV) [6].

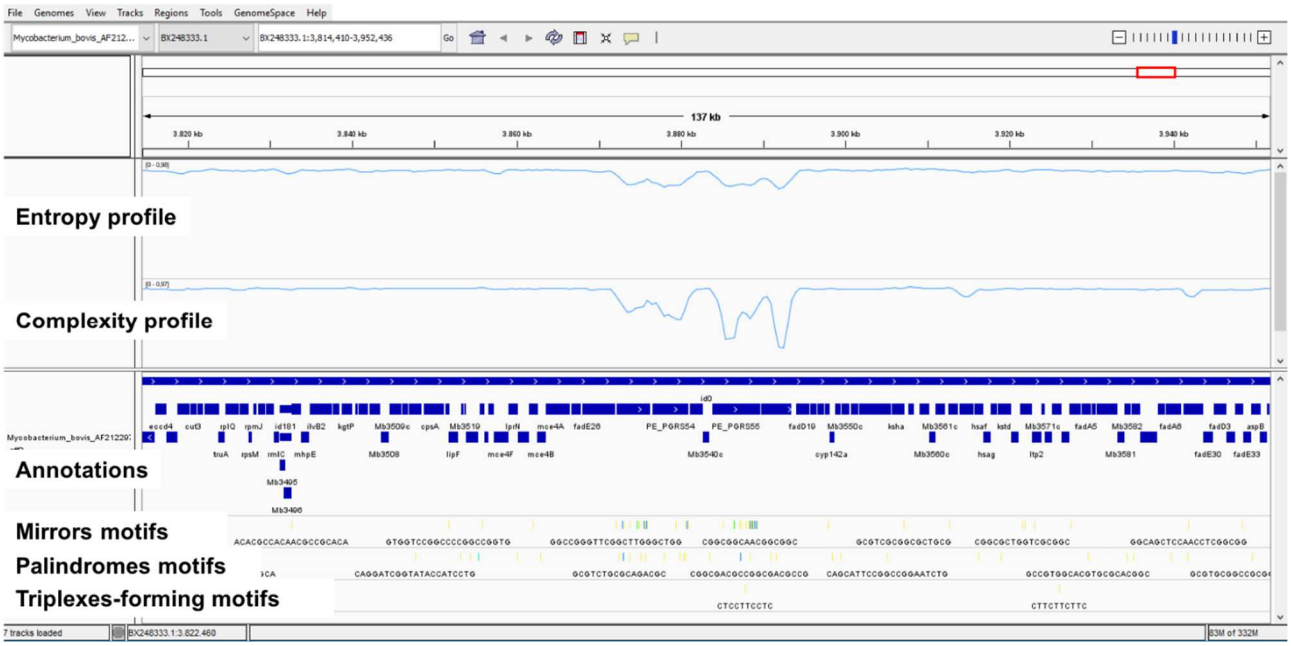
TO_WIG.PY

This parser can be used to better visualize the raw numerical output of the sequence Shannon entropy and linguistic complexity. This script creates a WIG format file that can be visualized using a genome browser such as IGV [6].

GENOME BROWSER VISUALIZATION

As an example of visualization, the snapshot below shows a zoomed region of the data obtained for the *Mycobacterium bovis AF2122/97* genome (see later on "Mycobacterium data analysis" section). The browser shows the entropy and complexity profiles, the track with the genomic annotations and the motifs detected by NeSSie.

Genome Browser



BENCHMARK

DETECTION OF MOTIFS WITH A MIRROR AND PALINDROMIC SYMMETRY

NeSSie can detect both exact and degenerate motifs with a mirror or a palindromic symmetry. To verify the correct functionality and performance of the tool we tested NeSSie on a custom dataset. The dataset, the results, and the command lines used for the analyses are available at the following URL: http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=nessie#Downloads

DATASET

We randomly generated a set of 1200 mirror and 1200 palindromic motifs ranging from 15bp to 30bp in length. Two thirds of the total motifs were degenerated with an increasing percentage of errors ranging from 0 to 20% with the aim to affect their perfect symmetry (i.e. introducing mismatches and insertions/deletions). For example, allowing up to 20% of sequence degeneracy in a motif of length 30 bp means that a maximum of 6 gaps or mismatches are allowed, disrupting the original symmetry. Different combinations of motif lengths and gaps/mismatches percentages were used as shown in Table 1. These motifs were inserted in random positions in both a repeated 5 Mb long sequence and the *E. Coli* genome (Accid: U00096.3), thus creating four different datasets: I) a repeated 'ACTG' sequence 5 Mb long with 1200 palindromic motifs, II) a repeated 'ACTG' sequence 5 Mb long with 1200 mirror motifs, III) *E. Coli* genome with 1200 palindromic motifs, IV) *E. Coli* genome with 1200 mirror motifs. The four generated sequences were built for different aims. I) and II) are 'ACTG' repeated sequences that do not intrinsically contain motifs with a mirror or palindromic symmetry. They were used to test if NeSSie is capable to detect only the inserted motifs without reporting false positive hits. III) and IV) are real sequences that originally contain symmetrical motifs, even if they are not described in the literature and consequently they cannot be evaluated for benchmarking purposes. These two sequences were used to test if NeSSie is still capable to detect all the inserted motifs also in a noisy context, where originally present motifs may interfere with the analysis.

Table 1 – Characteristics and number of the generated motifs.

	15 bp long	20 bp long	25 bp long	30 bp long	Total
Perfect motifs (exact symmetry)	100	100	100	100	400
Up to 15% degenerate	100	100	100	100	400
Up to 20% degenerate	100	100	100	100	400
					1200

RESULTS

The performance of the tool was evaluated based on the ability to detect the inserted motifs with the corresponding symmetry. A motif was considered found when NeSSie detected a hit within 15

bp from the insertion site. This permissive evaluation strategy is necessary because inserted motifs can be partly influenced by the nucleotides surrounding the insertion site. The results are shown in Table 2. NeSSie can detect all the inserted motifs in any analyzed sequence. Moreover, NeSSie does not detect false positive hits in the 'ACTG' sequences. These results confirm that the tool is able to perform an exhaustive search and detect both perfect and degenerate motifs, according to the defined parameters.

Table 2 – NeSSie results.

a - 'ACTG' 5Mb sequence	Perfect motifs (exact symmetry)	Up to 15% degenerate	Up to 20% degenerate
Search Mirror perfect	400	-	-
Search Mirror 15% degenerate	400	400	-
Search Mirror 20% degenerate	400	400	400
Search Palindrome perfect	400	-	-
Search Palindrome 15% degenerate	400	400	-
Search Palindrome 20% degenerate	400	400	400

b - *E. Coli* genome	Perfect motifs	Up to 15% degenerate	Up to 20% degenerate
Search Mirror perfect	400	-	-
Search Mirror 15% degenerate	400	400	-
Search Mirror 20% degenerate	400	400	400
Search Palindrome perfect	400	-	-
Search Palindrome 15% degenerate	400	400	-
Search Palindrome 20% degenerate	400	400	400

Columns contain the number of inserted motifs detected for the corresponding category (each category contains 400 motifs, see Table 1), while the rows contain the parameters used for the different analyses.

CONCLUSIONS

The assessment of NeSSie allowed us to evaluate not only the tool performance in the detection of sequence symmetries carrying high levels of degeneracy but also the computational time of a demanding genome wide search. On a randomly generated sequence of 180 Mb bases, NeSSie (run with the following parameters set: -k 15, -K 30, -t 20) completed the analyses in 1h:53m:05s using 3.2 Gb RAM on a workstation with 32 Gb RAM, Intel-Xeon E5530 @ 2.40GHz, mounting Debian 6.0.10. We also tried to perform a comparison with the tool Reputer [7], [8] that, to our knowledge, is one of the few available and working tools able to perform similar motif searches. It is worth pointing out that the algorithmic strategies are different. While NeSSie is developed to search for local symmetries, Reputer is developed to search for spaced repeats along the sequence. Unfortunately, this direct comparison was not possible probably due to the intrinsic limits of the suffix tree data structure in Reputer algorithm, which is known to be not efficient for the analyses of degenerate motifs. Indeed, we experienced a strong drop of computational efficiency when increasing the edit distance parameter (see Table 3).

Table 3 – Computational time for Reputer on a 10 Kb sequence. Edit distance correspondence with NeSSie -t parameter is the following: edit distance 5 corresponds to -t 25%, 4 -> 20% , 3 -> 15%, 2 -> 10%, 1 -> 5%.

	Analysis time
Perfect palindromic motifs of length 10	0 min 0.076 sec
Palindromic motifs of length 10, Edit distance 1	0 min 0.102 sec
Palindromic motifs of length 10, Edit distance 2	0 min 2.487 sec
Palindromic motifs of length 10, Edit distance 3	2 min 18.629 sec
Palindromic motifs of length 10, Edit distance 4	63 min 52.297 sec
Palindromic motifs of length 10, Edit distance 5	940 min 32.080 sec

DETECTION OF MOTIFS WITH A TRIPLEX FORMING POTENTIAL

NeSSie can detect motifs potentially associated with the formation of intramolecular DNA-triplexes (i.e. homo-purine motifs with a mirror symmetry). We performed a comparison with the tool Triplex [9], [10] that, to our knowledge, is the only available tool capable to detect both perfect and degenerate motifs potentially forming intramolecular triplexes in large DNA sequences. We also evaluated Triplexator [11] but a direct comparison was not possible because of the different aims and scope of the tool. While NeSSie and Triplex are developed to detect intramolecular DNA-triplexes, Triplexator is specifically developed to detect intermolecular triplexes and is not designed to detect local symmetries on the sequence. The dataset, the results and the command lines used to run the tools for the analyses are available with the supplementary materials that can be downloaded from: http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=nessie#Downloads

DATASET

To compare NeSSie and Triplex tools, we built different custom datasets. Each dataset contains 85000 sequences composed as follows: 170 'true' triplexes and 499 decoys generated from each real triplex used as seed. The 'true' triplexes are real motifs retrieved from the literature that were described in association with triplex formation. The complete list was obtained from supplementary data of Triplex paper [9]. The decoys are randomly created by introducing a certain amount of degeneration in the real motifs. Different amounts of degeneration were used to build different datasets to test the performances of the tools in different scenarios. We built 5 datasets that contain decoys with 10%, 20%, 25%, 30% and 40% of degeneration, respectively.

ANALYSES

The performances of the two tools were evaluated based on their ability to detect the real triplex-forming motifs with respect to the corresponding decoys. The tools were used to analyze the datasets and the results for each of the datasets were evaluated. Since the real triplex-forming motifs are very heterogeneous and different from one another, we individually analyzed each true motif with its corresponding decoy sequences. For each of the groups, the retrieved hits were

ordered by score and divided into 100 bins calculated based on the range of the scores (maximum score minus minimum score for the considered group). If multiple hits were found for the same sequence, only the best one was considered and used. We used a variable threshold to define which results to keep or discard. The threshold allows to progressively consider a smaller percentage of hits based on the score: precision and recall were calculated at different thresholds, starting from 100% (all results considered) and up to 1% (only the best 1% of scores were kept), using a step of 0.5. Using this approach, the precision-recall plots and the ROC curves were calculated for each of the methods on each of the datasets.

RESULTS

The results are shown in Figure 8 and Figure 9. When a very low degree of degeneration (10%) is considered, both tools fail to discriminate effectively between the original motifs and the decoys (data not shown). This is due to the relatively high similarity of true triplexes with their corresponding decoy set. Starting from 20% of degeneration, both tools increase their performance. NeSSie and Triplex perform similarly but NeSSie has a better precision and a better recall (Figure 8), with a lower number of false positives (Figure 9) compared to Triplex. This demonstrates that, though not highly specialized as Triplex tool, NeSSie is competitive and can detect triplex forming motifs with a slightly higher specificity and F-measure (see Table 4).

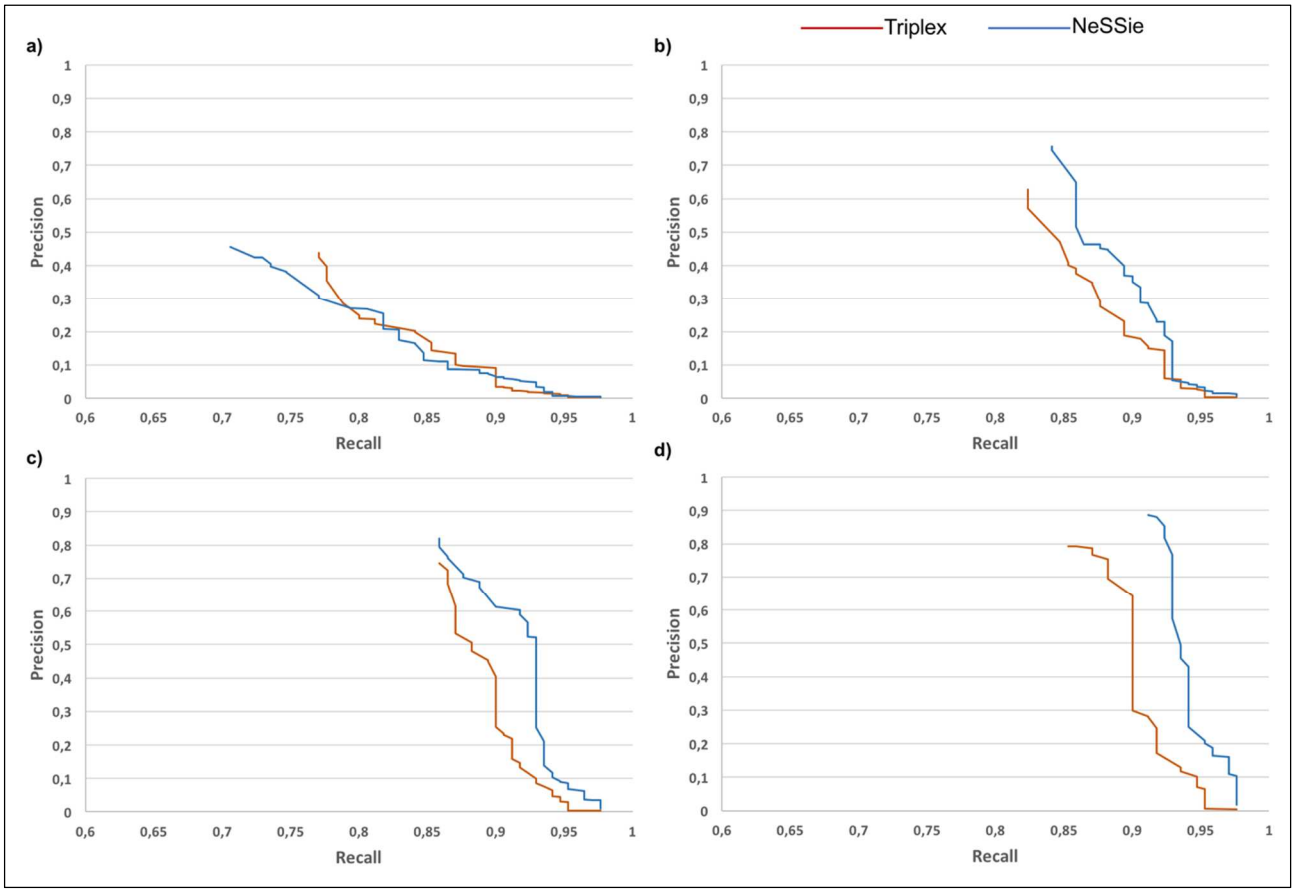


Figure 8 – Precision-recall plots. a) decoys with 20% of degeneration, b) decoys with 25% of degeneration, c) decoys with 30% of degeneration, d) decoys with 40% of degeneration.

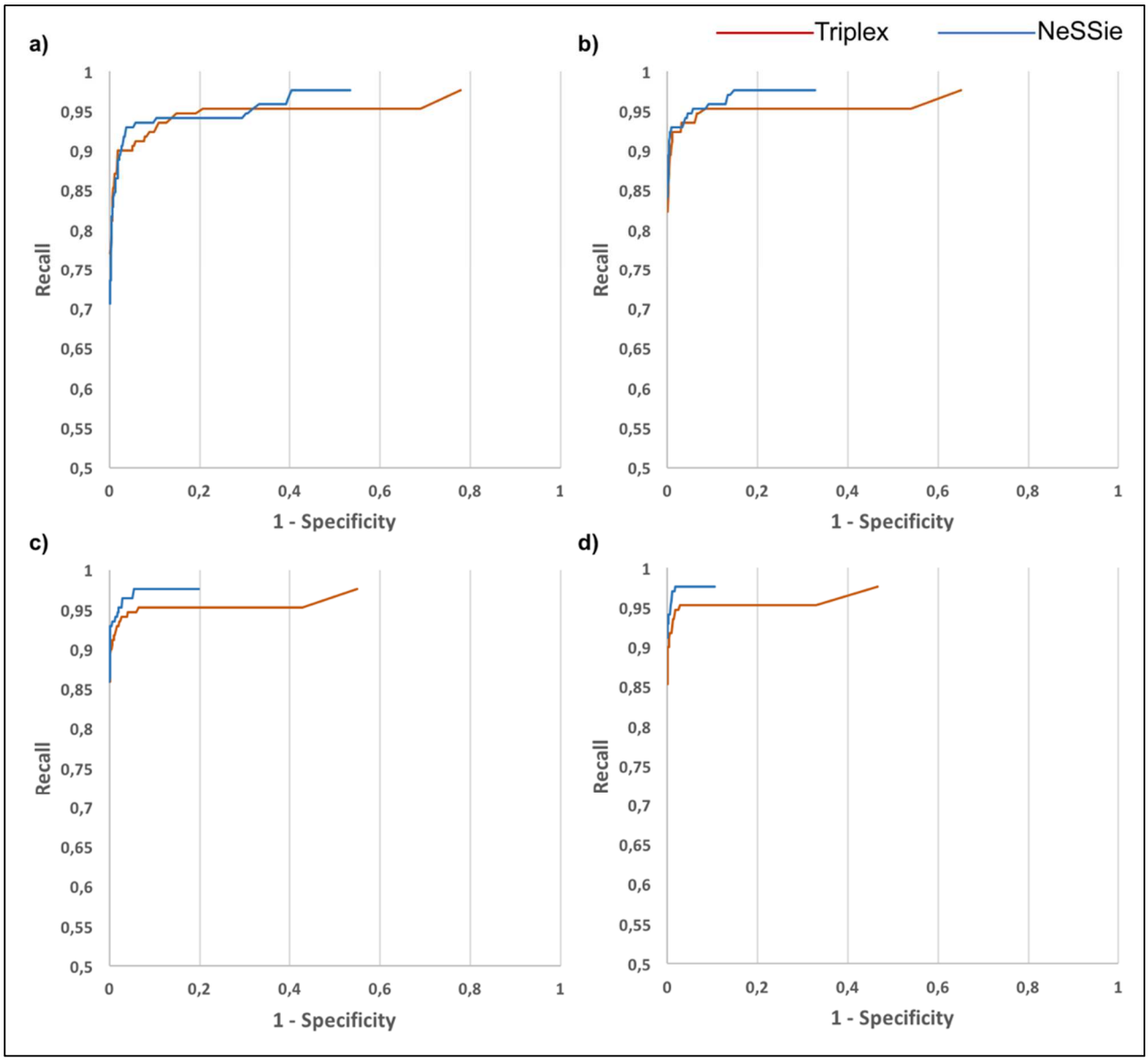


Figure 9 – ROC curves. a) decoys with 20% of degeneration, b) decoys with 25% of degeneration, c) decoys with 30% of degeneration, d) decoys with 40% of degeneration.

Table 4 - Maximum F-measure for the different datasets.

a – Triplex

Degeneration	F-measure	Threshold	Precision	Recall
10%	0.157937147	5	0.091503268	0.576470588
20%	0.558635394	8	0.43812709	0.770588235
25%	0.712468193	9	0.627802691	0.823529412
30%	0.797814208	9.5	0.744897959	0.858823529
40%	0.826815642	11	0.787234043	0.870588235

b – NeSSie

Degeneration	F-measure	Threshold	Precision	Recall
10%	0.174731183	2	0.113240418	0.382352941
20%	0.554272517	2	0.456273764	0.705882353
25%	0.796657382	2	0.756613757	0.841176471
30%	0.83908046	2	0.820224719	0.858823529
40%	0.899135447	11.5	0.881355932	0.917647059

COMPUTATIONAL TIME PERFORMACES

The performance of Triplex and NeSSie was compared using a randomly generated sequence 180 Mb long. The analysis was performed on a workstation (32 Gb RAM, Intel-Xeon E5530 @ 2.40GHz) mounting Debian 6.0.10. The results are show in Table 5.

Table 5 – Time and RAM usage for a 180 Mb sequence.

	Time (hr:min:sec)	RAM usage (Gb)
NeSSie	01:20:17	0.3
Triplex full analysis *	05:58:27	0.8
Triplex search only	00:30:02	0.5

* considers also the time used by R to convert and save the results into a fasta file.

MYCOBACTERIUM BOVIS ANALYSIS

We show an example of a potential application of NeSSie on the genome of *Mycobacterium bovis AF2122/97*. We combined the detection of mirror, palindromic and triplex forming motifs with the analysis of the sequence linguistic complexity and Shannon entropy. The dataset, the results and the command lines used for the analyses can be downloaded from : http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=nessie#Downloads

MOTIFS ANALYSES

Mirror and palindromic motifs in the range 10 bp – 40 bp were detected allowing up to a 5% of degeneration (gaps and mismatches), while for the triplex search a 5% of non-purine bases was allowed. These parameters led to the identification of 8796 motifs with a mirror symmetry, 8687 palindromic motifs and 77 triplexes forming motifs. The results obtained for mirror and palindromic motifs were further ranked by score and the lowest scoring motifs (short or highly degenerate) discarded. This led to a final pool of 968 motifs with a mirror symmetry and 907 motifs with a palindromic symmetry that represent the longest and the less degenerate motifs.

COMPLEXITY AND ENTROPY ANALYSES

The linguistic complexity was calculated on a sliding window 2 kb long using a 500 bp shift. For the Shannon entropy calculation, a different approach was used since to detect higher fluctuation in the measure a shorter window was used. Entropy was initially calculated using a sliding window of length 50 bp with a shift of 25 bp inside the main 2 kb sliding window, then an average score was calculated. A combined score obtained as the average of the linguistic complexity and Shannon entropy measures was also assigned to each 2 kb sliding window.

COMPREHENSIVE ANALYSIS

We analyzed the distribution and number of the identified motifs in the sliding windows coupled with their linguistic complexity and Shannon entropy. The motifs with a mirror symmetry were evaluated together with the triplex-forming motifs, since the latter represent a sub-class of mirror motifs. The partially overlapping motifs of the same type were considered together and counted as one, to avoid both biases and count overestimations. The windows were further grouped in bins based on the total number of contained motifs. The distributions of entropy, complexity and combined scores were analyzed for each bin (Figure 10, a-b-c). An average score was also calculated within each bin and for all the different measures (Figure 10, d). Only bins containing at least 15 sliding windows were considered in the analysis.

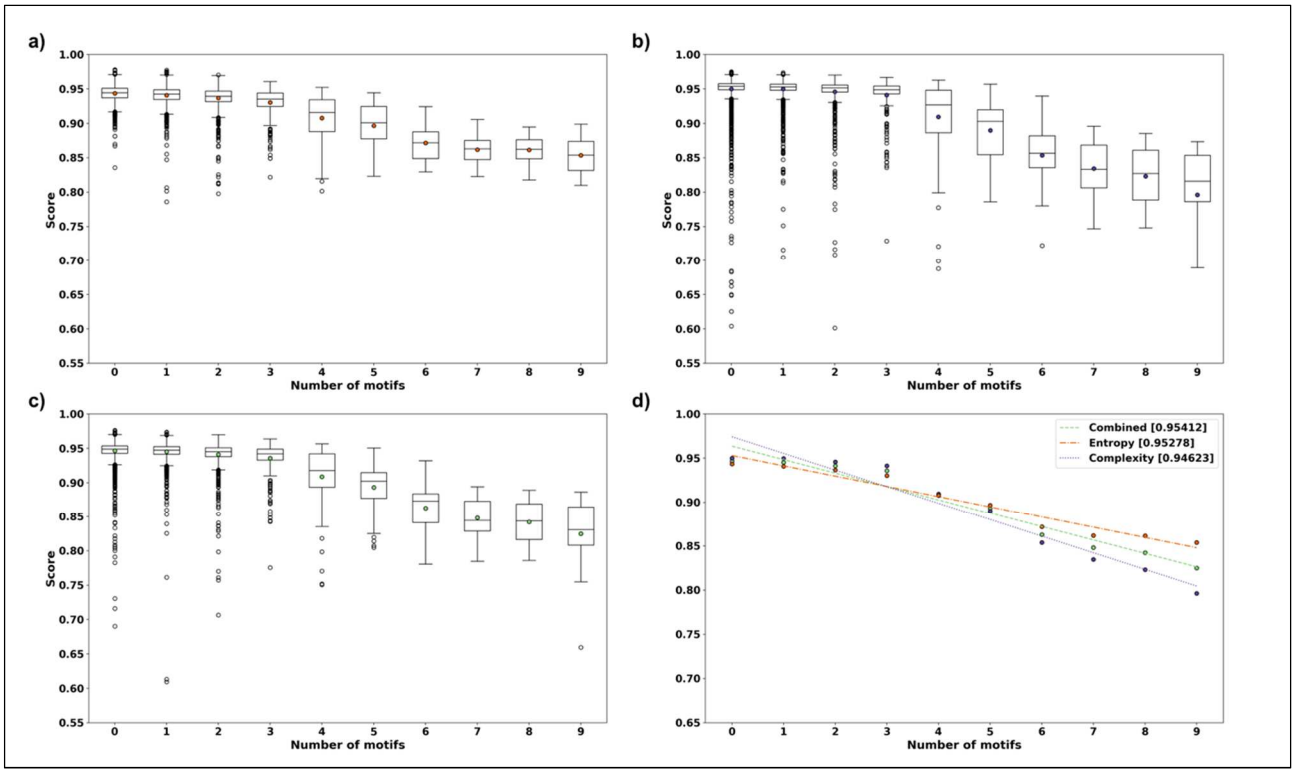


Figure 10 – Results for windows analysis. a) entropy scores distribution, b) complexity scores distribution, c) combined scores distribution, d) average scores plot with the corresponding trend-lines (r-squared reported in legend).

RESULTS

The analysis reveals that the highest scoring motifs with a mirror and palindromic symmetry are not randomly distributed across the genome but are clustered in regions characterized by a decrease in both Shannon entropy and linguistic complexity measures. It is possible to observe a correlation between the decrease of the entropy and complexity measures and the increase of the number of motifs in the window. Interestingly, the r-squared is slightly higher for the combined score than the single measures (Figure 10, d). We did not investigate further the putative cause and effect, if any, of the correlation among these different measures and motifs enrichment. A potential bias is due to the presence of tandem repeats (not shown) in these regions causing a drop of both Shannon entropy and linguistic complexity.

CONCLUSIONS

The genomic annotations of *M. b*ovis were integrated in the genome browser to better characterize the regions enriched in symmetrical motifs. We observed that the majority of these regions (data not shown) correspond and overlap with PE-PGRS genes [12]. PE-PGRS belong to a large family of glycine-rich proteins that are typical of several mycobacterial species. Although their function is still elusive, they seem to play an important role in the biology of the species conserving them. Given the peculiarity of their nucleotide composition, as revealed by NeSSie analysis, it could be interesting to further investigate why these genes are enriched in motifs with mirror and palindromic symmetries. The whole genome browser visualization data can be

downloaded from the following URL:
http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=nessie#genome_browser_visualization_example

REFERENCES

[1] S. B. Needleman and C. D. Wunsch, “A general method applicable to the search for similarities in the amino acid sequence of two proteins,” *J. Mol. Biol.*, vol. 48, no. 3, pp. 443–453, Mar. 1970.

[2] C. E. Shannon, “A mathematical theory of communication,” *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[3] T. Machado and J. A, “Shannon Entropy Analysis of the Genome Code,” *Mathematical Problems in Engineering*, 2012. [Online]. Available: https://www.hindawi.com/journals/mpe/2012/132625/. [Accessed: 25-Oct-2017].

[4] E. N. Trifonov, “Making sense of the human genome,” *Struct. Methods*, vol. 1, Human Genome Initiative and DNA Recombination, pp. 69–77, 1990.

[5] Y. L. Orlov and V. N. Potapov, “Complexity: an internet resource for analysis of DNA sequence complexity,” *Nucleic Acids Res.*, vol. 32, no. Web Server issue, pp. W628-633, Jul. 2004.

[6] J. T. Robinson *et al.*, “Integrative genomics viewer,” *Nat. Biotechnol.*, vol. 29, no. 1, pp. 24–26, Jan. 2011.

[7] S. Kurtz and C. Schleiermacher, “REPuter: fast computation of maximal repeats in complete genomes,” *Bioinforma. Oxf. Engl.*, vol. 15, no. 5, pp. 426–427, May 1999.

[8] S. Kurtz, J. V. Choudhuri, E. Ohlebusch, C. Schleiermacher, J. Stoye, and R. Giegerich, “REPuter: the manifold applications of repeat analysis on a genomic scale,” *Nucleic Acids Res.*, vol. 29, no. 22, pp. 4633–4642, Nov. 2001.

[9] M. Lexa, T. Martínek, I. Burgetová, D. Kopeček, and M. Brázdová, “A dynamic programming algorithm for identification of triplex-forming sequences,” *Bioinforma. Oxf. Engl.*, vol. 27, no. 18, pp. 2510–2517, Sep. 2011.

[10] J. Hon, T. Martínek, K. Rajdl, and M. Lexa, “Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences,” *Bioinforma. Oxf. Engl.*, vol. 29, no. 15, pp. 1900–1901, Aug. 2013.

[11] F. A. Buske, D. C. Bauer, J. S. Mattick, and T. L. Bailey, “Triplexator: Detecting nucleic acid triple helices in genomic and transcriptomic data,” *Genome Res.*, vol. 22, no. 7, p. 1372, Jul. 2012.

[12] L. S. Meena, “An overview to understand the role of PE_PGRS family proteins in Mycobacterium tuberculosis H37 Rv and their potential as new drug targets,” *Biotechnol. Appl. Biochem.*, vol. 62, no. 2, pp. 145–153, Apr. 2015.

G-quadruplex forming sequences in the genome of all known human viruses: a comprehensive guide

Presence and conservation of G4s in human viruses

Enrico Lavezzo[1¶], Michele Berselli[1¶], Ilaria Frasson[1¶], Rosalba Perrone[1], Giorgio Palù[1], Alessandra R. Brazzale[2*], Sara N. Richter[1*], Stefano Toppo[1*]

[1] Department of Molecular Medicine, University of Padova, Padova, Italy

[2] Department of Statistical Sciences, University of Padova, Padova, Italy

* Corresponding author.

Email: stefano.toppo@unipd.it (ST), sara.richter@unipd.it (SR), alessandra.brazzale@unipd.it (AB)

[¶]These authors contributed equally to this work.

Abstract

G-quadruplexes are non-canonical nucleic-acid structures that control transcription, replication, and recombination in organisms. G-quadruplexes are present in eukaryotes, prokaryotes, and viruses. In the latter, mounting evidence indicates their key biological activity. Since data on viruses are scattered, we here present a comprehensive analysis of potential quadruplex-forming sequences (PQS) in the genome of all known viruses that can infect humans. We show that occurrence and location of PQSs are features characteristic of each virus class and family. Our statistical analysis proves that their presence within the viral genome is orderly arranged, as indicated by the possibility to correctly assign up to two-thirds of viruses to their exact class based on the PQS classification. For each virus we provide: i) the list of all PQS present in the genome (positive and negative strands), ii) their position in the viral genome, iii) the degree of conservation among strains of each PQS in its genome context, iv) the statistical significance of PQS abundance. This information is accessible from a database to allow the easy navigation of the results: http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus. The availability of these data will greatly expedite research on G-quadruplex in viruses, with the possibility to accelerate finding therapeutic opportunities to numerous and some fearsome human diseases.

Author summary

G-quadruplexes are nucleic acid non-canonical structures that have been implicated in the regulation of different biological processes of many organisms. Their presence has been demonstrated also in several viral pathogens, providing new insights into viruses' biology and potentially serving as drug targets. Although experimental validation is needed to confirm the actual folding of G-quadruplexes, they can be inferred *in silico* directly from the nucleotide sequence. Several computational methods exist for this purpose, but they are all limited to the analysis of independent sequences. Since viral genomes can be highly variable, G-quadruplexes with important functional roles are expected to be conserved among strains and isolates belonging to the same viral species. Here we aimed at characterizing the potential quadruplex-forming sequences (PQS) content in the genome of viral human pathogens and assess their degree of conservation in each viral species. We demonstrate that many viruses possess more PQSs than

41 expected from their nucleotide composition and some of them are highly conserved within single viral
42 species, claiming some biological roles. We provide a website where the results of our analyses are
43 displayed for each virus with interactive graphics. This work is intended as a resource that can guide
44 scientists in the choice of the most promising candidates for functional characterization.

Introduction

G-quadruplexes (G4s) are nucleic-acid secondary structures that may form in single-stranded DNA and RNA G-rich sequences under physiological conditions [1]. Four Gs bind via Hoogsteen-type base-pairing to yield G-quartets: stacking of at least two G-quartets leads to G4 formation, through π-π interactions between aromatic systems of G-quartets. K^+ cations in the central cavity relieve repulsion among oxygen atoms and specifically support G4 formation and stability [2]. In the human genome, potential quadruplex-forming sequences (PQS) are clustered at definite genomic regions, such as telomeres, oncogene promoters, immunoglobulin switch regions, DNA replication origins and recombination sites [3]. In RNA, G4s and PQSs were mapped in mRNAs and in non-coding RNAs (ncRNAs) [4], such as long non-coding RNAs (lncRNAs) [5] and precursor microRNAs (pre-miRNAs) [6] indicating the potential of RNA G4s to regulate both pre- and post-transcriptional gene expression [7, 8].

Viruses are intracellular parasites that replicate by exploiting the cell replication and protein synthesis machineries. Viruses that infect humans are very diverse and, according to the Baltimore classification, they can be divided in seven groups based on the type of their genome and mechanism of genome replication: DNA viruses with 1) double-stranded (ds) and 2) single-stranded (ss) genome; RNA viruses with 3) ds genome, or ss genome with 4) positive (ssRNA (+)) or 5) negative (ssRNA (-)) polarity; 6) RNA or 7) DNA viruses with reverse transcription (RT) ability, whose genome is converted from RNA to DNA during the virus replication cycle (Table 1). Each of these classes possesses a peculiar replication cycle [9].

The presence of G4s in viruses and their involvement in virus key steps is increasingly evident in most of the Baltimore groups [10, 11]. In the dsDNA group, G4s were described in both *Herpesviridae* and *Papillomaviridae* families [12-20]. In ssDNA viruses, the presence of G4s was reported in the adeno-associated virus genome [21]. RNA G4s were described in the genomes of both ssRNA (+) (i.e. Zika, hepatitis C virus (HCV) [22, 23], and the severe acute respiratory syndrome (SARS) coronavirus [24, 25]) and ssRNA (-) viruses (i.e. Ebola virus [26]). A G4 was also detected in hepatis B virus (HBV) genome, the only member of dsDNA viruses with RT activity [27]. Finally, functionally significant G4s were identified both in the RNA and DNA proviral genome of the human immunodeficiency virus (HIV), a retrovirus belonging to group 6 (Table 1) [28-35], and [33, 34]in the LTR region of lentiviruses in general (ssRNA RT) [36].

Given this amount of scattered data, we here aimed at analyzing the presence of PQSs in the genome of all known viruses that can cause infections in humans. The analysis is performed at two distinct levels, globally for each viral genome and individually for each detected PQS. We asked the following: is the number of PQSs found in a viral genome simply due to chance, hence trivially reflecting genomic G/C content? And how much is each PQS conserved among the strains belonging to a viral species? To address these questions, we collected the whole viral genomes deposited in databanks, scanned them to detect all PQSs, and performed different statistical evaluations following the data analysis workflow shown in Fig 1. The detailed information on PQSs present in each human virus is available in an easily accessible web site with interactive graphics and genome browser visualization tools (http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus).

Fig 1. Example of virus classification in Families, Species and Strains with data analysis flowchart. The conservation analysis was performed among the strains within each virus species.

Results

Detection of G4 patterns in all known human viruses

All known viruses that cause infections in humans, according to the Viral Zone ExPASy web site (http://viralzone.expasy.org/all_by_species/678.html), were grouped in 7 classes according to Baltimore classification, which takes into account the viral genome nature: dsDNA, ssDNA, dsRNA, ssRNA(+), ssRNA(-), ssRNA(RT) and dsDNA(RT). Different replication strategies and structural similarities allow to further divide viruses in families (Table 1). The complete list of reference sequences for each virus included in the analyses is reported in S1 Table.

Table 1. **Virus families**.

Genome nature	**DNA**		**RNA**			**DNA and RNA**	
Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Genome type	**dsDNA**	**ssDNA**	**dsRNA**	**ssRNA (+)**	**ssRNA (-)**	**ssRNA (RT)**	**dsDNA (RT)**
Virus family	Herpes	Anello	Reo	Corona	Rhabdo	Retro	Hepadna
	Adeno	Parvo		Astro	Filo		
	Papilloma			Calici	Paramyxo		
	Polyoma			Flavi	Arena		
	Pox			Picorna	Bunya		
				Toga	Orthomyxo		

Virus families divided according to their genome and mechanism of replication. The suffix word "viridae" for each virus family has been omitted.

PQSs in viral genomes were searched by looking for the following patterns: [G(2)N(1-7)](3)G(2), [G(3)N(1-12)](3)G(3) and [G(4)N(1-12)](3)G(4), where both island and loop lengths were chosen to provide a comprehensive detection. We decided to expand the search to PQSs with very short islands and quite extended loops for the following reasons: first, the folding of PQS with GG-islands has been previously demonstrated in viruses [32]; second, since many viruses possess a RNA genome, and considering that RNA G4s are more stable than their DNA counterparts [37], PQSs with only two tetrads have a reasonable chance to fold in viral RNA genomes or in their intermediates. Finally, while long loops are known to destabilize G4 structures, their presence is anyway compatible with the folding of stable G4s at physiological temperature [38]. PQSs with bulged islands [39] and intermolecular G4s are not considered in the present study.

PQSs were searched in the positive and negative strand of each virus genome sequence, since both filaments are present and important in different stages of the viral replicative cycle of all virus classes. As the length of virus genomes greatly varies, i.e. from 235,646 nucleotides (nts) of the human cytomegalovirus (HCMV) to 1,682 nts of hepatitis delta virus (HDV), we reported the number of PQS independently of the genome length by normalizing their number per 1,000 nts (Fig 2). The PQS distribution for both the positive and negative strands is shown as a box plot for each Baltimore virus class,

whereas the PQS count for each virus within each class is shown as a dot besides the box plot (Fig 2). The negative strand of retroviruses (ssRNA (RT) viruses), ssDNA viruses and both strands of dsDNA viruses showed the largest presence of PQSs made of GG-, GGG- and GGGG-islands (box plots, Fig 2). Both strands of genomes of single virus families belonging to these groups and to ssRNA (+) and ssRNA (-) were enriched in PQSs of all G-islands types (dot plots, Fig 2). Conversely, dsRNA and dsDNA (RT) viruses notably lacked the presence of PQSs.

Fig 2. Box and whisker plots of PQSs in different virus classes. Each panel refers to the indicated type of G-island (GG, GGG, GGGG). The abundance of PQSs per 1 kb of viral genome is reported in the y-axis (for each viral species, the median value among all available strains is used) and the different virus categories in the x-axis. Boxplots are delimited by the first and third quartile and the straight and dotted lines drawn inside are the median and mean values, respectively, of the PQS distribution. The single observations are reported as dots close to the box plot. Whiskers delimit all the points that fall above/below the third/first quartile plus/minus 1.5 times the interquartile range (IQR). Orange and blue box plots refer to positive and negative strand respectively.

Then, we evaluated the conservation of PQSs among different strains of each viral species, hypothesizing that the presence of a conserved PQS within a less conserved genome environment could be an indication of a G4 with a biological function [40]. To allow for the evaluation of PQS conservation in the local context of viral genomes, we computed the "G4 scaffold conservation index" (G4_SCI) for each PQS in each virus species. This value measures the degree of conservation of G-islands that are necessary and sufficient to form a PQS: the higher the score, the higher the conservation of the PQS. An example of the results from such analysis is reported in Fig 3 for the lymphocytic choriomeningitis virus (segment S): all PQSs detected in the virus are plotted as vertical bars, the height and position of which represent the G4_SCI on the y-axis and the genome coordinates on the x-axis, respectively. In addition, the local sequence conservation (LSC) of the viral genome, calculated with a sliding window approach on all available viral sequences, is reported alongside as a red broken line. This visualization method allows the prompt identification of the presence, position, and conservation of G-islands within PQSs, together with

149 the overall local conservation of the genomic context. Moreover, the degree of conservation of the
150 connecting regions (loops) with respect to G-islands (the *loop_conservation* value) was calculated as the
151 difference between G4_SCI and LSC. Positive and negative *loop_conservation* scores indicate,
152 respectively, lower and higher conservation of connecting regions compared to the conservation of G-
153 islands. Values close to zero mean that both G-islands and connecting loops show the same level of
154 sequence conservation. In Fig 3, three PQSs formed by highly conserved GG-islands are shown for the S
155 segment of lymphocytic choriomeningitis virus, present in genomic regions both well and less well
156 conserved (Fig 3 at positions 1,790 in the positive strand, 1,760 and 2,680 in the negative strand). This
157 kind of analysis is available for all PQSs of all human virus species at
158 http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus (*loop_conservation* values are
159 included in tarballs downloadable for each viral class of the Baltimore classification, whereas each virus
160 species has a dedicated page displaying all graphical representations).

161

162 **Fig 3**. Conservation of PQSs and viral genomes. PQSs formed by GG-islands in the S segment of
163 lymphocytic choriomeningitis virus are shown. PQSs found in the positive and negative strands are
164 indicated as blue and orange vertical bars, respectively, while the height of bars represents the G4_SCI (the
165 conservation of G-islands). Local sequence conservation (LSC) of viral genomes is shown as a red broken
166 line. The x-axis indicates genome position, the y-axis the conservation %.

167

168 To assess the results, we retrieved from the literature all the available experimentally validated G4s
169 detected in human viruses. All patterns were confirmed also by our analysis and the complete list is
170 reported in S2 Table, together with the genomic coordinates of the predicted PQSs.

171 ## Statistical evidence of the presence of PQSs in the human virus genomes

172 G4 formation may be largely affected by G/C content, which greatly varies in viral genomes (from 76%
173 of Cercopithecine 2 herpes virus to 27% of Yaba like disease virus). Moreover, it has been shown that
174 some di- and trinucleotides are over- or under-represented in certain viruses [41, 42] and, in the context of

PQSs, this means that their abundance could be biased by unexpected frequencies of guanine homopolymers. G-island frequencies higher or lower than expected would lead to a potential over- or under-representation of PQSs, respectively.

To check whether the presence of PQSs was statistically relevant or whether it occurred by pure chance, we compared the results obtained from real viral genomes with those obtained by two different simulation strategies. The first one (single nucleotide assembling) assumes that the occurrence of each DNA base in the genome is independent [43]; the second (G-island reshuffling) considers that short sequences of a given length (k-mer) could be over- or under-represented in certain viral genomes [41, 42]. In the former case, sequences were generated with the same composition of nucleotides but different order with respect to references; in the latter, sequences were produced by reshuffling the positions of G-islands while keeping constant their number.

For each virus and simulation strategy, we produced 10,000 random sequences, which were screened with our PQSs detection pipeline. Real and simulated data were compared by computing a P-value, defined as twice the smaller proportion of simulated sequences that exhibit, respectively, a higher and lower count of PQSs as compared to the median value of all the available complete genome sequences for a certain virus. Hence, a P-value close to 1 means that the median PQS content in real viral sequences is not significant if compared to a random distribution; conversely, a P-value close to 0 means that PQS content is highly significant. This interpretation holds independently of the length of the genome and/or of the prevalence of either G/C bases or G-islands, as we compare the number of PQSs in a viral genome with the one we would expect in a simulated genome of the same length and of either the same base or G-island composition. To account for possible high discreteness of the data, a less conservative version of the P-value, called the mid-P value [44], was used. Segment diagrams of the mid-P values of the Baltimore grouped viruses are reported in S1 and S2 Figs [45]. The number of viruses whose median PQS count is significant at the 10% level is listed in Table 2 (virus names in S3 Table) with the indication of whether this median count is either higher or lower than the PQS count in simulated sequences.

Table 2. Relative abundance of viruses having a PQS content significantly different between real and simulated viral genomes.

	Number of viruses significantly different vs. randomization at single nucleotide level		Number of viruses significantly different vs. randomization at the G-island level		Number of viruses significantly different vs. both randomization	
G-island pattern	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**	**PQS more abundant in real sequences**	**PQS more abundant in simulated sequences**
GG (positive strand)	83/218 (38.1%)	9/218 (4.1%)	52/217 (24.0%)	4/217 (1.8%)	49/217 (22.6%)	3/217 (1.4%)
GG (negative strand)	83/187 (44.4%)	2/187 (1.1%)	67/187 (35.8%)	1/187 (0.5%)	59/186 (31.7%)	0 (0%)
GGG (positive strand)	41/78 (52.6%)	3/78 (3.8%)	40/75 (53.3%)	0 (0%)	34/74 (45.9%)	0 (0%)
GGG (negative strand)	32/68 (47.1%)	3/68 (4.4%)	32/69 (46.4%)	0 (0%)	28/68 (41.2%)	0 (0%)
GGGG (positive strand)	17/19 (89.5%)	0 (0%)	17/17 (100%)	0 (0%)	17/17 (100%)	0 (0%)
GGGG (negative strand)	23/28 (82.1%)	0 (0%)	23/25 (92.0%)	0 (0%)	23/25 (92.0%)	0 (0%)

The number of viruses where the amount of PQSs is significantly different at 10% level between real and simulated sequences is reported (with percentages in brackets). Values and percentages were calculated considering only viruses containing at least one PQS either in real or simulated sequences (this explains differences in denominators). The table reports significant values for either one of the two simulations (randomization of viral genomes at single nucleotide or at G-island levels) or both.

Our data show that most members of the dsDNA, ssDNA, and ssRNA (RT) present a highly significant content of PQSs formed by GG-, GGG- and/or GGGG-islands in one or both strands. ssRNA (-) and

ssRNA (+) classes are heterogeneous since some viruses are highly significant in any PQS category (from GG- to GGGG-islands), while others are not (see below). The presence of PQSs in members of the dsRNA group is notably less significant. Interestingly, few viruses display a smaller amount of PQSs than expected: both *Sagiyama virus* and *Human coronavirus HKU1* are depleted of PQSs belonging to GG-islands category in the positive genome strand when compared with both simulation strategies based on single nucleotide assembling and GG-island reshuffling. In addition, *Human parainfluenza virus 2* is poor of PQSs made of GG-island in the positive genome strand but is enriched in both GG- and GGG-type PQSs in the negative strand.

Overall, if we consider the viruses that contain at least one PQS in either the real or the simulated genomes, we observe that the increase in G-islands' length corresponds to a decrease in the absolute number of viruses containing PQSs, but it also corresponds to a dramatic increase in the fraction of them that is statistically significant.

By looking at the family level of viral classification, which is far more homogeneous than the Baltimore groups, some virus families emerge as prominently enriched in PQSs. Among them, *Herpesviridae* is not only the one with the highest PQS content, but most of its members display significantly more PQSs than expected in both genome strands and in all considered G-island lengths. Notably, some of the viruses belonging to *Herpesviridae* and showing the highest G/C content are statistically enriched in PQSs. This suggests that simply having a high G/C content is not a sufficient condition to justify the presence of such a high number of PQSs. Other viral families that are consistently enriched in PQSs are *Adenoviridae* and *Papillomaviridae*, especially in GG- (both strands) and GGG-island (positive strand) types. *Poxviridae* and *Parvoviridae* show an enrichment of GG-type PQSs in both genome strands, whereas the same pattern is enriched in the positive strand of all *Anelloviridae* members and in the negative strand of most *Paramyxoviridae* and *Retroviridae* viruses. All other families are generally not enriched in PQSs in any of the evaluated categories, with only a few exceptions that are listed in the following: L segments of *Lassa virus* and *Lymphocytic choriomeningitis virus* (*Arenaviridae*), *Wu* and *Merkel cell polyomaviruses* (*Polyomaviridae*), *Salivirus* (*Picornaviridae*), M and S segments of respectively *Crimean-Congo hemorrhagic fever virus* and *Rift Valley fever virus* (*Bunyaviridae*).

By comparing the results obtained independently from the two simulation strategies it is possible to draw additional conclusions. First, in most cases the results are concordant, meaning that both simulations

show similar trends in the statistical significance. Nonetheless, the overall number of viruses whose PQS content is significantly different with respect to simulated data is higher when real viral genomes are compared to those generated by single nucleotide assembling. This difference indicates that viral genome k-mer composition is indeed affecting the probability of randomly finding PQSs, at least in a proportion of viruses as shown in Fig 4: in the heatmaps, viruses that are significant in only one of the two simulations are reported for GG- and GGG-island patterns, whereas no such cases were found for GGGG-type PQSs.

Fig 4. Different results from single nucleotide (SN) and islands (ISL) reshuffling strategies. Heatmaps show all the viruses which are significant in only one of the two simulations or that obtain discordant results. Green and red boxes indicate that PQSs are more abundant in real and simulated genomes, respectively, with color intensity proportional to p-value size.

Finally, some remarkable exceptions exist where both simulations return a significant p-value, but with an opposite meaning. This is the case of two members of the *Poxviridae* family, namely *Molluscum contagiosum virus* and *Orf virus*, which are enriched in GG- and GGG-type PQSs in both strands of their genomes if compared with the islands reshuffling simulation but show the opposite behavior when compared with the single nucleotide assembling (they are also reported in Fig 4). While the full meaning of this observation is not clear to us, it seems that these viruses possess far less PQSs than they could have, but at the same time they are able to cluster their relatively few G-islands in more PQSs than expected.

Human virus PQSs position and overlap with genomic features

To check the prevalent positions of PQSs in virus genomes, we compared the coordinates of predicted PQSs with the available information regarding viral genome features. Genome coordinates were extracted for coding sequences (CDS), repeat regions (RR), 5'- and 3'-untranslated (UTR), and promoter regions. While CDS and RR are explicitly defined in RefSeq and GenBank databases, the annotation of UTRs and promoters is more inconsistent, being defined only for some viral species. For this reason, the annotations of genes and CDSs were exploited to indirectly extract the coordinates of 5'– and 3'–regulatory regions (see Materials and Methods for details). To determine the localization of PQSs in viral genomes, the

overlap extent between PQSs and genomic features was computed. Given the vast heterogeneity of the annotations reported in the feature fields, a manual revision was required to fix potential inconsistencies in annotations, regarding both keywords and coordinates. A revision was performed when possible, while controversial and uncertain annotations were not considered. These analyses are presented as bar charts for individual viral classes and G-island pattern types (GG-, GGG-, GGGG-island) (S3-S5 Figs). As regards the GGG-island type, the herpesvirus family of dsDNA viruses presents PQSs distributed along all the four identified genomic features, with a particularly high concentration in RR and, in some members, in the 5' – regulatory region. This feature is consistent with the reported extent of G4s in HSV-1, which are mainly clustered in the RR of the virus genome [12, 13]. Conversely, viruses belonging to the ssRNA (+) and ssRNA (-) classes show PQSs mainly grouped in CDS and in the 3'- and 5'-regulatory regions, respectively. HIV-1, belonging to ssRNA (RT) virus class, presents PQSs of the GGG-island type mainly in the RR and 3'-regulatory regions and in part in CDS. This distribution confirms previous data [28, 32]. Conversely, other retroviruses (ssRNA (RT)) such as HTLV-1 and HTLV-2, display PQSs in the CDS. Given the lower stringency of PQSs of the GG-island type, these are more widely distributed along the four identified genomic features, whereas the most stringent PQSs of the GGGG type, present only in herpesviruses (dsDNA) and HTLV-1 (ssRNA (RT)), show a clear-cut localization in the RR and CDS, respectively. These data indicate that the localization of PQSs in the viral genomes differs in virus classes.

The number and type of PQSs are characteristic of virus classes

In this line of thinking, we asked whether the observed number of PQSs, and more precisely its statistical significance with respect to the two random assembling scenarios, is representative for a specific Baltimore class. To answer this question, we checked whether it is possible to classify each virus to one of the six classes considered, that is, dsDNA, ssDNA, dsRNA, ssRNA (+), ssRNA (-) and ssRNA (RT), based on the information of how significant its median PQS counts are. We used a classifier built on multinomial logistic regression, as this method is both interpretable and robust to unbalanced group sizes as long as the group sizes are large enough. To avoid the latter drawback, we excluded from the model fit the hepatitis B virus, the only virus classified as dsDNA (RT), and the two unclassified Hepatitis delta and Hepatitis E viruses. Six features were used to classify the viruses, i.e. the six mid-P values (those calculated for GG-, GGG-, GGGG-, both in the positive and negative strand) which qualify the PQS content of the real viral

sequences. The values were multiplied by 1 or -1 depending on whether the median PQS count was over- or under-represented. Since real and corresponding simulated sequences contain the same base or G-islands composition, the classification model based on PQS content does not depend on the highly variable genome length and G/C content in the different virus classes but is specifically designed on the peculiar presence or absence of PQSs in each viral class. Furthermore, 34 viruses with no PQS count in all three G-island types in both the positive and negative strand and non-significant mid-P values at the 10% level were excluded from the analysis. We re-classified every viral genome used in our assessment using the discriminant function obtained from a leave-one-out analysis. This latter technique allowed us to accurately estimate how our classifier performs without the need to split our data into a training and a test set. The corresponding confusion matrix is given in Table 3 from where it is possible to extract the overall percentage of correct classifications that amount to 66.7% for the single nucleotide assembling model and 68.1% for the G-island reshuffling model. The agreement is good for the dsDNA, ssDNA, dsRNA, ssRNA (+) and ssRNA (-) classes. The two unclassified genomes of the Hepatitis delta and Hepatitis E viruses were classified as ssRNA (+).

Table 3. Confusion matrix for the semi-parametric classifier for G4 structure.

Sigle nucleotide assembling model		**Predicted class**						
		dsDNA	**ssDNA**	**dsRNA**	**ssRNA (+)**	**ssRNA (-)**	**ssRNA (RT)**	**Unclassified**
True class	**dsDNA**	33	0	0	1	4	1	0
	ssDNA	2	0	0	5	0	1	0
	dsRNA	0	0	13	0	5	0	0
	ssRNA (+)	4	0	0	45	13	0	0
	ssRNA (-)	3	0	8	18	48	0	0
	ssRNA (RT)	4	0	0	0	0	3	0
	Unclassified	0	0	0	2	0	0	0
G-island reshuffling model		**Predicted class**						
		dsDNA	**ssDNA**	**dsRNA**	**ssRNA (+)**	**ssRNA (-)**	**ssRNA (RT)**	**Unclassified**
True class	**dsDNA**	31	0	0	5	1	2	0
	ssDNA	4	0	0	2	2	0	0
	dsRNA	0	0	14	0	4	0	0
	ssRNA (+)	3	1	0	48	10	0	0
	ssRNA (-)	7	1	4	13	52	0	0
	ssRNA (RT)	3	0	0	3	1	0	0
	Unclassified	0	0	0	2	0	0	0

The number of viruses classified into the six Baltimore groups is shown, based on the two different simulation scenarios (single-nucleotide and G-island reshuffling). The classifier is based on a multinomial model and uses the one-sided mid-P values as features in combination with the information on whether the median PQS count is under- or over-represented.

Discussion

In this work we provide: i) the list of PQSs present in all human virus genomes (both positive and negative strands), ii) their position in the viral genome, iii) the degree of conservation of both G-islands and loops vs. the genome, iv) the statistical significance of PQS abundance in each virus. Our data show that viruses belonging to dsDNA, ssDNA, ssRNA (RT) and, to a less extent, ssRNA (+) and ssRNA (-) display the largest presence of GG-, GGG- and GGGG-type PQSs (box plots, Fig 2) and that the presence of PQSs in all these virus groups is statistically significant (segment diagrams, S1 and S2 Figs). Both results support a role of G4s in the virus biology: indeed, some G4s predicted in this work were already reported in viruses and were shown to possess specific and different functions.

We evidenced some general trends and exceptions that are worth noting if seen in comparative terms among all viruses considered in this study. Starting from the general features we noted: i) high G/C content is not sufficient per se to generate a high number of PQSs, as observed in G/C rich members of *Herpesviridae* family that are richer of PQSs than expected. ii) The statistical significance of PQSs found in real sequences tends to decrease when G-islands reshuffling (ISL) is compared with the corresponding single nucleotide assembling (SN), as is appreciable from the heatmap in Fig 4 (more intense color in the heatmap boxes). This suggests that short sequences of a given length (k-mer) could be over- or under-represented in certain viral genomes, as already reported in the literature [41, 42]. We observed that viral genomes enriched in PQSs also contain a higher number of G-islands than expected from mere nucleotide composition, especially evident in the GG-islands. iii) The unevenly distribution of PQSs can be used to classify membership of a virus in its corresponding category. This was not predictable *a priori* but up to two-thirds of unequivocal assignments suggest that for some viruses the PQS content works as a distinctive feature. iv) PQS localization shows differences in some virus classes, but this outcome is still incomplete due to lack of information in the databases about virus genomic features and partitioning into regulatory and coding regions.

Some other interesting observations are worth reporting as either special cases or exceptions. To start with, the ssRNA (-) group is the most heterogeneous one, since some viral species are significantly enriched in PQSs up to the most extended G-island type (GGGG), while others lack this feature. Surprisingly, two viruses of the dsDNA group, which was generally highly enriched in PQSs, show a

360 significantly lower presence of PQSs than expected in a random sequence with the same G/C content (SN,
361 S3 Table), even though the opposite result was observed vs. simulated genomes with the same G-islands
362 content as the real ones (ISL, S3 Table). These two viruses, i.e. *Molluscum contagiosum virus* and *Orf*
363 *virus*, are the only ones belonging to genera other than the *Orthopoxvirus* within the *Poxviridae* family that
364 cause skin lesions. Finally, dsRNA and dsDNA (RT) viruses are notably poor in PQSs and with mostly
365 null statistical significance; however, single PQSs are highly conserved (e.g. rotavirus a segment 6),
366 therefore still conveying potential biological interest.
367 These data indicate that PQSs are mainly present in ss-genome viruses, which in principle are more
368 suitable to fold into G4s since they do not require unfolding from a fully complementary strand. The major
369 exception to this evidence is the *Herpesviridae* family of dsDNA viruses. In this case, most PQSs reported
370 here and also previously described [12, 13] form in repeated regions. It is possible that repeated sequences
371 are more prone to alternative folding, as shown by several non-canonical structures that form in repeated
372 regions of the DNA [46-49]. However, for some herpesviruses many PQSs are also present in regulatory
373 regions, which may indicate yet undiscovered functional roles. To note that the investigation of PQSs was
374 performed on a maximum window of 52 nucleotides in the case of isolated G4s. Alternatively, when more
375 than four G-islands are found complying with the maximum distance allowed between consecutive islands,
376 this window is extended as long as the rules are satisfied, thus including multiple distinct PQSs or potential
377 isoforms. However, it is possible, especially for the ss genomes, that bases more distant in the primary
378 sequence interact among each other, therefore expanding the repertoire of G4 structures.
379 The significant enrichment of PQSs in many viruses with respect to the corresponding randomized
380 genomes is an indication that the clustering of G-islands did not occur by pure chance, suggesting a
381 specific biological role of the G4 structures [40]. Complementary to this, the analysis of the PQS
382 conservation highlights every PQS that is conserved among viral strains. Since one of the peculiarities of
383 viral genomes is their fast mutation rate, the strong conservation of a specific G-island pattern among
384 strains is per se an indication of the biological relevance of a PQS. In light of this, single conserved PQSs
385 in viruses that do not display statistically significant PQS enrichment may retain key functional roles. The
386 meaning of PQS conservation can have different explanations for the different viruses analyzed in this
387 study. Given the high variability in the number of full-genome sequences available for each species, a
388 more general evaluation of PQS conservation at higher taxonomic ranks (e.g. at the family level) could

have been more informative. Nonetheless, generating and analyzing whole-genome multiple alignments involving different viral species, even if belonging to the same family, is almost prohibitive given the huge variability that is usually present in their genome sequences. Hence, the conservation of each PQS has to be considered on a case by case basis, exploiting the visualization tools provided in the website. As an example, an interesting approach could be looking at the discrepancy between the conservation of G-islands and connecting loops (*loop_conservation*) as an additional indication on the likeliness of biological implications of a specific PQS. A positive *loop_conservation* value highlights G-islands more conserved than their connecting loops, suggesting that only the PQS scaffold is required for mechanisms that are important for the virus life cycle, while the loops are dispensable. Considering the high mutation rate of viruses, this type of conservation indicates sequences where G4 formation is most likely essential. Equally conserved G-islands and loops (*loop_conservation* value = 0) imply that both the PQS scaffold and connecting loops are potentially relevant for the virus and probably involve interactors that specifically recognize them. In this case, the high sequence conservation, especially in CDS, may depend on the required conservation of that peculiar gene product rather than the presence of a G4 structure. Nonetheless, the option of targeting these conserved G4s for therapeutic purposes remains unaltered and the availability of specific and conserved loops may only enhance the chance of finding selective ligands [50]. Therefore, from this point of view, the 'zero' class is the best scenario for the development of specific drugs. The "negative" *loop_conservation* value scenario is of less immediate interpretation: it is possible that false positive hits fall in this category as it is unexpected that G-islands are less conserved than their connecting loops.

The evidence provided here, the previous studies on G4s in viruses, and the possibility to correctly classify the majority of viruses based on their PQSs (Table 3) suggest that most of the virus classes adopted G4-mediated mechanisms to control their viral cycles.

Together with the associated database, which is projected to be periodically updated to keep up with the fast-growing list of novel sequenced viruses, this work offers comprehensive data to guide researchers in the choice of the most significant PQSs within a human virus genome of interest. Hopefully, this will accelerate research in this area with the identification of new G4-mediated mechanisms in viruses and the development of effective and innovative therapeutics.

Material and methods

PQS detection and evaluation of conservation

The complete list of viral species able to infect humans was retrieved from http://viralzone.expasy.org/all_by_species/678.html (accessed in April 2016) and, for each of them, all available complete genome sequences were downloaded from GenBank. Multiple alignments were built for each virus with usearch8 [51], using a permissive identity threshold (60%) to account for viral variability. Since in some cases nucleotide heterogeneity within viral species exceeded this value, multiple clusters of aligned sequences were obtained for some viruses, representing distinct genotypes. Considering the difficulty of obtaining high quality alignments beyond this limit of nucleotide similarity, all clusters obtained with this method were kept separate, manually assigned to specific genotypes and independently processed in the downstream analyses. One genome per each group of aligned sequences was chosen to serve as reference sequence, possibly belonging to the manually curated RefSeq database [https://www.ncbi.nlm.nih.gov/refseq/]. The complete list of selected reference sequences is reported in S1 Table.

PQSs were searched in all multiple-aligned nucleotide sequences with an in-house developed tool, as previously described [36, 52]. Briefly, a PQS was reported when at least four consecutive guanine islands (G-islands) were detected. If '*l*' is the minimum number of G in every G-island of a PQS and '*d*' is the maximum distance allowed between two consecutive G-islands, the following combinations of '*l*' and '*d*' were searched: $l = 2$ and $d = 7$; $l = 3$ and $d = 12$; $l = 4$ and $d = 12$. Patterns in the negative strands of viral genomes were searched by looking for cytosines (Cs) instead of Gs. The conservation of each PQS in the multiple aligned genomes of the viruses was determined by looking at the conservation not only of the G-islands but also their connecting loops. We computed different indexes to measure the nucleotide sequence conservation of viral genomes and PQSs:

i) G4_scaffold_conservation_index (*G4_SCI*): it is referred to the G-islands. For each virus and for every detected PQS, it is calculated as the percentage of independent genomes maintaining the corresponding G-islands:

$$G4\_SCI = \frac{N_{G\_islands}}{N_{tot}} * 100$$

where $N_{G\_islands}$ is the number of sequences possessing the G-islands in a certain genome position and N_{tot} is the total number of sequences available for the virus.

ii) Loop_conservation: it is the difference between G4_SCI and the local conservation of the viral sequence spanning the PQS (LSC_{G4}).

$$Loop\_conservation = G4\_SCI - LSC_{G4}$$

LSC_{G4} is calculated as the average of LSC windows overlapping the PQS. LSC measure is computed within a sliding window of fixed length (length 20, shift 10), averaging the conservation values of each position extracted from the multiple sequence alignments with Jalview [53]. They are formally defined as:

$$LSC = \frac{\sum_{i=1}^{20} c_{\max i}}{20}$$

$$LSC_{G4} = \frac{LSC_1 + \ldots + LSC_n}{n}$$

where $c_{max\ i}$ is the maximum conservation at position i of the multiple aligned sequences and n is the number of windows overlapping the PQS.The results of these analyses are presented individually for each virus and PQS (http://www.medcomp.medicina.unipd.it/main_site/doku.php?id=g4virus), together with the calculated profiles of simple linguistic complexity and Shannon entropy that can highlight other potential local features of the sequence (e.g. repeats and low complexity regions) [54]. All charts were generated with Plotly [https://plot.ly], exploiting Pandas [55] and Numpy Python libraries [56]. Multiple alignments are visualized with MSAViewer[57] and genomic features with JBrowse 1.15.0 [58]. Unless otherwise stated, analyses were conducted with ad hoc developed Python and Perl scripts, which are available in the website (scripts.tar.gz).

Evaluation of PQS conservation in real vs randomized viral sequences

To determine whether the presence of PQSs in a virus is a conserved feature or it is only a consequence of its nucleotide composition, simulated viral genomes were generated and compared with real data. Two different strategies were adopted to generate simulated data:

i) *Single nucleotide assembling (SN)*. A computational approach was adopted where, in analogy to Huppert and Balasubramanian [43], the viral genome was modelled as a multinomial stream based on the assumption that each DNA base is independent. These authors give an explicit solution for the prevalence of PQSs in the human genome as a function of *p(G)*, the probability of any base being G. In our approach,

we also accounted for the probability of cytosines (*p(C)*) and additionally assumed that adenine (A) and thymine (T) bases were equally likely to occur. As all four probabilities need to sum up to one, the statistical reference model is a multinomial distribution with probability vector *(p(G), p(C), p(A), p(T))*. We hence took as many independent draws from this multinomial distribution as the number of nucleotides in the reference viral genome (S1 Table). The probabilities *p(G)* and *p(C)* vary for each virus and reflect the prevalence of G and C bases present in that virus, while the remaining proportion is equally split to give *p(A)* and *p(T)*. For each virus, 10,000 independent sequences were produced *in silico* with this method; the 'sample' R command with replacement was used and the provided parameters were the genome length and the relative abundance of the four nucleotides in the real genomes.

ii) *G-islands reshuffling (ISL)*. For each virus, we generated a scaffold of length *n* made of only As, with *n* corresponding to the length of the virus reference genome; then, we replaced di-, tri-, or tetra-nucleotides, at random positions and without overlaps, with as many GG-, GGG-, GGGG-, CC-, CCC-, CCCC-islands as in the reference sequence, respectively. Overall, we generated 10,000 independent sequences for three different simulated datasets, one for each island length.

Statistical methods

The simulated sequences were scanned for the presence of PQSs as previously described. The 10,000 counts obtained for each simulation formed the empirical distribution for PQS prevalence under the hypothesis of random assembling of the genome in the SN and ISL models respectively. The mid-P value was calculated using a homemade function. The semiparametric classifier used to assign the virus to its exact class relying on its PQS content was based on the 'multinom' function of the R package 'nnet'.

PQSs position and overlap with genomic features

The feature tables containing viral genome annotations were downloaded from RefSeq or GenBank for all the reference sequences reported in S1 Table. Genome coordinates were extracted for coding sequences ('CDS'), repeat regions ('repeat_region'), 5'- and 3'-untranslated (UTR) and promoter regions. Given the annotation inconsistency of promoters and UTRs, two new feature categories were created, 5' – and 3' – regulatory regions that were defined by exploiting the annotation of genes and CDSs. We calculated boundaries for genes in the positive strand of viral genomes as follows: 5' – regulatory = $S_{gene} – S_{cds}$; 3' –

regulatory = E_{cds} - E_{gene}. For the genes in the negative strand of viral genomes we defined: 5' – regulatory = S_{cds} - S_{gene}; 3' – regulatory = E_{gene} - E_{cds}. S_{gene}, S_{cds}, E_{gene} and E_{cds} are the Start (S) and End (E) of genes and CDSs as extracted from the feature tables. These newly defined features contain both UTRs and promoters. Since the positive genomic strands are deposited in RefSeq for most of the viruses belonging to the ssRNA (-) class, the following sequences available as negative strands were converted into their inverse complement, together with the coordinates of their genomic features: Junin arenavirus segment S (NC_005081) and segment L (NC_005080), Lassa virus segment L (NC_004297), lymphocytic choriomeningitis virus segment S (GQ862982), Machupo virus segment S (AY924208) and L (AY624354), Pichinde virus segment S (NC_006447), Rift Valley fever virus segment S (NC_014395), and Toscana virus segment S (NC_006318). The overlap extent between PQSs and genomic features was computed by intersecting the genomic coordinates of each PQS with the genomic features extracted from the corresponding virus, and a positive count was recorded every time an overlap of at least one nucleotide was detected. Finally, to compare the enrichment in different feature classes, characterized by different sizes, a normalization step was performed. The total extension of each feature class (*i.e.* CDS, repeat_region, 5' – regulatory and 3' – regulatory) was calculated by summing the lengths of individual features. The total count of PQS overlapping a feature class was then divided by the total length of the class and multiplied by a factor 1,000 to obtain the number of PQS present every 1,000 nucleotides. This procedure was performed considering only the PQSs conserved in at least 80% of sequences for each viral species. All feature tables files were manually revised to fix inconsistencies in the use of keywords and coordinates.

References

1. Lipps HJ, Rhodes D. G-quadruplex structures: in vivo evidence and function. Trends Cell Biol. 2009;19(8):414-22. doi: 10.1016/j.tcb.2009.05.002. PubMed PMID: 19589679.
2. Sen D, Gilbert W. A sodium-potassium switch in the formation of four-stranded G4-DNA. Nature. 1990;344(6265):410-4. doi: 10.1038/344410a0. PubMed PMID: 2320109.
3. Maizels N, Gray LT. The G4 genome. PLoS Genet. 2013;9(4):e1003468. doi: 10.1371/journal.pgen.1003468. PubMed PMID: 23637633.
4. Rouleau S, Jodoin R, Garant JM, Perreault JP. RNA G-Quadruplexes as Key Motifs of the Transcriptome. Adv Biochem Eng Biotechnol. 2017. doi: 10.1007/10_2017_8. PubMed PMID: 28382477.
5. Jayaraj GG, Pandey S, Scaria V, Maiti S. Potential G-quadruplexes in the human long non-coding transcriptome. RNA Biol. 2012;9(1):81-6. doi: 10.4161/rna.9.1.18047. PubMed PMID: 22258148.
6. Mirihana Arachchilage G, Dassanayake AC, Basu S. A potassium ion-dependent RNA structural switch regulates human pre-miRNA 92b maturation. Chem Biol. 2015;22(2):262-72. doi: 10.1016/j.chembiol.2014.12.013. PubMed PMID: 25641166.
7. Agarwala P, Pandey S, Maiti S. The tale of RNA G-quadruplex. Org Biomol Chem. 2015;13(20):5570-85. doi: 10.1039/c4ob02681k. PubMed PMID: 25879384.
8. Cammas A, Millevoi S. RNA G-quadruplexes: emerging mechanisms in disease. Nucleic Acids Res. 2017;45(4):1584-95. doi: 10.1093/nar/gkw1280. PubMed PMID: 28013268.
9. Flint SJ, Racaniello VR, Glenn FR, Skalka AM, Enquist LW. Principles of Virology: Volume 1 Molecular Biology. 4th ed: ASM Press; 2015 August 17, 2015. 574 p.
10. Metifiot M, Amrane S, Litvak S, Andreola ML. G-quadruplexes in viruses: function and potential therapeutic applications. Nucleic Acids Res. 2014;42(20):12352-66. Epub 2014/10/22. doi: 10.1093/nar/gku999. PubMed PMID: 25332402.
11. Ruggiero E, Richter SN. G-quadruplexes and G-quadruplex ligands: targets and tools in antiviral therapy. Nucleic Acids Res. 2018. doi: 10.1093/nar/gky187. PubMed PMID: 29554280.
12. Artusi S, Nadai M, Perrone R, Biasolo MA, Palu G, Flamand L, et al. The Herpes Simplex Virus-1 genome contains multiple clusters of repeated G-quadruplex: Implications for the antiviral activity of a G-quadruplex ligand. Antiviral Res. 2015;118:123-31. Epub 2015/04/07. doi: 10.1016/j.antiviral.2015.03.016. PubMed PMID: 25843424.
13. Artusi S, Perrone R, Lago S, Raffa P, Di Iorio E, Palu G, et al. Visualization of DNA G-quadruplexes in herpes simplex virus 1-infected cells. Nucleic Acids Res. 2016;44(21):10343-53. doi: 10.1093/nar/gkw968. PubMed PMID: 27794039.
14. Gilbert-Girard S, Gravel A, Artusi S, Richter SN, Wallaschek N, Kaufer BB, et al. Stabilization of Telomere G-Quadruplexes Interferes with Human Herpesvirus 6A Chromosomal Integration. J Virol. 2017;91(14). doi: 10.1128/JVI.00402-17. PubMed PMID: 28468887.
15. Madireddy A, Purushothaman P, Loosbroock CP, Robertson ES, Schildkraut CL, Verma SC. G-quadruplex-interacting compounds alter latent DNA replication and episomal persistence of KSHV. Nucleic Acids Res. 2016;44(8):3675-94. doi: 10.1093/nar/gkw038. PubMed PMID: 26837574.
16. Murat P, Zhong J, Lekieffre L, Cowieson NP, Clancy JL, Preiss T, et al. G-quadruplexes regulate Epstein-Barr virus-encoded nuclear antigen 1 mRNA translation. Nat Chem Biol. 2014;10(5):358-64. Epub 2014/03/19. doi: 10.1038/nchembio.1479. PubMed PMID: 24633353.
17. Norseen J, Johnson FB, Lieberman PM. Role for G-quadruplex RNA binding by Epstein-Barr virus nuclear antigen 1 in DNA replication and metaphase chromosome attachment. J Virol. 2009;83(20):10336-46. Epub 2009/08/07. doi: 10.1128/jvi.00747-09. PubMed PMID: 19656898.
18. Tellam JT, Zhong J, Lekieffre L, Bhat P, Martinez M, Croft NP, et al. mRNA Structural constraints on EBNA1 synthesis impact on in vivo antigen presentation and early priming of CD8+ T cells. PLoS Pathog. 2014;10(10):e1004423. doi: 10.1371/journal.ppat.1004423. PubMed PMID: 25299404.

19. Lista MJ, Martins RP, Billant O, Contesse MA, Findakly S, Pochard P, et al. Nucleolin directly mediates Epstein-Barr virus immune evasion through binding to G-quadruplexes of EBNA1 mRNA. Nat Commun. 2017;8:16043. doi: 10.1038/ncomms16043. PubMed PMID: 28685753.
20. Tluckova K, Marusic M, Tothova P, Bauer L, Sket P, Plavec J, et al. Human papillomavirus G-quadruplexes. Biochemistry. 2013;52(41):7207-16. Epub 2013/09/21. doi: 10.1021/bi400897g. PubMed PMID: 24044463.
21. Satkunanathan S, Thorpe R, Zhao Y. The function of DNA binding protein nucleophosmin in AAV replication. Virology. 2017;510:46-54. doi: 10.1016/j.virol.2017.07.007. PubMed PMID: 28704696.
22. Fleming AM, Ding Y, Alenko A, Burrows CJ. Zika Virus Genomic RNA Possesses Conserved G-Quadruplexes Characteristic of the Flaviviridae Family. ACS Infect Dis. 2016;2(10):674-81. doi: 10.1021/acsinfecdis.6b00109. PubMed PMID: 27737553.
23. Wang SR, Min YQ, Wang JQ, Liu CX, Fu BS, Wu F, et al. A highly conserved G-rich consensus sequence in hepatitis C virus core gene represents a new anti-hepatitis C target. Sci Adv. 2016;2(4):e1501535. Epub 2016/04/07. doi: 10.1126/sciadv.1501535. PubMed PMID: 27051880.
24. Tan J, Vonrhein C, Smart OS, Bricogne G, Bollati M, Kusov Y, et al. The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. PLoS Pathog. 2009;5(5):e1000428. Epub 2009/05/14. doi: 10.1371/journal.ppat.1000428. PubMed PMID: 19436709.
25. Kusov Y, Tan J, Alvarez E, Enjuanes L, Hilgenfeld R. A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex. Virology. 2015;484:313-22. doi: 10.1016/j.virol.2015.06.016. PubMed PMID: 26149721.
26. Wang SR, Zhang QY, Wang JQ, Ge XY, Song YY, Wang YF, et al. Chemical Targeting of a G-Quadruplex RNA in the Ebola Virus L Gene. Cell Chem Biol. 2016;23(9):1113-22. doi: 10.1016/j.chembiol.2016.07.019. PubMed PMID: 27617851.
27. Biswas B, Kandpal M, Vivekanandan P. A G-quadruplex motif in an envelope gene promoter regulates transcription and virion secretion in HBV genotype B. Nucleic Acids Res. 2017. doi: 10.1093/nar/gkx823. PubMed PMID: 28981800.
28. Perrone R, Nadai M, Frasson I, Poe JA, Butovskaya E, Smithgall TE, et al. A dynamic G-quadruplex region regulates the HIV-1 long terminal repeat promoter. J Med Chem. 2013;56(16):6521-30. Epub 2013/07/20. doi: 10.1021/jm400914r. PubMed PMID: 23865750.
29. Perrone R, Butovskaya E, Daelemans D, Palu G, Pannecouque C, Richter SN. Anti-HIV-1 activity of the G-quadruplex ligand BRACO-19. J Antimicrob Chemother. 2014;69(12):3248-58. Epub 2014/08/12. doi: 10.1093/jac/dku280. PubMed PMID: 25103489.
30. Piekna-Przybylska D, Sullivan MA, Sharma G, Bambara RA. U3 region in the HIV-1 genome adopts a G-quadruplex structure in its RNA and DNA sequence. Biochemistry. 2014;53(16):2581-93. Epub 2014/04/17. doi: 10.1021/bi4016692. PubMed PMID: 24735378.
31. Amrane S, Kerkour A, Bedrat A, Vialet B, Andreola ML, Mergny JL. Topology of a DNA G-quadruplex structure formed in the HIV-1 promoter: a potential target for anti-HIV drug development. J Am Chem Soc. 2014;136(14):5249-52. Epub 2014/03/22. doi: 10.1021/ja501500c. PubMed PMID: 24649937.
32. Perrone R, Nadai M, Poe JA, Frasson I, Palumbo M, Palu G, et al. Formation of a unique cluster of G-quadruplex structures in the HIV-1 Nef coding region: implications for antiviral activity. PLoS One. 2013;8(8):e73121. doi: 10.1371/journal.pone.0073121. PubMed PMID: 24015290.
33. Lago S, Tosoni E, Nadai M, Palumbo M, Richter SN. The cellular protein nucleolin preferentially binds long-looped G-quadruplex nucleic acids. Biochim Biophys Acta. 2017;1861(5 Pt B):1371-81. doi: 10.1016/j.bbagen.2016.11.036. PubMed PMID: 27913192.
34. Scalabrin M, Frasson I, Ruggiero E, Perrone R, Tosoni E, Lago S, et al. The cellular protein hnRNP A2/B1 enhances HIV-1 transcription by unfolding LTR promoter G-quadruplexes. Sci Rep. 2017;7:45244. doi: 10.1038/srep45244. PubMed PMID: 28338097.

624 35. Perrone R, Doria F, Butovskaya E, Frasson I, Botti S, Scalabrin M, et al. Synthesis, Binding and
625 Antiviral Properties of Potent Core-Extended Naphthalene Diimides Targeting the HIV-1 Long
626 Terminal Repeat Promoter G-Quadruplexes. J Med Chem. 2015;58(24):9639-52. Epub 2015/11/26.
627 doi: 10.1021/acs.jmedchem.5b01283. PubMed PMID: 26599611.
628 36. Perrone R, Lavezzo E, Palu G, Richter SN. Conserved presence of G-quadruplex forming
629 sequences in the Long Terminal Repeat Promoter of Lentiviruses. Sci Rep. 2017;7(1):2018. doi:
630 10.1038/s41598-017-02291-1. PubMed PMID: 28515481.
631 37. Pandey S, Agarwala P, Maiti S. Effect of loops and G-quartets on the stability of RNA G-
632 quadruplexes. J Phys Chem B. 2013;117(23):6896-905. doi: 10.1021/jp401739m. PubMed PMID:
633 23683360.
634 38. Guedin A, Gros J, Alberti P, Mergny JL. How long is too long? Effects of loop size on G-
635 quadruplex stability. Nucleic Acids Res. 2010;38(21):7858-68. doi: 10.1093/nar/gkq639. PubMed
636 PMID: 20660477.
637 39. Mukundan VT, Phan AT. Bulges in G-quadruplexes: broadening the definition of G-
638 quadruplex-forming sequences. J Am Chem Soc. 2013;135(13):5017-28. doi: 10.1021/ja310251r.
639 PubMed PMID: 23521617.
640 40. Frees S, Menendez C, Crum M, Bagga PS. QGRS-Conserve: a computational method for
641 discovering evolutionarily conserved G-quadruplex motifs. Hum Genomics. 2014;8:8. doi:
642 10.1186/1479-7364-8-8. PubMed PMID: 24885782.
643 41. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R. Patterns of evolution and host gene
644 mimicry in influenza and other RNA viruses. PLoS Pathog. 2008;4(6):e1000079. doi:
645 10.1371/journal.ppat.1000079. PubMed PMID: 18535658.
646 42. Lobo FP, Mota BE, Pena SD, Azevedo V, Macedo AM, Tauch A, et al. Virus-host coevolution:
647 common patterns of nucleotide motif usage in Flaviviridae and their hosts. PLoS One.
648 2009;4(7):e6282. doi: 10.1371/journal.pone.0006282. PubMed PMID: 19617912.
649 43. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. Nucleic
650 Acids Res. 2005;33(9):2908-16. doi: 10.1093/nar/gki609. PubMed PMID: 15914667.
651 44. Armitage P, Berry G. Statistical Methods in Medical Research. Publications OBS, editor1994.
652 45. Chambers J, Cleveland W, Kleiner B, Tukey P. Graphical Methods for Data Analysis.
653 Wadsworth, editor1983.
654 46. Fry M, Loeb LA. The fragile X syndrome d(CGG)n nucleotide repeats form a stable
655 tetrahelical structure. Proc Natl Acad Sci U S A. 1994;91(11):4950-4. Epub 1994/05/24. PubMed
656 PMID: 8197163.
657 47. Sket P, Pohleven J, Kovanda A, Stalekar M, Zupunski V, Zalar M, et al. Characterization of
658 DNA G-quadruplex species forming from C9ORF72 G4C2-expanded repeats associated with
659 amyotrophic lateral sclerosis and frontotemporal lobar degeneration. Neurobiol Aging.
660 2015;36(2):1091-6. Epub 2014/12/03. doi: 10.1016/j.neurobiolaging.2014.09.012. PubMed PMID:
661 25442110.
662 48. Zhou B, Liu C, Geng Y, Zhu G. Topology of a G-quadruplex DNA formed by C9orf72
663 hexanucleotide repeats associated with ALS and FTD. Sci Rep. 2015;5:16673. Epub 2015/11/14. doi:
664 10.1038/srep16673. PubMed PMID: 26564809.
665 49. Reddy K, Zamiri B, Stanley SY, Macgregor RB, Jr., Pearson CE. The disease-associated
666 r(GGGGCC)n repeat from the C9orf72 gene forms tract length-dependent uni- and multimolecular
667 RNA G-quadruplex structures. J Biol Chem. 2013;288(14):9860-6. doi: 10.1074/jbc.C113.452532.
668 PubMed PMID: 23423380.
669 50. Parrotta L, Ortuso F, Moraca F, Rocca R, Costa G, Alcaro S, et al. Targeting unimolecular G-
670 quadruplex nucleic acids: a new paradigm for the drug discovery? Expert Opin Drug Discov.
671 2014;9(10):1167-87. doi: 10.1517/17460441.2014.941353. PubMed PMID: 25109710.
672 51. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics.
673 2010;26(19):2460-1. doi: 10.1093/bioinformatics/btq461. PubMed PMID: ISI:000282170000016.

674 52. Perrone R, Lavezzo E, Riello E, Manganelli R, Palu G, Toppo S, et al. Mapping and
675 characterization of G-quadruplexes in Mycobacterium tuberculosis gene promoter regions. Sci Rep.
676 2017;7(1):5743. doi: 10.1038/s41598-017-05867-z. PubMed PMID: 28720801.
677 53. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-a multiple
678 sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189-91. doi:
679 10.1093/bioinformatics/btp033. PubMed PMID: ISI:000265523300016.
680 54. Berselli M, Lavezzo E, Toppo S. NeSSie: a tool for the identification of approximate DNA
681 sequence symmetries. Bioinformatics. 2018. doi: 10.1093/bioinformatics/bty142. PubMed PMID:
682 29522153.
683 55. McKinney W. Data Structures for Statistical Computing in Python. Proceedings of the 9th
684 Python in Science Conference. 2010:51-6.
685 56. van der Walt S, Colbert SC, Varoquaux G. The NumPy Array: A Structure for Efficient
686 Numerical Computation. Computing in Science & Engineering. 2011;13(2):22-30. PubMed PMID:
687 ISI:000288053300003.
688 57. Yachdav G, Wilzbach S, Rauscher B, Sheridan R, Sillitoe I, Procter J, et al. MSAViewer:
689 interactive JavaScript visualization of multiple sequence alignments. Bioinformatics.
690 2016;32(22):3501-3. doi: 10.1093/bioinformatics/btw474. PubMed PMID: 27412096.
691 58. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome
692 browser. Genome Res. 2009;19(9):1630-8. doi: 10.1101/gr.094607.109. PubMed PMID: 19570905.

693

Supporting information

S1 Fig. PQS content in real vs simulated viral genomes (single nucleotide assembling). Segment diagrams of mid-P values obtained by comparing the PQS content detected in real and simulated viral genomes. Simulated viruses have the same nucleotide composition of the real ones, but different order. The three G-island types considered in the positive (+) and negative (-) strands of all human virus genomes are grouped in the 7 Baltimore classes. From left to right, each segment represents one of the three G-islands (GG, GGG, GGGG) in the positive (top half) and negative (bottom half) strands; the radius of a segment corresponds to 1 minus the mid-P value. Thus, full segments indicate highly significant PQSs, whereas null segments indicate non-significant PQSs, with respect to the random sequences.

S2 Fig. PQS content in real vs simulated viral genomes (G-island reshuffling). Segment diagrams of mid-P values obtained by comparing the PQS content detected in real and simulated viral genomes. Simulated viruses are obtained by reshuffling the positions of their GG-, GGG- or GGGG-islands. The three G-island types considered in the positive (+) and negative (-) strands of all human virus genomes are grouped in the 7 Baltimore classes. From left to right, each segment represents one of the three G-islands (GG, GGG, GGGG) in the positive (top half) and negative (bottom half) strands; the radius of a segment corresponds to 1 minus the mid-P value. Thus, full segments indicate highly significant PQSs, whereas null segments indicate non-significant PQSs, with respect to the random sequences.

S3 Fig. PQS overlap with genomic features – GG islands. Representative figure of GG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and regulatory regions at the 5' and 3' ends.

S4 Figure. PQS overlap with genomic features – GGG islands. Representative figure of GGG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and regulatory regions at the 5' and 3' ends.

S5 Fig. PQS overlap with genomic features – GGGG islands. Representative figure of GGGG-island PQSs that overlap with genomic features. The bar charts report the distribution of PQSs in genomic features where available and annotated in the database. The number of PQSs per 1kb is reported on the x-axis both for the positive (orange) and negative (blue) strands. The four features considered are coding sequences (CDS), repeat regions (RR), and regulatory regions at the 5' and 3' ends.

S1 Table. Accession numbers of reference sequences selected for each virus.

S2 Table. Experimentally validated G4s in human viruses.

S3 Table. List of viruses whose PQS content is significant at 10% with respect to randomized sequences.