



HAL
open science

Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases

Ole Tørresen, Bastiaan Star, Pablo Mier, Miguel Andrade-Navarro, Alex Bateman, Patryk Jarnot, Aleksandra Gruca, Marcin Grynberg, Andrey Kajava, Vasilis Promponas, et al.

► To cite this version:

Ole Tørresen, Bastiaan Star, Pablo Mier, Miguel Andrade-Navarro, Alex Bateman, et al.. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 2019, 47 (21), pp.10994-11006. 10.1093/nar/gkz841 . hal-03089273

HAL Id: hal-03089273

<https://hal.science/hal-03089273>

Submitted on 30 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases

Ole Kristian Tørresen¹, Bastiaan Star¹, Pablo Mier², Miguel A. Andrade-Navarro², Alex Bateman³, Patryk Jarnot⁴, Aleksandra Gruca⁴, Marcin Grynberg⁵, Andrey V. Kajava^{6,7}, Vasilis J. Promponas⁸, Maria Anisimova^{9,10}, Kjetill S Jakobsen¹, Dirk Linke^{11,*}

¹Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, NO-0316 Oslo, Norway

²Faculty of Biology, Johannes Gutenberg University Mainz, Hans-Dieter-Husch-Weg 15, 55128 Mainz, Germany

³European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton. CB10 1SD, UK

⁴Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

⁵Institute of Biochemistry and Biophysics PAS, Pawińskiego 5A, 02-106 Warsaw, Poland

⁶Centre de Recherche en Biologie cellulaire de Montpellier, UMR 5237 CNRS, Université Montpellier 1919 Route de Mende, CEDEX 5, 34293, Montpellier, France

⁷Institut de Biologie Computationnelle, 34095, Montpellier, France

⁸Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, PO Box 20537, CY 1678, Nicosia, Cyprus

⁹Institute of Applied Simulations, School of Life Sciences and Facility Management, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland

¹⁰Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland

¹¹Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, NO-0316 Oslo, Norway

* To whom correspondence should be addressed. Tel: +47-22857654; Email: dirk.linke@ibv.uio.no

ABSTRACT

The widespread occurrence of repetitive stretches of DNA in genomes of organisms across the tree of life imposes fundamental challenges for sequencing, genome assembly, and automated annotation of genes and proteins. This multi-level problem can lead to errors in genome and protein databases that are often not recognized or acknowledged. As a consequence, end users working with sequences with repetitive regions are faced with 'ready-to-use' deposited data whose trustworthiness is difficult to determine, let alone to quantify. Here, we provide a review of the problems associated with tandem repeat sequences that originate from different stages during the sequencing-assembly-annotation-deposition workflow, and that may proliferate in public database repositories affecting all downstream analyses. As a case study, we provide examples of the Atlantic cod genome, whose sequencing and assembly were hindered by a particularly high prevalence of tandem repeats. We complement this case study with examples from other species, where mis-annotations and sequencing errors have propagated into protein databases. With this review, we aim to raise the awareness level within the community of database users, and alert scientists working in the underlying workflow of database

1
2
3 creation that the data they omit or improperly assemble may well contain important biological
4 information valuable to others.
5

6 7 **GLOSSARY**

8
9 **aDNA:** Ancient DNA. DNA isolated from material that are up to several hundred thousand years old.

10
11 **Contigs:** Sequence assembled from shorter sequencing reads into a contiguous stretch of
12 nucleotides.
13

14 **de Bruijn graph:** One of two main computational approaches (the other is **OLC**) for the assembly of
15 sequencing reads into longer sequences such as contigs. Works by dividing reads into overlapping k -
16 mers. A graph is created with nodes corresponding to k -mers and directional edges connecting
17 overlapping nodes. A traversal of the graph can be output as contigs.
18

19 **GenBank:** One of several databases containing all publicly available DNA sequences.

20
21 **Homorepeat:** Also known as **homopolymer tract**, or **polyX** for amino acids, where X is the repeated
22 residue. A perfect **tandem repeat** with unit size one where all the nucleotides or amino acids are the
23 same.
24

25 **Interspersed repeat:** A motif or pattern that is found in multiple loci across a genome, such as
26 **transposable elements**. In contrast, a **tandem repeat** has the motif or pattern repeated in tandem at
27 one locus.
28

29 **K-mer:** A sequence of nucleotides that is k -residues long, such as a 31-mer with 31 nucleotides.

30
31 **LRR: Leucine rich repeats are amino acid motifs found in many different proteins, often repeated in**
32 **tandem.**
33

34 **NLR: Nod-like receptors are proteins involved in innate immune response and contains LRRs among**
35 **other domains.**
36

37 **OLC: Overlap-layout-consensus.** One of two main computational approaches (the other is **de Bruijn**
38 **graph**) for the assembly of sequencing reads into longer sequences such as contigs. Works by finding
39 common sequences in reads (overlaps), and creates a graph where the overlaps are nodes. Traversal
40 of the graph can be output as contigs.
41

42 **Polishing:** The act of mapping reads back to an assembly and recalling the consensus sequence.
43 This is a necessity for assemblies based on PacBio and/or Oxford Nanopore reads, and are often
44 performed in multiple rounds where at least the last couple are done with Illumina reads.
45

46 **Scaffolds:** Contains multiple contigs that are placed into proper order and orientation based on paired
47 reads or other positional information (linked reads, optical maps, linkage maps).
48

49 **Short tandem repeat (STR):** A **tandem repeat** with a unit size shorter than 10 nucleotides.

50
51 **Sequence Read Archive (SRA):** A database of sequencing data and alignment information from
52 high-throughput sequencing platforms such as Illumina, 454 and PacBio among others.
53

54 **Tandem repeat (TR):** A region of DNA or protein where a motif or pattern is repeated in tandem at
55 one locus. The motif or pattern has a size, which is usually called a unit size. For example, the tandem
56 repeat ACACACAC has a unit size of 2. This is in contrast to an **interspersed repeat** where the motif
57 or pattern is found in multiple loci across a genome.
58
59
60

1
2
3 **Transposable elements (TE):** A class of repetitive elements that often code for their own
4 propagation. Found across the genome as **interspersed repeats**.

5 **UniProtKB/Swiss-Prot:** A database of protein sequences that have been manually curated.

6 **VLRs: Variable lymphocyte receptors: immune genes found in jawless vertebrates, also containing**
7 **LRRs.**
8
9

10 11 12 **INTRODUCTION**

13
14 The availability of DNA and protein sequence data has revolutionized the way we study cellular,
15 molecular, physiological, evolutionary and developmental processes, allowing the association of
16 phenotypes with genotypes at a single nucleotide (or single amino acid) resolution. Researchers rely
17 on public sequence depositories and other databases for sharing their data, such as GenBank or
18 UniProt, and the content of these databases has grown exponentially in the last decades. While such
19 databases initially consisted predominantly of submissions of individual gene or protein sequences
20 that were carefully curated, large proportions of the content of genome and protein databases today
21 originate from different types of metagenome and genome sequencing and assembly projects.

22 GenBank, for example, included more than 2635 Gbp (billion base pairs) in its 2017 release number
23 221, of which 2242 Gbp (85%) originated from whole-genome shotgun sequencing (1). For an
24 informed use of such data, it is essential that end users understand the distinct contrast in quality
25 between individual, well-curated submissions and entries generated from automated sequence
26 annotation pipelines. The latter procedures can contain unrecognized errors.
27
28

29
30 Here, we argue that awareness of potential database errors is especially relevant with regards to
31 repetitive stretches of DNA, which can occur in both noncoding and coding regions of genomes. The
32 specific nature of this type of DNA sequences can introduce and propagate bias during multiple levels
33 of analyses, and resulting uncertainties and errors are automatically translated further into protein
34 sequences where they become impossible to recognize. Such issues may arise from problems
35 originating from DNA sequencing, from difficulties with assembling repetitive DNA regions and from
36 inaccuracies generated during the annotation process. The multiplicity of these error sources makes it
37 particularly difficult for researchers to understand and assess the bias that may be underlying the
38 sequences that they retrieve from public databases. As an example, in Table 1, we have listed the
39 total number of proteins in UniProtKB/Swiss-Prot that have changed the length of their repetitive
40 region from the first occurrence in the database to the latest – suggesting that errors in repetitive
41 region length have been identified and corrected. The average difference in length is 13.57 amino
42 acids, a substantial number. The 1669 proteins with differences in repeats (Table 1) are 6 % of all
43 proteins in the database that have a repetitive region (see Table 2). These numbers do not reflect a
44 true error rate but suggest that errors in repeat numbers and repeat length are frequent and might
45 often go unnoticed, especially in databases that are less well curated than UniProtKB/Swiss-Prot.
46
47
48
49
50
51
52
53
54
55

56 In this review, we discuss different types of sequencing and database errors, using prominent,
57 published examples where such errors have been found. We first provide a description of the different
58 types of repeats that occur on the DNA and protein level and an overview of DNA sequencing
59 technologies with their benefits and limitations. We then describe the genome assembly, annotation,
60

1
2
3 and database deposition processes, and then link these processes to the different types of errors that
4 may occur at different points in this workflow. We aim to alert the ever-growing community of database
5 end-users of these errors, and to raise awareness among the scientists working in the underlying
6 workflow of database creation, that data that they omit or improperly assemble may well contain
7 important biological information valuable to others.
8
9

10
11 *Repetitive elements in genomes.* Repetitive DNA occurs in all domains of life - Bacteria, Archaea and
12 Eukaryota - and can be grouped into two categories: interspersed repeats, such as transposable
13 elements occurring in multiple loci across the genome, and tandem repeats (TRs) that occur in a
14 single locus. In eukaryotes, repetitive DNA also occurs in specific chromosomal regions, such as the
15 (sub)telomeric regions (2, 3) and the centromeres (4). Transposable elements (TEs) are typically
16 several thousand base pairs (kbp) in size, and in eukaryotes their size can range from 100 base pairs
17 (bp) to 20 kbp (5). Large fractions of vertebrate genomes are filled with active and inactive fragments
18 of TEs, with more than 40% of the genome of zebrafish and more than a third of mammalian genomes
19 consist of TEs (6). Evolutionarily old TEs will accumulate mutations and will diverge from the original
20 sequence, and TEs can therefore lose their repetitive nature over time. In contrast, TRs may consist of
21 motifs as short as 1 bp, where the motif is repeated in tandem. Short tandem repeats (with a motif
22 shorter than 10 bp) were originally called microsatellites (7), longer tandem repeats (with a motif
23 between 10 to 100 bp) were called minisatellite DNA (8), and long tandem repeats (with a repeating
24 motif longer than 100 bp) were called satellite DNA (9). In eukaryotes (based on studies done on
25 metazoans, green algae, plants and yeast), the content of TRs with a unit size of 1-50 bp usually
26 varies between 2000 bp/Mbp and 55 000 bp/Mbp (corresponding to 0.2 to 5.5 % of the genome) (10,
27 11). Repeats also lead to significant intra-specific variation (i.e. variation between individuals of the
28 same species) (12, 13) as shown in a wide range of eukaryotes, for instance *Arabidopsis* (13, 14) and
29 *Drosophila* (15). Within humans, repeats outnumber the number of bases affected by SNP variation by
30 an order of magnitude (4-5 fold) (16). Intra-specific variation poses its own intrinsic challenges for
31 instance when sequencing samples from pooled individuals (17).
32
33

34 Short tandem repeats (STR) are less prevalent in bacteria compared to eukaryotes – presumably due
35 to the typically compact bacterial genomes – but nonetheless regularly occur in bacterial coding
36 regions (18).
37

38 TEs can cause “breakage” of a continuous assembly and lead to assembly collapse, where the
39 number of copies of a repeat found in a genome assembly is lower than the true number, but the
40 relatively large and often evolutionary divergent TEs are unlikely to greatly affect the accuracy of
41 sequencing, assembly and annotation of individual protein-coding regions. While such TEs might
42 sometimes insert themselves into gene regions, the disruptive effects of multiple kbps of sequence
43 inserted into coding regions likely make these events extremely rare. In contrast, TRs are usually
44 much shorter, and can often be in frame in coding regions; therefore, we mainly focus on the problems
45 caused by this class of repeats on the sequencing, assembly, annotation and database deposition
46 processes.
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 *Short and long tandem repeats in coding sequences.* TRs are found in both non-coding and coding
4 genomic regions, and the latter make repeated sequences also ubiquitous in proteomes. Conservative
5 estimates suggest that TRs are present in at least one third of human protein sequences and in half of
6 the protein sequences of the unicellular malaria parasite *Plasmodium falciparum* and the mold
7 *Dictyostelium discoideum* (19, 20). In UniProtKB/Swiss-Prot, 5 % of all proteins have a repetitive
8 region (see Supplementary Material and Table 2). The TR regions come in various flavors; from single
9 amino acid repeats (homorepeats) to the repetition of homologous domains of 100 or more residues
10 (21, 22). TRs with short repetitive units are more frequent than those with long repetitive units (19, 23,
11 24), and repeats are more frequent in Eukaryota compared to Bacteria and Archaea (Table 2). With
12 their highly mutable nature, the presence of variable TRs in coding sequences may directly lead to an
13 increase in protein variation and modification, which is particularly relevant for functional and
14 evolutionary studies (25, 26).

15
16 Tri-nucleotide repeats in coding regions may result in amino acid homorepeats (or polyX). These are
17 widely distributed in all branches of the tree of life and in many protein types (27). Like other TRs,
18 homorepeats can be important for function and their length variation is modulated by selection, as has
19 been demonstrated for many protein families (28). In particular, the expansion of CAG repeats that
20 translate to polyglutamine tracts (polyQ) have been widely studied. These polyQ stretches seem to be
21 advantageous for function in protein interactions. When the length of the repeats is too long, the
22 resulting proteins can aggregate and cause disease, leading to selection against further repeat
23 expansion (29). Dedicated databases and resources have been developed to list and characterize
24 amino acid homorepeats of all types (30, 31).

25
26 Approximately half of the TR regions in proteins may be naturally unfolded (32-34), while the other
27 half of these repetitive regions folds with a plethora of shapes and functions (35, 36). Their protein
28 structures can be subdivided into five major classes: (i) crystalline aggregates formed by regions with 1
29 or 2 residue long repeats, (ii) fibrous structures stabilized by interchain interactions with 3-7 residue
30 repeats, (iii) structures with the repeats of 5–40 residues dominated by solenoid proteins, (iv) “closed”
31 (not elongated) structures with 30-60 residue long repeats and, finally, (v) “beads on a string” structures
32 with typical size of repeats over 50 residues, which are already large enough to fold independently into
33 stable domains (35, 36). When studying repetitive protein structures, it is essential that the underlying
34 sequence information is accurate, not only regarding the type of repeats, but also the exact repeat unit
35 number, as the latter will for example influence the length of protein fibres or the curvature of solenoid
36 proteins. Unexpectedly high conservation of TR repeat unit number and order has been reported for
37 proteins from species separated by long evolutionary time (23, 37). This implies that negative selective
38 pressures act on TRs to preserve important protein functions. The same studies suggest that
39 diversifying selective pressures may play equally important role in function of TR-containing proteins.
40 For example, leucine-rich repeats can be both conserved and play role in adaptation (37-39). Indeed,
41 consistent with this premise, TRs are frequently found in virulence factors of pathogens, toxins,
42 allergens, amyloidogenic proteins and other disease-related sequences. Fast-evolving repeat regions
43 might confer variation to the surface proteins of pathogens allowing them to escape the host defense
44 systems (40, 41). Moreover, there is an increasing amount of evidence for a causal relationship between
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 mutations in TR regions and human-inherited genetic disorders (42). All these examples show that
4 errors in databases are not only an academic problem but also pose risks in analyses of medically
5 relevant data.
6

7 In the following sections, we discuss different problems that occur in today's sequence databases.
8 All these problems originate directly or indirectly from the sequencing and assembly process, and all
9 relate to repeats on the DNA level, leading to fundamental errors in the final database entries.
10
11

12 **SEQUENCING AND GENOME ASSEMBLY ARE AFFECTED BY TANDEM REPEATS**

13
14 *High-throughput sequencing technologies.* High-throughput sequencing technologies remain under
15 fast development and several types of technology have been or are currently available. Each of these
16 technologies has its own distinct features that influence their ability to characterize repeats. In the
17 Sanger sequencing technology era, each read was accompanied by a fluorescent peak trace
18 chromatogram. This enabled researchers to double-check whether or not the correct base was
19 incorporated in a position, which could be helpful in troublesome regions such as repeats. While
20 similar information is available for high-throughput sequencing technologies, usually encoded as
21 quality scores, the massive amounts of data produced makes it infeasible to manually check the
22 quality of individual bases.
23
24

25 The most widely-used technology is the Illumina sequencing platform (43). This technology has a
26 relatively low sequencing error rate (<0.1 %) (44), and errors are mainly due to substitution errors.
27 Nonetheless, Illumina reads are relatively short (< 250 bp), which is a limiting factor since many repeat
28 regions are longer than the length of the read. This technology is therefore not able to fully resolve
29 such longer repeats.
30
31

32 Platforms with significantly longer read length comprise the Single Molecule Real Time Sequencing
33 from Pacific Biosystems ("PacBio") (45) and Nanopore Sequencing from Oxford Nanopore
34 Technologies ("Nanopore") (46). The longer read lengths (1-100+ kbp, usually 10-40 kbp) can
35 successfully span longer stretches of repetitive DNA such as TRs and TEs. Both platforms, however,
36 have high single-pass error-rates (11-15 % for PacBio (47), similar for Nanopore (48)). The majority of
37 these errors consist of insertion and deletions (indels), leading to additional or fewer nucleotides
38 compared to the actual genomic sequence. These error rates can be addressed by more sequencing
39 data (to a higher coverage), which will allow for better error correction during assembly. This effort
40 comes at considerable additional economic costs, which can be up to an order of magnitude more
41 expensive than Illumina sequencing.
42
43

44 A discontinued platform is the Roche/454 pyrosequencing technology. Producing reads up to 1000
45 bp, the 454 technology had difficulty with accurately sequencing homopolymers, leading to indel errors
46 in such regions (49). Albeit 454 finds nearly no use for whole-genome sequencing today, data
47 obtained from this technology still constitutes a considerable part of the DNA and protein sequence
48 databases, being the platform with the second most entries in SRA still today (see Supplementary
49 Material). The Ion Torrent system is similar to the Roche/454, and also has similar issues with indels
50 (50). The relatively long read lengths of these technologies have benefits for crossing repeat regions,
51
52
53
54
55
56
57
58
59
60

1
2
3 yet this advantage is somewhat negated by their inability to correctly assess longer (more than 4-5
4 nucleotides) stretches of homopolymers (51).

5
6 It is clear from descriptions above that in a perfect world, all sequence data generated would consist
7 of high-coverage, long-range PacBio or Nanopore sequencing as a basis, with some Illumina data for
8 error correction. Yet, the short Illumina reads are economical, accurate and can resolve most parts of
9 any genome, which includes most coding regions and degraded TEs. The economy and utility of the
10 Illumina platform is the main reason why so many genomes have been and are still sequenced by that
11 technology, even though PacBio and Nanopore sequencing would technically yield more complete
12 genome assemblies. Given the widespread use of Illumina technology, genome assemblies and
13 databases are currently likely biased against longer TRs in that many of them do not get incorporated
14 into assembled sequences. How this impacts or biases protein databases cannot be quantified, but
15 individual examples show that especially data from short-read technologies must be taken with care
16 when working with repeat proteins; we show some of these examples in detail further below. We do
17 know that large fractions of proteins in protein databases do contain short TR regions (5 % in
18 UniProtKB/Swiss-Prot, Table 2) and that some of these have had changes in their TR region length
19 from one 'version' of the protein to another (Table 1). Taken together, it is likely that protein databases
20 underrepresent TRs and that many of the TRs that are in these databases are not correct.
21
22
23
24
25
26
27
28

29 *Genome assembly methods.* The process of genome assembly creates a tentative reconstruction of a
30 complete genome based on information found in the sequencing reads and possibly other sources of
31 information, such as linkage maps. There are two major approaches for genome assembly, the "*de*
32 *Bruijn graph*" and "*overlap/layout/consensus (OLC) methods*" and these differ significantly in how
33 repeats get resolved during the assembly process.
34
35

36 The *de Bruijn graph* method uses subsequences (*k*-mers) found in the reads and creates a graph
37 where each node represents a fixed-length sequence (*k*-mer), and the edges connect two *k*-mers with
38 *k*-1 bp sequence in common (which can be found in multiple reads) (52). This graph is then parsed,
39 and depending on implementation, contigs (contiguous sequence based on consensus sequence from
40 the reads) and scaffolds (contigs ordered and oriented based on paired read information) are
41 generated. For the *de Bruijn* approach, the length of an entire repeat region has to be shorter than the
42 *k*-mer (which is usually between 21 and 96, with 31 often used as the default setting) to be properly
43 resolved. For instance, the *de Bruijn graph*-based assembler ALLPATHS-LG collapses all repeats
44 equal to or longer than 96 to 96, its *k*-mer size, in its first processing stages (53), but the repeats can
45 be expanded later in the assembly process. Newer implementations of the *de Bruijn* approach, such
46 as SPAdes (54) and SKESA (55), use multiple *k*-mers to better assemble low sequence coverage
47 regions and repeats. However, neither are designed to assemble larger (such as plant or vertebrate)
48 genomes.
49
50
51
52
53
54

55 One implementation of the *OLC method* was *Celera Assembler*, which was used to assemble the
56 *Drosophila* genome in 2000 (56), the first whole genome shotgun sequencing project of a multicellular
57 organism. This approach works by first detecting overlap between all sequencing reads, then creating
58 a graph based on the overlaps, simplifying and traversing the graph, before outputting so-called
59
60

1
2
3 unitigs (sequences that are either unique in the genome or are collapsed, repeated sequence where
4 repeats occurring in multiple locations in a genome are all found on top of each other in one
5 sequence), based on a multiple sequence alignment from the overlaps (57). Because the overlap step
6 compares each read to all other reads, computational demand can be high (certainly higher than the
7 *de Bruijn* method), but it is reduced with fewer but longer reads because fewer overlaps need to be
8 computed. The overlap step can also tolerate mismatches and indels between the reads, and
9 therefore performs well with longer reads even if these are error-prone. The unitigs are further
10 categorized into unique and repeat unitigs, before they are ordered and oriented into scaffolds based
11 on information from paired reads (if included in the assembly). The *OLC method* can resolve those
12 repeats that are shorter than the read length, and it is not limited by any *k*-mer size as the *de Bruijn*
13 method. Before the availability of long reads such as PacBio and Nanopore, the shorter Illumina reads
14 were usually assembled with the *de Bruijn* method because *OLC* can be computationally demanding.
15 Now, with long reads decreasing in cost, most genome sequencing projects utilize these and
16 assemble them with an assembler implementing *OLC*. This will lead to more complete genomes being
17 published, with more repeats resolved.
18
19
20
21
22
23
24
25

26 *Repeat content and fragmented assemblies.* While the choice of best-practice sequencing methods
27 and assembly approaches can be used to minimize the effects of repeats, their amount, length,
28 localization and sequence identity constitute key limitations to obtaining a complete and contiguous
29 genome assembly (58). TE content is likely the largest factor contributing to fragmented genome
30 assemblies (59). This holds for both assemblies based on Illumina and for PacBio reads, but the
31 problem is larger for assemblies with shorter reads. TE content is part of the reason why larger
32 genomes are harder to assemble, since it is highly correlated with genome size (6, 60). While TEs
33 might induce gaps in the genome assembly, the effects of TRs are harder to quantify. It is not
34 completely clear how PacBio reads handle long STR regions. In one study (61), the authors
35 investigated how PacBio reads handled different STRs, and showed that less than 50 % of reads
36 called the correct length of a STR consisting of 30xAC, most likely due to polymerase slippage errors.
37 This observation partly contradicts the notion that long reads might be the solution to resolving
38 repetitive regions (see conclusions section). However, such slippage problems appear limited to
39 extreme examples, and overall, PacBio-based assemblies using *OLC* should be more accurate than
40 Illumina-based assemblies with regards to STRs (62).
41
42
43
44
45
46
47
48

49 **EXAMPLES OF REPEAT-DRIVEN ERROR PROLIFERATION**

50
51 *Tandem repeats cause sequencing and genome assembly challenges.* Significant variation in the
52 natural abundance of TRs exists in different organisms which complicates assembly procedures and
53 the development of adequate algorithms that perform well in all cases. Atlantic cod (*Gadus morhua*)
54 has been identified as a vertebrate species with an exceptionally high occurrence of STRs (63, 64), in
55 particular AC dinucleotide repeats (62, 65). The high abundance of these repeats has caused several
56 complications, both from a laboratory and bioinformatic perspective, and on the level of DNA and
57 (translated) protein sequences. The first *de novo* assembly (gadMor1) of the Atlantic cod genome was
58
59
60

1
2
3 based on 454 sequencing data (66) and resulted in a fragmented assembly with many gaps. More
4 than 30% of the contig edges contained an STR and nearly a quarter of the gaps in scaffolds were
5 flanked by STRs (Supplementary Note 7 in (66)), indicating that these STRs strongly affected the
6 successful assembly into more contiguous genomic regions. By incorporating PacBio reads, an
7 updated assembly (gadMor2; (62)) yielded an improved continuity, allowing a more in-depth
8 quantification of these repeats. For instance, the antifreeze glycoproteins were completely missing in
9 the gadMor1 assembly (67), while they are found in gadMor2 (see section '*Tandem repeats can hinder*
10 *proper gene annotation*' below). While it is well established that repeats in general can hinder genome
11 assembly, there is little discussions about TRs in particular in the literature besides the example
12 above. For instance, in a discussion regarding fragmented genome assemblies of plants, the authors
13 do discuss briefly the role of TEs in the fragmentation of the assemblies, but never mention TRs in the
14 same setting (68). When discussing repeat content, they only mention TEs. They further mention long
15 reads as the main aid in generating more complete genome assemblies.

16
17 The prolific STR occurrence in Atlantic cod may also interfere with PCR amplification, often an
18 essential step for creating sequencing libraries. Ancient DNA (aDNA) sequencing data from historic
19 Atlantic cod specimens contained inflated STR abundances (up to 35 percent), which is far beyond the
20 naturally observed levels (65). This inflation can be suppressed by a reduced number of amplification
21 cycles and by the inclusion of synthesized dinucleotide repeat oligonucleotides during amplification.
22 These data indicate that a biased amplification reaction, whereby repeats "*self-prime*" during PCR,
23 leads to artificially high levels of AC and AG repeats. Although this *self-priming* appears to be
24 particularly problematic in cod – likely due to its high content of repeats with relatively low sequence
25 complexity (65) – this process also explains the typical PCR fragmentation patterns observed when
26 using transcript-activator like effector (TALE) technology (69). This highlights the propensity of
27 repetitive DNA to interfere with amplification in a variety of protocols and conditions.

28
29 *Tandem-repeated gene families causing assembly collapse.* Gene family expansions often originate
30 from a gene locus being replicated in tandem, giving rise to two or more (almost) identical copies of a
31 gene that can be regarded in essence as a long tandem repeat (70). Over time, these two copies can
32 evolve independently, resulting in two genes with different function (neofunctionalization) or two genes
33 with different expression patterns (subfunctionalization). One such example is the α - and β -globin
34 clusters in vertebrates, where multiple globin genes are found in tandem in each cluster, and where
35 the different genes are expressed at different stages during the development (71). In teleost fishes,
36 the two chromosomal regions are inhabited by different numbers of α - and β -genes, reflecting
37 functional diversity (72). For instance, the different numbers of hemoglobin genes in codfishes are
38 suggested to reflect the depth the different species are found at (i.e. a temperature-variation proxy)
39 (73). Another gene family that greatly expanded in teleost fish are the nod-like receptor (NLR) genes
40 (74, 75), genes encoding proteins active in the innate immune system. It is not completely clear why
41 this class of genes are expanded, but since they are involved in pathogen recognition the expansion
42 might correspond to novel pathogen environments (75). In most teleost species, there does not seem
43 to be a clear pattern to the genomic distribution of these genes (74), and although in many cases
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 occurring as clustered (tandem) repeats they are also spread across the genome similar to
4 transposable elements. Most notably, this multiplicity of similar sequences can cause local genome
5 assembly collapse (i.e. the repeated genes are so similar that they collapse into one gene/region
6 displaying much higher coverage than the rest of the genome) and annotation problems (i.e.
7 annotated as a single gene while in reality multiple, or the genes might be hidden from annotation
8 because the software register them as repeats). This problem can be illustrated by different releases
9 of the zebrafish genome. In previous versions of this genome assembly (i.e. Zv6) the *NLR* genes were
10 more or less collapsed. However, zebrafish assembly GRCz10 was created with substantial efforts in
11 BAC and fosmid clones to close gaps, which enabled researchers to show that 159 of the 368
12 identified *NLR* genes are present as TRs on the long arm of chromosome 4 (76). As a further
13 complicating repeat-issue they occur interspersed with Zn-finger genes and arranged irregularly. The
14 specific organization of the *NLR* and *Zn-finger* genes is likely the result of multiple different local
15 duplications. The repeated nature of this huge genomic architecture makes it difficult to be confident
16 that all the genes have been properly assembled and annotated, even with manual annotation and
17 curation (76).

18
19
20
21
22
23
24
25 Many immune genes such as NLRs contain leucine rich repeats (LRRs) (77). These are tandem
26 repeats at the amino acid level, but not necessary at the nucleotide level. In jawless vertebrates the
27 variable lymphocyte receptors (VLRs), another class of immune genes, also contain LRRs (78). In
28 lamprey there are three *VLR* genes that each have multiple LRR-encoding modules in their vicinity.
29 Together they can encode several hundreds of different proteins (78). During lymphocyte
30 development, the *VLR* gene region is reorganised, ending up with the incorporation of several of the
31 surrounding LRR modules. Different lymphocytes have different organisations of their *VLR* gene. In
32 the sea lamprey assembly the *VLRC* gene is not complete and is found together with 182 different
33 LRR donor genomic cassettes on 24 scaffolds (79). It is likely that the nature of these LRR cassettes
34 make them hard to assemble properly, but this is not fully clear from the literature (79). An improved
35 genome assembly of sea lamprey including PacBio reads has recently been published (80), but it
36 remains to be seen if that assembly would resolve these complicated regions better.

37
38
39
40
41
42 Long tandem repeats (LTRs) are often associated with protein-coding regions, and can include
43 duplicated genes as well as duplicated (or otherwise multiplied) domains within a protein-coding gene.
44 They are affected by the filtering and masking operations during genome assembly. A problem occurs
45 when the read length of the sequencing method is shorter than the LTR – in this case, repeat numbers
46 can be massively misjudged. In the case of protein-coding regions, this has direct effects on the
47 interpretation of biological function. LTRs are not uncommon in structural proteins on cell surfaces,
48 and in pathogenicity factors of bacteria, parasites, and viruses. As an example, Wrobel et al. (81) have
49 shown that in the fish pathogen *Yersinia ruckeri*, a surface adhesin involved in biofilm formation called
50 IIm has >20 Ig-like domains repeated in tandem that are identical even on the DNA level (repeat
51 length ~300 bp). Repeat numbers vary slightly from strain to strain, but in this case only PacBio-based
52 genomes show the correct number of repeats (Figure 1). Deposited genomes based on short-read
53 methods show underestimated repeat numbers (by a factor of 4 to 5). The fact that the underestimated
54 repeat number is an approximation made during genome assembly is not visible in the deposited
55
56
57
58
59
60

1
2
3 genome data. In a very similar example, Franzén et al. find that in the human and animal parasite
4 Giardia, variable surface proteins (VSPs) are difficult to sequence using 454 sequencing. Using this
5 technology, only a few genes could be assembled due to their highly repetitive nature (82). From other
6 experiments (including some re-sequencing using different technologies), the authors estimate that ca.
7 300 of these repetitive surface proteins should exist in the genome. In yeast, a large set of LTR
8 proteins are included in flocculation (self-adhesion), a process important in biotechnology for removal
9 of the yeast cells by sedimentation or filtration. These *flo* genes are often truncated in deposited
10 genomes, but it is possible that in many cases, this is due to sequencing and assembly issues, and
11 that in reality, these genes are intact in many of the sequenced strains (83). In primates, filaggrin
12 protein is a component of the skin, and the underlying genes have copy number variations between
13 different species (84). The gene contains multiple copies (10-12) of a repeat that is 972-975
14 nucleotides long. Here, researchers found incomplete versions of the gene for chimpanzee, gorilla,
15 orangutan and macaque in the NCBI database, but were able to reconstruct the complete genes by
16 using a combination of PacBio and Illumina sequencing (84), again showing the importance of the
17 choice of sequencing technology. One extreme example of a LTR is *Pseudomonas koreensis* P19E3
18 where a 70 kbp repeat could not be resolved by PacBio sequencing reads (85). However, by utilizing
19 very long reads from Oxford Nanopore in addition to PacBio and Illumina sequences, the researchers
20 were able to properly resolve this LTR (85). Even in cases such as this, researchers may take different
21 approaches to representing the sequence within the database. Guo et al. (86, 87) identified a 37 kbp
22 repeat in the *Marinomonas primoryensis* ice binding protein (MplBP) but were unable to sequence
23 through the region with PacBio sequencing. Based on pulsed-field gel electrophoresis they estimated
24 that is contained about 120 copies of a 104 amino acid. When submitting the protein sequence, they
25 deposited two sequences, one for the amino terminal side of the repeats and one for the carboxy
26 terminal side of the repeats. In other cases such as the sequence determination of the R28 protein
27 from *Streptococcus pyogenes* (88) the authors determined the sequence of the terminal repeats as
28 well as random internal repeats derived from PCR and based on the estimated size of the PCR
29 product of the complete repeat region deposited a full length sequence with an assumption that every
30 repeat was identical.

31
32
33 It is worth noting that repeat numbers within coding regions may vary within a single bacterial
34 colony, potentially leading to another level of complication when estimating repeat numbers. This
35 effect is called hypervariable copy number variation; an example is the SasG protein from
36 *Staphylococcus aureus* strain NCTC 8325 which contains eight identical 128 amino acid B repeats.
37 Roche and colleagues found that PCR of the full length SasG gene led to a ladder of products differing
38 in size by the 400 bp repeat size (89). Individual bands were gel purified and used as a new template
39 for PCR and in each case only a single band was identified demonstrating that the different size
40 products were not due to mis-priming of the repeat DNA during amplification.

41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 **ANNOTATION OF FUNCTION CAN BE AFFECTED BY TANDEM REPEATS**

57
58 *Annotation of repeats.* The task of accurate characterization of TRs should not rely on just one
59 method. This is because the statistical error rates and power of TR prediction vary extensively for
60

1
2
3 different repeat types and different methods - due to fundamental differences in prediction
4 methodology and method assumptions (24). For example, the Tandem Repeats Finder program
5 appears to be very conservative and has a very low power of predicting diverged repeats (Figure 3 in
6 Schaper et al. 2012). As a result, the agreement of TR annotations by different methods is low, since
7 different methods achieve optimal power for different subsets of TR space (in terms of TR unit length,
8 repeat number and unit similarity). Indeed, testing four selected popular TR finders, Schaper et al.
9 (2012) reported that 89% of TRs were found by only one program, <1% were found by three and only
10 0.2% by all four programs. To improve the accuracy and power of TR annotation, it is advisable to use
11 a proper statistical framework combined with a meta-approach that employs several repeat prediction
12 methods, followed by subsequent filtering of false positives using rigorous statistical tests (90).
13 Currently, such procedure can be implemented using the Tandem Repeat Annotation Library (TRAL)
14 (91). The TRAL library can be easily included in developing new pipelines for genome assembly and
15 repeat annotation. Further, TRAL allows for evolutionary analyses of the annotated repeats, such as
16 evaluating whether a TR region may be under selection.

17
18 A genome assembly is most useful when different features such as genes, TEs and other repeats
19 are annotated with their precise location on a scaffold/chromosome and with a unique identifier. This
20 can then provide essential background information for further experiments on gene expression or
21 function, for example when investigating the difference in gene expression between two experimental
22 set-ups with RNA-Seq (92). We often distinguish between structural annotation, specifying all the
23 genes with their intron and exon structure, and functional annotation of genes and their properties
24 (including individual function (e.g. for enzymes) or function in more complex pathways (e.g. in
25 signaling)) (93, 94). A key issue is the typical workflow of annotation in semi-automated pipelines. The
26 annotation process starts with identifying as many repetitive elements as possible, possibly by creating
27 a custom-made repeat library using both homology-based and *de novo* tools (95). Complete TEs often
28 contain genes that are used to facilitate transposition and are often considered less important when
29 investigating a particular species compared to the specific genes of that species. Repeat libraries are
30 thus used to mask the repeats, making annotation of the genes of the species under investigation
31 easier, but removing information related to genes found in transposable elements. TEs and TRs are
32 usually masked. The reason for masking repeats is that *ab initio* gene prediction programs such as
33 AUGUSTUS (96) or GeneMark (97) need to be trained, i.e., optimized for the specific species with
34 regards to codon bias and splicing signals, and this training can be biased by repeats. Evidence for
35 actively expressed genes can be added in the form of transcriptome data assembled by Trinity (98) or
36 StringTie (99), or with the full-length transcripts generated by PacBio Iso-Seq (100). The transcriptome
37 data is often crucial, since it - of the methods mentioned here - alone provide concrete evidence for
38 the presence of the particular genes of a species, and not just assumed via prediction or mapping of
39 proteins. Non-redundant protein databases such as UniProtKB/Swiss-Prot (101) can be included as
40 the basis for annotation, ideally complemented by specific databases of well-annotated proteins from
41 closely related species. All this information can then be integrated by using a program such as
42 MAKER (102, 103) or EVM (104). This approach provides a set of predicted transcripts and proteins,
43 together with a GFF (General Feature Format) track with positions of all the annotated features,
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3 describing their properties. The predicted proteins can be searched using InterProScan (105) to
4 classify proteins to different molecular functions, biological processes and pathways. Since such
5 annotation is likely to be performed on assemblies where biologically relevant repetitive sequences
6 have been removed from the data already, it may generate serious problems. The most important is
7 the risk of removal of vital information about the genome from the final annotation. Consequently, if a
8 TR makes up a large part of an exon or a whole gene, that exon or gene would not be properly
9 annotated.
10
11
12

13
14 *Tandem repeats can hinder gene annotation.* While the process above can already accidentally filter
15 out genes with repetitive regions, the more detailed annotation process can add another level of
16 problems. Specifically, homology search methods such as BLAST usually have built-in filters that
17 hinder alignment to low complexity regions (which often exist as part of repetitive regions or are
18 repetitive regions) (106), and are not adapted to accurately align homologous sequences with different
19 numbers of TR units.
20
21
22

23 Therefore, the annotation process is often just a rough overview of the different genes, repeats and
24 other features in the species of interest, and may not be sufficient for investigations into gene families
25 that are particularly interesting for a researcher. Manual inspection, re-annotation and re-alignment are
26 often necessary for troublesome gene families. One such gene family is the anti-freeze proteins, in
27 particular the anti-freeze glycoproteins (AFGPs) of nototheniid fishes and codfishes (107, 108). In
28 nototheniids the AFGPs consist of a repeated pattern of Thr-Ala(/Pro)-Ala, and in codfishes it
29 sometimes is represented by Arg-Ala(/Pro)-Ala (108). The repeated nature of these gene families
30 requires manual annotation, and this was used in a comparative survey of AFGPs in nototheniid
31 fishes and codfishes (109). Indeed, the automated annotation of the Atlantic cod genome masked
32 these genes as repeats and they would not have been properly characterized without careful
33 investigation using BLAST (109). These genes were not properly assembled in the first version of the
34 Atlantic cod genome (66), but were in the second version created with PacBio reads (62, 109).
35
36
37
38
39

40 Detection of genuine gene fusion events has been reported long before the first complete genomes
41 became available (110, 111), but beyond that point they have been proven instrumental in detecting
42 gene/protein associations with high specificity (112, 113). Repeats may artificially cause gene fusion
43 events, when genes/proteins that are encoded as distinct units in the genome under study (possibly in
44 distant loci or even in different chromosomes). More specifically, in the case where the 5' and 3'
45 termini of two gene loci share a similar repeat or low complexity pattern, there is an increased
46 probability that genome assemblers can erroneously detect an overlap, thus artificially fusing these
47 genes into a single entity. There are known cases where similar repeat regions in adjacent genes can
48 lead to recombination-driven gene fusion (114), but with short sequence reads, assembly errors can
49 arguably lead to 'artificially' fused genes (as detailed above). Such erroneous gene calls may (i)
50 become the cause of downstream gene-prediction or annotation errors, (ii) generate false positive
51 predictions for gene/protein associations and (iii) hinder large-scale genome evolution studies (115,
52 116).
53
54
55
56
57
58
59
60

1
2
3 *Databases, submission and curation.* DNA and protein sequences are routinely submitted to online
4 repositories that make these data available to the public. This is a largely unsupervised process and
5 there is usually little or no post-submission curation of the data. For nucleotide sequences, submitters
6 must only ensure that the submission adheres to various formatting and data standards, and the
7 archival database will make various automated checks of the data and metadata. Problems such as
8 misassembly and contamination are not investigated. At the protein level, the UniProt database takes
9 predicted sequences from nucleotide entries and places them within the UniProtKB/TrEMBL portion of
10 the database with no further quality control. The RefSeq database, at least for bacterial genomes,
11 ignores the submitted protein sequences and runs their own bespoke PGAP pipeline - this leads to a
12 more consistent set of protein sequences and annotations. Only the manually reviewed section of
13 UniProt, UniProtKB/Swiss-Prot allows for corrections to be made to protein sequences and curators
14 will merge multiple entries from UniProtKB/TrEMBL, thus improving the likelihood of identifying the
15 fully correct protein sequence. But even when manually curated, it is difficult to assess whether or not
16 a protein contains the correct number of a repeated pattern or amino acid, and whether errors have
17 occurred in the underlying DNA sequencing process. The difficulty of identifying and classifying DNA
18 tandem repeats, in addition to their extreme variation from species to species, as well as within
19 populations, has promoted the development of specialized bioinformatic algorithms and databases
20 dedicated to repeat detection and characterization.

21
22
23
24
25
26
27
28
29 The first database on human repetitive DNA elements, including TRs, was developed in 1992 (117),
30 eventually becoming RepBase (118). Widespread genome sequencing further fueled the development
31 of specialized resources (both methods for detecting repeats and repeat databases). The parallel
32 development of general and specialized resources related to DNA tandem repeats, has been crucial to
33 the increased awareness of their widespread distribution and has been instrumental for their use both
34 in basic and applied science. With over 50 TR detectors available, equally numerous repeat sequence
35 databases exist today whose data is constantly used in practical applications like agriculture, medicine
36 and forensics. Examples include the Human Genome Browser at UCSC (119), the STRBase (120)
37 maintained by the National Institute of Standards and Technology (NIST, Maryland, US) or the
38 Tandem Repeats Database (TRDB; (121)). Some of these databases have specific applications. For
39 instance, the STRBase has a focus on human STRs whereas the TRDB was developed as a
40 workbench for sequence analyses. Other specialized databases have been developed recently in this
41 regard (e.g. (122-126)), starting off from human-centered research questions and expanding to
42 examples of many other species, such as the tobacco plant (127), *Trichophytum rubrum*, a fungus
43 causing skin disease (128), or the Cannabis plant to characterize the origin of hemp seeds (US
44 Cannabis DNA database; (129)). Despite this diversity, the majority of these databases rely on the
45 results of well-established automated bioinformatic approaches such as the Tandem Repeats Finder
46 (TRF) program (130) or RepeatMasker (118) to characterize repeat content. Especially the use of
47 RepeatMasker as *the* preferred software to identify and mask repeats, (<http://www.repeatmasker.org/>),
48 has allowed the standardized treatment of raw genomic sequences and reproducibility of protocols for
49 the establishment of these databases. However, using RepeatMasker and TRF on their own might not
50
51
52
53
54
55
56
57
58
59
60

1
2
3 be enough to accurately characterize all TRs, and using a meta-approach such as TRAL (mentioned
4 above) would likely lead to better annotation of TRs in both proteins and DNA.
5
6

7 **CONCLUSIONS**

8
9 Both short and long repeat regions in genomes convey important biological functions; but as they
10 cause significant technical problems with DNA sequencing, genome assembly, and gene and genome
11 annotation, they often include significant errors, or are even omitted from datasets in public databases.
12 Researchers with an interest in the function of such repeats may not be fully aware of the multi-level
13 complexities and use genome data without questioning its quality. It is possible but not well
14 documented that numerous publications on repeat numbers, gene duplications or recombination
15 events are based on erroneous data and thus might include wrong evolutionary or functional
16 conclusions. There is no easy solution to this issue and the key purpose of this article is to raise the
17 awareness to the problem, especially amongst end-users of genome and protein databases, but
18 likewise amongst the researchers working on sequencing, assembly and annotation projects that are
19 often not fully aware of the biological importance of the repeat regions that they mis-sequence, mask,
20 or remove. It would be beneficial if deposited data included qualitative and quantitative information on
21 the type of sequencing methods used, the quality of the assembly and of the annotation. We strongly
22 encourage the use of long-read sequencing technologies to better capture the tandem repeats at the
23 sequencing and assembly stages. Specifically, we urge researchers to aim for a sequencing strategy
24 similar to what has been decided for the Vertebrate Genome Project (not published, but partly
25 described in (131) and on [https://www.rockefeller.edu/research/vertebrate-genomes-
26 project/technology-pipeline-and-policies/](https://www.rockefeller.edu/research/vertebrate-genomes-project/technology-pipeline-and-policies/)), and for Earth Biogenome Project (132). This sequencing
27 strategy should in most cases lead to chromosome level genome assemblies for eukaryotes, where
28 there are few gaps in the sequence and most repeats are resolved. For prokaryotes, substantial
29 coverage in PacBio reads (60x), plus some Illumina reads (50x) and some coverage in very long
30 Nanopore reads as described earlier would likely lead to complete prokaryote genome assemblies
31 (85). It is important that more than one round of polishing with Illumina reads are performed on the
32 assemblies, as that reduces any issues that might stem from the long reads (133, 134). The
33 combination of long and short reads has been shown to be beneficial for resolving tandem repeats in
34 genomes (135), and it should create a better foundation for characterizing large gene families that
35 might be underreported. **Recent technological advances by PacBio have enabled circular consensus
36 sequencing of both RNA and DNA, resulting in long (>10 kbp), highly accurate (99.8 %) reads (136).
37 Wide-spread adoption of these technologies should address most of the issues raised here.** While
38 best-practice methods and quality control can improve new datasets that are made available to the
39 research community, it is less clear how to manage the many problems found in existing, deposited
40 data. More work should go into identifying such issues. It would be of great help if databases would
41 allow user comments to deposited items, to alert other users of the problems and to avoid the
42 reiteration of mistakes and misinterpretations. We expect that the wide-spread adaptation of such
43 recommendations is improved by an increased awareness of the challenges associated with TRs
44 within the community of database creators and end-users.
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ACKNOWLEDGEMENTS

The idea for this article was developed during two consecutive meetings of the EU COST-Action BM1405 “Non-globular proteins: From sequence to structure, function and application in molecular physiopathology”. This work is also supported by Research Council of Norway (grant # 251076) to KSJ and by institutional funds of the University of Oslo, Faculty of Mathematics and Natural Sciences, to DL and BS.

REFERENCES

1. Benson,D.A., Cavanaugh,M., Clark,K., Karsch-Mizrachi,I., Ostell,J., Pruitt,K.D. and Sayers,E.W. (2018) GenBank. *Nucleic Acids Res*, **46**, D41–D47.
2. Blackburn,E.H. and Gall,J.G. (1978) A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in Tetrahymena. *J Mol Biol*, **120**, 33–53.
3. Riethman,H., Ambrosini,A. and Paul,S. (2005) Human subtelomere structure and variation. **13**, 505–515.
4. Mehta,G.D., Agarwal,M.P. and Ghosh,S.K. (2010) Centromere identity: a challenge to be faced. *Mol Genet Genomics*, **284**, 75–94.
5. Kidwell,M.G. (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, **115**, 49–63.
6. Chalopin,D., Naville,M., Plard,F., Galiana,D. and Volf,J.-N. (2015) Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol Evol*, **7**, 567–580.
7. Litt,M. and Luty,J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *The American Journal of Human Genetics*, **44**, 397–401.
8. Jeffreys,A.J., Wilson,V. and Thein,S.L. (1985) Hypervariable ‘minisatellite’ regions in human DNA. *Nature*, **314**, 67–73.
9. Vergnaud,G. and Denoeud,F. (2000) Minisatellites: mutability and genome architecture. *Genome Res*, **10**, 899–907.
10. Mayer,C., Leese,F. and Tollrian,R. (2010) Genome-wide analysis of tandem repeats in *Daphnia pulex* - a comparative approach. *BMC Genomics*, **11**, 277.
11. Zhao,Z., Guo,C., Sutharzan,S., Li,P., Echt,C.S., Zhang,J. and Liang,C. (2014) Genome-wide analysis of tandem repeats in plants and green algae. *G3*, **4**, 67–78.
12. Gymrek,M. (2017) A genomic view of short tandem repeats. *Current Opinion in Genetics & Development*, **44**, 9–16.
13. DeBolt,S. (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol Evol*, **2**, 441–453.
14. Press,M.O., McCoy,R.C., Hall,A.N., Akey,J.M. and Queitsch,C. (2018) Massive variation of short tandem repeats with functional consequences across strains of Arabidopsis thaliana. *Genome Res*, **28**, 1169–1178.

15. Chakraborty,M., VanKuren,N.W., Zhao,R., Zhang,X., Kalsow,S. and Emerson,J.J. (2018) Hidden genetic variation shapes the structure of functional elements in *Drosophila*. *Nat Genet*, **50**, 20–25.
16. Consortium,T.1.G.P. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.
17. Futschik,A. and Schlötterer,C. (2010) The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, **186**, 207–218.
18. Zhou,K., Aertsen,A. and Michiels,C.W. (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiology Reviews*, **38**, 119–141.
19. Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1999) A census of protein repeats. *J Mol Biol*, **293**, 151–160.
20. Pellegrini,M. (2015) Tandem Repeats in Proteins: Prediction Algorithms and Biological Role. *Front. Bioeng. Biotechnol.*, **3**, 1536.
21. Heringa,J. (1998) Detection of internal repeats: how common are they? *Curr Opin Struct Biol*, **8**, 338–345.
22. Andrade,M.A., Ponting,C.P., Gibson,T.J. and Bork,P. (2000) Homology-based method for identification of protein repeats using statistical significance estimates. *J Mol Biol*, **298**, 521–537.
23. Schaper,E., Gascuel,O. and Anisimova,M. (2014) Deep Conservation of Human Protein Tandem Repeats within the Eukaryotes. **31**, 1132–1148.
24. Schaper,E., Kajava,A.V., Hauser,A. and Anisimova,M. (2012) Repeat or not repeat?—Statistical validation of tandem repeat prediction in genomic sequences. **40**, 10005–10017.
25. Kushwaha,A.K. and Grove,A. (2013) C-terminal low-complexity sequence repeats of *Mycobacterium smegmatis* Ku modulate DNA binding. *Bioscience reports*, **33**, 175–184.
26. Radó-Trilla,N. and Albà,M. (2012) Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. *BMC Evol Biol*, **12**, 155–10.
27. Jorda,J. and Kajava,A.V. (2010) Protein Homorepeats: Sequences, Structures, Evolution, and Functions. *Advances in Protein Chemistry and Structural Biology*, **79**, 59–88.
28. Mularoni,L., Ledda,A., Toll-Riera,M. and Albà,M.M. (2010) Natural selection drives the accumulation of amino acid tandem repeats in human proteins. *Genome Res*, **20**, 745–754.
29. Mier,P. and Andrade-Navarro,M.A. (2018) Glutamine Codon Usage and polyQ Evolution in Primates Depend on the Q Stretch Length. *Genome Biol Evol*, **10**, 816–825.
30. Mier,P. and Andrade-Navarro,M.A. (2017) dAPE: a web server to detect homorepeats and follow their evolution. *Bioinformatics*, **33**, 1221–1223.
31. Lobanov,M.Y., Sokolovskiy,I.V. and Galzitskaya,O.V. (2014) HRaP: database of occurrence of HomoRepeats and patterns in proteomes. *Nucleic Acids Res*, **42**, D273–D278.
32. Tompa,P. (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays*, **25**, 847–855.
33. Simon,M. and Hancock,J.M. (2009) Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biol*, **10**, R59.
34. Jorda,J., Xue,B., Uversky,V.N. and Kajava,A.V. (2010) Protein tandem repeats – the more perfect, the less structured. *The FEBS Journal*, **277**, 2673–2682.

- 1
- 2
- 3 35. Kajava,A.V. (2012) Tandem repeats in proteins: From sequence to structure. *Journal of Structural*
- 4 *Biology*, **179**, 279–288.
- 5
- 6 36. Paladin,L., Hirsh,L., Piovesan,D., Andrade-Navarro,M.A., Kajava,A.V. and Tosatto,S.C.E. (2017)
- 7 RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein
- 8 structures. *Nucleic Acids Res*, **45**, D308–D312.
- 9
- 10 37. Schaper,E. and Anisimova,M. (2015) The evolution and function of protein tandem repeats in
- 11 plants. **206**, 397–410.
- 12
- 13 38. Kajava,A.V., Anisimova,M. and Peeters,N. (2008) Origin and Evolution of GALA-LRR, a New
- 14 Member of the CC-LRR Subfamily: From Plants to Bacteria? *PLoS ONE*, **3**, e1694.
- 15
- 16 39. Szalkowski,A.M. and Anisimova,M. (2013) Graph-based modeling of tandem repeats improves
- 17 global multiple sequence alignment. *Nucleic Acids Res*, **41**, e162–e162.
- 18
- 19 40. Verstrepen,K.J., Jansen,A., Lewitter,F. and Fink,G.R. (2005) Intragenic tandem repeats generate
- 20 functional variability. *Nat Genet*, **37**, 986–990.
- 21
- 22 41. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution.
- 23 *Trends in Genetics*, **22**, 253–259.
- 24
- 25 42. Sutherland,G.R. and Richards,R.I. (1995) Simple tandem DNA repeats and human genetic
- 26 disease. *Proc Natl Acad Sci USA*, **92**, 3636–3641.
- 27
- 28 43. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P.,
- 29 Evers,D.J., Barnes,C.L., Bignell,H.R., *et al.* (2008) Accurate whole human genome sequencing
- 30 using reversible terminator chemistry. *Nature*, **456**, 53–59.
- 31
- 32 44. Glenn,T.C. (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*,
- 33 **11**, 759–769.
- 34
- 35 45. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B.,
- 36 *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–
- 37 138.
- 38
- 39 46. Olasagasti,F., Lieberman,K.R., Benner,S., Cherf,G.M., Dahl,J.M., Deamer,D.W. and Akeson,M.
- 40 (2010) Replication of individual DNA molecules under electronic control using a protein nanopore.
- 41 *Nat Nanotechnol*, **5**, 798–806.
- 42
- 43 47. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics, Proteomics &*
- 44 *Bioinformatics*, **13**, 278–289.
- 45
- 46 48. Weirather,J.L., de Cesare,M., Wang,Y., Piazza,P., Sebastiano,V., Wang,X.-J., Buck,D. and
- 47 Au,K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore
- 48 Technologies and their applications to transcriptome analysis. *F1000Research*, **6**, 100.
- 49
- 50 49. Balzer,S., Malde,K., Lanzén,A., Sharma,A. and Jonassen,I. (2010) Characteristics of 454
- 51 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–5.
- 52
- 53 50. Bragg,L.M., Stone,G., Butler,M.K., Hugenholtz,P. and Tyson,G.W. (2013) Shining a Light on Dark
- 54 Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comp Biol*, **9**, e1003031.
- 55
- 56 51. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct Comparisons of
- 57 Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA
- 58 Sample. *PLoS ONE*, **7**, e30087.
- 59
- 60 52. Zerbino,D.R. and Birney,E. (2008) Velvet: Algorithms for de novo short read assembly using de
- Bruijn graphs. *Genome Res*, **18**, 821–829.

- 1
2
3 53. Gnerre,S., Maccallum,I., Przybylski,D., Ribeiro,F.J., Burton,J.N., Walker,B.J., Sharpe,T., Hall,G.,
4 Shea,T.P., Sykes,S., *et al.* (2011) High-quality draft assemblies of mammalian genomes from
5 massively parallel sequence data. *Proc Natl Acad Sci USA*, **108**, 1513–1518.
6
7 54. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M.,
8 Nikolenko,S.I., Pham,S., Prjibelski,A.D., *et al.* (2012) SPAdes: A New Genome Assembly
9 Algorithm and Its Applications to Single-Cell Sequencing. <https://home.liebertpub.com/cmb>, **19**,
10 455–477.
11
12 55. Souvorov,A., Agarwala,R. and Lipman,D.J. (2018) SKESA: strategic k-mer extension for
13 scrupulous assemblies. *Genome Biol*, **19**, 1–13.
14
15 56. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A.,
16 Mobarry,C.M., Reinert,K.H., Remington,K.A., *et al.* (2000) A whole-genome assembly of
17 *Drosophila*. *Science*, **287**, 2196–2204.
18
19 57. Miller,J.R., Koren,S. and Sutton,G.G. (2010) Assembly algorithms for next-generation sequencing
20 data. *Genomics*, **95**, 315–327.
21
22 58. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing:
23 computational challenges and solutions. *Nature Rev Genet*, **13**, 36–46.
24
25 59. Sotero-Caio,C.G., Platt,R.N., Suh,A. and Ray,D.A. (2017) Evolution and diversity of transposable
26 elements in vertebrate genomes. *Genome Biol Evol*, **9**, 161–177.
27
28 60. Elliott,T.A. and Gregory,T.R. (2015) What's in a genome? The C-value enigma and the evolution of
29 eukaryotic genome content. *Philos Trans R Soc Lond, B, Biol Sci*, **370**, 20140331.
30
31 61. Liljegen,M.M., de Muinck,E.J. and Trosvik,P. (2016) Microsatellite Length Scoring by Single
32 Molecule Real Time Sequencing – Effects of Sequence Structure and PCR Regime. **11**,
33 e0159232.
34
35 62. Tørresen,O.K., Star,B., Jentoft,S., Reinar,W.B., Grove,H., Miller,J.R., Walenz,B.P., Knight,J.,
36 Ekholm,J.M., Peluso,P., *et al.* (2017) An improved genome assembly uncovers prolific tandem
37 repeats in Atlantic cod. *BMC Genomics*, **18**, 95.
38
39 63. Adams,R.H., Blackmon,H., Reyes-Velasco,J., Schield,D.R., Card,D.C., Andrew,A.L.,
40 Waynewood,N. and Castoe,T.A. (2016) Microsatellite landscape evolutionary dynamics across
41 450 million years of vertebrate genome evolution. *Genome*, **59**, 295–310.
42
43 64. Jiang,Q., Li,Q., Yu,H. and Kong,L. (2014) Genome-wide analysis of simple sequence repeats in
44 marine animals—a comparative approach. **16**, 604–619.
45
46 65. Star,B., Hansen,M.H., Skage,M., Bradbury,I.R., Godiksen,J.A., Kjesbu,O.S. and Jentoft,S. (2016)
47 Preferential amplification of repetitive DNA during whole genome sequencing library creation from
48 historic samples. *Sci Technol Archaeol Res*, **2**, 36–45.
49
50 66. Star,B., Nederbragt,A.J., Jentoft,S., Grimholt,U., Malmstrøm,M., Gregers,T.F., Rounge,T.B.,
51 Paulsen,J., Solbakken,M.H., Sharma,A., *et al.* (2011) The genome sequence of Atlantic cod
52 reveals a unique immune system. *Nature*, **477**, 207–210.
53
54 67. Zhuang,X., Yang,C., Fevolden,S.-E. and Cheng,C.-H. (2012) Protein genes in repetitive
55 sequence—antifreeze glycoproteins in Atlantic cod genome. *BMC Genomics*, **13**, 293.
56
57 68. Belser,C., Istace,B., Denis,E., Dubarry,M., Baurens,F.-C., Falentin,C., Genete,M., Berrabah,W.,
58 Chèvre,A.-M., Delourme,R., *et al.* (2018) Chromosome-scale assemblies of plant genomes using
59 nanopore long reads and optical maps. *Nature Plants* 2018 4:2, **4**, 879–887.
60

- 1
2
3 69. Hommelsheim,C.M., Frantzeskakis,L., Huang,M. and Ülker,B. (2014) PCR amplification of
4 repetitive DNA: a limitation to genome editing technologies and many other applications. *Sci Rep*,
5 **4**, 5052.
6
7 70. Hurler,M. (2004) Gene Duplication: The Genomic Trade in Spare Parts. *PLoS Biol*, **2**, e206.
8
9 71. Hardison,R.C. (2012) Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor Perspectives in*
10 *Medicine*, **2**, a011627–a011627.
11
12 72. Opazo,J.C., Butts,G.T., Nery,M.F., Storz,J.F. and Hoffmann,F.G. (2013) Whole-genome
13 duplication and the functional diversification of teleost fish hemoglobins. *Mol Biol Evol*, **30**, 140–
14 153.
15
16 73. Baalsrud,H.T., Voje,K.L., Tørresen,O.K., Solbakken,M.H., Matschiner,M., Malmstrøm,M.,
17 Hanel,R., Salzburger,W., Jakobsen,K.S. and Jentoft,S. (2017) Evolution of Hemoglobin Genes in
18 Codfishes Influenced by Ocean Depth. *Sci Rep*, **7**, 7956.
19
20 74. Tørresen,O.K., Briec,M.S.O., Solbakken,M.H., Sørhus,E., Nederbragt,A.J., Jakobsen,K.S.,
21 Meier,S., Edvardsen,R.B. and Jentoft,S. (2018) Genomic architecture of haddock
22 (*Melanogrammus aeglefinus*) shows expansions of innate immune genes and short tandem
23 repeats. *BMC Genomics*, **19**, 240.
24
25 75. Stein,C., Caccamo,M., Laird,G. and Leptin,M. (2007) Conservation and divergence of gene
26 families encoding components of innate immune response systems in zebrafish. *Genome Biol*, **8**,
27 R251.
28
29 76. Howe,K., Schiffer,P.H., Zielinski,J., Wiehe,T., Laird,G.K., Marioni,J.C., Soylemez,O.,
30 Kondrashov,F. and Leptin,M. (2016) Structure and evolutionary history of a large family of NLR
31 proteins in the zebrafish. *Open Biol*, **6**, 160009–224.
32
33 77. Matsushima,N., Takatsuka,S., Miyashita,H. and Kretsinger,R.H. (2019) Leucine Rich Repeat
34 Proteins: Sequences, Mutations, Structures and Diseases. *PPL*, **26**, 108–131.
35
36 78. Boehm,T., McCurley,N., Sutoh,Y., Schorpp,M., Kasahara,M. and Cooper,M.D. (2012) VLR-Based
37 Adaptive Immunity. <https://doi.org/10.1146/annurev-immunol-020711-075038>, **30**, 203–220.
38
39 79. Das,S., Hirano,M., Aghaallaei,N., Bajoghli,B., Boehm,T. and Cooper,M.D. (2013) Organization of
40 lamprey variable lymphocyte receptor C locus and repertoire development. *Proc Natl Acad Sci*
41 *USA*, **110**, 6043–6048.
42
43 80. Smith,J.J., Timoshevskaya,N., Ye,C., Holt,C., Keinath,M.C., Parker,H.J., Cook,M.E., Hess,J.E.,
44 Narum,S.R., Lamanna,F., *et al.* (2018) The sea lamprey germline genome provides insights into
45 programmed genome rearrangement and vertebrate evolution. *Nat Genet*, **50**, 270–277.
46
47 81. Wrobel,A., Ottoni,C., Leo,J.C., Gulla,S. and Linke,D. (2018) The repeat structure of two
48 paralogous genes, *Yersinia ruckeri* invasin (*yrInv*) and a '*Y. ruckeri* invasin-like molecule', (*yrIIm*)
49 sheds light on the evolution of adhesive capacities of a fish pathogen. *Journal of Structural*
50 *Biology*, **201**, 171–183.
51
52 82. Franzen,O., Jerlström-Hultqvist,J., Castro,E., Sherwood,E., Ankarklev,J., Reiner,D.S., Palm,D.,
53 Andersson,J.O., Andersson,B. and Svärd,S.G. (2009) Draft Genome Sequencing of *Giardia*
54 *intestinalis* Assemblage B Isolate GS: Is Human Giardiasis Caused by Two Different Species?
55 *PLOS Pathogens*, **5**, e1000560.
56
57 83. Khatri,I., Tomar,R., Ganesan,K., Prasad,G.S. and Subramanian,S. (2017) Complete genome
58 sequence and comparative genomics of the probiotic yeast *Saccharomyces boulardii*. *Sci Rep*, **7**,
59 371.
60
61 84. Romero,V., Hosomichi,K., Nakaoka,H., Shibata,H. and Inoue,I. (2017) Structure and evolution of
the filaggrin gene repeated region in primates. *BMC Evol Biol*, **17**, 10.

- 1
2
3 85. Schmid,M., Frei,D., Patrignani,A., Schlapbach,R., Frey,J.E., Remus-Emsermann,M.N.P. and
4 Ahrens,C.H. (2018) Pushing the limits of de novo genome assembly for complex prokaryotic
5 genomes harboring very long, near identical repeats. *Nucleic Acids Res*, **46**, 8953–8965.
6
7 86. Guo,S., Stevens,C.A., Vance,T.D.R., Olijve,L.L.C., Graham,L.A., Campbell,R.L., Yazdi,S.R.,
8 Escobedo,C., Bar-Dolev,M., Yashunsky,V., *et al.* (2017) Structure of a 1.5-MDa adhesin that
9 binds its Antarctic bacterium to diatoms and ice. *Sci Adv*, **3**, e1701440.
10
11 87. Guo,S., Garnham,C.P., Whitney,J.C., Graham,L.A. and Davies,P.L. (2012) Re-evaluation of a
12 bacterial antifreeze protein as an adhesin with ice-binding activity. *PLoS ONE*, **7**, e48805.
13
14 88. Stålhammar-Carlemalm,M., Areschoug,T., Larsson,C. and Lindahl,G. (1999) The R28 protein of
15 *Streptococcus pyogenes* is related to several group B streptococcal surface proteins, confers
16 protective immunity and promotes binding to human epithelial cells. *Mol Microbiol*, **33**, 208–219.
17
18 89. Roche,F.M., Massey,R., Peacock,S.J., Day,N.P.J., Visai,L., Pietro Speziale, Lam,A., Pallen,M. and
19 Foster,T.J. (2003) Characterization of novel LPXTG-containing proteins of *Staphylococcus aureus*
20 identified from genome sequences. *Microbiology*, **149**, 643–654.
21
22 90. Anisimova,M., Pečerska,J. and Schaper,E. (2015) Statistical Approaches to Detecting and
23 Analyzing Tandem Repeats in Genomic Sequences. **3**.
24
25 91. Schaper,E., Korsunsky,A., Messina,A., Murri,R., Pečerska,J., Stockinger,H., Zoller,S., Xenarios,I.
26 and Anisimova,M. (2015) TRAL: Tandem repeat annotation library. *Bioinformatics*,
27 10.1093/bioinformatics/btv306.
28
29 92. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A.,
30 Szczesniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X., *et al.* (2016) A survey of best practices for
31 RNA-seq data analysis. *Genome Biol*, **17**, 13.
32
33 93. Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Rev*
34 *Genet*, **13**, 329–342.
35
36 94. Hoff,K.J. and Stanke,M. (2015) Current methods for automated annotation of protein-coding
37 genes. *Current Opinion in Insect Science*, **7**, 8–14.
38
39 95. Bergman,C.M. and Quesneville,H. (2007) Discovering and detecting transposable elements in
40 genome sequences. *Briefings in Bioinformatics*, **8**, 382–392.
41
42 96. Stanke,M., Diekhans,M., Baertsch,R. and Haussler,D. (2008) Using native and syntenically
43 mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.
44
45 97. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y.O. and Borodovsky,M. (2005) Gene identification
46 in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**, 6494–6506.
47
48 98. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L.,
49 Raychowdhury,R., Zeng,Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data
50 without a reference genome. *Nat Biotechnol*, **29**, 644–652.
51
52 99. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression
53 analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc*, **11**, 1650–
54 1667.
55
56 100. Gonzalez-Garay,M.L. (2016) Introduction to Isoform Sequencing Using Pacific Biosciences
57 Technology (Iso-Seq). In *Transcriptomics and Gene Regulation*, Translational Bioinformatics.
58 Springer, Dordrecht, Dordrecht, Vol. 9, pp. 141–160.
59
60 101. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*, **43**, D204–
12.

- 1
2
3 102. Campbell,M.S., Law,M., Holt,C., Stein,J.C., Moghe,G.D., Hufnagel,D.E., Lei,J.,
4 Achawanantakun,R., Jiao,D., Lawrence,C.J., *et al.* (2014) MAKER-P: a tool kit for the rapid
5 creation, management, and quality control of plant genome annotations. *Plant Physiol*, **164**, 513–
6 524.
7
8 103. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database
9 management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.
10
11 104. Haas,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and
12 Wortman,J.R. (2008) Automated eukaryotic gene structure annotation using EVIDENCEModeler
13 and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.
14
15 105. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J.,
16 Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification.
17 *Bioinformatics*, **30**, 1236–1240.
18
19 106. Mier,P., Paladin,L., Tamana,S., Petrosian,S., Hajdu-Soltész,B., Urbanek,A., Gruca,A.,
20 Plewczynski,D., Grynberg,M., Bernadó,P., *et al.* (2019) Disentangling the complexity of low
21 complexity proteins. *Briefings in Bioinformatics*, **27**, 331.
22
23 107. Chen,L., DeVries,A.L. and Cheng,C.-H.C. (1997) Evolution of antifreeze glycoprotein gene from a
24 trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci USA*, **94**, 3811–3816.
25
26 108. Chen,L., DeVries,A.L. and Cheng,C.-H.C. (1997) Convergent evolution of antifreeze
27 glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc Natl Acad Sci USA*, **94**, 3817–
28 3822.
29
30 109. Baalsrud,H.T., Tørresen,O.K., Hongrø Solbakken,M., Salzburger,W., Hanel,R., Jakobsen,K.S.
31 and Jentoft,S. (2017) De novo gene evolution of antifreeze glycoproteins in codfishes revealed by
32 whole genome sequence data. *Mol Biol Evol*, **35**, 593–606.
33
34 110. Zakin,M.M., Duchange,N., Ferrara,P. and Cohen,G.N. (1983) Nucleotide sequence of the metL
35 gene of Escherichia coli. Its product, the bifunctional aspartokinase ii-homoserine dehydrogenase
36 II, and the bifunctional product of the thrA gene, aspartokinase I-homoserine dehydrogenase I,
37 derive from a common ancestor. *J Biol Chem*, **258**, 3028–3031.
38
39 111. Ferone,R. and Roland,S. (1980) Dihydrofolate reductase: thymidylate synthase, a bifunctional
40 polypeptide from Crithidia fasciculata. *Proc Natl Acad Sci USA*, **77**, 5802–5806.
41
42 112. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined
43 algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
44
45 113. Enright,A.J., Iliopoulos,I., Kyripides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for
46 complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
47
48 114. Zhao,X., Oh,S.-H., Coleman,D.A. and Hoyer,L.L. (2011) ALS51, a newly discovered gene in the
49 Candida albicans ALS family, created by intergenic recombination: analysis of the gene and
50 protein, and implications for evolution of microbial gene families. *FEMS Immunol Med Microbiol*,
51 **61**, 245–257.
52
53 115. Nagy,A., Szláma,G., Szarka,E., Trexler,M., Bányai,L. and Patthy,L. (2011) Reassessing domain
54 architecture evolution of metazoan proteins: major impact of gene prediction errors. *Genes*
55 (*Basel*), **2**, 449–501.
56
57 116. Promponas,V.J., Iliopoulos,I. and Ouzounis,C.A. (2015) Annotation inconsistencies beyond
58 sequence similarity-based function prediction – phylogeny and genome structure. *Standards in*
59 *Genomic Sciences* 2015 10:1, **10**, 108.
60
61 117. Jurka,J., Walichiewicz,J. and Milosavljevic,A. (1992) Prototypic sequences for human repetitive
62 DNA. *J Mol Evol*, **35**, 286–291.

- 1
2
3 118. Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements.
4 *Trends Genet*, **16**, 418–420.
5
6 119. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M. and Haussler, D.
7 (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.
8
9 120. Ruitberg, C.M., Reeder, D.J. and Butler, J.M. (2001) STRBase: a short tandem repeat DNA
10 database for the human identity testing community. *Nucleic Acids Res*, **29**, 320–322.
11
12 121. Gelfand, Y., Rodriguez, A. and Benson, G. (2007) TRDB--the Tandem Repeats Database. *Nucleic*
13 *Acids Res*, **35**, D80–7.
14
15 122. Hussing, C., Bytyci, R., Huber, C., Morling, N. and Børsting, C. (2018) The Danish STR sequence
16 database: duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit. *Int. J.*
17 *Legal Med.*, **18**, 100.
18
19 123. Adnan, A., Zhan, X., Kasim, K., Rakha, A. and Xin, X.J. (2018) Population data and phylogenetic
20 structure of Han population from Jiangsu province of China on GlobalFiler STR loci. *Int. J. Legal*
21 *Med.*, **132**, 1301–1304.
22
23 124. Ossowski, A., Piatek, J., Parafiniuk, M., Pudlo, A., Pepinski, W., Skawronska, M., Szeremeta, M.,
24 Niemcunowicz-Janica, A. and Soltyszewski, I. (2017) Genetic variation of 15 autosomal STRs in a
25 population sample of Bedouins residing in the area of the Fourth Nile Cataract, Sudan. *Anthropol*
26 *Anz*, **74**, 263–268.
27
28 125. Kim, E.H., Lee, H.Y., Kwon, S.Y., Lee, E.Y., Yang, W.I. and Shin, K.-J. (2017) Sequence-based
29 diversity of 23 autosomal STR loci in Koreans investigated using an in-house massively parallel
30 sequencing panel. *Forensic Science International: Genetics*, **30**, 134–140.
31
32 126. Pamjav, H., Fóthi, Á., Fehér, T. and Fóthi, E. (2017) A study of the Bodrogeköz population in north-
33 eastern Hungary by Y chromosomal haplotypes and haplogroups. *Mol Genet Genomics*, **292**,
34 883–894.
35
36 127. Wang, X., Yang, S., Chen, Y., Zhang, S., Zhao, Q., Li, M., Gao, Y., Yang, L. and Bennetzen, J.L.
37 (2018) Comparative genome-wide characterization leading to simple sequence repeat marker
38 development for *Nicotiana*. *BMC Genomics*, **19**, 500.
39
40 128. Franco, M.E., Bitencourt, T.A., Marins, M. and Fachin, A.L. (2017) In silico characterization of
41 tandem repeats in *Trichophyton rubrum* and related dermatophytes provides new insights into
42 their role in pathogenesis. *Database (Oxford)*, **2017**, 1.
43
44 129. Houston, R., Birck, M., LaRue, B., Hughes-Stamm, S. and Gangitano, D. (2018) Nuclear,
45 chloroplast, and mitochondrial data of a US cannabis DNA database. *Int. J. Legal Med.*, **132**,
46 713–725.
47
48 130. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids*
49 *Res*, **27**, 573–580.
50
51 131. Teeling, E.C., Vernes, S.C., Dávalos, L.M., Ray, D.A., Gilbert, M.T.P., Myers, E. Bat1K Consortium
52 (2018) Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes
53 for All Living Bat Species. *Annu. Rev. Anim. Biosci.*, **6**, 23–46.
54
55 132. Lewin, H.A., Robinson, G.E., Kress, W.J., Baker, W.J., Coddington, J., Crandall, K.A., Durbin, R.,
56 Edwards, S.V., Forest, F., Gilbert, M.T.P., *et al.* (2018) Earth BioGenome Project: Sequencing life
57 for the future of life. *Proc Natl Acad Sci USA*, **115**, 201720115–4333.
58
59 133. Koren, S., Phillippy, A.M., Simpson, J.T., Loman, N.J. and Loose, M. (2019) Reply to ‘Errors in long-
60 read assemblies can critically affect protein prediction’. *Nat Biotechnol*, **30**, 1.

- 1
2
3 134. Watson, M. and Warr, A. (2019) Errors in long-read assemblies can critically affect protein
4 prediction. *Nat Biotechnol*, **37**, 124.
5
6 135. Weissensteiner, M.H., Pang, A.W.C., Bunikis, I., Höjjer, I., Vinnere-Pettersson, O., Suh, A. and
7 Wolf, J.B.W. (2017) Combination of short-read, long-read and optical mapping assemblies reveals
8 large-scale tandem repeat arrays with population genetic implications. *Genome Res*, **27**,
9 gr.215095.116–708.
10
11 136. Wenger, A.M., Peluso, P., Rowell, W.J., Chang, P.-C., Hall, R.J., Concepcion, G.T., Ebler, J.,
12 Fungtammasan, A., Kolesnikov, A., Olson, N.D., *et al.* (2019) Accurate circular consensus long-
13 read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*,
14 **74**, 1–8.
15
16
17
18

19 **Table 1. Summary of proteins from UniProtKB/Swiss-Prot where the length of repetitive region has**
20 **changed between different versions of the database.**

| Proteins (n) | Proteins with different sequence between versions (n) | Proteins with different repetitive region lengths (n) | Average / standard deviation of the length of repetitive regions in original version of the sequence ¹ | Average / standard deviation of the length of repetitive regions in the version 2018_06 of the sequence ¹ | Average / standard deviation of the difference in lengths of repetitive regions ¹ |
|--------------|---|---|---|--|--|
| 554241 | 74434 | 1669 | 31.14 / 72.09 | 35.20 / 84.08 | 13.57 / 45.69 |

21
22
23
24
25
26
27
28
29
30
31
32
33 ¹ measured in amino acid residues
34
35

36 **Table 2. Differences of repetitive region lengths in evolutionarily distinct groups of organisms.**

| Database name | Number of proteins | Number of proteins with STRs | % of proteins with STRs | Median ¹ | Average ¹ | Standard deviation ¹ | Number of clusters ² |
|------------------------------|--------------------|------------------------------|-------------------------|---------------------|----------------------|---------------------------------|---------------------------------|
| UniProtKB/Swiss-Prot (total) | 554241 | 28003 | 5.05% | 14.75 | 15.14 | 3.69 | 6237 |
| Archaea | 19525 | 351 | 1.80% | 10.71 | 10.63 | 1.27 | 45 |
| Bacteria | 333691 | 6794 | 2.04% | 17.38 | 17.45 | 2.66 | 1048 |
| Euk: Fungi | 33613 | 3996 | 11.89% | 13.46 | 13.79 | 3.65 | 893 |
| Euk: | 27607 | 3372 | 12.21% | 17.34 | 18.62 | 7.95 | 812 |

| | | | | | | | |
|-----------------|-------|------|-------|-------|-------|------|------|
| Invertebrata | | | | | | | |
| Euk: Vertebrata | 18292 | 1461 | 7.99% | 13.66 | 13.90 | 2.42 | 1801 |
| Euk: Plants | 42101 | 3601 | 8.55% | 12.51 | 12.82 | 2.98 | 795 |
| Viruses | 16852 | 889 | 5.28% | 14.07 | 14.15 | 2.57 | 203 |

¹repetitive region length, measured in amino acid residues

²Clustering was used to define repeat classes. **Should a protein contain three different, co-localized STRs, the clustering method will produce 6 clusters: three with regular STRs and three with fused repeats. See also** supplementary material for more information.

Figure 1. DNA alignment of a ~39 kbp-long DNA region containing the *yrllm* gene and flanking CDS in *Y. ruckeri* genomes deposited in GenBank. Each CDS is indicated by a yellow arrow, with the percentage of sequence identity to CSF007-82 reported inside the arrow. *yrllm* consists of an array of tandemly repeated, identical Ig-like domains (in red) and in addition of Ig-like domains of lower pairwise sequence similarity (in orange). It is usually capped by a C-type lectin domain (CTL, in green). The dashed lines indicate gaps in the DNA alignment. In strain 150 the grey box indicates a contig break in the assembly. The asterisk (*) indicates assemblies generated through PacBio SMRT sequencing. Note that the other assemblies have significant lower repeat numbers, suggesting that the repeats were not found using short-read sequencing technologies. Modified with permission from Wrobel et al 2018.

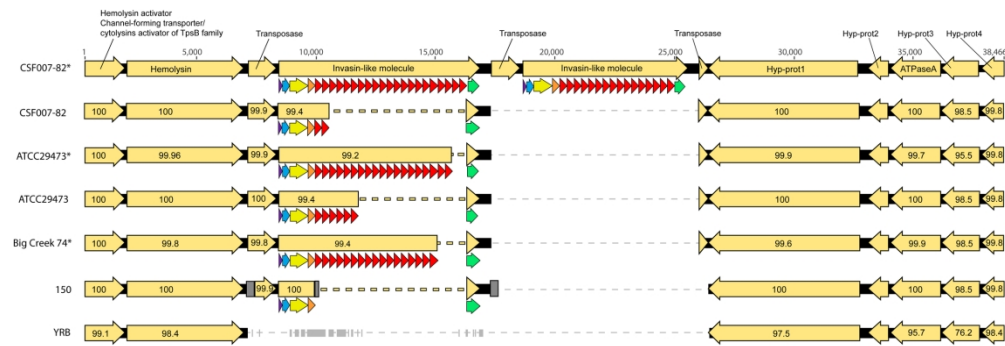


Figure 1. DNA alignment of a ~39 kbp-long DNA region containing the yrIIm gene and flanking CDS in *Y. ruckeri* genomes deposited in GenBank. Each CDS is indicated by a yellow arrow, with the percentage of sequence identity to CSF007-82 reported inside the arrow. yrIIm consists of an array of tandemly repeated, identical Ig-like domains (in red) and in addition of Ig-like domains of lower pairwise sequence similarity (in orange). It is usually capped by a C-type lectin domain (CTL, in green). The dashed lines indicate gaps in the DNA alignment. In strain 150 the grey box indicates a contig break in the assembly. The asterisk (*) indicates assemblies generated through PacBio SMRT sequencing. Note that the other assemblies have significant lower repeat numbers, suggesting that the repeats were not found using short-read sequencing technologies. Modified with permission from Wrobel et al 2018.

200x68mm (300 x 300 DPI)

Supplementary material

Number of entries of 454 sequences in SRA

In the NCBI SRA database (<https://www.ncbi.nlm.nih.gov/sra/>), a total of approximately $3 \cdot 10^5$ entries are returned with the following query:

```
"ls454"[Platform]
```

```
Total number of SRA entries is approximately  $6 \cdot 10^5$ . (query: ("pacbio smrt"[Platform] OR "bgiseq"[Platform] OR "capillary"[Platform] OR "complete genomics"[Platform] OR "helicos"[Platform] OR "illumina"[Platform] OR "ion torrent"[Platform] OR "ls454"[Platform] OR "oxford nanopore"[Platform] OR "pacbio smrt"[Platform])).
```

The vast majority are sequences derived from the Illumina platform ($>5 \cdot 10^6$ entries), with the 454 platform being the second most numerous.

Sequencing errors in proteins

Methods

To analyse the number and length distribution of short tandem repeats we used two different approaches. In the first one, we checked how the number of repeats and their lengths changed in time when the new versions of sequences were submitted into UniProtKB/Swiss-Prot [1] database. In the second one, we checked if the distribution of the number of the same type of repeats in protein families can be questionably diverse. For both purposes, we analysed sequences from the UniProtKB/Swiss-Prot database using version 2018_06.

To identify and cluster repetitive regions we used our unpublished method (Jarnot, P., Ziemaska-Legińska, J., Grynberg, M., & Gruca, A., *in preparation*). It finds strings of repeats composed of one amino acid (homorepeats) or a few amino acids (STRs). This algorithm identifies and retrieves also imperfect tandem repeats from protein sequences by scanning all sequences in a database. Imperfect repeats mean that the algorithm allows for insertions in between repeats and mutations of amino acids within the repetitive region. The minimum length of homorepeats identified by the method is 6 and the minimum number of repeats in an STRs is 3. The position of a tandem repeat and the information about the type of repeat are collected for further analysis.

The clustering phase uses repeats and their assigned ‘types’ (classification) found during identification of tandem repeats. Additionally, one type of repeat can be followed by another type of repeat in the protein sequence, defined as ‘fused repeats’. During clustering, fused repeats are also taken into account. Please note that if a protein contains three different STRs which are placed next to each other, then the method will produce 6 clusters: three with regular STRs and three with fused repeats.

In the first part of the analyses, we investigated lengths of repetitive regions between different versions of the same protein sequences (available at uniprot.org). For each sequence from the UniProtKB/Swiss-Prot we retrieved the latest (version 2018_06) and the first version. We aligned the sequences using KAlign [2] with default parameters and trimmed non-repeated parts of the sequences where the two versions differed (for example overhangs). This left common parts and STRs (Figure 1). We then retrieved STRs from these sequences to finally analyse the difference in length of these repetitive regions.

```

MESQQDEAVQTKGASTSSDAQDQGAEKGAKNKTTEATEGPTSEPPLSGPGRLLKKTAMKLF
MESQQDEAVQTKGASTSSDAQDQGAEKGAKNKTTEATEGPTSEPPLSGPGRLLKKTAMKLF

GGKKGICTLPSFFGGGRSKGSGKVSSKSLNKS KTHDGLSEASQGPEDVVEETDLSTPL
GGKKGICTLPSFFGGGRSKGSGKVSSKSLNKS KTHDGLSEASQGPEDVVEETDLSTPL

SKSSAQFPSSQSANGALEIGSKHKTSGTEAIEKAGVEKVPVSVHKPKKSLKSFFSSIRRHR
SKSSAQFPSSQSANGALEIGSKHKTSGTEAIEKAGVEKVPVSVHKPKKSLKSFFSSIRRHR

KGKTSQADQSVPGAKELEGARTRSEHEVSSISLPSSEEIFRDTRKENAKPQDAPGPKMSP
KGKTSQADQSVPGAKELEGARTRSEHEVSSISLPSSEEIFRDTRKENAKPQDAPGPKMSP

AQVHFSPTTEKAACKNPEKLTRTCASEFMQPKPVLEGGSLLEPHTSETEGKVVAGEVNPP
AQVHFSPTTEKAACKNPEKLTRTCASEFMQPKPVLEGGSLLEPHTSETEGKVVAGEVNPP

NGPVGQQLSLLFGDVTSLKSFDSLTCGDI IAEQDMSMTDSMASGGQRANRDGTRKSSC
NGPVGQQLSLLFGDVTSLKSFDSLTCGDI IAEQDMSMTDSMASGGQRANRDGTRKSSC

LVTYQGGGEEMALPDDDDNDDEEEEEEEEEKKKKKKKKKKKKKKK-----
LVTYQGGGEEMALPDDDDNDDEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEELEDEEEVKDG

-----
EENDDLEYLWASAQIYPRFNMNLGYHTAISPSHQGYMLLDPVQSYPNLGLGELLTPQSDQ

-----
QESAPNSDEGYDSTTPGFEDDSGEALGLAHRDCLPRDSYSGDALYEFYEPDSDLHSPP

-----
GDDCLYDLRGRNSEMLDFFLNLEPFSSRPPGAMETEEERLVTIQKQLLYWELRREQREAQ

```

 EACAREAHAREAYARDTHHTRESYGRNVRARETQALEAHSQEGRVQETKVRQEKPALEYQM

 RPLGPSVMGLVAGTSGGSQTSHRGTTSAFPATSSSEPDWRDFRPLEKRFEGTCSKKDQST

 CLMQLFQSDAMFEPDMQEQANFGGSPRKAYPSYSPPEEPEEEEEKEGNATVSFSQALVEF

 TSNGNLFTSMSYSSDSDSSFTQNLPELPPMVTFDIADVERDGEKCEENPEFNDEDLTA

 SLEAFELGYHKKHAFNSYHSRFYQGLPWGVSSLPRYLGLPGVHPRPPAAMALNRRSRSL

 DNAESLELELSSSHLAQGYMESDELQAHQEDSDEEGEEEEGEWGRDSPLSLYTEPPGVYD

 WPPWAHCPLPVGPGGLAWMSPNQLYEPFNQSSYVQATCCVPPVAMPVSVPGRTPGDSVSQL

 ARPShLPLPMGPCYNLQSQASQSGRAKPRDVLLPVDEPSCSSISGANSQSQAQKPVGITHG

 IPQLPRVRPEPFQLPNHYRASNLDSLKERGEQGASLSTSYSSTAMGNLAK

Supplementary Figure 1. Identification of repeats differences in sequences of APC membrane recruitment protein 1, *Mus musculus* (Mouse) (Q7TS75). The blue part is common to both sequences and this part is analysed. Red part is omitted.

The second part of our analyses focused on describing the differences in STRs length in specific protein families. For that purpose, we divided the UniProtKB/Swiss-Prot database into following taxonomies: Archaea, bacteria, fungi, invertebrates, vertebrates, plants and viruses. In the next step, we retrieved STRs from each sub-database and clustered them by type of repeats and families. Then we generated statistics for each cluster in order to find differences in lengths in repetitive regions in the same families for particular taxonomies.

Results

We found that in the UniProtKB/Swiss-Prot database 1669 (0.3%) proteins have differences in repetitive regions between the first and the last (current) submitted version of the protein sequence. These regions vary in length and quantity of repeats. The average absolute difference is 13.57 amino acids. The average length of the repetitive region in the first version of sequences is 31.14 whereas in the current version it is 35.2. The results of our analysis are summarised in Table 1.

While analysing the distribution of STRs in protein families for specific taxonomies we found out that 12.21% of invertebrate proteins contain short tandem repeats, especially PolyQ and PolyN, and many of them are characterized by a large variation in length within the same family. For instance the paralogs of probable serine/threonine-protein kinase dyrk1 (Q76NV1, Q54V83, *Dictyostelium discoideum*) are quite similar in case of high complexity regions, however PolyN repetitive regions in the first protein which is positioned in range 107-276 is almost 4 times longer and more regular, i.e., fewer insertions and mutations, than the corresponding region in the second protein (10-53).

Another example is the pair: probable basic-leucine zipper transcription factor O (Q54GH0, *Dictyostelium discoideum*) and the CCAAT/enhancer-binding (Q02638, *Drosophila virilis*) proteins. If we align both sequences using MUSCLE tool [2], it reveals that PolyQ region (28-76) is over twice longer in the first protein than in the second protein (47-69).

We discovered that another group of organisms that are also abundant in STRs are fungi. 11.89% of fungi proteins contain STRs. In contrast to invertebrates, differentiation in fungi is more visible in non-homopolymeric repeats. For instance the protein sequences of DNA-directed RNA polymerase II subunit rpb1 from *Schizosaccharomyces pombe* (P36594) and RNA polymerase II subunit rpb1 from *Encephalitozoon cuniculi* (Q8SSC4) possess recurring regions consisting of the YSPTSPSYS subsequence at the C-terminus. This region occurs in the ranges 1553-1752 and 1466-1572, respectively, therefore the STR in the *S. pombe* sequence is almost twice as long as in its *E. cuniculi* counterpart. Significant difference in length can also be observed in proteins described as Mediator of RNA polymerase II transcription subunit 15 (Q75BI6 and Q9Y808) from *Ashbya gossypii* and *S. pombe*, which have glutamine homorepeats at ranges 282-365 and 256-289, respectively. Therefore, this first STR is 252% longer than the same region in the second protein.

Short tandem repeats in vertebrates are even more complex than in fungi, even if only about 8% of proteins contain STRs. Histone-lysine N-methyltransferase 2D (Q6PDK2, *Mus musculus*) which was added to UniProt database November 30, 2010 contains significantly more homorepeats of glutamine than histone-lysine N-methyltransferase 2C (Q8BRH4, *Mus musculus*) which was added to the database in October 10, 2003. Overall, there seems to be more STRs in more recent SwissProt additions.

Length variation of STRs in Archaea is very low. That is because proteins of these organisms are rarely composed of STRs. About 1.8% of proteins contain STRs. STRs in Archea proteins are mostly composed of A, E, Q, G, K amino acids.

Summary

In this research, we have shown that with new methods of sequencing, the number of repetitive regions in proteins changed significantly as well as the length of these regions.

Proteins in the same families share similar biological function. It has been shown that repetitive regions can have crucial functions in proteins [3]. These functions are related to the length of repetitive regions, therefore if a specific repetitive region has an important function in a protein, then the length of this repetitive region should not vary a lot within the protein family. Here we have shown some cases where the length of repetitive regions in the same family varies significantly.

Conclusion

By analysis of the different versions of the same protein sequences submitted to UniProtKB/Swiss-Prot database, we have shown that with the improvement of sequencing methods numbers of repeats and their lengths may change significantly. Additionally, we analysed the differences between the distribution of STRs length in specific protein families for particular taxonomies. Our results show that repetitive regions in the same taxon and family may vary significantly. These statements lead us straight to hypothesise that there are still many repetitive regions in UniProtKB/Swiss-Prot database which are erroneously sequenced.

References

1. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45: D158-D169 (2017)
2. Timo Lassmann and Erik L.L. Sonnhammer (2005). Kalign - an accurate and fast multiple sequence alignment algorithm. *BMC Bioinformatics* 6:298
3. Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5):1792-1797.
4. Hamm D.C., Bondra E.R., Harrison M.M. Transcriptional activation is a conserved feature of the early embryonic factor zelda that requires a cluster of four zinc fingers for DNA binding and a low-complexity activation domain. *The Journal of Biological Chemistry.* 2015;290(6):3508-3518.