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.

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

Ole Kristian Tørresen[1], Bastiaan Star[1], Pablo Mier[2], Miguel A. Andrade-Navarro[2], Alex Bateman[3], Patryk Jarnot[4], Aleksandra Gruca[4], Marcin Grynberg[5], Andrey V. Kajava[6,7], Vasilis J. Promponas[8], Maria Anisimova[9,10], Kjetill S Jakobsen[1], Dirk Linke[11,*]

[1]Centre for Ecological and Evolutionary Synthesis, Department of Biosciences, University of Oslo, NO-0316 Oslo, Norway
[2]Faculty of Biology, Johannes Gutenberg University Mainz, Hans-Dieter-Husch-Weg 15, 55128 Mainz, Germany
[3]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton. CB10 1SD, UK
[4]Institute of Informatics, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[5]Institute of Biochemistry and Biophysics PAS, Pawińskiego 5A, 02-106 Warsaw, Poland
[6]Centre de Recherche en Biologie cellulaire de Montpellier, UMR 5237 CNRS, Universite Montpellier 1919 Route de Mende, CEDEX 5, 34293, Montpellier, France
[7]Institut de Biologie Computationnelle, 34095, Montpellier, France
[8]Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, PO Box 20537, CY 1678, Nicosia, Cyprus
[9]Institute of Applied Simulations, School of Life Sciences and Facility Management, Zurich University of Applied Sciences (ZHAW), Wädenswil, Switzerland
[10]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[11]Section 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

creation that the data they omit or improperly assemble may well contain important biological information valuable to others.

GLOSSARY

aDNA: Ancient DNA. DNA isolated from material that are up to several hundred thousand years old.
Contigs: Sequence assembled from shorter sequencing reads into a contiguous stretch of nucleotides.
de Bruijn graph: One of two main computational approaches (the other is **OLC**) for the assembly of sequencing reads into longer sequences such as contigs. Works by dividing reads into overlapping *k*-mers. A graph is created with nodes corresponding to *k*-mers and directional edges connecting overlapping nodes. A traversal of the graph can be output as contigs.
GenBank: One of several databases containing all publicly available DNA sequences.
Homorepeat: Also known as **homopolymer tract,** or **polyX** for amino acids, where X is the repeated residue. A perfect **tandem repeat** with unit size one where all the nucleotides or amino acids are the same.
Interspersed repeat: A motif or pattern that is found in multiple loci across a genome, such as **transposable elements**. In contrast, a **tandem repeat** has the motif or pattern repeated in tandem at one locus.
***K*-mer:** A sequence of nucleotides that is *k*-residues long, such as a 31-mer with 31 nucleotides.
LRR: Leucine rich repeats are amino acid motifs found in many different proteins, often repeated in tandem.
NLR: Nod-like receptors are proteins involved in innate immune response and contains LRRs among other domains.
OLC: Overlap-layout-consensus. One of two main computational approaches (the other is **de Bruijn graph**) for the assembly of sequencing reads into longer sequences such as contigs. Works by finding common sequences in reads (overlaps), and creates a graph where the overlaps are nodes. Traversal of the graph can be output as contigs.
Polishing: The act of mapping reads back to an assembly and recalling the consensus sequence. This is a necessity for assemblies based on PacBio and/or Oxford Nanopore reads, and are often performed in multiple rounds where at least the last couple are done with Illumina reads.
Scaffolds: Contains multiple contigs that are placed into proper order and orientation based on paired reads or other positional information (linked reads, optical maps, linkage maps).
Short tandem repeat (STR): A **tandem repeat** with a unit size shorter than 10 nucleotides.
Sequence Read Archive (SRA): A database of sequencing data and alignment information from high-throughput sequencing platforms such as Illumina, 454 and PacBio among others.
Tandem repeat (TR): A region of DNA or protein where a motif or pattern is repeated in tandem at one locus. The motif or pattern has a size, which is usually called a unit size. For example, the tandem repeat ACACACAC has a unit size of 2. This is in contrast to an **interspersed repeat** where the motif or pattern is found in multiple loci across a genome.

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

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

VLRs: Variable lymphocyte receptors: immune genes found in jawless vertebrates, also containing LRRs.

INTRODUCTION

The availability of DNA and protein sequence data has revolutionized the way we study cellular, molecular, physiological, evolutionary and developmental processes, allowing the association of phenotypes with genotypes at a single nucleotide (or single amino acid) resolution. Researchers rely on public sequence depositories and other databases for sharing their data, such as GenBank or UniProt, and the content of these databases has grown exponentially in the last decades. While such databases initially consisted predominantly of submissions of individual gene or protein sequences that were carefully curated, large proportions of the content of genome and protein databases today originate from different types of metagenome and genome sequencing and assembly projects. GenBank, for example, included more than 2635 Gbp (billion base pairs) in its 2017 release number 221, of which 2242 Gbp (85%) originated from whole-genome shotgun sequencing (1). For an informed use of such data, it is essential that end users understand the distinct contrast in quality between individual, well-curated submissions and entries generated from automated sequence annotation pipelines. The latter procedures can contain unrecognized errors.

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

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

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

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

TEs can cause “breakage” of a continuous assembly and lead to assembly collapse, where the number of copies of a repeat found in a genome assembly is lower than the true number, but the relatively large and often evolutionary divergent TEs are unlikely to greatly affect the accuracy of sequencing, assembly and annotation of individual protein-coding regions. While such TEs might sometimes insert themselves into gene regions, the disruptive effects of multiple kbps of sequence inserted into coding regions likely make these events extremely rare. In contrast, TRs are usually much shorter, and can often be in frame in coding regions; therefore, we mainly focus on the problems caused by this class of repeats on the sequencing, assembly, annotation and database deposition processes.

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

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

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

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

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

SEQUENCING AND GENOME ASSEMBLY ARE AFFECTED BY TANDEM REPEATS

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

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

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

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

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

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

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

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

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

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

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

EXAMPLES OF REPEAT-DRIVEN ERROR PROLIFERATION

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

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

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

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

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

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

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

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

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

ANNOTATION OF FUNCTION CAN BE AFFECTED BY TANDEM REPEATS

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

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

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

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

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

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

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

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

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

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

CONCLUSIONS

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

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 Volff,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.

35. Kajava,A.V. (2012) Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology*, **179**, 279–288.

36. Paladin,L., Hirsh,L., Piovesan,D., Andrade-Navarro,M.A., Kajava,A.V. and Tosatto,S.C.E. (2017) RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Res*, **45**, D308–D312.

37. Schaper,E. and Anisimova,M. (2015) The evolution and function of protein tandem repeats in plants. **206**, 397–410.

38. Kajava,A.V., Anisimova,M. and Peeters,N. (2008) Origin and Evolution of GALA-LRR, a New Member of the CC-LRR Subfamily: From Plants to Bacteria? *PLoS ONE*, **3**, e1694.

39. Szalkowski,A.M. and Anisimova,M. (2013) Graph-based modeling of tandem repeats improves global multiple sequence alignment. *Nucleic Acids Res*, **41**, e162–e162.

40. Verstrepen,K.J., Jansen,A., Lewitter,F. and Fink,G.R. (2005) Intragenic tandem repeats generate functional variability. *Nat Genet*, **37**, 986–990.

41. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics*, **22**, 253–259.

42. Sutherland,G.R. and Richards,R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc Natl Acad Sci USA*, **92**, 3636–3641.

43. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R., *et al.* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

44. Glenn,T.C. (2011) Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, **11**, 759–769.

45. Eid,J., Fehr,A., Gray,J., Luong,K., Lyle,J., Otto,G., Peluso,P., Rank,D., Baybayan,P., Bettman,B., *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

46. Olasagasti,F., Lieberman,K.R., Benner,S., Cherf,G.M., Dahl,J.M., Deamer,D.W. and Akeson,M. (2010) Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat Nanotechnol*, **5**, 798–806.

47. Rhoads,A. and Au,K.F. (2015) PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics*, **13**, 278–289.

48. Weirather,J.L., de Cesare,M., Wang,Y., Piazza,P., Sebastiano,V., Wang,X.-J., Buck,D. and Au,K.F. (2017) Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, **6**, 100.

49. Balzer,S., Malde,K., Lanzén,A., Sharma,A. and Jonassen,I. (2010) Characteristics of 454 pyrosequencing data--enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–5.

50. Bragg,L.M., Stone,G., Butler,M.K., Hugenholtz,P. and Tyson,G.W. (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. *PLoS Comp Biol*, **9**, e1003031.

51. Luo,C., Tsementzi,D., Kyrpides,N., Read,T. and Konstantinidis,K.T. (2012) Direct Comparisons of Illumina vs. Roche 454 Sequencing Technologies on the Same Microbial Community DNA Sample. *PLoS ONE*, **7**, e30087.

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.

53. Gnerre,S., Maccallum,I., Przybylski,D., Ribeiro,F.J., Burton,J.N., Walker,B.J., Sharpe,T., Hall,G., Shea,T.P., Sykes,S., *et al.* (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA*, **108**, 1513–1518.

54. Bankevich,A., Nurk,S., Antipov,D., Gurevich,A.A., Dvorkin,M., Kulikov,A.S., Lesin,V.M., Nikolenko,S.I., Pham,S., Prjibelski,A.D., *et al.* (2012) SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *https://home.liebertpub.com/cmb*, **19**, 455–477.

55. Souvorov,A., Agarwala,R. and Lipman,D.J. (2018) SKESA: strategic k-mer extension for scrupulous assemblies. *Genome Biol*, **19**, 1–13.

56. Myers,E.W., Sutton,G.G., Delcher,A.L., Dew,I.M., Fasulo,D.P., Flanigan,M.J., Kravitz,S.A., Mobarry,C.M., Reinert,K.H., Remington,K.A., *et al.* (2000) A whole-genome assembly of Drosophila. *Science*, **287**, 2196–2204.

57. Miller,J.R., Koren,S. and Sutton,G.G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics*, **95**, 315–327.

58. Treangen,T.J. and Salzberg,S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Rev Genet*, **13**, 36–46.

59. Sotero-Caio,C.G., Platt,R.N., Suh,A. and Ray,D.A. (2017) Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol Evol*, **9**, 161–177.

60. Elliott,T.A. and Gregory,T.R. (2015) What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. *Philos Trans R Soc Lond, B, Biol Sci*, **370**, 20140331.

61. Liljegren,M.M., de Muinck,E.J. and Trosvik,P. (2016) Microsatellite Length Scoring by Single Molecule Real Time Sequencing – Effects of Sequence Structure and PCR Regime. **11**, e0159232.

62. Tørresen,O.K., Star,B., Jentoft,S., Reinar,W.B., Grove,H., Miller,J.R., Walenz,B.P., Knight,J., Ekholm,J.M., Peluso,P., *et al.* (2017) An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. *BMC Genomics*, **18**, 95.

63. Adams,R.H., Blackmon,H., Reyes-Velasco,J., Schield,D.R., Card,D.C., Andrew,A.L., Waynewood,N. and Castoe,T.A. (2016) Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome*, **59**, 295–310.

64. Jiang,Q., Li,Q., Yu,H. and Kong,L. (2014) Genome-wide analysis of simple sequence repeats in marine animals—a comparative approach. **16**, 604–619.

65. Star,B., Hansen,M.H., Skage,M., Bradbury,I.R., Godiksen,J.A., Kjesbu,O.S. and Jentoft,S. (2016) Preferential amplification of repetitive DNA during whole genome sequencing library creation from historic samples. *Sci Technol Archaeol Res*, **2**, 36–45.

66. Star,B., Nederbragt,A.J., Jentoft,S., Grimholt,U., Malmstrøm,M., Gregers,T.F., Rounge,T.B., Paulsen,J., Solbakken,M.H., Sharma,A., *et al.* (2011) The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, **477**, 207–210.

67. Zhuang,X., Yang,C., Fevolden,S.-E. and Cheng,C.-H. (2012) Protein genes in repetitive sequence—antifreeze glycoproteins in Atlantic cod genome. *BMC Genomics*, **13**, 293.

68. Belser,C., Istace,B., Denis,E., Dubarry,M., Baurens,F.-C., Falentin,C., Genete,M., Berrabah,W., Chèvre,A.-M., Delourme,R., *et al.* (2018) Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nature Plants 2018 4:2*, **4**, 879–887.

69. Hommelsheim,C.M., Frantzeskakis,L., Huang,M. and Ülker,B. (2014) PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. *Sci Rep*, **4**, 5052.

70. Hurles,M. (2004) Gene Duplication: The Genomic Trade in Spare Parts. *PLoS Biol*, **2**, e206.

71. Hardison,R.C. (2012) Evolution of Hemoglobin and Its Genes. *Cold Spring Harbor Perspectives in Medicine*, **2**, a011627–a011627.

72. Opazo,J.C., Butts,G.T., Nery,M.F., Storz,J.F. and Hoffmann,F.G. (2013) Whole-genome duplication and the functional diversification of teleost fish hemoglobins. *Mol Biol Evol*, **30**, 140–153.

73. Baalsrud,H.T., Voje,K.L., Tørresen,O.K., Solbakken,M.H., Matschiner,M., Malmstrøm,M., Hanel,R., Salzburger,W., Jakobsen,K.S. and Jentoft,S. (2017) Evolution of Hemoglobin Genes in Codfishes Influenced by Ocean Depth. *Sci Rep*, **7**, 7956.

74. Tørresen,O.K., Brieuc,M.S.O., Solbakken,M.H., Sørhus,E., Nederbragt,A.J., Jakobsen,K.S., Meier,S., Edvardsen,R.B. and Jentoft,S. (2018) Genomic architecture of haddock (*Melanogrammus aeglefinus*) shows expansions of innate immune genes and short tandem repeats. *BMC Genomics*, **19**, 240.

75. Stein,C., Caccamo,M., Laird,G. and Leptin,M. (2007) Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. *Genome Biol*, **8**, R251.

76. Howe,K., Schiffer,P.H., Zielinski,J., Wiehe,T., Laird,G.K., Marioni,J.C., Soylemez,O., Kondrashov,F. and Leptin,M. (2016) Structure and evolutionary history of a large family of NLR proteins in the zebrafish. *Open Biol*, **6**, 160009–224.

77. Matsushima,N., Takatsuka,S., Miyashita,H. and Kretsinger,R.H. (2019) Leucine Rich Repeat Proteins: Sequences, Mutations, Structures and Diseases. *PPL*, **26**, 108–131.

78. Boehm,T., McCurley,N., Sutoh,Y., Schorpp,M., Kasahara,M. and Cooper,M.D. (2012) VLR-Based Adaptive Immunity. *https://doi.org/10.1146/annurev-immunol-020711-075038*, **30**, 203–220.

79. Das,S., Hirano,M., Aghaallaei,N., Bajoghli,B., Boehm,T. and Cooper,M.D. (2013) Organization of lamprey variable lymphocyte receptor C locus and repertoire development. *Proc Natl Acad Sci USA*, **110**, 6043–6048.

80. Smith,J.J., Timoshevskaya,N., Ye,C., Holt,C., Keinath,M.C., Parker,H.J., Cook,M.E., Hess,J.E., Narum,S.R., Lamanna,F., *et al.* (2018) The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat Genet*, **50**, 270–277.

81. Wrobel,A., Ottoni,C., Leo,J.C., Gulla,S. and Linke,D. (2018) The repeat structure of two paralogous genes, *Yersinia ruckeri* invasin (*yrInv*) and a '*Y. ruckeri* invasin-like molecule', (*yrIlm*) sheds light on the evolution of adhesive capacities of a fish pathogen. *Journal of Structural Biology*, **201**, 171–183.

82. Franzen,O., Jerlström-Hultqvist,J., Castro,E., Sherwood,E., Ankarklev,J., Reiner,D.S., Palm,D., Andersson,J.O., Andersson,B. and Svärd,S.G. (2009) Draft Genome Sequencing of Giardia intestinalis Assemblage B Isolate GS: Is Human Giardiasis Caused by Two Different Species? *PLOS Pathogens*, **5**, e1000560.

83. Khatri,I., Tomar,R., Ganesan,K., Prasad,G.S. and Subramanian,S. (2017) Complete genome sequence and comparative genomics of the probiotic yeast *Saccharomyces boulardii*. *Sci Rep*, **7**, 371.

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.

85. Schmid,M., Frei,D., Patrignani,A., Schlapbach,R., Frey,J.E., Remus-Emsermann,M.N.P. and Ahrens,C.H. (2018) Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *Nucleic Acids Res*, **46**, 8953–8965.

86. Guo,S., Stevens,C.A., Vance,T.D.R., Olijve,L.L.C., Graham,L.A., Campbell,R.L., Yazdi,S.R., Escobedo,C., Bar-Dolev,M., Yashunsky,V., *et al.* (2017) Structure of a 1.5-MDa adhesin that binds its Antarctic bacterium to diatoms and ice. *Sci Adv*, **3**, e1701440.

87. Guo,S., Garnham,C.P., Whitney,J.C., Graham,L.A. and Davies,P.L. (2012) Re-evaluation of a bacterial antifreeze protein as an adhesin with ice-binding activity. *PLoS ONE*, **7**, e48805.

88. Stålhammar-Carlemalm,M., Areschoug,T., Larsson,C. and Lindahl,G. (1999) The R28 protein of Streptococcus pyogenes is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells. *Mol Microbiol*, **33**, 208–219.

89. Roche,F.M., Massey,R., Peacock,S.J., Day,N.P.J., Visai,L., Pietro Speziale, Lam,A., Pallen,M. and Foster,T.J. (2003) Characterization of novel LPXTG-containing proteins of Staphylococcus aureus identified from genome sequences. *Microbiology*, **149**, 643–654.

90. Anisimova,M., Pečerska,J. and Schaper,E. (2015) Statistical Approaches to Detecting and Analyzing Tandem Repeats in Genomic Sequences. **3**.

91. Schaper,E., Korsunsky,A., Messina,A., Murri,R., Pečerska,J., Stockinger,H., Zoller,S., Xenarios,I. and Anisimova,M. (2015) TRAL: Tandem repeat annotation library. *Bioinformatics*, 10.1093/bioinformatics/btv306.

92. Conesa,A., Madrigal,P., Tarazona,S., Gomez-Cabrero,D., Cervera,A., McPherson,A., Szcześniak,M.W., Gaffney,D.J., Elo,L.L., Zhang,X., *et al.* (2016) A survey of best practices for RNA-seq data analysis. *Genome Biol*, **17**, 13.

93. Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Rev Genet*, **13**, 329–342.

94. Hoff,K.J. and Stanke,M. (2015) Current methods for automated annotation of protein-coding genes. *Current Opinion in Insect Science*, **7**, 8–14.

95. Bergman,C.M. and Quesneville,H. (2007) Discovering and detecting transposable elements in genome sequences. *Briefings in Bioinformatics*, **8**, 382–392.

96. Stanke,M., Diekhans,M., Baertsch,R. and Haussler,D. (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, **24**, 637–644.

97. Lomsadze,A., Ter-Hovhannisyan,V., Chernoff,Y.O. and Borodovsky,M. (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res*, **33**, 6494–6506.

98. Grabherr,M.G., Haas,B.J., Yassour,M., Levin,J.Z., Thompson,D.A., Amit,I., Adiconis,X., Fan,L., Raychowdhury,R., Zeng,Q., *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, **29**, 644–652.

99. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc*, **11**, 1650–1667.

100. Gonzalez-Garay,M.L. (2016) Introduction to Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq). In *Transcriptomics and Gene Regulation*, Translational Bioinformatics. Springer, Dordrecht, Dordrecht, Vol. 9, pp. 141–160.

101. UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res*, **43**, D204–12.

102. Campbell,M.S., Law,M., Holt,C., Stein,J.C., Moghe,G.D., Hufnagel,D.E., Lei,J., Achawanantakun,R., Jiao,D., Lawrence,C.J., *et al.* (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol*, **164**, 513–524.

103. Holt,C. and Yandell,M. (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, **12**, 491.

104. Haas,B.J., Salzberg,S.L., Zhu,W., Pertea,M., Allen,J.E., Orvis,J., White,O., Buell,C.R. and Wortman,J.R. (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*, **9**, R7.

105. Jones,P., Binns,D., Chang,H.-Y., Fraser,M., Li,W., McAnulla,C., McWilliam,H., Maslen,J., Mitchell,A., Nuka,G., *et al.* (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics*, **30**, 1236–1240.

106. Mier,P., Paladin,L., Tamana,S., Petrosian,S., Hajdu-Soltész,B., Urbanek,A., Gruca,A., Plewczynski,D., Grynberg,M., Bernadó,P., *et al.* (2019) Disentangling the complexity of low complexity proteins. *Briefings in Bioinformatics*, **27**, 331.

107. Chen,L., DeVries,A.L. and Cheng,C.-H.C. (1997) Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci USA*, **94**, 3811–3816.

108. Chen,L., DeVries,A.L. and Cheng,C.-H.C. (1997) Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc Natl Acad Sci USA*, **94**, 3817–3822.

109. Baalsrud,H.T., Tørresen,O.K., Hongrø Solbakken,M., Salzburger,W., Hanel,R., Jakobsen,K.S. and Jentoft,S. (2017) De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol Biol Evol*, **35**, 593–606.

110. Zakin,M.M., Duchange,N., Ferrara,P. and Cohen,G.N. (1983) Nucleotide sequence of the metL gene of Escherichia coli. Its product, the bifunctional aspartokinase ii-homoserine dehydrogenase II, and the bifunctional product of the thrA gene, aspartokinase I-homoserine dehydrogenase I, derive from a common ancestor. *J Biol Chem*, **258**, 3028–3031.

111. Ferone,R. and Roland,S. (1980) Dihydrofolate reductase: thymidylate synthase, a bifunctional polypeptide from Crithidia fasciculata. *Proc Natl Acad Sci USA*, **77**, 5802–5806.

112. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

113. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

114. Zhao,X., Oh,S.-H., Coleman,D.A. and Hoyer,L.L. (2011) ALS51, a newly discovered gene in the Candida albicans ALS family, created by intergenic recombination: analysis of the gene and protein, and implications for evolution of microbial gene families. *FEMS Immunol Med Microbiol*, **61**, 245–257.

115. Nagy,A., Szláma,G., Szarka,E., Trexler,M., Bányai,L. and Patthy,L. (2011) Reassessing domain architecture evolution of metazoan proteins: major impact of gene prediction errors. *Genes (Basel)*, **2**, 449–501.

116. Promponas,V.J., Iliopoulos,I. and Ouzounis,C.A. (2015) Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure. *Standards in Genomic Sciences 2015 10:1*, **10**, 108.

117. Jurka,J., Walichiewicz,J. and Milosavljevic,A. (1992) Prototypic sequences for human repetitive DNA. *J Mol Evol*, **35**, 286–291.

118. Jurka,J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet*, **16**, 418–420.

119. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res*, **12**, 996–1006.

120. Ruitberg,C.M., Reeder,D.J. and Butler,J.M. (2001) STRBase: a short tandem repeat DNA database for the human identity testing community. *Nucleic Acids Res*, **29**, 320–322.

121. Gelfand,Y., Rodriguez,A. and Benson,G. (2007) TRDB--the Tandem Repeats Database. *Nucleic Acids Res*, **35**, D80–7.

122. Hussing,C., Bytyci,R., Huber,C., Morling,N. and Børsting,C. (2018) The Danish STR sequence database: duplicate typing of 363 Danes with the ForenSeq™ DNA Signature Prep Kit. *Int. J. Legal Med.*, **18**, 100.

123. Adnan,A., Zhan,X., Kasim,K., Rakha,A. and Xin,X.J. (2018) Population data and phylogenetic structure of Han population from Jiangsu province of China on GlobalFiler STR loci. *Int. J. Legal Med.*, **132**, 1301–1304.

124. Ossowski,A., Piatek,J., Parafiniuk,M., Pudlo,A., Pepinski,W., Skawronska,M., Szeremeta,M., Niemcunowicz-Janica,A. and Soltyszewski,I. (2017) Genetic variation of 15 autosomal STRs in a population sample of Bedouins residing in the area of the Fourth Nile Cataract, Sudan. *Anthropol Anz*, **74**, 263–268.

125. Kim,E.H., Lee,H.Y., Kwon,S.Y., Lee,E.Y., Yang,W.I. and Shin,K.-J. (2017) Sequence-based diversity of 23 autosomal STR loci in Koreans investigated using an in-house massively parallel sequencing panel. *Forensic Science International: Genetics*, **30**, 134–140.

126. Pamjav,H., Fóthi,Á., Fehér,T. and Fóthi,E. (2017) A study of the Bodrogköz population in north-eastern Hungary by Y chromosomal haplotypes and haplogroups. *Mol Genet Genomics*, **292**, 883–894.

127. Wang,X., Yang,S., Chen,Y., Zhang,S., Zhao,Q., Li,M., Gao,Y., Yang,L. and Bennetzen,J.L. (2018) Comparative genome-wide characterization leading to simple sequence repeat marker development for Nicotiana. *BMC Genomics*, **19**, 500.

128. Franco,M.E., Bitencourt,T.A., Marins,M. and Fachin,A.L. (2017) In silico characterization of tandem repeats in Trichophyton rubrum and related dermatophytes provides new insights into their role in pathogenesis. *Database (Oxford)*, **2017**, 1.

129. Houston,R., Birck,M., LaRue,B., Hughes-Stamm,S. and Gangitano,D. (2018) Nuclear, chloroplast, and mitochondrial data of a US cannabis DNA database. *Int. J. Legal Med.*, **132**, 713–725.

130. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*, **27**, 573–580.

131. Teeling,E.C., Vernes,S.C., Dávalos,L.M., Ray,D.A., Gilbert,M.T.P., Myers,E.Bat1K Consortium (2018) Bat Biology, Genomes, and the Bat1K Project: To Generate Chromosome-Level Genomes for All Living Bat Species. *Annu. Rev. Anim. Biosci.*, **6**, 23–46.

132. Lewin,H.A., Robinson,G.E., Kress,W.J., Baker,W.J., Coddington,J., Crandall,K.A., Durbin,R., Edwards,S.V., Forest,F., Gilbert,M.T.P., *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc Natl Acad Sci USA*, **115**, 201720115–4333.

133. Koren,S., Phillippy,A.M., Simpson,J.T., Loman,N.J. and Loose,M. (2019) Reply to 'Errors in long-read assemblies can critically affect protein prediction'. *Nat Biotechnol*, **30**, 1.

134. Watson,M. and Warr,A. (2019) Errors in long-read assemblies can critically affect protein prediction. *Nat Biotechnol*, **37**, 124.

135. Weissensteiner,M.H., Pang,A.W.C., Bunikis,I., Höijer,I., Vinnere-Pettersson,O., Suh,A. and Wolf,J.B.W. (2017) Combination of short-read, long-read and optical mapping assemblies reveals large-scale tandem repeat arrays with population genetic implications. *Genome Res*, **27**, gr.215095.116–708.

136. Wenger,A.M., Peluso,P., Rowell,W.J., Chang,P.-C., Hall,R.J., Concepcion,G.T., Ebler,J., Fungtammasan,A., Kolesnikov,A., Olson,N.D., *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, **74**, 1–8.

Table 1. Summary of proteins from UniProtKB/Swiss-Prot where the length of repetitive region has 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

[1] measured in amino acid residues

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

[1]repetitive region length, measured in amino acid residues

[2]Clustering 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 (CTLD, 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.

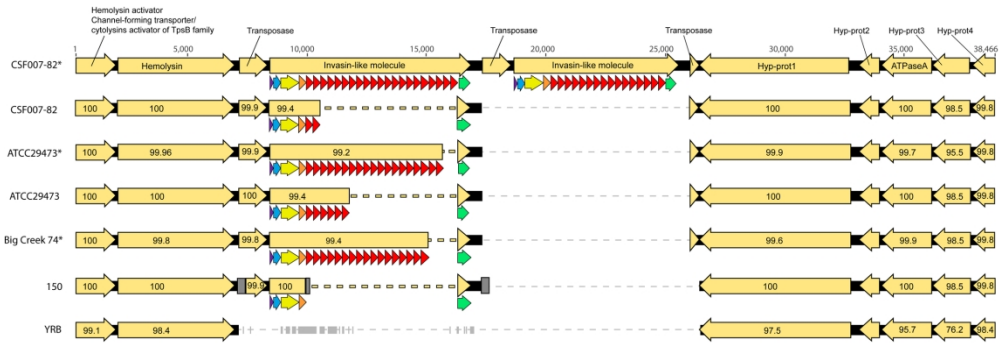

Figure 1. DNA alignment of a ~39 kbp-long DNA region containing the yrIlm gene and flanking CDS in Y. ruckeri genomes deposited in GenBank. Each CDS is indicated by a yellow arrow, with the percentage of sequence identity to CSF007-82 reported inside the arrow. yrIlm consists of an array of tandemly 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 (CTLD, 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 $3*10^5$ entries are returned with the following query:

"ls454"[Platform]

Total number of SRA entries is approximately $6*10^5$. (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 (>$5*10^6$ 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., Ziemska-Legięcka, 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.

```
MESQQDEAVQTKGASTSSDAQDQGAEKGAKNKTTEATEGPTSEPPLSGPGRLKKTAMKLF
MESQQDEAVQTKGASTSSDAQDQGAEKGAKNKTTEATEGPTSEPPLSGPGRLKKTAMKLF

GGKKGICTLPSFFGGGRSKGSGKVSSKKSLNKSKTHDGLSEASQGPEDVVIEETDLSTPL
GGKKGICTLPSFFGGGRSKGSGKVSSKKSLNKSKTHDGLSEASQGPEDVVIEETDLSTPL

SKSSAQFPSSQSANGALEIGSKHKTSGTEAIEKAGVEKVPSVHKPKKSLKSFFSSIRRHR
SKSSAQFPSSQSANGALEIGSKHKTSGTEAIEKAGVEKVPSVHKPKKSLKSFFSSIRRHR

KGKTSGADQSVPGAKELEGARTRSHEHVSSISLPSSEEIFRDTRKENAKPQDAPGPKMSP
KGKTSGADQSVPGAKELEGARTRSHEHVSSISLPSSEEIFRDTRKENAKPQDAPGPKMSP

AQVHFSPTTEKAACKNPEKLTRTCASEFMQPKPVLEGGSLEEPHTSETEGKVVAGEVNPP
AQVHFSPTTEKAACKNPEKLTRTCASEFMQPKPVLEGGSLEEPHTSETEGKVVAGEVNPP

NGPVGDQLSLLFGDVTSLKSFDSLTGCGDIIAEQDMDSMTDSMASGGQRANRDGTKRSSC
NGPVGDQLSLLFGDVTSLKSFDSLTGCGDIIAEQDMDSMTDSMASGGQRANRDGTKRSSC

LVTYQGGGEEMALPDDDDNDDEEEEEEEEKKKKKKKKKKKKKKK----------------
LVTYQGGGEEMALPDDDDNDDEEEEEEEEEEEEEEEEEEEEEEEEEELLEDEEEVKDG

------------------------------------------------------------
EENDDLEYLWASAQIYPRFNMNLGYHTAISPSHQGYMLLDPVQSYPNLGLGELLTPQSDQ

------------------------------------------------------------
QESAPNSDEGYYDSTTPGFEDDSGEALGLAHRDCLPRDSYSGDALYEFYEPDDSLEHSPP

------------------------------------------------------------
GDDCLYDLRGRNSEMLDPFLNLEPFSSRPPGAMETEEERLVTIQKQLLYWELRREQREAQ
```

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-hompolymeric 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.