Computational Biology and Chemistry 61 (2016) 258–269

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

From NGS assembly challenges to instability of fungal mitochondrial genomes: A case study in genome complexity

Elizabeth Misas[a,b], José Fernando Muñoz[a,b], Juan Esteban Gallo[a,c], Juan Guillermo McEwen[a,d], Oliver Keatinge Clay[a,e,*]

[a] Cellular & Molecular Biology Unit, Corporación para Investigaciones Biológicas, Medellín, Colombia
[b] Institute of Biology, Universidad de Antioquia, Medellín, Colombia
[c] Doctoral Program in Biomedical Sciences, Universidad del Rosario, Bogotá, Colombia
[d] School of Medicine, Universidad de Antioquia, Medellín, Colombia
[e] School of Medicine and Health Sciences, Universidad del Rosario, Bogotá, Colombia

ARTICLE INFO

Article history:
Received 31 July 2015
Received in revised form 3 February 2016
Accepted 16 February 2016
Available online 21 February 2016

Keywords:
Next generation sequencing
Repetitive DNA
Fungal mitochondria
Genome assembly
DNA secondary structure

ABSTRACT

The presence of repetitive or non-unique DNA persisting over sizable regions of a eukaryotic genome can hinder the genome's successful *de novo* assembly from short reads: ambiguities in assigning genome locations to the non-unique subsequences can result in premature termination of contigs and thus over-fragmented assemblies. Fungal mitochondrial (mtDNA) genomes are compact (typically less than 100 kb), yet often contain short non-unique sequences that can be shown to impede their successful *de novo* assembly *in silico*. Such repeats can also confuse processes in the cell *in vivo*. A well-studied example is ectopic (out-of-register, illegitimate) recombination associated with repeat pairs, which can lead to deletion of functionally important genes that are located between the repeats. Repeats that remain conserved over micro- or macroevolutionary timescales despite such risks may indicate functionally or structurally (e.g., for replication) important regions. This principle could form the basis of a mining strategy for accelerating discovery of function in genome sequences. We present here our screening of a sample of 11 fully sequenced fungal mitochondrial genomes by observing where exact *k*-mer repeats occurred several times; initial analyses motivated us to focus on 17-mers occurring more than three times. Based on the diverse repeats we observe, we propose that such screening may serve as an efficient expedient for gaining a rapid but representative first insight into the repeat landscapes of sparsely characterized mitochondrial chromosomes. Our matching of the flagged repeats to previously reported regions of interest supports the idea that systems of persisting, non-trivial repeats in genomes can often highlight features meriting further attention.

1. Introduction

Repeats along a chromosome of a eukaryotic genome can severely inhibit the success of *de novo* or reference assemblies of the chromosome's sequence from short reads *in silico*, i.e., the stability of the assembly process. This problem is especially pronounced in next-generation sequencing (NGS) pipelines when the short (currently ≤300 bp) reads obtained are then piped through a de Bruijn graph-based *de novo* assembly program (Compeau et al., 2011). Such programs consider in a first instance the NGS reads' subsequences of a fixed length *k* (*k*-mers, words of length *k*), and use them to build up and analyze a quasi-flow on a de Bruijn or de Bruijn-like graph. Concurrently or in a second instance, such programs may then also integrate the information of the full read sequences *per se* and/or, where the reads are paired-end, the pairing information.

A good example of a genome or chromosome that illustrates the repeat assembly problem is a familiar and long-studied mitochondrial genome of a unicellular fungus, the S288C strain of baker's yeast, *Saccharomyces cerevisiae*. In a previous study (Muñoz et al., 2014), we fed subsequences of this completely sequenced mitochondrial genome, i.e., ideal, short, paired *in silico*-generated subsequences of read length 100 bp, without any sequencing errors and without any competing nuclear genomic DNA, to a widely used NGS assembly program. We confirmed what is perhaps already obvious on fundamental grounds: over a large range of *k* value

Abbreviations: NGS, next generation sequencing; mt, mitochondrial; mtDNA, mitochondrial DNA; nt, nucleotide; bp, base pair; kb, kilobase pair; Mb, megabase pair; *ori*, origin of replication; *orf*, open reading frame.

* Corresponding author.
E-mail address: oliver.clay@gmail.com (O.K. Clay).

http://dx.doi.org/10.1016/j.compbiolchem.2016.02.016
1476-9271/

choices ranging up to 63 bp, repeated sequences from the eight *ori* (origin of replication) regions consistently prevented full assembly: the *ori*'s were precisely the positions where the contigs or scaffolds could no longer be extended. In addition to fundamental (i.e., assembler-independent) limits that repeats can impose on the assembly process, there are also assembler-specific problems or complexities that can cause unexpected profiles of the assembly variation as one changes the *k* value (Gallo et al., 2014).

Persistent repeats preventing full *de novo* assembly, i.e., presenting limits to the contiguity of scaffolds/contigs that can be achieved using typical NGS pipelines, can be interpreted as compromising the *stability of the assembly process in silico*. The same repeats that can prevent a satisfactory assembly, such as the *ori* repeats in the mitochondrial genome of baker's yeast, can also compromise the *stability of the genome itself in vivo* or *in vitro*, i.e., they can destabilize processes within the cell and affect the structural stability of the genome (Bernardi, 2005). Just as a *de novo* short-read assembly program can get confused if it encounters a sequence of moderate length that is repeated (i.e., not unique), and may then terminate contig extension because it is not clear where the sequence continues, so a process of the cell such as recombination can apparently get confused at the same places in a mitochondrial genome because of ambiguity, and then create a deletion mutant in which DNA located between the two repeats is lost (such as in the well-studied petite-colony mutants in baker's yeast; Marotta et al. (1982)). In the human mitochondrial genome, a particularly common block deletion that entails clinical consequences, called the common deletion, deletes almost 5 kb. This may be a result of two copies of a 13-bp sequence at the ends of the deleted segment (Samuels et al., 2004), or the cause may be secondary structures co-localizing with those two copies rather than the copies' sequences themselves, as has been proposed based on deletional spectra (Guo et al., 2010).

In the present study, we focused on mitochondrial genomes of 11 unicellular fungi that appear to have been completely and reliably sequenced. We used a simple strategy, namely a search for recurring oligonucleotides (*w*-mers) of a fixed length, to screen for presence of possibly longer sequences that are repeated many times within the genome; an ultimate goal, towards which this study can offer only a first start, would be to understand the repeats and repeat systems in those and other mitochondrial genomes in the light of their possible function and evolution.

The working hypothesis motivating our study was that precisely the risks taken by a mitochondrion that continues to maintain sizable repeats, persisting in numerous copies that are interspersed around its relatively small genome, could often be a clue signaling functionally and/or structurally important sequence features at or flanking the repeats' genomic locations. Indeed, in the absence of any benefit, one might expect that natural selection will tend to eliminate variants exhibiting risky repetition, and to favor variants exhibiting no or only benign repetition. Such a principle, if consolidated by testing, could then be applied to efficiently 'mine for meaning', in a high throughput fashion, across mitochondrial or even nuclear genomes. As a start, we went through the steps of systematically screening our 11 chosen fungal mitochondrial genomes.

2. Materials and methods

2.1. Definition of mitochondrial genome

In this study we use the term "mitochondrial genome" as synonymous with "mitochondrial DNA genome" or "mtDNA genome". Most genome reports use this definition. However, some authors include also nuclear-encoded mitochondrial genes in the definition. Thus, Wallace (2013) defines the mitochondrial genome via the role, not the physical location, of the DNA and its genes, so that the "mitochondrial genome consists of thousands of copies of the maternally inherited mtDNA plus between 1000 and 2000 nDNA genes".

2.2. Species and strains

An objective of the present study was to gain a first insight into the presence and types of repeats in fungal mitochondrial genomes that can be highlighted using a simple strategy, namely a search for persistently recurring oligonucleotides of a fixed length. For reasons given below, we focused primarily on 17-mers. Since repeats can be the most difficult sequences to assemble from reads, we took care to select only fungi represented by a complete mitochondrial genome sequence that appeared reliable, with no obvious gaps that might correspond to unsequenced repeats. We therefore did not include partially sequenced/assembled genomes, in order to avoid mistaking 'absence of evidence' for 'evidence of absence' of repeats.

We sampled transversally across fungi associated with animals and well-studied model fungi, but did not include, for example, non-model fungi associated exclusively with plants, and our list of 11 mtDNA sequences does not include all complete mitochondrial genome sequences of such fungi now available (see, e.g., van de Sande, 2012; Joardar et al., 2012).

BLASTN searches showed that a short segment of the mitochondrial genome sequence of *Paracoccidioides brasiliensis* Pb18 was of non-fungal origin, and it was removed.

2.3. '17 × 4' criterion for persistent repeats

For the study conducted here, unless otherwise mentioned, we chose a pilot criterion that considers *persistent* repeats to be those sequences of length at least 17 bp that are repeated at least 4 times in a given mitochondrial genome. Fourfold repeats of sequences longer than 17 bp always correspond to two or more overlapping 17-bp repeats and are therefore also represented.

We justify our choice of 17 bp by our interest in (a) risk of genome-destabilizing ectopic recombination,[1] (b) risk of assembly-destabilizing repetition, and (c) secondary structures possibly associated with repeats. In these contexts, we note that (a) experimental, quantitative recombination studies consistently suggest that a perfectly repeated 17-mer should generally pose a risk of ectopic recombination, when the repeats occur within short distances similar to those found within a circular fungal mitochondrial genome; (b) some NGS short read assembly programs, such as SOAPdenovo2, offer *k*-mer size choices starting at around 17 bp; and (c) 17 bp is approximately the size of a conceivable, minimal compact stemloop/hairpin with flanks (e.g., flank 2 bp + stem 5 bp + loop 3 bp + stem 5 bp + flank 2 bp; see also Forsdyke (1995) and related oligonucleotide symmetry searches in (Bultrini et al., 2003; Baisnee et al., 2002)). We also note that a recent analysis of frequently repeated words in the human nuclear genome focused on words of a similar length, 15–16 bp (Zahradnik et al., 2014).

[1] We note that the mitochondrial processes that typically interest us here do not involve routine crossovers between two homologous chromosomes during meiosis, so it is not clear if borrowed terms such as 'ectopic', 'illegitimate', or 'out-of-register' recombination really make sense for intra-chromosomal recombination of (haploid) mitochondrial DNA. It is not evident what a corresponding 'non-ectopic' or 'in-register' recombination would be in such a context, as there is no expected, natural, routine or programmed 'in-register' mtDNA recombination event in the mitochondrion's life cycle to which one could refer as an obvious standard. We will however still borrow traditional metaphors such as 'ectopic' to refer to intra- or inter-chromosomal recombination between non-overlapping regions that have high sequence similarity.