Contents lists available at ScienceDirect

Fungal Genetics and Biology

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

Genomic CG dinucleotide deficiencies associated with transposable element hypermutation in Basidiomycetes, some lower fungi, a moss and a clubmoss

A. John Clutterbuck*

Wolfson Link Building, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow G12 8QQ, Scotland, UK

ARTICLE INFO

Keywords:
RIP
CpG deficiency
Basidiomycete genomes
Hypermutation
DNA methylation
Transposable elements

ABSTRACT

Many Basidiomycete genomes include substantial fractions that are deficient in CG dinucleotides, in extreme cases amounting to 70% of the genome. CG deficiency is variable and correlates with genome size and, more closely, with transposable element (TE) content. Many species have limited CG deficiency; it is therefore likely that there are other mechanisms that can control TE proliferation. Examination of TEs confirms that C-to-T transition mutations in CG dinucleotides may comprise a conspicuous proportion of differences between paired elements, however transition/transversion ratios are never as high as those due to RIP in some Ascomycetes, suggesting that repeat-associated CG mutation is not totally pervasive. This has allowed gene family expansion in Basidiomycetes, although CG transition differences are often prominent in paired gene family members, and are evidently responsible for destruction of some copies. A few lower fungal genomes exhibit similar evidence of repeat-associated CG mutation, as do the genomes of the two lower plants *Physcomitrella patens* and *Selaginella moellendorffii*, in both of which mutation parallels published methylation of CHG as well as CG nucleotides. In Basidiomycete DNA methylation has been reported to be largely confined to CG dinucleotides in repetitive DNA, but while methylation and mutation are evidently associated, it is not clear which is cause and which effect.

1. Introduction

RIP was discovered experimentally in the filamentous Ascomycete *Neurospora crassa*, where it was shown to introduce multiple C → T transition mutations into repeat sequences, such as TEs or experimental duplications. This process occurred during the sexual cycle and in *N. crassa* was found to preferentially target CA dinucleotides (reviewed by Selker (1990, 2002)). Subsequently, similar experimental results have been obtained in a number of other Ascomycetes (Hane et al., 2015), and examination of dinucleotide frequency distribution[1] (DFD) in 49 filamentous Ascomycete genomes suggests that RIP is widespread, leading to accumulation of conspicuous genomic fractions depleted for different cytosine dinucleotides, not just CAs. The affected dinucleotides vary between species, but in each case correspond to those most frequently involved in transition differences between members of TE families (Clutterbuck, 2011) where, in extreme cases, transitions outnumber transversions by more than 1000 to one.

In studying filamentous Ascomycetes Clutterbuck (2011) employed genomes of a Zygomycete, a Basidiomycete and Ascomycete yeasts as controls in which evidence of RIP was not expected, and indeed was not found: all genomic dinucleotide frequencies followed simple curves, generally corresponding to slightly skewed normal distributions. However, evidence of RIP-like hypermutation has been reported in members of the Pucciniomycotina subphylum of Basidiomycetes (Hood et al., 2005; Johnson et al., 2010; Horns et al., 2012; Amselem et al., 2015) and in other fungi: *Ganoderma* spp. (Zhu et al., 2015), *Ustilago hordei* (Laurie et al., 2012). In the present study genomes of a wider range of Basidiomycetes and other fungi have been tested for DFD anomalies, revealing prominent CG deficiencies in many of them.

Prompted by the findings of Zemach et al. (2010) that cytosines in repeat sequences are methylated in Basidiomycetes and some lower plants, genomes of these species, along with those of a number of other organisms, were also examined for DFD anomalies and mutation patterns in repeated elements.

2. Materials and methods

2.1. Genomes

Basidiomycete and lower fungal genomes are listed in

* Address: Flat B/R, 43 Partickhill Road, Glasgow G11 5BY, Scotland, UK.
E-mail address: jclutterbuck481@btinternet.com.

[1] Abbreviations: DFD = Dinucleotide Frequency Distribution. μCG/μCH = the relative frequencies of CG and CH dinucleotides in the deRIP consensus that have suffered, in either of the paired elements, from C→T transitions on either strand.

http://dx.doi.org/10.1016/j.fgb.2017.04.005
Received 3 February 2017; Received in revised form 10 April 2017; Accepted 17 April 2017
Available online 21 April 2017

Supplementary Table S1, arranged in taxonomic groupings: Hibbett et al. (2007, 2014), Matheny et al. (2006), Binder and Hibbett (2006: Boletales), Binder et al. (2010, 2013: Jaapiales etc, Polyporales), Riley et al. (2014), Nagy et al. (2016). Most fungal and plant genomes used were downloaded from the Mycocosm and Phytozome portals of the Department of Energy Joint Genome Institute (JGI), http://genome.jgi.doe.gov (4 January 2017), (Grigoriev et al., 2014), or the Broad Institute of Harvard and MIT http://www.broadinstitute.org/ (15 May 2015). Other sources: *Ustilago hordei* IBIS: Institute of bioinformatics and systems biology, Helmholtz Zentrum münchen http://www.helmholtz-muenchen.de/ibis/institute/about-us/index.html (4 January 2017); GaLuDB: *Ganoderma lucidum* genome database http://www.medfungi.org/20141010/galu (24 June 2015), and NCBI: National Center for Biotechnology Information http://www.ncbi.nlm.nih.gov (4 January 2017).

Di- and trinucleotides are written here as e.g. CG and TCG respectively, and complementary oligonucleotides as e.g. CA|TG and TCG|CGA. IUPAC ambiguity codes are also used: H = A, C or T; W = A or T; N = any base.

Genomes were scanned in 200 bp or longer windows for nucleotide frequencies using the modifications of the dedicated Perl script DFDscan.pl (Clutterbuck, 2011), and dinucleotide frequency distribution (DFD) anomalies were quantified using Perl normalfit.pl to fit a series of overlapping normal curves to observed distributions.

Sets of transcripts or TE sequences were scanned after concatenation by removal of FASTA headings. For whole genomes complementary dinucleotide frequencies are generally superimposable and dinucleotide frequency data are presented as the mean of the two e.g. CA|TG. Following the same procedure for concatenated gene or TE sets has the effect of levelling the bias that results from reading only the coding strand.

2.2. Transposable elements

Some TE sequences were obtained from published sources, but most were identified by BLAST analysis, using as queries predicted protein sequences from *Aspergillus nidulans* TEs (Clutterbuck et al., 2008, Supplementary files on disc, sequences also deposited at RepBase: Jurka et al., 2005): Mariner-6_AN-p, Gypsy-1_AN-p2, I-1_AN-p1 + p2 (a LINE-like element) and Helitron-1_ANp. Further TE family members were identified by BLASTn with the DNA sequences of primary sequences found. Elements are designated according to the BLAST query by which they were detected; their nature has not generally been investigated further, although where it was, the initial identity was confirmed. Alignments were tested and ORFs identified using GeneJockeyII for Macintosh. A further Perl script was used to scan paired elements for codon position of mutations. Where an ORF was evidently subject to a frameshift mutation, this was corrected before further analysis, and where no single ORF could be identified the most frequently mutated position was assumed to be codon position 3. As far as possible, pairs of elements in two, or sometimes more, different TE superfamilies (Kapitonov and Jurka, 2008) were sampled from each genome. Note that in most species the first two alignable elements to be detected in any superfamily were chosen for comparison, but in the earliest genomes examined more extensive families were first aligned and pairs of elements of interest were then selected from these. The elements examined should not therefore be considered as a random sample.

Pairs of TEs and sequences of retrieved gene families were examined using variants of the Perl script TE-RIPtest.pl (Clutterbuck, 2011), which returns the following data: CG content of each element and its relation to that expected on the basis of C and G content of the element, the ratio of transition to transversion differences between the two elements (ts/tv). CN deficiencies can also be seen as resulting from C → T transitions, and were counted relative to the frequency of each CN dinucleotide available for mutation in a de-RIP consensus that models a hypothetical ancestral element before hypermutation (Cambareri et al., 1998); for paired elements this is equivalent to comparison with unmutated sites in the other member of the pair. Transition mutations are counted on both strands, i.e. G → A transitions in the available top strand are treated as C → T mutations on the lower strand with flanking bases according to strand polarity. CG dinucleotides available for mutation are counted as double sites. Mutational preference for CG over CH sites is expressed as μCG/μCH[1]: the relative frequencies of CG and CH dinucleotides in the deRIP consensus that have suffered in either of the paired elements from C → T transitions on either strand.

2.3. Relevant proteins

DNA methyltransferases and RNAi-related proteins were identified by keyword or GO term annotation at the genome source, or in a few cases by BLAST searches. Tet genes were detected by tblastn with seven sequences of Basidiomycete TET proteins reportedly carried by TEs (Iyer et al., 2014): GIs: 315464672 (*Sporisorium reilianum*), 328858153 (*Melampsora laricis-populina*), 331223829 and 413160178 (*Puccinia graminis*), 170102298 (*Laccaria bicolor*), 299752738 (*Coprinopsis cinerea*), 426195020 (*Agaricus bisporus*), 393232900 (*Auricularia subglabra*.

3. Results

3.1. Basidiomycete fungi

3.1.1. Genomic dinucleotide frequency distribution (DFD) anomalies

Many Basidiomycete genomes display dinucleotide frequency anomalies, comparable to the varied proportions of cytosine dinucleotide deficiencies seen in Ascomycetes, but here largely confined to CGs. As an example, Fig. 1a shows frequency curves representing all 16 dinucleotides for *Gymnopus luxurians,* where only the CG curve stands out as including a substantial depleted component. Fig. 1b illustrates the curve fitting procedure for measurement of the proportion of the genome affected by CG depletion: a single normal distribution (normal-2) corresponds to the main portion of the curve, while normal distribution 1 represents the CG-depleted fraction, positive values of which add up to 18.8% of the genome. In some cases more than one overlapping normal distribution is required to cover either fraction and commonly a minor additional curve is required to accommodate very CG-rich sequences, e.g. Supplementary Fig. S1a (see also Huff and Zilberman, 2014). The full list of genomes studied and the size of their CG deficient fractions is given in Supplementary Table S1.

CG-depleted fractions vary widely among the 92 genomes studied (Fig. 2), 11 species show no deficiency, a further 51 have small (< 5% of genome) but distinctive CG-specific deficiencies e.g. *Agaricus bisporus*: see Supplementary Fig. S1a), and in 20 genomes, listed in Table 1, more than 10% of the genome is CG-deficient. Heavily CG depleted genomes are taxonomically well distributed, but are most conspicuous in the Boletales, Pucciniales and Agaricales (Supplementary Table S1).

The extent of CG loss (as opposed to the proportion of the genome affected) in the depleted fraction of each genome has not been quantified since the CG contents of the undamaged TE sequences are unknown, but in many genomes the depleted fraction peaks at zero, i.e. the majority of scanned windows in the depleted fraction have no CG dinucleotides, while other genomes CG depletion is less severe e.g. in *Serpula lacrymans* and *Melampsora laricis-populina* the deficient fraction peaks at higher frequencies (Table 1, column 6; see also Supplementary Fig. S1c and d). C → T transitions in CG dinucleotides are expected to create equal excesses in the CA and TG distributions as reported for the Ascomycete *Uncinocarpus reesei* (Zemach et al., 2010) and the Basidiomycete *Ustilago hordei* (Laurie et al., 2012), but the effect on frequency of these dinucleotides is to shift the mean without affecting the shape of the distribution.