Fungal Genetics and Biology xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Fungal Genetics and Biology

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

Mind the gap; seven reasons to close fragmented genome assemblies

Bart P.H.J. Thomma [a,*,1], Michael F. Seidl [a], Xiaoqian Shi-Kunne [a], David E. Cook [a], Melvin D. Bolton [b], Jan A.L. van Kan [a], Luigi Faino [a,1]

[a] Laboratory of Phytopathology, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands
[b] United States Department of Agriculture, Agricultural Research Service, Northern Crop Science Laboratory, Fargo, ND 58102-2765, USA

ARTICLE INFO

Article history:
Received 22 July 2015
Revised 27 August 2015
Accepted 28 August 2015
Available online xxxx

Keywords:
Fungal genome
Next-generation sequencing
Assembly
Transposable element
Repeat

ABSTRACT

Like other domains of life, research into the biology of filamentous microbes has greatly benefited from the advent of whole-genome sequencing. Next-generation sequencing (NGS) technologies have revolutionized sequencing, making genomic sciences accessible to many academic laboratories including those that study non-model organisms. Thus, hundreds of fungal genomes have been sequenced and are publically available today, although these initiatives have typically yielded considerably fragmented genome assemblies that often lack large contiguous genomic regions. Many important genomic features are contained in intergenic DNA that is often missing in current genome assemblies, and recent studies underscore the significance of non-coding regions and repetitive elements for the life style, adaptability and evolution of many organisms. The study of particular types of genetic elements, such as telomeres, centromeres, repetitive elements, effectors, and clusters of co-regulated genes, but also of phenomena such as structural rearrangements, genome compartmentalization and epigenetics, greatly benefits from having a contiguous and high-quality, preferably even complete and gapless, genome assembly. Here we discuss a number of important reasons to produce gapless, finished, genome assemblies to help answer important biological questions.

1. Introduction

Research into the biology of filamentous microbes has greatly benefited from the advent of whole-genome sequencing. Initially, genomes were sequenced with Sanger sequencing, requiring considerable investments of time, labor and costs. However, since the completion of the 12 Mb whole-genome sequence of the yeast *Saccharomyces cerevisiae* in 1996 (Goffeau et al., 1996), the 40 Mb whole-genome sequence of the first filamentous fungus *Neurospora crassa* in 2003 (Galagan et al., 2003), and the 65–95 Mb whole-genome sequences of the first two oomycetes that belong to the genus *Phytophthora* (Tyler et al., 2006), next-generation sequencing (NGS) technologies have increased the speed and scalability of genomic sequencing at a significantly reduced cost. These technological improvements have revolutionized biological research, making genomic sciences accessible to many academic laboratories including those that study non-model organisms. Thus, today hundreds of fungal genomes have been sequenced and are publically available.

NGS technologies routinely produce gigabases of sequence output in limited time. They are typically divided into short-read technologies that produce DNA sequence reads of up to 500 bp and long-read technologies that produce reads of >1 kb, the first typically being rather accurate (one sequence error per 1 kb on average) while the latter technologies are presently still quite error-prone (10–20 sequence errors per 100 bases on average) (Metzker, 2009). Long-read technologies, such as single-molecule real-time (SMRT) sequencing and Nanopore sequencing (Ashton et al., 2015; Huddleston et al., 2014; Laszlo et al., 2014; Powers et al., 2013), greatly facilitate the assembly of reads into contiguous sequences that ideally comprise complete chromosomes. Long-read sequencing is particularly beneficial because longer reads promote the identification of unique reads that map to unique sites. However, long-read sequencing technologies are still relatively expensive and therefore not yet routinely used for *de novo* genome sequencing. Consequently, eukaryotic genome sequencing initiatives are presently still largely based on short-read sequencing strategies and typically yield considerably fragmented genome assemblies (Koboldt et al., 2013).

Many genome sequencing projects aim to identify the genic coding regions of a particular genome, as this information greatly facilitates biological research aimed at characterizing the

* Corresponding author.
E-mail address: bart.thomma@wur.nl (B.P.H.J. Thomma).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.fgb.2015.08.010
1087-1845/

physiology or molecular biology of particular cellular processes. Therefore, many researchers are satisfied with a discontinuous whole-genome sequence, as long as the genic space has been sufficiently covered. However, DNA sequences that do not encode proteins, that were initially regarded as junk DNA, are increasingly linked to important traits controlling the life-style, adaptability and evolution of many organisms (Bickhart and Liu, 2014; Raffaele and Kamoun, 2012; Seidl and Thomma, 2014). Thus, there is increasing interest to assemble whole genomes beyond just the protein-coding regions. Here, we discuss seven types of genetic elements or phenomena, the study of which greatly benefits from having contiguous and high-quality, preferably even complete and gapless, genome assemblies.

2. Reason 1: Identification of life-style determining "effector" genes

Species are continuously evolving in dynamic environments in order to maintain or increase their fitness, relying on a plethora of mechanisms to generate genetic variation (Seidl and Thomma, 2014). For microbial plant pathogens it is well established that secreted effector molecules play crucial roles to support host colonization (Chisholm et al., 2006; de Jonge et al., 2011; Jones and Dangl, 2006; Thomma et al., 2011). However, other types of symbionts such as endophytes and mutualists employ effectors to establish host interactions in addition to pathogens (Rovenich et al., 2014). Conceivably, as symbionts establish their interactions in environments that are inhabited by other microbes, effectors may not only act to modulate putative hosts but also microbiome co-inhabitants. Moreover, effectors are likely to be employed by species that do not establish host interactions in order to manipulate their environment and establish themselves (Kombrink and Thomma, 2013). Indeed, the saprophyte *Verticillium tricorpus* has an effector repertoire resembling that of its pathogenic sister species *Verticillium dahliae*, both in size and type of effectors (Seidl et al., 2015). Thus, it is becoming increasingly apparent that secreted effectors play decisive roles in niche establishment of fungi with diverse life styles (Rovenich et al., 2014), and reliable calling of putative effector genes in genomics projects is important to study fungal biology.

Most effectors characterized to date are small cysteine-rich secreted proteins that lack significant sequence homology to characterized proteins and protein domains that can be used to reliably identify them. Whereas gene prediction programs often poorly predict effector genes due to their relatively small size and lack of homology or domains, reliable prediction of effector gene complements is even more obscured by the general observation that they often reside in plastic, fast-evolving genomic areas, such as subtelomeric regions. The finding that effectors reside in such compartments is no coincidence, as pathogen effectors are typically subject to strong selective forces such as those imposed by host immune systems. This incites selection for accelerated evolution through loss or modification, but also gain in case the selection pressure is relieved. This accelerated evolution can be established particularly in dynamic genomic compartments (Dong et al., 2015; Raffaele and Kamoun, 2012; Seidl and Thomma, 2014), fostering highly diversified effector complements between microbial lineages. This genomic plasticity is often established through repetitive elements, such as transposable elements (TEs), which can influence local plasticity through several mechanisms. TEs can change position, which may induce gene knockout, modulate gene regulation, or cause double-strand DNA breaks by TE excision. More generally, repetitive elements can play a passive role by acting as a substrate for recombination. Thus, optimal prediction of full effector catalogs requires accurate assembly of notoriously plastic genomic regions that are currently often poorly assembled (Fig. 1).

3. Reason 2: Identification of clusters of co-regulated genes

Gene clusters can be broadly defined as the close linkage of at least two genes that participate in the same metabolic or developmental pathway. Particularly in fungi, gene clusters have been shown to be involved in diverse roles such as nutrient use, mating type, pathogenicity, and the production of secondary metabolites (SMs) (Bolton et al., 2014; Bolton and Thomma, 2008; Keller and Hohn, 1997). Fungal SMs are generally not essential for normal growth, but are important in specific niches or developmental stages. However, due to their commercial importance and defined roles in pathogenicity of human, animal, and plant hosts, fungal SMs have received considerable attention (Keller et al., 2005). So far, only approximately 25% of all easily predictable SM clusters have SM final compounds assigned to them (Galazka and Freitag, 2014).

Fungal SM classes are generally defined by a multi-domain keystone enzyme in their respective biosynthetic pathways. Prominent examples include polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs), hybrid NRPS–PKS enzymes, prenyltransferases, and terpene cyclases associated with the production of polyketides, non-ribosomal peptides, NRPS–PKS hybrids, indole alkaloids, and terpenes, respectively (Keller et al., 2005). In addition, genes involved with the modification of intermediates formed from such enzymes are typically found in contiguous gene clusters that may also include transport-related genes involved with SM efflux (Martín et al., 2005) and/or pathway-specific transcription factors that control expression of the cluster (Fernandes et al., 1998; Hohn et al., 1999). Several hypotheses exist for the maintenance of SM genes as clusters. For one, unlinked SM pathway genes are at a greater risk for dissociation during meiotic recombination (Galazka and Freitag, 2014) or chromosomal rearrangements (de Jonge et al., 2013). Furthermore, clustering may promote horizontal transfer of an entire cluster. For example, a functional 23 gene sterigmatocystin biosynthesis cluster spanning 57 kb was likely horizontally transferred from *Aspergillus* to *Podospora anserina*, expanding the metabolic diversity of the latter species (Slot and Rokas, 2011). Additionally, gene clustering may facilitate strict coordination of gene expression, which may be particularly important during the biosynthesis of SMs that have potentially toxic intermediates (McGary et al., 2013).

SM gene clusters typically exhibit co-regulation. Although many SM clusters contain a pathway-specific transcription factor, its expression is often not sufficient for cluster expression since SM clusters can be enveloped within transcriptionally-silent heterochromatin that must be restructured before expression is possible. This is likely directly related to the tendency of SM clusters to be located in subtelomeric chromosomal regions where such chromatin modifiers impact transcription of clustered genes (Palmer and Keller, 2010). Moreover, subtelomeric regions housing SM clusters tend to be enriched in repetitive elements. Interestingly, repetitive elements play a role in the expression of the *Aspergillus nidulans* penicillin gene cluster since removal of repetitive DNA on either side of the cluster resulted in a reduction in penicillin production (Shaaban et al., 2010).

Thus, like effectors SM clusters also tend to reside in plastic, fast-evolving genomic areas, such as subtelomeric regions that are enriched in repetitive elements. For full appreciation of the potential of a particular microbe to produce SMs, accurate assembly and annotation of SM clusters is required (Fig. 1A), which requires accurate assembly of typically poorly assembled repeat-rich regions.

Please cite this article in press as: Thomma, B.P.H.J., et al. Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet. Biol. (2015), http://dx.doi.org/10.1016/j.fgb.2015.08.010