The *Candida dubliniensis* Genome Project

John Gamble[1], David Harris[1], David Saunders[1], Hubert Renauld[1], Gary Moran[2], Carol Munro[3], Derek Sullivan[4], Neil Gow[3] and Matt Berriman[1]
[1]Pathogen Sequencing Unit, Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK. [2]School of Dental Science, University of Dublin, Trinity College, Dublin, Ireland.
[3]University of Aberdeen, Institute of Medical Sciences, Aberdeen, UK. [4]Dublin Dental School & Hospital, Trinity College, Dublin, Ireland.

1. Introduction

Candida dubliniensis is the most recently discovered member of the genus *Candida*. In phylogenetic terms, it is more closely related to *C. albicans* than any other member of the genus. However, in terms of pathogenicity, it is ranked only fifth or sixth in terms of serious invasive disease (*C. albicans* being first). This suggests that a comparative analysis of the two genomes should define those genes most responsible for virulence in *Candida* species in general, and *C. albicans* in particular, and may suggest novel targets for therapeutic intervention. To this end, we have sequenced the genome of the commonly studied CD36 strain of *C. dubliniensis*.

2. The *Candida dubliniensis* genome assembly

The finished *C. dubliniensis* genome assembly contains 298,000 sequencing reads, assembled into eight chromosomes (1-7 and R), representing an 11-fold average coverage. The haploid genome size is 14,617,683 bp. For comparison purposes, the assembly was based on the karyotype of the well-characterized *C. albicans* strain SC5314. *In vivo*, however, the *C. dubliniensis* karyotype is highly complex, showing multiple chromosomal rearrangements[1]. The difference between the *in silico* genome assembly and its *in vivo* counterpart is illustrated below.

Fig 1a: Representation of the *Candida dubliniensis* genome assembly *in silico*

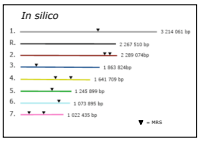
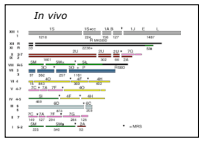


Fig 1b: Illustration of the *in vivo* karyotype of *Candida dubliniensis* strain CD36

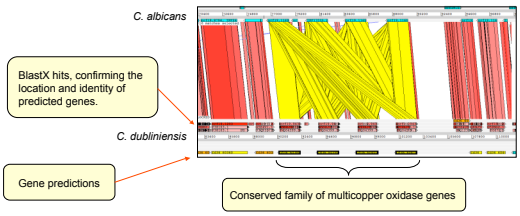


In the above figures, each chromosome has been colour-coded in order to illustrate the extent of chromosomal rearrangement *in vivo*. Note that the *in silico* assembly represents the haploid genome, whereas the *in vivo* karyotype is diploid. Figure 1b is from Magee *et al.* (2008).

3. Annotation of the *C. dubliniensis* genome

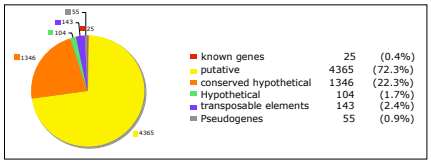
The Artemis Comparison Tool (ACT) allows sequence comparisons between large genomic sequences to be visualized intuitively[2]. In the analysis shown here, chromosome 6 from *C. albicans* and from *C. dubliniensis* were compared using TBlastX and the output visualised in ACT. The coloured bands, representing regions of sequence similarity, allowed identification of conserved genes between the two *Candida* species (Fig. 2a). The figure shows that synteny is largely conserved, and illustrates the utility of ACT for identifying gene families.

Fig 2a. TBlastX comparison of *C. albicans* and *C. dubliniensis* chromosome 6.



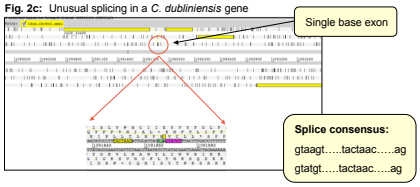
The identity of predicted genes was confirmed by BlastX analysis of each chromosome sequence against the UniProt database. This was confirmed by Fasta analysis of each predicted CDS feature against UniProt. In this way, 6,038 CDS features were annotated, of which more than 4,000 can be putatively identified.

Fig. 2b: Annotated genes in the *C. dubliniensis* genome



Approximately 6% of the total genes annotated are spliced, some of them showing highly unusual splicing patterns. This putative glucose transporter gene on chromosome R has four exons, one of which consists of a single nucleotide.

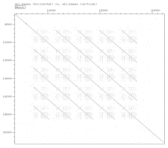
Fig. 2c: Unusual splicing in a *C. dubliniensis* gene



4. The *C. dubliniensis* Major Repeat Sequence (MRS)

The Major Repeat Sequence (MRS) is a feature unique to the genomes of *C. albicans* and *C. dubliniensis*[1], and may contribute to karyotypic variation in these species by acting as a 'hot-spot' for chromosomal translocation[4]. A gene-poor region of chromosome 2 was found to contain a cluster of five SfiI restriction sites, and to contain a CDS feature which, in *C. albicans*, was described as being located in the MRS RB2 region. Dot-plot analysis (Dotter) of this region showed that it contained five tandem repeat units, as shown below, in Fig. 3a.

Fig. 3a:



- The repeat-containing block of sequence is 9959 bp long.
- It contains four tandem repeats of 2106 bp each, plus one partial repeat of 1535 bp.
- Each repeat unit contains a single SfiI site.
- This is likely to be the RPS "core" of the MRS region.

A single RPS unit from this region was Blasted against each *C. dubliniensis* chromosome in turn, in order to locate the other MRS regions. As shown in Figure 1a (left), each chromosome, except chromosome R, carries at least one MRS element.

Alignment of 10 kb of sequence upstream and downstream of the core RPS unit from each MRS showed that the conserved flanking HOK and RB2 units previously noted in *C. albicans*[5] are also present in *C. dubliniensis* (Fig 3b). Thus, the MRS of *C. dubliniensis* and *C. albicans* appear to share the same basic structure and organization. An example of a fully annotated MRS (from chromosome 2) is shown below (Fig 3c).

Fig. 3b

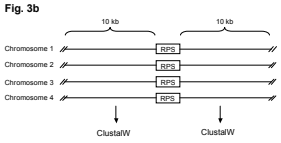
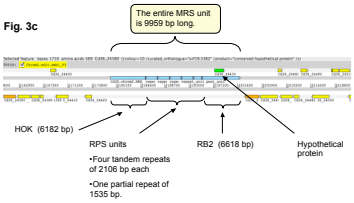


Fig. 3c



5. Comparison with *C. albicans*

We are currently comparing the annotated genomes of *C. dubliniensis* and *C. albicans*, in order to understand the molecular basis of their differing virulence properties. Preliminary analysis has revealed differences in gene content and arrangement, two examples of which are shown here.

Fig 4a. The biotin synthesis genes BIO3 and BIO4 are missing from *C. dubliniensis*

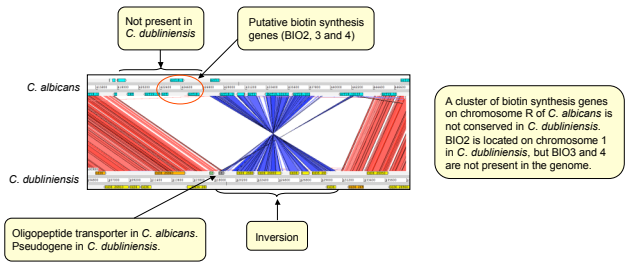
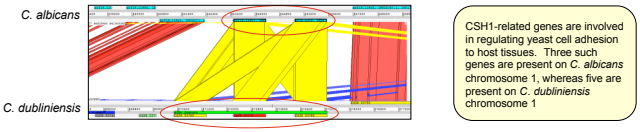


Fig 4b. Expansion of CSH1-related gene cluster on *C. dubliniensis* chromosome 1



6. *Candida dubliniensis* genome assembly in GeneDB

The *C. dubliniensis* genome sequence and all annotated gene models can be accessed from GeneDB (www.genedb.org).

Specific genes and their products are listed alphabetically, or can be searched for directly, and sequence files are available for download.

Fig. 5a

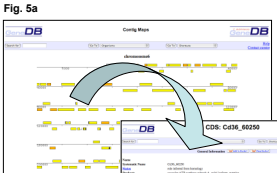
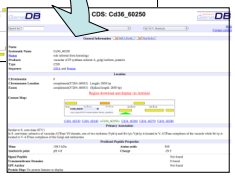


Fig. 5b



Gene maps of each chromosome are available (Fig. 5a). Full annotation for a given gene can be obtained by double-clicking on it within the chromosome map window (Fig. 5b).

References

1. Magee, BB. *et al.* (2008). Fungal Genetics and Biology 45: 338-350.
2. Carver, TJ. *et al.* (2005). Bioinformatics 21 (16): 3422-3423.
3. Singleton, DR. *et al.* (2001). J. Bacteriol. 183(12): 3582-3588.
4. Lephart, PR. *et al.* (2005). Eukaryotic Cell 4(4): 733-741.
5. Chindamporn, A. *et al.* (1998). Microbiology 144: 849-857.

Acknowledgements

The *Candida dubliniensis* genome sequencing project is funded by The Wellcome Trust (WTSI core funding).

The authors wish to acknowledge the contribution of the numerous staff at their respective institutions who have contributed to the sequencing and preliminary analysis of the *C. dubliniensis* genome.