

# 3

## Fungal Genomics

David Fitzpatrick and Edgar Mauricio Medina Tovar

### 3.1 Introduction

Genomics is defined as the study of an organism's complete genome sequence. The first complete (nonviral) genome to be sequenced was the bacterium *Haemophilus influenzae* in 1995. Today, more than 1300 bacterial genomes have been sequenced. Baker's yeast (*Saccharomyces cerevisiae*) was the first eukaryote to have its genome completely sequenced (released in 1996). Since then, over 250 eukaryote genomes have been completed, including our own (in 2001). Because of their relatively small genome size, roles as human/crop pathogens and importance in the field of biotechnology, 102 fungal genomes (Table 3.1) have been sequenced to date, accounting for approximately 40 % of all available eukaryotic genomic data. Some species, such as *S. cerevisiae* have actually had multiple strains sequenced. Presently, the majority (78 %) of fungal species that have been sequenced belong to the Ascomycota phylum; furthermore, there is a significant bias towards species that are pathogens of humans. Reduced costs and recent improvements associated with new sequencing technologies (Section 3.2) should mean that a wider range of evolutionarily, environmentally and biotechnologically interesting organisms will become available in the coming years.

This abundance of genomic data has moved the fungal kingdom to the forefront of eukaryotic genomics. While some of the species sequenced are closely related, others have diverged 1 billion years ago. This enables us to use fungi to study evolutionary mechanisms associated with eukaryotic genome structure, organization and content. Furthermore, doing a direct comparison between two or more closely related pathogenic and nonpathogenic species, a process called comparative genomics (Section 3.4), permits us to locate metabolic pathways or genes associated with virulence.

**Table 3.1** List of fungal genomes currently sequenced.

Species	Lineage	No. genes
<i>Allomyces macrogynus</i>	Chytridiomycetes	17 600
<i>Batrachochytrium dendrobatidis</i>	Chytridiomycetes	8 732
<i>Spizellomyces punctatus</i>	Chytridiomycetes	8 804
<i>Mucor circinelloides</i>	Zygomycota	10 930
<i>Phycomyces blakesleeanus</i>	Zygomycota	16 528
<i>Rhizopus oryzae</i>	Zygomycota	17 459
<i>Alternaria brassicicola</i>	Ascomycota	10 688
<i>Ashbya gossypii</i>	Ascomycota	4 717
<i>Aspergillus carbonarius</i>	Ascomycota	11 624
<i>Aspergillus clavatus</i>	Ascomycota	9 120
<i>Aspergillus flavus</i>	Ascomycota	12 587
<i>Aspergillus fumigatus</i>	Ascomycota	9 887
<i>Aspergillus nidulans</i>	Ascomycota	10 560
<i>Aspergillus niger</i>	Ascomycota	8 592
<i>Aspergillus oryzae</i>	Ascomycota	12 063
<i>Aspergillus terreus</i>	Ascomycota	10 406
<i>Blastomyces dermatitidis</i>	Ascomycota	9 522
<i>Botryotinia cinerea</i>	Ascomycota	16 448
<i>Candida albicans</i>	Ascomycota	6 205
<i>Candida dubliniensis</i>	Ascomycota	5 928
<i>Candida glabrata</i>	Ascomycota	5 202
<i>Candida guilliermondii</i>	Ascomycota	5 920
<i>Candida lusitanae</i>	Ascomycota	5 941
<i>Candida parapsilosis</i>	Ascomycota	5 823
<i>Candida tropicalis</i>	Ascomycota	6 258
<i>Chaetomium globosum</i>	Ascomycota	11 124
<i>Coccidioides immitis</i>	Ascomycota	10 654
<i>Coccidioides posadasii</i>	Ascomycota	10 124

**Table 3.1** (Continued)

Species	Lineage	No. genes
<i>Cochliobolus heterostrophus</i>	Ascomycota	9 633
<i>Cryphonectria parasitica</i>	Ascomycota	11 184
<i>Debaryomyces hansenii</i>	Ascomycota	6 272
<i>Fusarium graminearum</i>	Ascomycota	13 321
<i>Fusarium oxysporum</i>	Ascomycota	17 608
<i>Fusarium verticillioides</i>	Ascomycota	14 195
<i>Histoplasma capsulatum</i>	Ascomycota	9 251
<i>Kluyveromyces lactis</i>	Ascomycota	5 076
<i>Kluyveromyces waltii</i>	Ascomycota	5 350
<i>Lachancea thermotolerans</i>	Ascomycota	5 091
<i>Lodderomyces elongisporus</i>	Ascomycota	5 799
<i>Magnaporthe grisea</i>	Ascomycota	11 109
<i>Microsporium canis</i>	Ascomycota	8 765
<i>Microsporium gypseum</i>	Ascomycota	8 876
<i>Mycosphaerella fijiensis</i>	Ascomycota	10 313
<i>Mycosphaerella graminicola</i>	Ascomycota	10 933
<i>Nectria haematococca</i>	Ascomycota	15 707
<i>Neosartorya fischeri</i>	Ascomycota	10 403
<i>Neurospora crassa</i>	Ascomycota	9 908
<i>Neurospora discreta</i>	Ascomycota	9 948
<i>Neurospora tetrasperma</i>	Ascomycota	10 640
<i>Paracoccidioides brasiliensis</i>	Ascomycota	9 136
<i>Penicillium chrysogenum</i>	Ascomycota	12 773
<i>Penicillium marneffeii</i>	Ascomycota	10 638
<i>Pichia pastoris</i>	Ascomycota	5 040
<i>Pichia stipitis</i>	Ascomycota	5 807
<i>Podospora anserina</i>	Ascomycota	10 601
<i>Pyrenophora tritici-repentis</i>	Ascomycota	12 169

(continued)

**Table 3.1** (Continued)

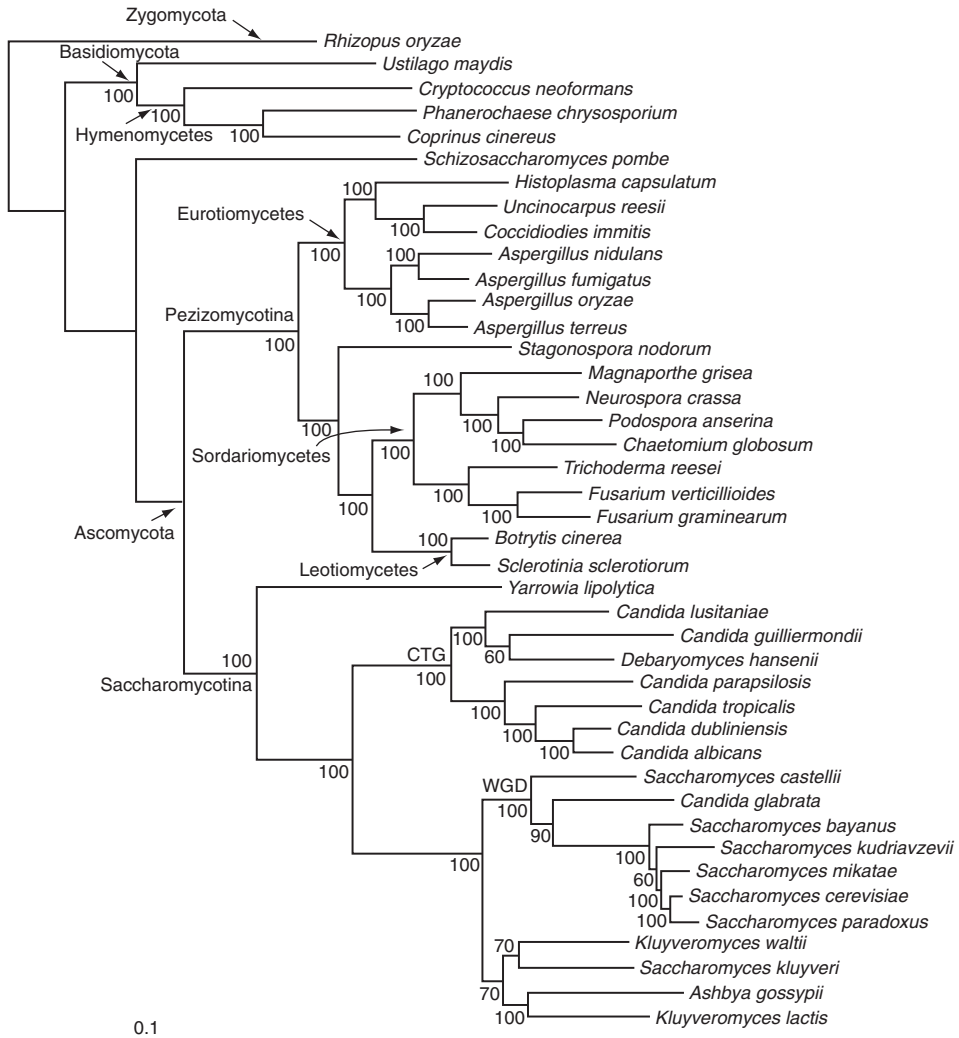
Species	Lineage	No. genes
<i>Saccharomyces bayanus</i>	Ascomycota	9 417
<i>Saccharomyces castelli</i>	Ascomycota	4 677
<i>Saccharomyces cerevisiae</i>	Ascomycota	5 885
<i>Saccharomyces kluyveri</i>	Ascomycota	5 321
<i>Saccharomyces kudriavzevii</i>	Ascomycota	3 768
<i>Saccharomyces mikatae</i>	Ascomycota	9 057
<i>Saccharomyces paradoxus</i>	Ascomycota	8 955
<i>Schizosaccharomyces cryophilus</i>	Ascomycota	5 057
<i>Schizosaccharomyces japonicus</i>	Ascomycota	4 814
<i>Schizosaccharomyces octosporus</i>	Ascomycota	4 925
<i>Schizosaccharomyces pombe</i>	Ascomycota	5 010
<i>Sclerotinia sclerotiorum</i>	Ascomycota	14 522
<i>Sporotrichum thermophile</i>	Ascomycota	8 806
<i>Stagonospora nodorum</i>	Ascomycota	16 597
<i>Talaromyces stipitatus</i>	Ascomycota	13 252
<i>Thielavia terrestris</i>	Ascomycota	9 815
<i>Trichoderma atroviride</i>	Ascomycota	11 100
<i>Trichoderma reesei</i>	Ascomycota	9 143
<i>Trichoderma virens</i>	Ascomycota	11 643
<i>Trichophyton equinum</i>	Ascomycota	8 560
<i>Trichophyton rubrum</i>	Ascomycota	8 625
<i>Trichophyton tonsurans</i>	Ascomycota	8 230
<i>Uncinocarpus reesii</i>	Ascomycota	7 798
<i>Vanderwaltozyma polyspora</i>	Ascomycota	5 367
<i>Verticillium alboatrum</i>	Ascomycota	10 220
<i>Verticillium dahliae</i>	Ascomycota	10 535
<i>Yarrowia lipolytica</i>	Ascomycota	6 448
<i>Zygosaccharomyces rouxii</i>	Ascomycota	4 991

**Table 3.1** (Continued)

Species	Lineage	No. genes
<i>Agaricus bisporus</i>	Basidiomycota	10 438
<i>Coprinopsis cinerea</i>	Basidiomycota	13 394
<i>Cryptococcus gattii</i>	Basidiomycota	6 210
<i>Cryptococcus neoformans</i>	Basidiomycota	6 967
<i>Heterobasidion annosum</i>	Basidiomycota	12 299
<i>Laccaria bicolor</i>	Basidiomycota	19 036
<i>Moniliophthora perniciosa</i>	Basidiomycota	13 560
<i>Phanerochaete chrysosporium</i>	Basidiomycota	10 048
<i>Pleurotus ostreatus</i>	Basidiomycota	11 603
<i>Postia placenta</i>	Basidiomycota	9 113
<i>Schizophyllum commune</i>	Basidiomycota	13 181
<i>Serpula lacrymans</i>	Basidiomycota	14 495
<i>Tremella mesenterica</i>	Basidiomycota	8 313
<i>Melampsora laricis-populina</i>	Basidiomycota	16 831
<i>Puccinia graminis</i>	Basidiomycota	20 566
<i>Sporobolomyces roseus</i>	Basidiomycota	5 536
<i>Malassezia globosa</i>	Basidiomycota	4 286
<i>Ustilago maydis</i>	Basidiomycota	6 522

### 3.1.1 The Fungal Kingdom

Fungi are eukaryotic organisms (contain a nucleus and membrane-bound organelles) and form one of the kingdoms of life. They lack chlorophyll and are saprobic (live on dead organic matter). Traditionally, fungi were thought to be closely related to plants; however, recent phylogenetic studies have shown that fungi are more closely related to animals than plants. Fungi are ubiquitous and can be beneficial (useful in biotechnology), harmful (cause disease) or mutualistic (symbionts with plants). The exact number of fungal species is unknown, but it is estimated to be 1.5 million. Phylogenetic analysis (Figure 3.1) has revealed that there are four distinct phyla within the fungal kingdom; they are the Chytridiomycota, Zygomycota, Ascomycota and Basidiomycota.



**Figure 3.1** Fungal phylogeny based on 42 complete genomes. The Ascomycota, Basidiomycota and Zygomycota make up three of the four fungal phyla and are present as monophyletic clades.

The Chytridmycota (chytrids) is the only fungal phylum to produce zoospores and requires water for their dispersal. They are an ancient group of organisms and are thought to have changed little since fungi first diverged from the last common ancestor of all eukaryotes. Most chytrids live in soil or freshwater, although some are found in marine environments, where they have important roles in the decomposition of organic matter. The chytrid *Batrachochytrium dendrobatidis* has been shown to be responsible for a disease in amphibians called chytridiomycosis, which is responsible for declining frog populations in tropic regions.

The Zygomycota reproduce sexually and form thick-walled sexual spores called zygospores. Zygomycetes are morphologically diverse and account for 1 % of all described fungal species. They are also the most ecologically diverse phyla of fungi, living as saprophytes on dung, fruit and soil. They can also be found in the gut of arthropods and some are pathogens of plants, animals and other fungi. Some well-known members include *Mucor* and *Rhizopus* species, which cause bread mould and fruit rots respectively.

The Ascomycota is the largest fungal phylum, accounting for approximately 65 % of all known fungal species. The distinguishing feature of this phylum is an ascus. The ascus is the sexual spore-bearing cell where meiosis followed by one round of mitosis occurs to generate eight (or a multiple of eight) ascospores. Ascospores have thick walls and, therefore, are resistant to adverse conditions; but under favourable conditions they will germinate to form a haploid fungus. Three subphyla have been described in the Ascomycota; they are the subphyla Saccharomycotina, Pezizomycotina and Taphrinomycotina. The Saccharomycotina lack an ascoma, resulting in naked asci, and include important species such as *S. cerevisiae* (brewer's yeast) and *Candida albicans* (human pathogen). Members of the Pezizomycotina include all filamentous fungi (moulds) and include species such as *Aspergillus fumigatus* (human pathogen) and *Penicillium chrysogenum* (produces penicillin antibiotic). The Taphrinomycotina phylum includes many diverse morphologies, including the fission yeast form of *Schizosaccharomyces pombe*.

The Basidiomycota accounts for approximately 35 % of the known fungal species. A number of Basidiomycetes are instantly recognizable, as they produce elaborate fruit bodies, including puffballs and mushrooms. Well-known edible Basidiomycota mushrooms include *Agaricus bisporus* (common mushroom) and *Pleurotus ostreatus* (oyster mushroom). The ability to degrade lignin (found in plant cell wall) by certain members (e.g. *Armillaria mellea*) of the Basidiomycota is significant, as few microbes have this ability. Fungi that can degrade lignin are interesting in a biotechnological sense, as they have the potential to detoxify and delignify waste products.

## 3.2 Genome Sequencing

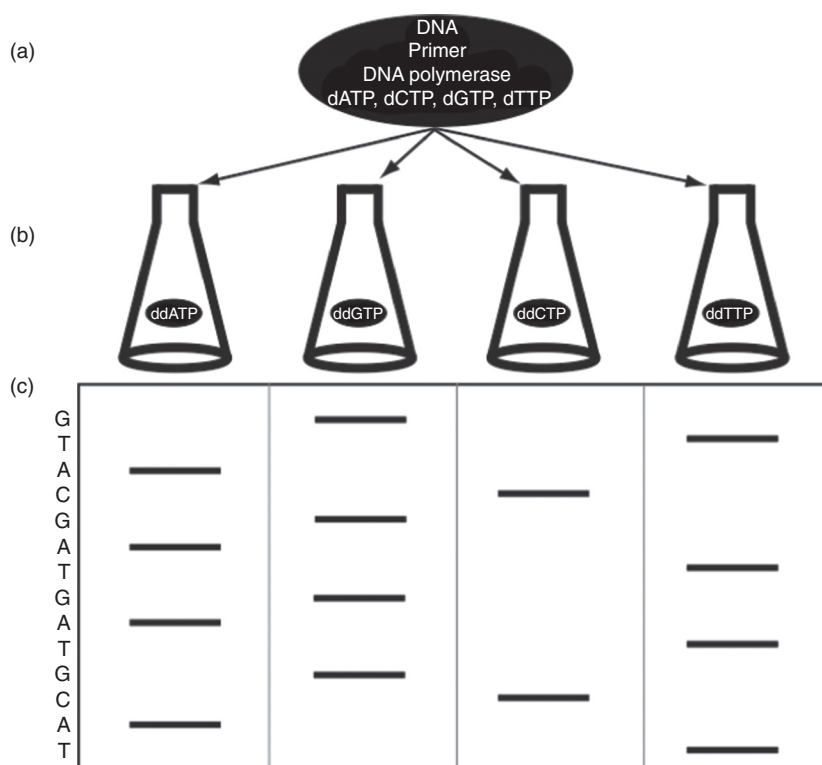
### 3.2.1 Sanger Sequencing

Fredrick Sanger won the Nobel Prize in 1958 for determining the amino acid sequence of the protein insulin. After this, he turned his attention to developing sequencing methods for RNA and DNA. In 1977 Sanger published a method for DNA sequencing commonly referred to as dideoxy sequencing (or chain termination) and won his second Nobel Prize for this work in 1980.

Sanger sequencing relies on DNA polymerase (a replication enzyme) to synthesize a new strand of DNA, which in turn can reveal the sequence of the

target DNA strand. DNA polymerase replicates a new DNA strand complementary to a piece of single-stranded DNA, by linking the 5'-hydroxyl end of a free nucleotide to the 3'-OH group of the nucleotide at the end of a primer. A primer is a small piece of single-stranded DNA that can hybridize to one strand of the template DNA and be extended by successive additions of nucleotides. As well as a supply of nucleotide triphosphates (dNTPs), the Sanger method also requires 2',3'-dideoxynucleotide triphosphates (ddNTPs) in small quantities relative to dNTPs. ddNTPs contain no reactive 3'-OH and, therefore, terminate DNA synthesis once they are incorporated into the primer extension.

A typical reaction mixture contains dATP, dTTP, dCTP, dGTP and one ddNTP (ddGTP, for example). Primer extension continues until an unmatched nucleotide is paired with a complementary ddNTP. Many fragments, each ending with a ddNTP of varying length, are produced from such a reaction (Figure 3.2).



**Figure 3.2** Schematic of the Sanger sequencing method. (a) Four separate DNA extension reactions are carried out. Materials required include single-stranded DNA, DNA polymerase, DNA primers and all four dNTPs. One of the dNTPs is radioactively labelled to enable visualization in (c). (b) Each of the four reactions contains a different dideoxynucleoside triphosphate (ddNTP). Synthesis continues until a ddNTP is incorporated, terminating extension reaction. (c) Products are separated based on size on a polyacrylamide gel and the sequence can be determined.

Radioactive sulfur or phosphorus isotopes are incorporated into the newly synthesized DNA template via labelled dNTPs, therefore making all fragments detectable by radiography. Fragments can then be separated based on length with polyacrylamide gel electrophoresis and the sequence can be determined (Figure 3.2). To determine the relative position of all four nucleotides it is necessary to run four reactions (each with a different ddNTP) in parallel (Figure 3.2).

### 3.2.2 Next-Generation Sequencing

When Fred Sanger and co-workers first developed their enzyme-based chain termination method for DNA sequencing they could not have predicted the massive advances in sequencing technology that have taken place in recent years. Next-generation sequencing (NGS) refers to novel commercial technologies which make it possible to generate millions of sequence reads (hundreds of base pairs in length) in a single sequencing reaction. With respect to fungi, NGS has been used to resequence targeted strains (such as *S. cerevisiae*), sequence *de novo* genomes (such as the xylose fermenter *Pichia stipitis*), analyse transcriptomes and characterize fungi in environmental samples. In the following two sections we will examine two of the most popular NGS platforms: Roche/454 GS FLX pyrosequencer and the Illumina genome analyser.

#### 3.2.2.1 Roche/454 GS FLX Pyrosequencer

The first commercial NGS was introduced in 2004 by 454 Life Sciences (now Roche Diagnostics). It utilizes pyrosequencing, a technique that ultimately emits light (using the firefly enzyme luciferase) after each incorporation of a nucleotide by DNA polymerase. With the latest instrument and sequencing kits and reagents it is possible to generate more than 1 million reads (average length 400 bases) in a single 10 h run.

The Roche/454 sequencer has three basic steps: single-stranded template DNA library preparation, emulsion-based clonal amplification of the library and sequencing by synthesis.

The DNA library preparation stage fragments sample DNA into small single-stranded DNA fragments (300–800 base pairs). Universal adapters specific for 3' and 5' ends are added to each fragment. Each universal adapter is 44 bases in length and consists of a 20-base PCR primer, a 20-base sequencing primer and an initiating 4-base (TCGA) sequence.

For the clonal amplification stage the single-stranded DNA library is mixed with small DNA capture beads (~35 µm in size). The beads contain one of the adapter primers and ligate a single-stranded DNA library fragment. The ratio of capture beads to library DNA is chosen to ensure that each bead binds a single DNA fragment. The bead-bound library complexes are emulsified with amplification reagents, resulting in microreactors containing just one bead with one unique sample–library fragment. In parallel, each library fragment

is amplified using thermal cycling within its own microreactor. The end result is several million copies of unique library DNA per bead. At the end of this phase the emulsion is broken down. DNA-positive beads are enriched and deposited onto a PicoTiterPlate (PTP) (a solid surface containing wells (~44  $\mu\text{m}$ )). The DNA-positive beads are overlaid with packing and enzyme (luciferase and sulfurylase) beads.

The final step is sequencing by synthesis. Nucleotides are flown across the PTP sequentially in a fixed order. Nucleotides that are complementary to the template strand are incorporated by the DNA polymerase, extending the DNA strand. If the template DNA contains three adjacent guanines (G), three cytosines (C) will be incorporated into the sequencing strand. As incorporation of nucleotides occurs at different rates, strands extend at different rates. Nucleotide incorporation generates free pyrophosphate, which is converted to ATP by the sulfurylase enzyme. ATP results in the oxidation of luciferin by the enzyme luciferase and light whose intensity is proportional to the number of bases incorporated is emitted. Light photons are captured by a charge-coupled-device camera and signal intensity per nucleotide is used to determine the sequence of template DNA.

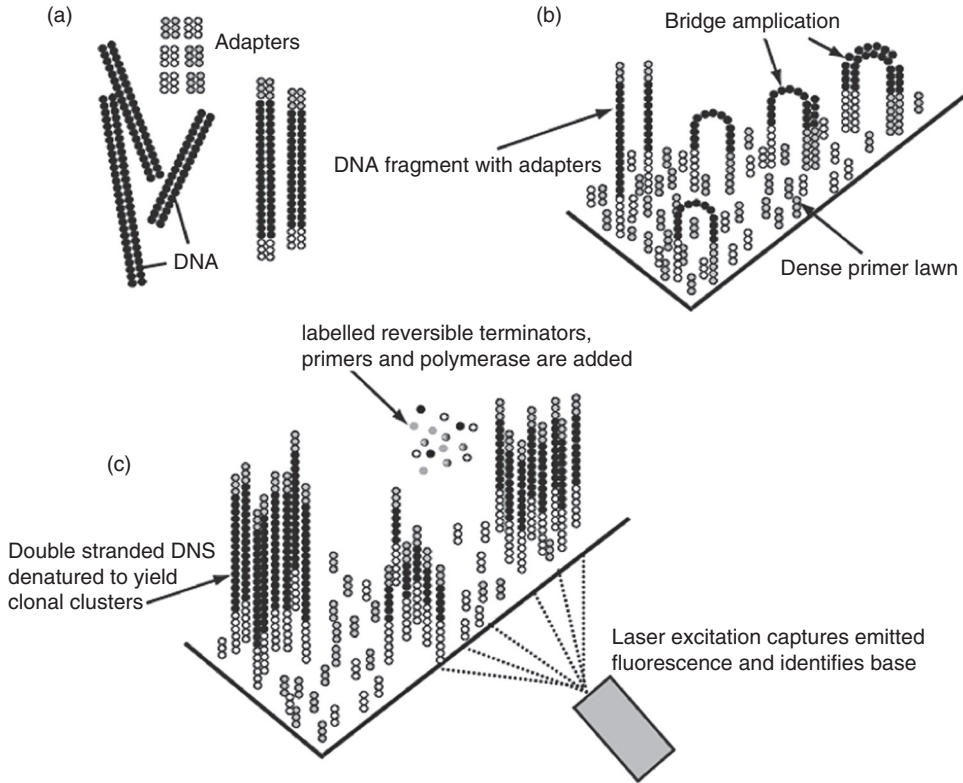
### 3.2.2.2 *Illumina Genome Analyser*

The Illumina genome analyser (IGA) was released in 2006 and currently can output 300 gigabases in a single run. Currently, read lengths (up to 150 bp) are shorter than those of 454 sequencing. The major advantage of the IGA over other sequencing platforms is the quantity of data produced at low cost. The sequencing process used by the IGA has three main steps: DNA library preparation, generation of clonal clusters and sequencing by synthesis.

The DNA library is prepared by fragmenting sample DNA by sonication (ultrasound) or nebulization (vaporizing) and sequencing adapters are ligated to the fragments (Figure 3.3a).

Clonal clusters are then generated by immobilizing sequencing templates on a flow cell. The flow cell is composed of silica and has eight lanes running lengthways. Separate DNA libraries can be loaded into different lanes, enabling eight individual sequencing runs per slide. Adapter-ligated template is pumped into the flow cell and template DNA is captured by forward/reverse 'lawn' primers that are covalently linked to the flow cell. Free ends of DNA template attach to lawn primers forming U-shaped bridges (Figure 3.3b). Unlabelled nucleotides are added and solid-phase bridge amplification occurs, resulting in double-stranded clonal clusters. Reverse strands are removed from double-stranded DNA and sequencing primers are hybridized to free 3' ends; this step ensures all clusters are sequenced in the same direction from the same end. The flow cell is then transferred to the IGA for sequencing.

IGA sequencing by synthesis involves the incorporation of fluorescent terminator deoxynucleoside triphosphate (dNTP) (Figure 3.3c). During each sequencing



**Figure 3.3** Schematic of the Illumina genome analyser sequencing technology. (a) DNA is fragmented and adapters are ligated to both ends of the fragments. (b) Single-stranded fragments bind randomly to the surface of the flow cell; see main text. (c) Sequencing by synthesis; see main text.

cycle, DNA polymerase incorporates a single dNTP to each of the growing nucleic acid chains (Figure 3.3). After each cycle, the IGA images the relevant fluorescent dye identifying the base and then cleaves the terminator dye so that addition of the next nucleotide can proceed. The sequence lengths of all clusters are identical as they are governed by the number of cycles (nucleotide incorporation, imaging and cleavage) undertaken.

## 3.3 Bioinformatics Tools

### 3.3.1 Locating Homologues

Sequence similarity searches are an essential component of genomic studies. They allow researchers to identify homologues, conserved structural motifs and help assign putative functions to unannotated genes in *de novo* genomes. The

**Table 3.2** Useful online resources.

Database	URL address
FGI	<a href="http://www.broadinstitute.org/annotation/fungi/fgi/index.html">http://www.broadinstitute.org/annotation/fungi/fgi/index.html</a>
SGD	<a href="http://www.yeastgenome.org/">http://www.yeastgenome.org/</a>
CGD	<a href="http://www.candidagenome.org/">http://www.candidagenome.org/</a>
AspGD	<a href="http://www.aspgd.org/">http://www.aspgd.org/</a>
CADRE	<a href="http://www.cadre-genomes.org.uk/">http://www.cadre-genomes.org.uk/</a>
<i>Aspergillus fumigatus</i> database	<a href="http://old.genedb.org/genedb/asp/">http://old.genedb.org/genedb/asp/</a>
CandidaDB	<a href="http://genodb.pasteur.fr/cgi-bin/WebObjects/CandidaDB">http://genodb.pasteur.fr/cgi-bin/WebObjects/CandidaDB</a>
NCBI	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
Fungal Genomes Central	<a href="http://www.ncbi.nlm.nih.gov/projects/genome/guide/fungi/">http://www.ncbi.nlm.nih.gov/projects/genome/guide/fungi/</a>
Wellcome trust Sanger Institute	<a href="http://www.sanger.ac.uk">http://www.sanger.ac.uk</a>
EMBL	<a href="http://www.embl.de/">http://www.embl.de/</a>
DDBJ	<a href="http://www.ddbj.nig.ac.jp/">www.ddbj.nig.ac.jp/</a>
SWISSPROT	<a href="http://expasy.org/sprot/">http://expasy.org/sprot/</a>
Fungal Tree of Life	<a href="http://aftol.org/">http://aftol.org/</a>
CGOB	<a href="http://cgob.ucd.ie/">http://cgob.ucd.ie/</a>

last 15 years has seen an exponential increase in the quantity of genetic data available in public databases such as NCBI (Table 3.2). To utilize this deluge of genetic data, it is imperative we have efficient similarity search techniques.

### 3.3.1.1 *Global and Local Alignments*

The methods used to infer homology can be categorized into two main types. A global alignment attempts to align two sequences over their entire length. Global sequence alignment works best when the sequences being compared are approximately the same length and highly similar. A local alignment, on the other hand, attempts to align two sequences at regions where high similarity is observed instead of trying to align the entire length of the two sequences being compared. Local alignments are usually more meaningful than their global counterparts, as they align conserved domains that may be important functionally even though the matched region is only a small proportion of the entire sequence length.

Needleman and Wunsch first implemented dynamic programming in 1970 to align sequences globally. Dynamic programming is a computational technique that determines the highest scoring alignment between two sequences. Smith and Waterman later adapted the Needleman and Wunsch method to align sequences locally. Both methods utilize a scoring matrix where rows and columns correspond to the bases/amino acids being aligned. The first row and column of the matrix are filled with zeroes; the remaining cells are filled iteratively with values dependent on neighbouring cell values. If a matrix cell corresponds to an identical base/residue, a match score is added to the score from the neighbouring diagonal square. Alternatively, the maximum score is determined from cells above by adding a gap penalty. Gap penalties are generally negative numbers. Local alignments are produced by starting at the highest scoring cell in the matrix and following a trace path to a cell that scores a zero.

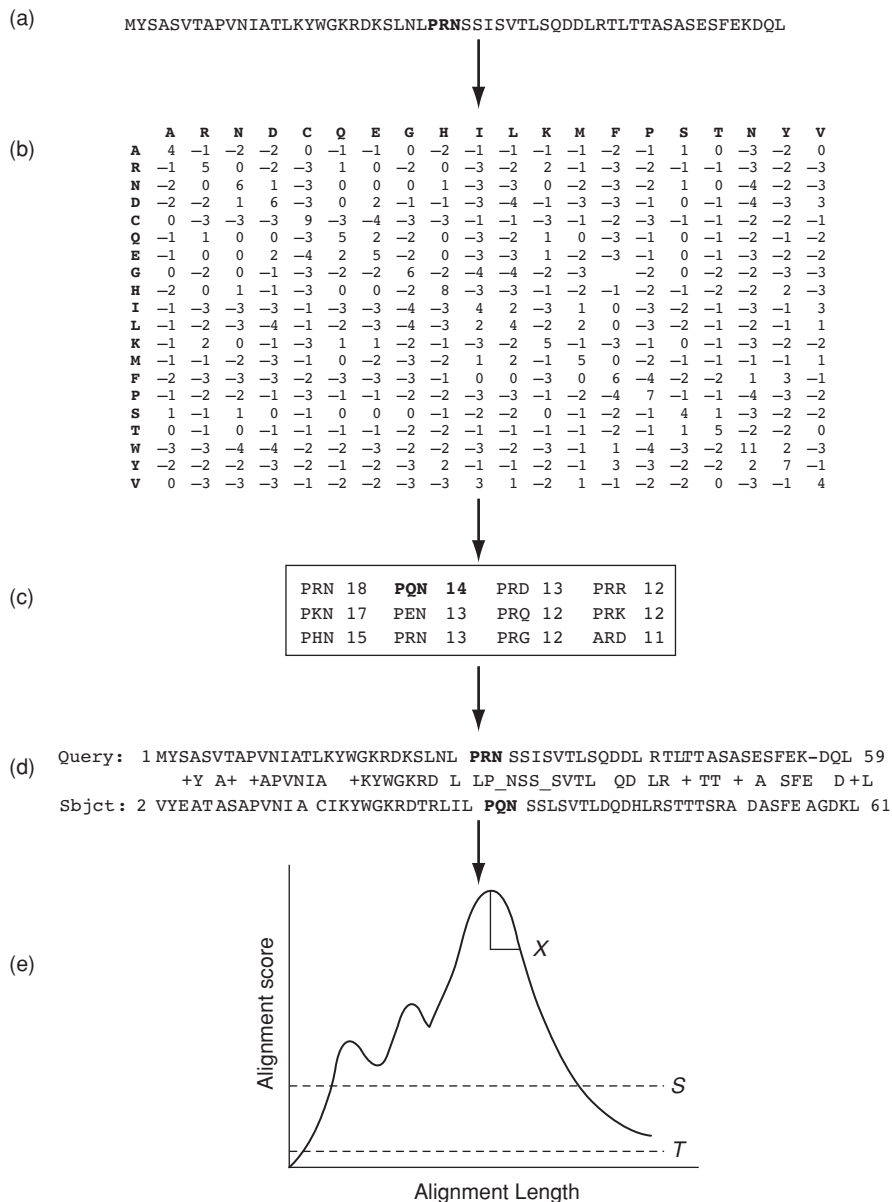
### 3.3.1.2 BLAST

The basic local alignment search tool (BLAST) is the most commonly used method for locating homologues in a sequence database. BLAST is both sensitive and efficient at locating regions of sequence similarity between nucleotide or protein sequences.

The BLAST algorithm begins by ‘seeding’ the search with a small subset of letters (query word) from the query sequence (Figure 3.4). The query word and related words (where conservative substitutions have been introduced) are located. All words are scored by a scoring matrix and this yields the ‘neighbourhood’ (Figure 3.4). BLAST uses a neighbourhood threshold  $T$  to determine which words are closely related to the original query word. Increasing the value of  $T$  implies that only closely related sequences are considered, while decreasing it allows for distantly related sequences to be considered.

The original query word is aligned to a word above the neighbourhood threshold (Figure 3.4). The BLAST algorithm then proceeds to extend the alignment in both directions, tracking the alignment score by addition of matches, mismatches and gaps. The maximal length of the alignment is determined by the number of positions aligned versus the cumulative score of the alignment. The alignment extension continues until the number of mismatches starts to decrease the cumulative score of the alignment; if this decrease is large enough (above a predefined value  $X$ , Figure 3.4), the alignment procedure ceases and the resultant alignment is called the high-scoring segment pair (HSP). A score threshold is defined by the algorithm, and if the HSP clears this score then the alignment is reported in the BLAST result file.

Finally, the biological significance of an HSP is determined. BLAST uses the  $E$ -value to calculate the number of HSPs that that would have a score greater than  $S$  by chance alone. Lower values of  $E$  imply greater biological significance; in essence,  $E$  can infer whether the HSP is a false positive.



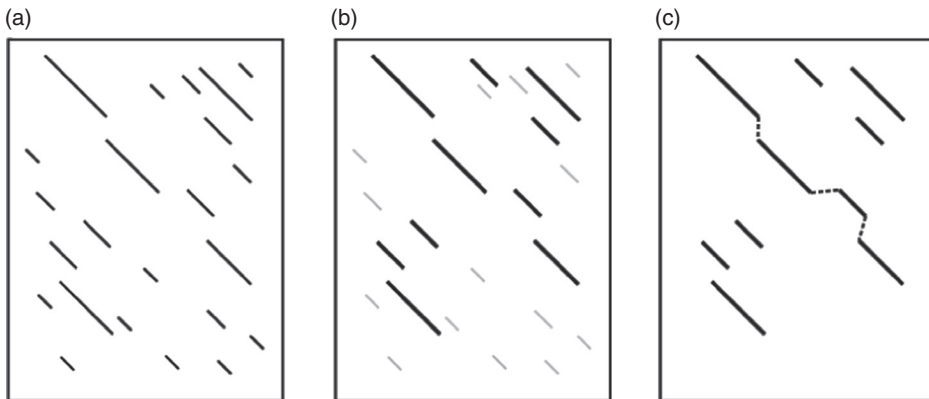
**Figure 3.4** Steps taken by the BLAST algorithm when searching a database. The query sequence (a) is compared with a scoring matrix (b) and scores for query words of a given length (three in this case) are calculated (c), query words greater than a certain threshold  $T$  are used to search the database. (d) The algorithm attempts to extend the alignment either side of the query word that has a hit in the target database. (e) Extension continues until the alignment score falls off more than the allowable significance decay  $X$ .

### 3.3.1.3 FASTA

Like BLAST, FASTA is a program for rapid alignment of pairs of protein or DNA sequences and was the first widely used algorithm utilized for database similarity searching.

FASTA begins by locating subsequences above a particular word length from the database sequence to subsequences of the query sequence. In FASTA, the word length parameter is termed  $ktup$  and it is equivalent to  $W$  in BLAST searches. FASTA generates diagonal lines on a dotplot where residues match up (Figure 3.5). The FASTA algorithm next locates diagonal regions in the alignment matrix that contain as many  $ktup$  matches as possible with short distances separating them (Figure 3.5). The top 10 highest scoring diagonal regions are retained and correspond to high scoring local alignments that do not contain gaps.

FASTA then determines which of the adjacent diagonals can be joined together thereby increasing the overall length of the alignment. For each diagonal that are connected a joining penalty is invoked and the overall score is determined by addition of the net scores of individual diagonals minus the joining penalties. The score of the enlarged diagonals is referred to as  $initn$ . All enlarged diagonals are ranked based on score and the highest scoring ones are aligned optimally using a local alignment strategy. Finally, FASTA assesses the significance of the alignment by randomly generating sequences of similar length and composition as the query sequences and calculates the probability that an alignment would be seen by random chance.



**Figure 3.5** Steps taken by the FASTA algorithm when searching a database. (a) Words common to the query and target sequence are located. FASTA connects words close to one another and these are represented by diagonal lines. (b) The top 10 diagonals are selected for further analysis. (c) Diagonals are aligned optimally using a local alignment strategy.

### 3.3.2 Multiple Sequence Alignment

Multiple sequence alignment (MSA) is a method that allows us to infer the interrelationships between DNA or protein families. While pairwise alignments are useful for locating homologues in databases and illustrating conservation between two sequences, they are not as informative as MSA. MSA has the ability to locate conserved residues/domains amongst thousands of sequences that can provide insights into important evolutionary and physiochemical processes. MSA is the first step in phylogenetic analysis and is commonly used when designing primers for DNA amplification.

Multiple sequence alignment is much more computationally intensive and difficult than the pair-wise strategy employed by BLAST and FASTA. One of the most commonly used MSA algorithms is CLUSTAL, and it utilizes progressive alignment to efficiently align all sequences of interest. CLUSTAL follows three steps.

- 1 An initial assessment of how closely related different sequences are to one another by performing pair-wise alignments.
- 2 A guide tree is generated based on the pairwise alignment scores.
- 3 Sequences are aligned progressively guided by the phylogenetic tree. Closely related sequences are aligned first, and then additional sequences and groups are aligned.

CLUSTAL refines its progressive alignments by implementing a number of alignment penalties. For example, gap insertion and extension penalties exist to reflect that the chances of a gap within a hydrophilic region are more likely, as these are generally loops or random coil regions where gaps are more common. Similarly, residue-specific penalties are enforced so that domains that are rich in glycine are more likely to have an adjacent gap than positions that are rich in valine, for example.

### 3.3.3 Gene Ontology

When the first comparison between two complete eukaryotic genomes (yeast and nematode worm) was performed, researchers were surprised to discover a high proportion of genes displayed orthology between these two distantly related organisms (diverged ~1.6 billion years ago). Orthologues are genes that are derived from a common ancestor and commonly have the same function. Following from this, knowledge of the biological role of an orthologue in one species can be used to illuminate the putative function of the other orthologue. However, organizing biological data from multiple species databases is a major

challenge and is made harder when different databases use different terminologies to describe the same process.

To overcome these difficulties, the Gene Ontology (GO) Consortium was set up in 2000 with the goal of producing a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism. Ontology terms provide a framework for storing and querying different databases using the same search terms. The GO Consortium provides detailed annotations for 12 important model organisms (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Dictyostelium discoideum*, *Drosophila melanogaster*, *Escherichia coli*, *Gallus gallus*, *Mus musculus*, *Rattus norvegicus*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*). Collectively, those 12 species are referred to as the GO reference genomes. The GO consists of over 26 000 terms arranged in three branches.

- 1 **Cellular component:** an individual component of a cell, but part of some larger object, such as an anatomical structure (nuclear membrane, for example).
- 2 **Biological process:** describes broad biological goals, such as mitosis or purine metabolism.
- 3 **Molecular function:** describes the roles carried out by individual gene products, examples include transcription factors and DNA binding.

The annotation of newly sequenced fungi can be greatly accelerated by comparisons to the GO reference genomes. *De novo* genes can be assigned putative functions based on sequence similarity to existing genes in one of the model organisms. The fact that two of the model organisms are fungi (*S. cerevisiae* and *S. pombe*) makes the GO resource highly applicable to genome annotation in newly sequenced fungal genomes.

## 3.4 Comparative Genomics

### 3.4.1 Gene Families Associated with Disease

Comparative genomic analyses have shown that certain gene families are important for virulence in some fungal species. For example, a comparative analysis of 34 fungal genomes identified gene families that are specific to fungal plant pathogens (*Botryotinia cinerea*, *Ashbya gossypii*, *Magnaporthe grisea*, *Sclerotinia sclerotiorum*, *Stagonospora nodorum*, *Ustilago maydis* and *Fusarium graminearum*). These families have expanded in terms of number (through duplication) during the evolution of phytopathogens. The same study also predicted the set of secreted proteins encoded by each phytopathogen and located gene families that were significantly enriched in the secretome (proteins secreted

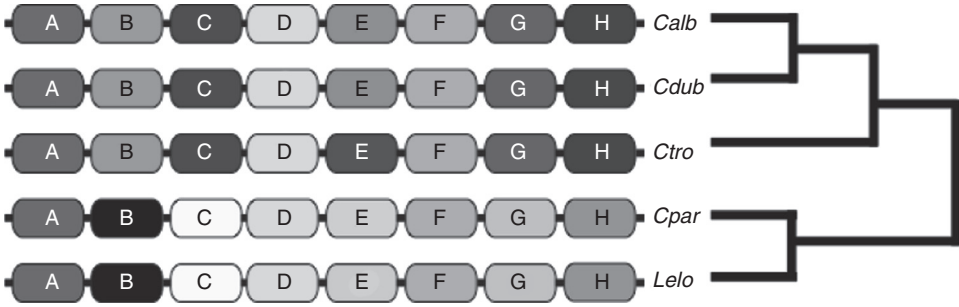
from a cell) of these species. Not surprisingly, many of the protein families identified are associated with pathogenic processes such as plant cell-wall degradation and biosynthesis of toxins. Similarly, the complete genome sequence of the corn smut fungus *U. maydis* uncovered a large set of secreted proteins, many of which are arranged in clusters. These genes account for ~20 % of all proteins secreted from *U. maydis*. Furthermore, the deletion of individual clusters seriously affects virulence, implicating the importance of these extracellular proteins.

In *Candida* species there are a number of gene families that are particularly enriched in highly pathogenic species (*C. albicans*, *C. parapsilosis* and *C. tropicalis*) compared with nonpathogenic species. For example, comparative analysis has shown that the Hyr/Iff proteins are present in all *Candida* species, but they are present in large numbers in the *Candida* pathogens (11, 17 and 18 copies respectively). Members of this family are components of the cell wall and, based on sequence similarity, are known to be evolving rapidly. They most likely play a role in host/fungal recognition, as rapid evolution of cell-wall proteins is a common escape mechanism employed by microbial pathogens. Another family enriched in pathogenic *Candida* species is the agglutinin-like sequence (Als) family. ALS genes are well characterized in *C. albicans* and are important for adhesion to host cells, plastic surfaces and biofilm formation. The ability to bind to plastic surfaces is a major problem in a hospital setting, as it allows *Candida* species to enter the blood stream via medical devices, such as intravenous drips; similarly reduced susceptibility to antifungal drugs is observed when *Candida* species grow as biofilms.

### 3.4.2 Synteny

The term synteny was traditionally used by geneticists to indicate the presence of two or more loci on the same chromosome. In the postgenomic era the concept of synteny has been expanded to address the relative order of genes on chromosomes that share a common evolutionary history. Two regions are considered syntenic if multiple consecutive genes are found in a conserved order between the two genomes under consideration (Figure 3.6). Synteny between two species may break down due to genome rearrangements in one or both species since they last shared a common ancestor.

Comparative fungal genomic analyses have shown that syntenic structure is generally conserved between very closely related fungal species, but it reduces as species become more distantly related. For example, the main subdivisions of Saccharomycotina yeasts share minimal synteny conservation between one another. However, large syntenic blocks are observed when members of the same subdivision are compared. For example, an analysis of nine *Candida* genomes (using the *Candida* genome order database (CGOB, Table 3.2)) showed they shared a high proportion of syntenic blocks. Conservation of gene order between



**Figure 3.6** Synteny of eight orthologues in five *Candida* species. Genes A, D and F are syntenic in all species; this is represented by conservation of colour between them. Genes B, C, G and H are syntenic in *C. albicans*, *C. dubliniensis* and *C. tropicalis*. Orthologues of B, C, G and H are present in *C. parapsilosis* and *L. elongisporus* and are syntenic with one another, although they are located in different genomic locations relative to the first three species. The degree of synteny between the five species closely matches the phylogeny of these species.

closely related species correlates with phylogenetic analyses (Figure 3.6). Other studies have shown that homologous chromosomes between the *Saccharomyces* ‘*sensu stricto*’ species are almost collinear, differing only by a small number of translocations and large inversions (segment of DNA is reversed). A comparison of two Basidiomycete genomes (*Coprinopsis cinerea* and *Laccaria bicolor*) showed they share extensive regions of synteny. The largest syntenic blocks occur in regions with low meiotic recombination rates and contain no transposable elements (cause translocations).

### 3.4.3 *In Silico* Metabolic Analysis

The availability of fungal genomes allows us to compare and contrast the metabolic repertoire of different species *in silico*. Detailed information from a metabolic pathway in one organism allows us to locate similarities or differences in another. Comparative metabolic analysis provides us with insights into potential disease mechanisms between pathogenic and nonpathogenic fungal species. Similarly, they also allow us to investigate the metabolic differences that allow one species to break down a particular substrate while another cannot.

Comparative studies of fungal species have shown that the genomic locations of certain genes are not random. For example, an analysis of the *S. cerevisiae* genome revealed that there is a significant tendency for genes from the same metabolic pathway to cluster in its genome. Similarly, genes involved in secondary metabolism are often clustered in the genomes of filamentous fungi (such as *Aspergilli* species).

An analysis of synteny between nine *Candida* genomes showed that approximately 20 % of metabolic pathways analysed display some evidence of clustering (lie within 10 genes of one another). One of the clustered pathways is involved in the metabolism of *N*-acetylglucosamine (Nag) to fructose-6-phosphate. It had initially been proposed that the ability of pathogenic strains of *Candida* to utilize Nag as alternative carbon sources is an important virulence factor. The three genes involved in the conversion of Nag to fructose-6-phosphate are hexokinase kinase (HXK1), Nag-6-phosphate deaminase (NAG1) and Nag-6-phosphate deacetylase (DAC1). These act sequentially on Nag and are present in *C. albicans* in a cluster termed the Nag regulon. Synteny analysis showed that the Nag regulon is conserved in nearly all *Candida* species. The conservation of the Nag regulon in pathogens like *C. albicans*, *C. tropicalis* and *C. parapsilosis* and nonpathogens such as *C. dubliniensis*, *Lodderomyces elongisporus* and *Debaryomyces hansenii* suggests that the ability to utilize Nag is not a virulence factor. *S. cerevisiae* is missing the Nag regulon and cannot utilize Nag; however, it has been shown that expression of *C. albicans* NAG genes in *S. cerevisiae* enables it to utilize Nag.

### 3.4.4 Horizontal Gene Transfer

Horizontal gene transfer (HGT) is the exchange of genes between different strains or species. HGT introduces new genes into a recipient genome that are either homologous to existing genes or belong to entirely new sequence families. Bacterial genomic sequencing has revealed that HGT is prominent in bacterial evolution and has been linked to the acquisition of drug resistance and the ability to catabolize certain amino acids that are important virulence factors. There are numerous methods to detect genes that have been transferred horizontally into a genome, including locating genes with an atypical base or codon usage pattern. Another approach is to perform a similarity search of candidate genes against a database and locate unexpected top database matches. These approaches have the advantage of speed and automation, but do not have a high degree of accuracy. Some notable flaws with the similarity-based approach of detecting HGT were brought to attention when the initial publication of the human genome reported that there were 223 genes that have been transferred from bacterial pathogens to humans. These findings were based on top hits from a BLAST search, but subsequent phylogenetic analyses showed these genes were not recently transferred from bacterial species through HGT. Indeed, the most convincing method to detect HGT is by phylogenetic inference. Topological disagreement (incongruence) between trees inferred for one gene family and that inferred for another can often be parsimoniously explained only by invoking HGT.

The process of gene transfer has been assumed to be of limited significance to fungi. However, the availability of fungal genome data (Table 3.1) and subsequent comparative genomic analyses are showing the importance of HGT in the

genome evolution of fungi. For example, *S. cerevisiae* has acquired 13 genes (from bacteria) via HGT since it diverged from its close relative *A. gossypii* (Figure 3.1). This number corresponds to a small minority of the *S. cerevisiae* genome (less than 1%). However, these 13 genes have contributed to important functional innovations, including the ability to synthesize biotin, to grow under anaerobic conditions and to utilise sulfate from several organic sources. Other documented examples of HGT in fungi include the acquisition of bacterial metabolic genes by *C. parapsilosis* and the acquisition of a toxin gene (ToxA) by *Pyrenophora tritici-repentis* from *Stagonospora nodorum* resulting in *Pyrenophora* infestations of wheat.

Unlike prokaryotes, the mechanisms of gene transfer into fungi are poorly understood. To date, no DNA uptake mechanism has been identified. Interkingdom conjugation between bacteria and yeast has been observed, however, and *S. cerevisiae* is transformant competent under certain conditions. However, HGT is probably facilitated by the fact that fungi are saprobes that live in close proximity with other organisms.

## 3.5 Genomics and the Fungal Tree of Life

### 3.5.1 Phylogenetics

The goal of phylogenetics is to arrange a set of populations, species, individuals or genes into a logical arrangement that infers the evolutionary relationships amongst them. Evolutionary relationships infer the historical development of species and are usually presented as an evolutionary tree (Figure 3.1). Traditional methods of fungal systematics, such as vegetative cell morphology, sexual states, physiological responses to fermentation and growth tests, can assign fungal species to particular genera and families. The fungal fossil record is poor, however, and fungi exhibit few morphological characters; therefore, an alternative approach is desirable. Fungal sequence data (RNA, DNA and protein) have been used successfully to infer evolutionary relationships amongst species. In many cases, aligned sequences (Section 3.3.2) are processed as a distance matrix. Species that are most closely related will have a small distance, while distantly related species will have a larger distance measure. Phylogenetic algorithms such as UPGMA, minimum evolution and neighbour joining are used to represent distance matrices as phylogenetic trees.

The choice of phylogenetic markers for inferring the fungal tree of life is a contentious issue. Ideally, a phylogenetic marker should be ubiquitous throughout the species under consideration, present in single copy, have slowly evolving sites and be unlikely to undergo HGT. For this reason a significant majority of accepted relationships between fungal organisms are determined using 18S ribosomal DNA. However, single-gene analyses are dependent on the phylogenetic markers having an evolutionary history that reflects that of the entire organism,

an assumption which is frequently violated. Also, individual genes contain a limited number of sites and, in turn, limited resolution. An alternative approach to single-gene phylogenies are multigene phylogenies. These attempt to combine all available phylogenetic markers. There are two commonly used methods to do this: concatenated multigene phylogeny reconstruction and supertree analysis.

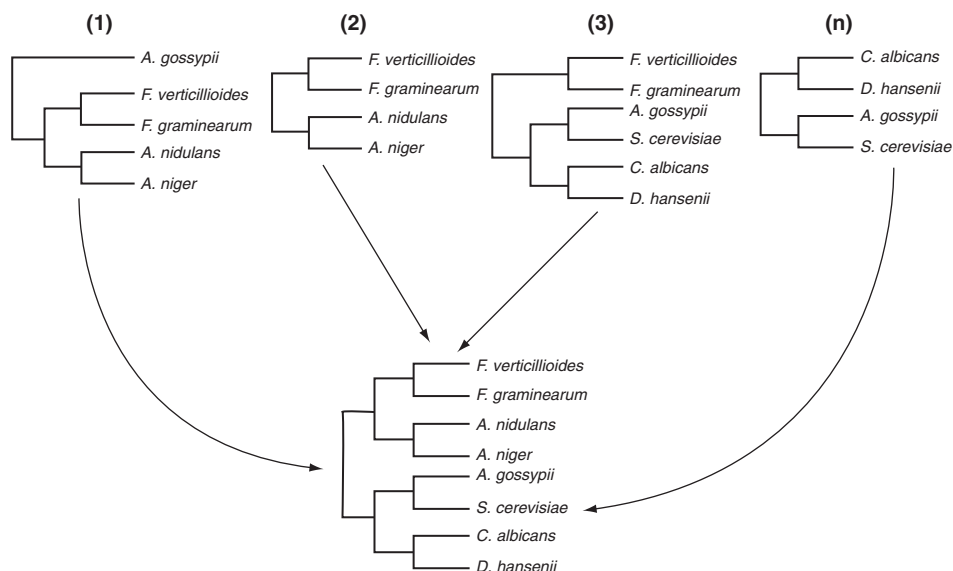
### 3.5.1.1 *Concatenated Multigene Phylogenies*

Multigene concatenation essentially appends many aligned genes together to give a large super alignment. Combining the data increases its informativeness, helps resolve nodes, basal branching and improve phylogenetic accuracy. Numerous species phylogenies have been derived by concatenation of universally distributed genes. Recently, the Fungal Tree of Life consortium (Table 3.2) used six housekeeping genes (18S rRNA, 28S rRNA, 5.8S rRNA, elongation factor 1-alpha and two RNA polymerase II subunits (RPB1 and RPB2)) from 199 fungal species to reconstruct the evolutionary history of the fungal kingdom. As well as showing the evolutionary history of all fungal phyla, this analysis showed that the loss of spore flagella from early diverging fungi (similar to extant chytrids) coincided with the development of novel spore dispersal mechanisms leading to the diversification of terrestrial fungi.

### 3.5.1.2 *Supertrees*

Supertree methods take all input trees and generate a single representative species phylogeny (Figure 3.7). Individual input trees are derived from single genes. Comparative fungal genomic analyses have shown that less than 1 % of all fungal genes are universally distributed. This situation implies that when we reconstruct multigene phylogenies we are ignoring 99 % of the genes found in fungi. Ideally we would use 100 % of the gene data. Supertree methods enable us to do this.

Supertree methods generate a phylogeny from a set of input trees that possess fully or partially overlapping sets of taxa (Figure 3.7). Therefore, supertree methods take as input a set of phylogenetic trees and return a phylogenetic tree that represents the input trees. This type of analysis yields a phylogeny that maximizes the number of genes used and, therefore, is truly representative of the entire genome. A supertree analysis of 42 complete fungal genomes identified 4805 individual gene families (Figure 3.1). Individual phylogenies for each gene family were reconstructed and the complete set was summarized by supertree techniques. This analysis showed that within the Saccharomycotina a monophyletic (single) clade containing *C. albicans* and close relatives is evident (Figure 3.1). Species within this clade translate the codon CTG as serine rather than leucine. A second monophyletic clade (Figure 3.1) containing genomes that have undergone a whole-genome duplication (*S. cerevisiae* and close relatives) is



**Figure 3.7** Representation of supertree reconstruction. Supertree methods take individual gene trees and express them as a single representative phylogeny. Thousands of trees (expressed as  $n$ ) can be used as input for supertree techniques.

also evident. Supertree techniques are becoming more popular in phylogenetic analysis and will be useful in reconstructing the fungal tree of life as additional fungal genomes become available.

## 3.6 Online Fungal Genomic Resources

### 3.6.1 NCBI: Fungal Genomes Central

The National Centre for Biotechnology Information (NCBI; Table 3.2) is a resource that houses public sequence databases (such as GenBank), supporting bibliographic/biological annotation, bioinformatics tools and associated applications.

The NCBI has a dedicated fungal genome section (Fungal Genomes Central (FGC)). The FGC has direct links to the 309 fungal genome sequencing projects currently underway or completed, as well as links to raw sequence reads generated by fungal sequencing centres such as the BROAD and Wellcome Trust Sanger Institute (Table 3.2). The complete sequences of 80 fungal mitochondria are also available for download. A fungal BLAST search utility is present enabling users to search genes of interest against all available fungal genomic data housed by the NCBI. A system for automated detection of homologues amongst

completely sequenced fungal genomes is available; more than 9000 fungal gene families are currently catalogued. Finally, the FGC has a taxonomy search engine containing the names and phylogenetic lineages of all fungal organisms that have molecular data in NCBI databases.

### 3.6.1.1 *Saccharomyces, Candida and Aspergillus Genome Databases*

The *Saccharomyces* Genome Database (SGD) (Table 3.2) went online in 1997 and is a specialized database dedicated specifically to *S. cerevisiae*. It is housed at the Stanford Human Genome Center and currently receives over 200 000 database hits a week. SGD provides users with access to the complete *S. cerevisiae* genome, its genes and their products, mutant phenotypes and the literature supporting these data. It should be noted that SGD is not a primary sequence database (contains information of the sequence alone) but instead collects DNA and protein sequence information from primary providers (such as GenBank, EMBL, DDBJ and SwissProt (Table 3.2)) and assembles all available information into datasets that are useful for molecular biologists. Therefore, SGD is considered a composite database, as it amalgamates a variety of different primary database sources and cuts out the need to search multiple resources.

SGD is highly annotated, and supporting literature linked to each gene is curated by dedicated SGD curators. Weekly automated searches of PubMed locate literature associated with *S. cerevisiae* genes or products and these are refined by curators who assign a given publication with appropriate genes. SGD provides an excellent text-based search interface that allows users to search by gene name, gene information, protein information, author name or full text. Amongst other things, SGD also allows users to perform BLAST database searches, view yeast metabolic pathways, search yeast-specific literature, view gene expression data from multiple microarray studies and view genes relative positions on chromosomes.

SGD organizes gene information around locus pages. The gene name and associated systematic name are shown at the top of each locus page. Information about the feature of the gene is also given; genes can be 'verified', meaning there is experimental evidence to show they are expressed or 'uncharacterized' implying a lack of experimental evidence. The 'description' section details important information known about the gene and associated products. Each gene product is assigned gene ontology terms that describe its molecular functions, location within the cell and putative biological processes in which it participates. A mutant phenotype section is also visible on the locus page. This section lists the type of mutation and any corresponding observable phenotype. Links to sequence information and literature describing the gene of interest are also available from the locus page.

The *Candida* Genome Database (CGD) went online in 2005 and is the central resource for researchers studying *Candida* pathogenesis and genetics. Before the launch of CGD, three independent web sites contained information about the

*Candida* genome sequence and associated gene products. The Stanford Genome Technology Center (Table 3.2) sequenced and distributed the genome; CandidaDB (Table 3.2) contained annotated genes for early assemblies of *C. albicans*, as did the *Candida* Working Annotation group (Table 3.2). The information available in these three web sites was initially pooled together and has subsequently been expanded on. CGD is based on the SGD framework; therefore, the software, user interfaces and data structure in CGD are identical to those described for the SGD above.

The *Aspergillus* Genome Database (AspGD, Table 3.2) went online in 2009 and is an online genomic resource for scientists studying the genetics and molecular biology of *Aspergilli* species. Currently, there are a number of databases containing information for multiple *Aspergillus* genomes. For example the Central *Aspergillus* Data Repository (CADRE) database (Table 3.2) contains clinical and patient-oriented information, the *Aspergillus* genome site at the Broad Institute (Table 3.2), the *Aspergillus fumigatus* database (Table 3.2) and also other web sites that focus on sequencing projects of one or several *Aspergillus* species. AspGD aims to link the resources of these individual databases and complement them by implementing in-depth manual curation of the primary scientific literature associated with the data. As with the CGD, AspGD is based on the SGD framework described above. AspGD is currently focusing on high-quality curation of *A. nidulans*, the best-characterized species of the *Aspergilli*, but will add information for other *Aspergilli* species (*A. fumigatus*, *A. flavus*, *A. oryzae*, *A. niger*, *A. clavatus*, *A. terreus* and *Neosartotya fischeri*) in the near future.

### 3.6.2 The Fungal Genome Initiative

The Fungal Genome Initiative (FGI, Table 3.2) is a partnership between the Broad Institute of Harvard and MIT and the broad fungal research community. It is directed by a steering committee of fungal geneticists and biologists who back in 2000 realized the slow pace of fungal genome sequencing projects was a major barrier to fungal biomedical and evolutionary research. A strategy where multiple organisms would be sequenced simultaneously as part of a cohesive strategy instead of individual ad hoc projects was conceived. The primary selection criteria for sequencing were (a) the importance of the organism in human health and commercial activities, (b) the value of the organism as a tool for studies of fungal diversity and comparative genomics and (c) presence of genetic resources and an established research community.

Initially, 15 fungi were selected for sequencing covering three broad aspects of fungal diversity:

- 1 Medical; e.g. *Rhizopus oryzae*, cause of mucormycosis.
- 2 Commercial; e.g. *Aspergillus flavus*, source of aflatoxin in food.
- 3 Evolutionary; e.g. *Neurospora discreta*, study of population genetics.

A subsequent 48 genomes have been targeted. In order to utilize the strengths of a comparative approach, strategic clusters were chosen (*Candida*, *Cryptococcus*, *Aspergillus*, *Fusarium*, *Histoplasma*, *Coccidioides*, *Penicillium*, *Neurospora*, *Schizosaccharomyces*, *Puccinia* and *Schizosaccharomycetes*).

To date, over 50 fungal genomes have been sequenced by the FGI and all sequence data is can be accessed publicly via the FGI homepage (Table 3.2). Information pertaining to genome maps and basic statistics about genome size, gene density and GC content are available. As well as providing standard database search tools such as BLAST, the FGI also incorporates text-based searches that allow researchers to locate genes based on a particular function or protein domain. Users can also select closely related organisms (from the same cluster) and view synteny maps.

### 3.7 Conclusion

The majority of fungi that have been sequenced to date are important biological pathogens (*C. albicans*, *A. fumigatus* and *Cryptococcus neoformans*, for example) or helpful species involved in brewing/fermentation (*S. cerevisiae* and *Aspergillus niger*, for example). Because of this, there is an unintentional bias in terms of the phylogenetic distribution of species sequenced. Owing to falling sequencing costs and their relatively small genome size, a deluge of fungal genomic data from all fungal phyla is expected in the years ahead. This data will allow us to address many new questions about fungal evolution and pathogenicity and will undoubtedly help uncover novel proteins with medical and biotechnological potential.

### Revision Questions

- Q 3.1** List the four major phyla of the fungal kingdom.
- Q 3.2** Outline the steps taken when sequencing DNA using the Sanger method.
- Q 3.3** What is next-generation sequencing?
- Q 3.4** What information is contained in online databases such as the *Saccharomyces* genome database (SGD), *Candida* genome database (CGD) and the *Aspergillus* database (AspGD)?
- Q 3.5** List the main difference between a global and local alignment.
- Q 3.6** What is a phylogenetic supertree?
- Q 3.7** What is the gene ontology (GO)? List and explain the three main branches of GO.

- Q 3.8** Explain how comparative genomics has the potential to uncover the genetic basis of disease.
- Q 3.9** What is gene synteny?
- Q 3.10** What is HGT?

## Further Reading

### Journal Articles

- Altschul, S.F., Gish, W., Miller, W. *et al.* (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Fitzpatrick, D.A., Logue, M.E., Stajich, J.E. and Butler, G. (2006) A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. *BMC Evolutionary Biology*, **6**(1), 99.
- Fitzpatrick, D.A., O’Gaora, P., Byrne, K.P. and Butler, G. (2010) Analysis of gene evolution and metabolic pathways using the *Candida* Gene Order Browser. *BMC Genomics*, **10–11**(1), 290.
- Hall, C. and Dietrich, F.S. (2007) The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics*, **177**(4), 2293–2307.
- The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**(1), 25–29.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

### Books

- Janitz, M. (2008) *Next-Generation Genome Sequencing*, Wiley-Blackwell.
- Baxevanis, A.D. and Ouellette, B.F. (2005) *Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins*, 3rd edn, Wiley-Interscience.