Swansea University
Prifysgol Abertawe

1

2 Medical School

3 Submitted to Swansea University in fulfilment of the requirements
4 for the Degree of MSc in Medical and Health Care Studies by
5 Research

6

7 Genomic interrogation of *Candida*
8 *albicans* with relation to reproductive
9 health and fertility

10

11 Luke Golby

12 2021

13

Summary

Candida albicans is a commensal yeast that can colonize a variety of host-associated niches including the human urogenital tract. It is the most common cause of fungal infections both superficial and systemic. Fungal infections, including vulvovaginal candidiasis, have been heavily implicated as a multifaceted cause in human infertility with host immune effects and microbiome alterations being other influencing factors. Previous work investigated the prevalence and diversity of a number of *Candida albicans* isolates sourced from individuals with differing fertility statuses using MLST-based methods. This current study aimed to use comparative genomic methods to investigate at whole genome level the previously described isolates in combination with database genomes to identify if genes or genetic variants display an association with the ability to colonize certain niches. Pangenome construction and enrichment analysis of database *C. albicans* assemblies showed an enrichment of virulence genes with the core genome. A genome wide association study of the Swansea isolates and a large dataset originating from NCBI's sequence read archive (SRA) identified 35 variants significantly associated with isolation from the female reproductive tract which. These variants presented enrichment for functions related to antifungal resistance and hyphal growth. Together, these variants may influence the ability for a strain to persist within the female reproductive tract and to be capable of causing recurring vulvovaginal candidiasis thus potentially influencing fertility. These results offer ideal targets for further study from a genomic perspective to explore their ecological presence within the organism's natural environment and further as targets for phenotypic investigations. The outcomes of which can be used to better our understanding of how *C. albicans* can influence reproductive health and wellbeing.

DECLARATION

This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed..... [redacted]

Date 23/9/2021

STATEMENT 1

This thesis is the result of my own investigations, except where otherwise stated. Where correction services have been used, the extent and nature of the correction is clearly marked in a footnote(s).

Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed .... [redacted]

Date 23/9/2021

STATEMENT 2

I hereby give consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed.... … [redacted]

Date 23/09/2021

88 Contents Page

Chapter 1 – General Introduction

1.1 - *Candida* Biology

1.1.1 - Cell Morphology

Candida albicans is often described as a dimorphic fungus due to its ability to grow as both as unicellular yeast and multicellular hyphal or filamentous cells, however it does have several further morphological phenotypes including opaque, gastrointestinally-induced transition (GUT) and pse udohyphal (*1, 2*). Differences between these morphologies are described in table 1.

Table 1: Description of the main *Candida albicans* morphological phenotypes, adapted from Noble, 2016.

Morphological Phenotype	Cell Shape	Unicellular / Multicellular	Special Morphological Features	Special Functions	Host Interactions
Yeast	Round-to-Oval	Unicellular	N/A	Conventional Biofilm Formation	Bloodstream Virulence, Mouth, Vagina, Skin and Gastrointestinal Tract Commensalism
Hypha	Tube	Multicellular	N/A	Thigmotropism, Conventional Biofilm Formation	Induced Endocytosis, Penetration of Host Epithelial Cells, Mouth, Vagina and Bloodstream Virulence
Pseudohypha	Elongated Ellipsoid	Multicellular	Indented Cell-Cell Junctions	Conventional Biofilm Formation	Mouth, Vagina and Bloodstream Virulence
Opaque	Ellipsoid	Unicellular	Surface Pimples	Parasexual Reproduction Competent	High Fitness in Neonatal Mouse Skin Colonization Models
GUT	Ellipsoid	Unicellular	N/A	Unknown	High Fitness in Mouse Gastrointestinal Commensalism Models

248

249 The three primary *Candida albicans* morphologies: yeast, pseduohypha and hypha have all been
250 implicated in pathogenesis however, morphogenesis to the hyphal form is required for disseminated
251 infections (*1*).

252 GUT cells are specially adapted for survival in the gastrointestinal tract, living at a relatively high
253 abundance compared to other unicellular *Candida albicans* morphologies without harming the host
254 and outcompeting other morphologies (*1*). Transition to the GUT phenotype is driven by increased
255 expression of *WOR1*, which is induced by transition through the gastrointestinal tract (*1*). The distal
256 gastrointestinal tract is depleted of glucose which is absorbed in the proximal small bowel (*3*). This
257 requires changes to *Candida albicans* metabolism to optimise utilization of the nutrients available
258 (*1*). GUT cells achieve this by altering their metabolism, with downregulation of the glucose
259 utilization pathway, and an upregulation of the N-acetylglucosamine metabolism, a glucose
260 derivative found in host mucin, and short chain fatty acid metabolism, which are produced by gut
261 bacteria following fermentation of indigestible carbohydrates, both of which are more common in
262 the distal human gastrointestinal tract (*3-5*). A decrease in iron uptake is also seen as the high
263 availability of iron throughout the large bowel and restriction helps defend against iron-related
264 toxicity (*6*).

265 **1.1.2 - White-Opaque Switching and Parasexual Cycle**

266 *Candida albicans* is usually heterozygous at the mating type locus (MTL) but through mitotic
267 recombination or through loss of one copy of chromosome 5 and duplication of the other, can yield
268 homozygous as either MTLa/a or MTLα/α (*7*). This can cause a switch from the normal 'white' yeast
269 morphology to the elongated 'opaque' morphology (*7*). White cells are seen to be spherical and have
270 bright, raised colonies while opaque cells are more elongated and have darker and flatter colonies
271 (*8*). Switching between the white and opaque phenotypes is a rare event however it can be induced
272 by environmental signals (*9*). Opaque cells are capable of parasexual reproduction with other
273 opaque cells and shows differential expression of genes compared with white cells which alters their
274 ability to colonize different niches, virulence and ability to evade the immune system (*10-12*). During
275 oropharyngeal candidiasis opaque cells are seen to be either cleared from the oropharynx or switch
276 to the white phenotype due to the opaque phenotypes inability to invade the oral epithelial cells
277 (*13*). This is potentially because of reduced expression of *ALS3*, an invasin that is required for
278 *Candida albicans* to actively penetrate epithelial cells (*13*).
279 Parasexual reproduction involves conjugation of two cells resulting in a tetraploid zygote that
280 undergoes chromosomal loss until reaching a near diploid state. This often results in high levels of

homozygosity and aneuploidy (*14*). This loss of chromosomes in a random process that produces a variety of aneuploid intermediates. In some cases this can lead to complex genome architectures and karyotypes, with simultaneous occurrence of disomic, trisomic and tetratsomic chromosomes (*15*). This process enables diversification of the genome when subjected to stress by revealing new combinations of recessive traits through loss of heterozygosity events, while this can be adverse to the individual the diversification may enable survival through the stressful conditions (*16*).

1.1.3 - Symbiotic Relationships

Candida albicans is an opportunistic fungus, capable of existing as a commensal or a pathogenic organism within a human host (*2*). Within a healthy human microbiome *C. albicans* exists as a member of a complex community of microorganisms, however when dysbiosis occurs, and C albicans becomes dominant, infection occurs (*17*). Host factors are an important component in establishing the relationship between *Candida albicans* and its host, with a host genetic susceptibility to *Candida* playing an important role (*18*). Defects in the innate immune response and cellular immunity can cause susceptibility to infection with different mutations giving susceptibility to infection in different sites (*17*). Defects in the *STAT1* (Signal transducer and activator of transcription 1) gene which codes for a transcription factor involved in mediating the cellular response to interferons, cytokines and growth factors can be linked to chronic mucocutaneous candidiasis (*19*). Of particular importance is a defect in IL-17 signalling which is essential in control of *Candida albicans* in the oropharynx and this defect is complicit in increased susceptibility to oropharyngeal candidiasis (*20*).

While host factors are an important component in this relationship a number of fungal factors, such as morphology, adhesin expression, cell wall proteins candidalysin expression can also influence the role of *C. albicans* within the host (*17*).

1.2 - *Candida albicans* Niches

1.2.1 - Female Reproductive Tract

The upper female reproductive tract consist of the fallopian tubes, uterus and endocervix while the lower reproductive tract comprises the ectocervix and the vagina (*21*).

The healthy female genital microbiome is a dynamic environment consisting of a range of microorganisms, mostly consisting of a significant number of *Lactobacillus* species (*22*).The protective role these *lactobacilli* provide is vital for sexual and reproductive health (*23*). The most common cause of change in this environment is bacterial vaginosis (BV), a condition in which *Lactobacillus* sp. are no longer dominate within the microbial community but instead a variety of obligate anaerobic bacteria usually found in small quantities become more established (*24*). The

314 second most common of infectious vaginitis is vulvovaginal candidiasis (VVC), a similar alteration in
315 the microbiome caused by *Candida* species overgrowth and inflammation (*25*). Thus, destabilisation
316 of eubiosis by means of reduced *Lactobacilli* species results in increases of local pH and reduction of
317 antimicrobial products produced by *Lactobacilli*

318 Of course, eubiosis maintenance is not limited to the microbes present, to ensure the female
319 reproductive tract can maintain a healthy microenvironment capable of allowing endogenous
320 vaginal floral growth while preventing invasion by harmful pathogens cyclic hormone levels control
321 the vaginal mucosa (*26, 27*). The presence of neutrophils also helps to protect against pathogenic
322 microorganisms (*26*).

323 **1.2.2 - Male Urogenital Tract**

324 The male urogenital tract is formed by the combination of the male urinary and reproductive
325 systems. The urinary system consists of the kidneys, bladder, ureters and urethra while the
326 reproductive tract consists of the testicles, epididymis, vas deferens, seminal vesicles, prostate
327 glands and penis.
328 Seminal fluid provides ideal conditions for survival and transport of bacteria, viruses, parasites and
329 fungi (*28-30*). Fungi within the urogenital tract can cause a variety of infections such as urethritis and
330 ulcers by species such as *Tinea corporis* and *Malassezia* species (*31, 32*). Of the fungal infections,
331 only candidiasis is considered to be sexually transmitted, while the infection is more common in
332 women, men can act as a reservoir for infection through transmission of pathogenic or AMR
333 resistant strains (*2, 30*).
334 While the lower male urogenital tract shows a diverse range of microorganisms the upper male
335 urogenital tract is typically absent of microorganisms, except in cases of infection. Sexually
336 transmitted diseases and circumcision can both impact the makeup of the lower genital tract
337 microbiome offering the environment for opportunist pathogens to cause infection (*33*).

338 **1.2.3 - Human Gastro-intestinal Tract**

339 The human gastro-intestinal tract is made up of the upper tract, the mouth, oesophagus, stomach
340 and duodenum while the lower tract consists of cecum, colon, rectum and anus.
341 Microbiota in the gastro-intestinal tract benefit the host through a range of physiological functions
342 such as shaping the intestinal epithelium, metabolism of short chain fatty acids and protection
343 against pathogens (*34-36*). They are known to have major effect on host health either causing or
344 perpetuating diseases such as….
345 The majority of microorganisms in the gastro-intestinal tract belong to the *Proteobacteria*,
346 *Firmicutes*, *Actinobacteria* and *Bacteroidetes* phyla (*37*). *Candida albicans* is also present, often as a

commensal organism and its ability to colonize is impacted by the host immune system, bacterial competitors, and their own gene expression (*38, 39*).

The gastro-intestinal mycobiome is less stable than the microbiome in general, however *Candida* species are found to be one of the more commonly detected fungi probably due to the *Candida albicans* GUT morphology being specially adapted to survival in the gastro-intestinal tract by increasing expression of genes associated with fatty acid metabolism and N-acetylglucosamine (*5, 40*). However, many microbiome studies highlight the impact of bacterial communities on gut health with the description of the role of fungi such as Candida lest often described.

1.3 - *Candida albicans* and Human Health

1.3.1 - Role as a Commensal

Candida albicans are a common member of the human mycobiome, acquired at or near birth, across a range of body sites (*41*). As an opportunistic pathogen, infection can be common, but they are also documented to provide some essential benefits, particularly in development of the immune system and protection from pathogens (*42, 43*). In order to successfully colonize a human host *Candida albicans* must be able to adhere to epithelial and mucosal surfaces, prevent strong immune responses and outcompete or co-inhabit with other members of the host's microbiome (*41*). The morphology of *Candida albicans* can be an important factor in determining whether it acts as a commensal or a pathogen, with yeast cells usually being commensal and hyphal cell types often being pathogenic, however this is not always the case (*1*). For instance, commensalism in the gut is promoted by *SFU1*, a gene part of a unique iron utilization system that restricts uptake of iron, which in the gut can reach toxic levels for *Candida albicans* (*6*). This highlights the complex network of phenotypes and of course genotypes that may determine the ability for colonisation and indeed pathogenic ability.

The role of adhesins in colonization by *Candida albicans* is still not well known, however the agglutinin-like sequence (*ALS*) gene family, a group of GPI-anchored proteins with adhesive properties, is the best studied (*41, 44*). The gene family has eight members (*ALS1-7 and ALS9*) that are expressed in both the yeast and hyphal morphologies (*41*). These adhesins have been seen to have a complex role with deletion of *ALS2* and *ALS3* causing a decrease in adhesion while deletion of *ALS5-7* caused increased adhesion (*45-47*). Alongside the *ALS* family a hypha specific adhesin, Hyphal Wall Protein 1 (*HWP1*), is also seen to be highly expressed during colonization and its deletion causes a significant decrease in virulence (*48*). Adhesive properties can be altered by changes to *Candida albican's* morphological state, with the hyphal form being both more adhesive and more virulent (*41*).

380 *Candida albicans* is believed to play a role in providing protection for the host against bacterial
381 pathogens with *Candida albicans* colonised mice being more likely to survive infection by *Clostridium*
382 *difficile* than those without *C. albicans* (*43*). However, the exact reason for this is not fully
383 understood. *Candida albicans* has also been seen to inhibit *Pseudomonas aeruginosa* pathogenicity
384 despite not hindering its ability to colonize. It has been hypothesised that this is achieved through
385 supressing expression of siderophore encoding genes (*49*).

386

387 **1.3.2 – *Candida albicans* as a Pathogen**

388 Despite displaying protective functions against bacterial pathogenesis, *C. albicans* itself can cause
389 infections. Infections are in two main categories, superficial infections of mucosal surfaces and life-
390 threatening systemic infections (*50*). The latter highlighted by documented rates of mortality as a
391 result of C albicans infection as high as 40% in those who are immunocompromised and those
392 receiving immunosuppressants (*51*).

393 Infection can often be elicited as a consequentially to the use of invasive medical devices such as
394 intravenous lines and catheters with approximately 50% of catheters used representing a site of
395 infection. These allow for many of the human defences such as the mucosal surfaces to be bypassed
396 and facilitates systemic infections through the blood stream (*51, 52*).

397 Morphology is an important factor in the ability of *Candida albicans* to act as a pathogen with
398 different morphologies showing different interactions with the host and different expression of
399 virulence factors (*1*). Hyphae in particular show unique expression of several adhesins, tissue-
400 degrading enzymes and antioxidant defence proteins (*53-56*). The increased virulence of hyphal
401 morphology compared to other morphologies make them the dominant cell type seen in superficial
402 *Candida albicans* infections such as vulvovaginal and oropharyngeal candidiasis (*57, 58*). Hyphae can
403 also penetrate epithelial cells using a combination of physical pressure and secreted enzymes while
404 the yeast morphotype can only colonize the surface of the epithelium (*57, 59*)

405 While differences in morphology are a major factor in the ability of *Candida albicans* to infect the
406 surface tissues the main three morphologies are all present in cases of disseminated candidiasis and
407 the ability to switch between these cell types is vital for virulence in bloodstream models (*60, 61*).

408 Biofilms are communities of microorganisms that often form on solid surfaces or at liquid-air
409 interfaces in a range of environments, including within humans. These communities display different
410 characteristics to planktonic cells (*52*). The *Candida albicans* biofilm consists of the yeast,

411 psuedohyphal and hyphal cell types with an extracellular matrix (*62, 63*). One of the most important
412 roles of a biofilm is to protect cells against environmental damage, both physical and chemical.
413 Within a host-pathogen context these biofilms can often be as a shield against the offensive immune
414 system (*52, 64*). Resistance to antimicrobial agents within a biofilm is another survival and
415 persistence route. Primarily due to the upregulation of two major classes efflux pumps within the
416 extracellular matrix, the ATP-binding cassette superfamily and the major facilitator class (*65, 66*), the
417 extracellular matrix also acts as a physical barrier to contribute to drug resistance (*67*). Thus, the
418 afore mentioned medically implanted devices provide ideal environments for persistence through
419 surface biofilm formation (*52*).

420 *Candida albicans* can utilise two methods to invade host cells, induced endocytosis and active
421 penetration (*68*). Induced endocytosis involves the expression of invasins that mediate binding to
422 cadherins on host cells triggering engulfment of the fungal cell by the host cell. This process has
423 been seen in multiple cell types including dead cells indicating this is a passive process (*69, 70*).
424 Active penetration is unique to hyphal cells and can occur either by invasion of the epithelial cell or
425 by passing through intercellular junctions between the epithelial cells (*70, 71*). Once through this
426 barrier systemic infection becomes a possibility.

427 **1.3.3 – Host immune response**

428 The relationship between a host and *Candida albicans* is most significantly influenced by the host's
429 innate and adaptive immune responses. Healthy immune systems can control the growth of *Candida*
430 *albicans*, however when the immune system is compromised these restrictions are removed
431 allowing for uncontrolled growth. This can lead to invasion of the mucosal surface by the hyphae
432 leading to infection, causing damage to the underlying tissue and potentially dissemination
433 throughout the entire host (*51*). A key component of the host immune-microbe relationship in a site-
434 specific manner is the gut microbiome. As previously documented, dysbiosis of the gut microbiome
435 as a result of external environmental changes (antibiotic use) or host immune changes offers
436 capacity for *C. albicans* as well as other microbes to proliferate resulting in community changes (*72*).
437 In consequence, the conventionally immune modulating microbiome, now suffering dysbiosis,
438 further perpetuates an altered immune response which in term augments pathogenesis (*72*).

439 Dendritic cells act as a link between the innate and adaptive immune responses and initiate the
440 adaptive immune response against *Candida albicans*. This is done by the presentation of antigens to
441 immature T-cells by the dendritic cells through the use of T-cell receptors (*51*). Antigens are
442 obtained after immature dendritic cells are recruited to the site of an infection by chemokines and
443 anti-microbial peptides (*73, 74*). *Candida albicans* is then recognised through interactions between

444 pattern recognition receptors (PRRs) on the dendritic cell surface and pathogen-associated
445 molecular patterns (PAMPs) present on the fungal cell wall (*75*). These PRRs recognise highly
446 conserved structures that are part of the fungal cell wall such as *N*-linked and *O*-linked mannans and
447 β-glucans (*76, 77*). After detection, the fungal cells are phagocytosed and surface proteins are
448 processed into antigenic proteins which are assembled onto major histocompatibility complex
449 (MHC) class II molecules (*51, 78*). These antigens are then both presented to memory T-cells at the
450 site of infection and to naïve T-cells in the lymph nodes. T-cell receptors can interact with the MHC
451 class II and through the secretion of cytokines lead to activation and differentiation of the naïve T-
452 cells into specialised T-cells (*51*). Different dendritic cell subsets have been seen to cause different
453 T-cell responses with Langerhans cells promoting a Th17 response but not a $CD8^+$ T-cell response,
454 while $Langerin^+$ cells cause a Th1 and $CD8^+$ response while inhibiting a Th17 response, resulting in
455 mixed dendritic cell populations giving a non-redundant response (*51, 79*).

456 $CD4^+$ and $CD8^+$ T-cells are both involved in the immune response to *Candida albicans* after activation
457 by dendritic cells (*51*). While $CD8^+$ T-cells are shown to inhibit *Candida albicans* hyphal growth the
458 main mechanism in which the adaptive immune system responds is through $CD4^+$ T-cells generating
459 a T-helper response (*51, 80*). The importance of this response is seen in HIV/AIDS patients commonly
460 develop oropharyngeal candidiasis because of the lack of $CD4^+$ T-cell protection (*81*). There are four
461 subsets of T-helper cells, Th1, Th2, Th17 and Treg, and the subset is dictated by the
462 microenvironment and the cytokines present when dendritic cells interact with the naïve $CD4^+$ T-
463 cells. Th1 and Th17 responses are the most important in fungal clearance at the mucosal surfaces
464 while the Th2 response is more associated with increased growth and dissemination of *Candida*
465 *albicans* (*51*).

466 **1.4 - Infertility**

467 Infertility is the inability to establish a clinical pregnancy after 12 months of regular, unprotected
468 sexual intercourse or the impairment of an individual to reproduce. Infertility is defined as occurring
469 over a restricted time period while sterility is a permeant state of infertility. Infertility can be
470 primary, an individual who has never been involved in a clinical pregnancy and is classified as
471 infertile, and secondary, where the individual meets the criteria of infertility but has previously been
472 involved in a clinical pregnancy (*82*).

473 Infertility effects approximately 8-12% of couples worldwide with rates varying based on gender and
474 geographical region (*83*). Infertility prevalence varies around the world with rates varying from 14%
475 in developed western countries to 30% in some developing regions in women (*84*). Men contribute
476 to 50% of infertility cases and are wholly responsible for 20-30% of infertility cases (*85*). Secondary

477 infertility is the most common form of infertility, particularly in the developing world where there
478 are high rates of unsafe abortions and poor maternity care which result in reproductive tract
479 infections (*86, 87*).

480 Infertility is effected by three major factors, time of unwanted non-conception, the age of the
481 female partner and disease-related infertility (*88*). Disease related infertility can affect either gender
482 or be gender specific, a list of the diseases that affect fertility is shown in table 2.

Table 2: Lists of diseases that effect fertility and the gender they effect. Adapted from Borght and Wyns, 2018.

Disease related infertility in both genders	Disease related infertility in women	Disease related infertility in men
Hypogonadotrophic hypogonadism	Premature ovarian insufficiency	Testicular deficiency
Hyperprolactinemia	Polycystic ovary syndrome	Post-testicular impairment
Disorders of ciliary function	Endometriosis	
Cystic fibrosis	Uterine fibroids	
Infection	Endometrial polyps	
Systemic conditions		
Lifestyle related factors		

483

484 **1.4.1 - Female Genital Tract Infections and Infertility**

485 The healthy vaginal microbiome is primarily dominated by *Lactobacillus* species however when the
486 vaginal microbiome is altered it can lead to a range of adverse conditions including infertility (*89*).
487 The most common cause of change in this environment is bacterial vaginosis (BV), a condition in
488 which *Lactobacillius* sp. are no longer dominate but instead a number of anaerobic bacteria usually
489 found in small quantities become more established (*24*). The second most common of infectious
490 vaginitis is vulvovaginal candidiasis (VVC), a similar alteration in the microbiome caused by *Candida*
491 species overgrowth and inflammation (*25*). Previous research has already established a link between
492 abnormal vaginal microflora, BV and infertility (*90, 91*). A possible correlation between the absence
493 of BV infections and pregnancy was also observed. It was theorised that this could be due to
494 pregnancy acting as a protective factor due to hormonal changes favouring *Lactobacilli* colonization,
495 or that rate of conception is higher in the absence of infection (*90*), potentially a combination of the
496 two. Further, it has been demonstrated that a non-*Lactobacillus* dominated vaginal microbiome is
497 associated with significant decreases in both pregnancy chance and healthy pregnancies (*92*).

498 Female mice models have shown that *Candida albicans* present in the reproductive tract caused
499 sperm agglutination and immobilization while also not causing any signs of histopathological
500 changes to the reproductive organs (*93*).

501 **1.4.2 - Male Genital Tract Infections and Infertility**

502 Microbes have been shown to have a major effect on sperm function through a range of
503 mechanisms including sperm-bacteria cellular interactions leading to agglutination, secretion of
504 bacterial membrane proteins that alter motility and sperm ultrastructure and in the production of
505 reactive oxygen species (*94-96*). *Candida* infection of the urogenital tracts has been shown to cause
506 male infertility through a decrease in spermatozoa motility and azoospermia (*97, 98*). Male fertility
507 has also been negatively affected by *C. albicans* specifically through morphological damage to
508 spermatozoa, most commonly through breakdown of the acrosome leading to its complete loss (*99,*
509 *100*). Aggregation of spermatozoa have also been observed when incubated with *C. albicans in vitro*
510 (*30*). This is possibly due to mannan on the surface of many *Candida* species that mainly consists of
511 mannose residues. Spermatozoa have a corresponding receptor for this carbohydrate offering a
512 binding ability which negatively affects seminal parameters (*30*). Infections of the male reproductive
513 tract are shown to be associated with a decreased reproductive capacity and the severity of this
514 impact is often variable as the reproductive microbiome is dynamic (*101*). This highlights the role of
515 both partners in fertility and brings to focus how the reproductive tract microbiota could be
516 visioned.

517 **1.5 - Vulvovaginal Candidiasis**

518 **1.5.1 - Prevalence and Symptoms**

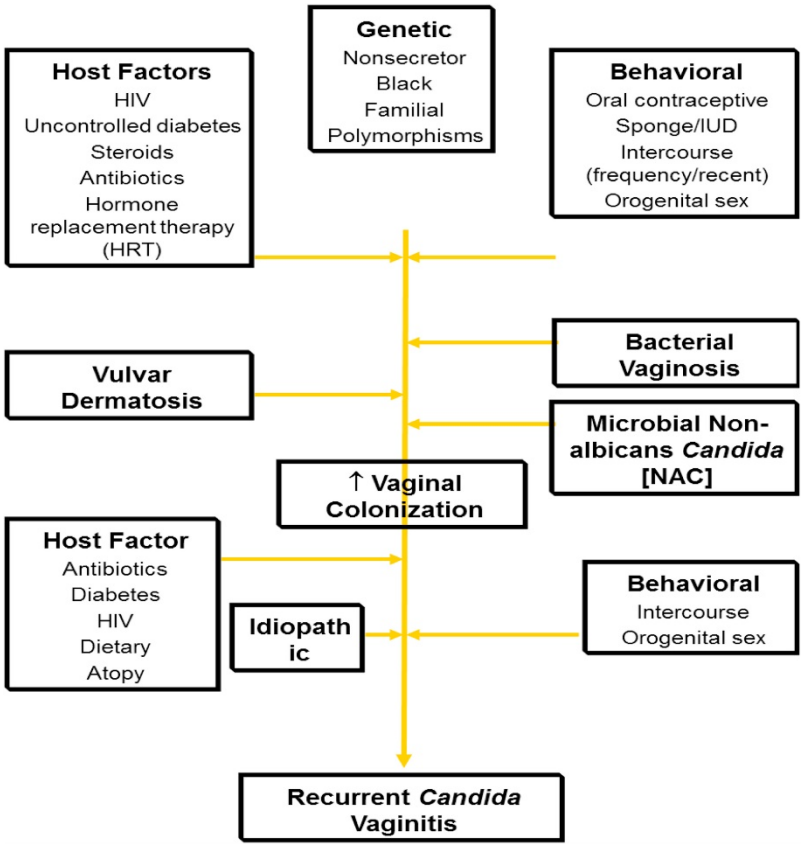
519 Approximately 70% of women will suffer from VVC in their lifetimes with 50% suffering from a
520 second infection after anti-fungal treatment, usually by the same strain as the original infection
521 (*102*). There have also been cases of recurrent episodes of VVC separated by asymptomatic periods
522 known as recurrent vulvovaginal candidiasis (RVVC) which effects approximately 138 million women
523 worldwide each year (*18, 103*). *Candida albicans* is the most common species seen in VVC being
524 responsible for over 90% of acute cases of VVC and 85-90% of cases of RVVC (*25, 102*). To be defined
525 as recurrent vulvovaginal candidiasis an individual must have 3 or more acute episodes of VVC within
526 a 12 month period (*18*). In some women asymptomatic colonization of *Candida* can occur for years
527 while symbiosis with the vaginal microbiome is maintained. Only on a breakdown of this
528 relationships will acute symptomatic VVC occur (*18*), the reasoning for the dysbiosis can be varied.

529 Symptoms of VVC can include an odourless white vaginal discharge, itching and irritation around the
530 vagina and soreness or stinging during urination and sexual intercourse (*104*). These symptoms are
531 often most prominent just before the menstrual period and often a history of similar symptoms.
532 These symptoms can often overlap with the symptoms of other common reproductive tract
533 infections so additional testing, such as pH testing or a whiff test, are required to confirm a diagnosis
534 (*105, 106*).

535 **1.5.2 - Predisposing Factors of VVC and RVVC**

536 *Candida* blastospores will migrate into the vestibule and vagina from the lower gastrointestinal tract
537 and adhere to the vaginal epithelial cells. While this usually happens in low numbers and the
538 *Candida* exist in symbiosis with the vaginal microbiota, a breakdown in this relationship results in
539 acute VVC (*18*). Increased *Candida* colonization is one factor in increased susceptibility to RVVC with
540 a number of host factors, as shown in Figure 1 also influencing the risk of developing RVVC (*18*).

541
542
543
544
545
546
547
548
549
550
551
552
553
554
555



556 Figure 1 - Host, behavioural and genetic factors that increase the risk of developing recurrent vulvovaginal candidasis. Image obtained from Sobel, 2016.

557 RVVC can be categorised as either primary or secondary RVVC. Primary RVVC refers to cases of RVVC
558 where secondary factors are not apparent and genetic factors are believed to be the most dominant
559 causal factor. Secondary RVVC can be linked to other triggerable factors however in most cases
560 there will still be a link to genetic influences (*18*). These genetic influences are seen with the links

561 between increased RVVC susceptibility and ethnicity and blood group (*107, 108*). Links have also
562 been seen between several genes and RVVC such as dectin and *TLR2* although the reason for
563 increased susceptibility is unknown and probably due to the high level of confounding factors (*18,*
564 *109*).

565 **1.5.3 - Treatment Options**

566 While VVC can be treated there is no guaranteed permanent cure due to the influence of genetic
567 factors causing increased susceptibility to recurrent infection. The current method for reducing the
568 risk of RVVC where there is no secondary stimuli triggering infections is a long-term course of
569 suppressive anti-fungal agents, usually a regime of fluconazole (*18, 110, 111*). Where other treatable
570 risk factors are identified, such as use of an oral contraceptive, this can be treated in tandem with
571 antifungal agents however, the efficacy of these additional steps is mostly anecdotal (*112*).

572 Currently the RVVC fluconazole treatment involves 150mg every 72 hours over 3 doses followed by a
573 weekly 150mg dose for a period of at least 6 months. This therapy has shown to be safe, affordable
574 and effective, with episodes of symptomatic vaginitis being approximately 5%. Ending this regime
575 will result in a return of RVVC in the first 4 months in up to 50% of individuals (*110, 111*). Of course, a
576 major factor in the success of AM treatment is the sensitivity profile of the infection causing *Candida*
577 species whereby resistance and the mechanism of which is currently well known.

578 **1.6 - Antimicrobial Resistance**

579 **1.6.1 - Triazole Action and Resistance**

580 Triazoles are heterocyclic compounds consisting of a five membered ring of two carbon atoms and
581 three nitrogen atoms, the most commonly used in anti-fungal treatments is fluconazole (*113*)
582 Fluconazole is a first-generation triazole derived from ketoconazole, another prominent anti-fungal
583 agent in animal models. Fluconazole was first synthesized as a replacement for imidazole which
584 would be better metabolized and water soluble (*114*). Fluconazole is one of the most commonly
585 used anti-fungal agents however it is limited in its range of action compared to other anti-fungals,
586 with it primarily being active against yeasts but inactive against filamentous fungi. It is very
587 commonly used to treat oropharyngeal, oesophageal and disseminates candidiasis (*115*). The
588 structure of fluconazole is shown in Figure 2.

589

590

591

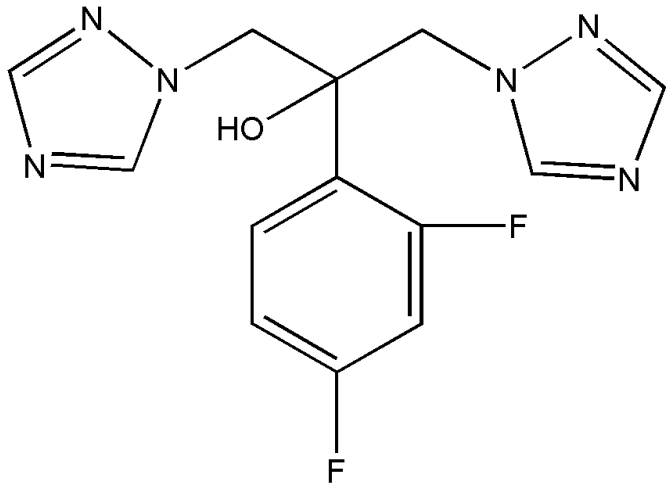


Figure 2: The structure of fluconazole, a synthetic triazole consisting of two triazole groups at position 1 and 3 and by a diflufophenyl group at position 2. Image adapted from PubChem using Chemdraw.

Triazoles act by inhibiting cytochrome P450, a haemoprotein involved in drug metabolism. Within fungi, 14-α sterol demethylase (CYP51) is the main target for triazole action, preventing synthesis of ergosterol, an important component in the fungal cell membrane (*116, 117*). This leads to an accumulation of methylsterols in the fungal cell membrane resulting in either cell death or inhibition of growth (*115, 118*).

Resistance to triazoles can occur due to modifications to the CYP51 gene or by independent mechanisms (*119*). Changes to the CYP51 gene usually involve point mutations that cause amino acid substitutions, resulting in reducing access by the triazoles, such as substitutions to M220 and G54, or reducing the ability of the triazole to bind to the haem, as seen in the G448S substitution (*120, 121*). The CYP51 promoter can also contain variant sites resulting in increased expression, this is usually accompanied by point mutations within the coding region of the gene itself. Both a 34bp tandem repeat in the promoter and a L98H amino acid substitution and a 46bp tandem repeat with multiple point mutations resulting in multiple amino acid substitutions (Y121F and T289A) (*122, 123*). CYP51-independent mechanisms to resistance have also been seen with upregulation of ABC transporters showing an association with azole resistance (*124*). In *Candida albicans* upregulation of the *CDR1* and *CDR2* genes by *TAC1* showed an association with resistance to multiple antifungal drugs including azole resistance (*124*).

1.6.2 - Amphotericin B Action and Resistance

Amphotericin B (AmB) is a polyene antimicrobial synthesized in the polyketide biosynthetic pathway in bacteria naturally but it can also be produced synthetically (*125*). Its structure is shown in Figure 3.

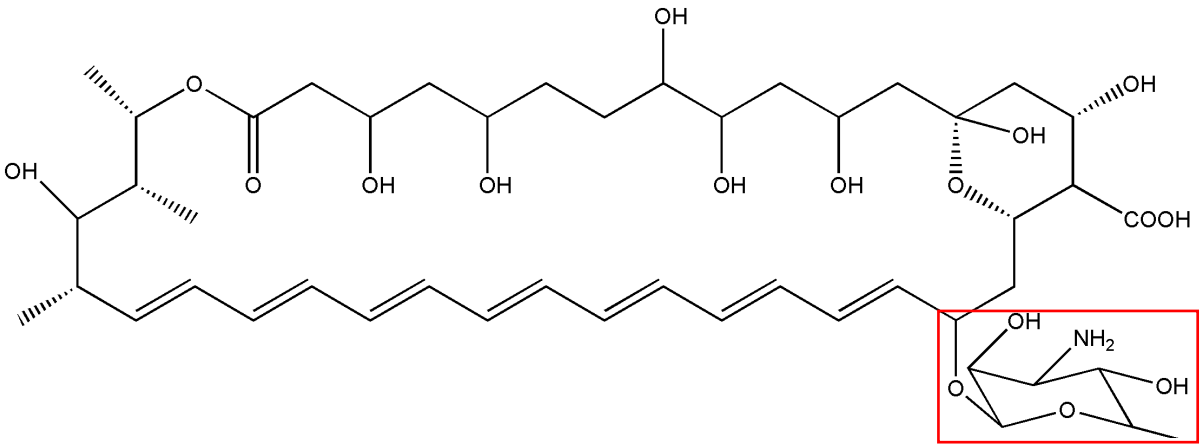


Figure 3: The structure of amphotericin B, a polyene structure with a mycosamine group that binds to ergosterol allowing for the formation of pores in fungal membranes. This group is highlighted by the red box. Image adapted from PubChem using Chemdraw.

AmB is used in treating invasive and systemic fungal infections including those of *Candida* as well as cases of leishmaniases *(125).* AmB has two effects on fungal cells, binding to sterols, primarily ergosterol, and induction of oxidative damage (*126*). AmB binds to sterols in the fungal cell wall through its hydrophobic domain allowing multimeric pores to be formed. These pores allow the movement of small cations into the fungal cell causing a depletion of intracellular ions, leading to cell death (*126*). AmB also has an effect simply by sequestering ergestorol and preventing it from being used in its various processes (endocytosis, vacuole fusion and cell membrane protein stabilisation), enhancing its fungicidal activity through this multi-mode of action (*127, 128*).

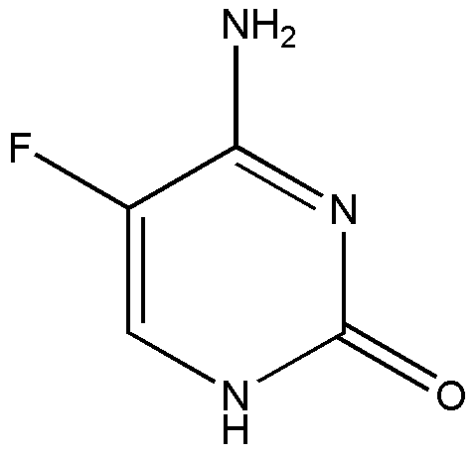
AmB has also shown to have fungicidal activity when its ability to form pores is impaired suggesting a different mechanism for killing fungal cells (*129*). This is through the induction of oxidative damage through the direct production of free radicals (*130*). The mechanism by which AmB oxidative damage causes an antifungal effect is unknown however it has been shown that AmB can act both as an auto oxidizer, but it can also act as an antioxidant (*131, 132*).

Resistance to AmB can come about due to decreases in ergosterol content and a build-up of sterol intermediates as well as changes to the fungal cell wall of cells in biofilms (*133, 134*). This can cause cross resistance with azoles as well as they both act on ergosterol (*135*). Resistance can also come about after exposure to fluconazole that gives resistance to oxidative stress (*136*). This resistance is associated with increased expression of *ERG* genes, stress genes and decreased expression of mitochondrial enzymes, indicating that AmB resistance could be associated with decreased mitochondrial activity and reactive oxygen species production (*137*).

1.6.3 - 5-Flucytosine Action and Resistance

655 5-Flucytosine (5-FC) is a synthetic organofluorine with a 5 substituted fluorine that is metabolised in
656 the pyrimidine salvage pathway (*138*). Its structure is shown in Figure 4.

657
658
659
660
661
662
663
664



665
666 Figure 4: Structure of 5-Fluctyosine, an organofluorine compound with a fluorine substituted at position 5. Deamination of the compound when taken in by the cells is required for antifungal activity. Image adapted from PubChem using Chemdraw.

667

668 5-FC is the standard antifungal drug of use alongside amphotericin B in treatment of cryptococcal
669 meningitis and in invasive and life-threatening *Candida* infections (*138*). While 5-FC has no antifingal
670 activity itself when taken in by a cell and metabolised into 5-flurouracil (5-FU) which has two
671 mechanisms of antifungal activity (*138, 139*). The first of these mechanisms involves the conversion
672 of 5-FU into 5-fluorouridine triphosphate (FUTP) which then replaces uridylic acid within fungal tRNA
673 and inhibiting protein synthesis (*138*). The other mechanism involved the conversion of 5-FU into 5-
674 fluorodeoxyuridine monophosphate (FdUMP) by uridine monophosphate pyrophosphorylase.
675 FdUMP is an inhibitor of thymidylate synthetase, an important enzyme in the synthesis of thymidine,
676 leading to inhibition of DNA synthesis (*139*).

677 Two mechanism for 5-FC resistance have been identified, a decrease in uptake and metabolism of 5-
678 FU due to mutations causing deficiencies in the required enzymes and through increased production
679 of pyrimidines that compete with the 5-FC metabolites, reducing their effect (*140, 141*) Of these the
680 most commonly seen form of acquired resistance is due to defective uridine monophosphate
681 pyrophosphorylase preventing the conversion of 5-FU into FdUMP (*142*).

682 **1.7 - *Candida albicans* Genome**

683 **1.7.1 - Description of the Genome and Karyotype**

684 *Candida albicans* is a diploid organism with 8 pairs of chromosomes ranging in size from 1 to 3.5
685 megabases with an overall haploid genome size of approximately 15 Mb and average GC content of

686 approximately 33.5%, depending on strain (*143, 144*). The size, centromere location and percentage
687 of genome of each chromosome is shown in table 3. *Candida* is a member of the highly diverse
688 *Saccharomycotina* sub-phylum. *Saccharomycotina* consists of eight major clades of which *C. albicans*
689 belongs to the CTG clade, a selection of organisms in which the CUG codon codes for serine in place
690 of leucine (*145*).

Table 3: Chromosome-specific information for *Candida albicans.* Adapted from (Jones,2004; van het Hoog,2007)

Chromosome	Approximate Size (Mbases)	Centromere Location (bp)	Percentage of Genome
1	3.1	1,561,879	21
2	2.3	1,924,378	16
3	1.8	816,770	12
4	1.7	1,000,800	12
5	1.2	465,800	8
6	1.1	975,879	7
7	1	423,765	7
R	2.5	1,748,965	17

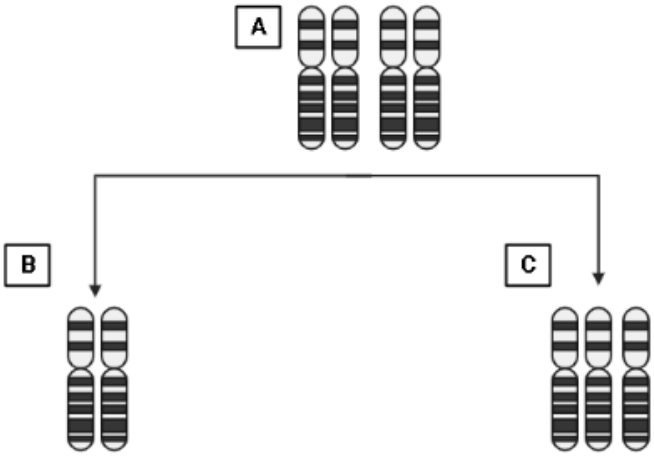
691 **1.7.2 – Aneuploidy in *C. albicans***

692 *C. albicans* shows a high degree of genome plasticity and karyotypic variability with chromosomal
693 rearrangements commonly occurring in response to a variety of stresses. Copy number variation
694 (CNV) has been detected for small chromosomal regions such as individual genes, or indeed the
695 entire chromosome, the latter of which is referred to as aneuploidy (*14, 143, 146*). In *C. albicans*
696 aneuploidy has been detected in all eight chromosomes, arising due to errors in DNA replication or
697 in the replication machinery (*14*). While aneuploidy is usually a disadvantage for an individual, under
698 cellular stress it can confer an advantage, as is commonly seen in response to anti-fungal agents in
699 which case trisomy of a chromosome is most described (Figure 3) (*147*). However, little has been
700 documented that relates to occurrence in any linage or niche. The rational for a selective advantage
701 caused by aneuploidy has been suggested to relate to gene copy number whereby an increase in
702 copy number of genes coding for products related to managing stress or resistance responses means
703 a proportional increase in expression of these genes (*14*)

704 Monosomy of chromosomes only causes a twofold decrease in gene transcription in approximately
705 15% of genes while approximately 40% of genes are maintained at the disomic level indicating a

706 cellular mechanism for transcription homeostasis (*148*). A small number of genes can even be
707 excessively upregulated in monosomy. Similar regulation of transcription is also in trisomy with 25%
708 of trisomic genes only being expressed at the disomic levels (*149*).

709
710
711
712
713
714
715
716



717
718 Figure 5: Example of aneuploidy and how it occurs in *Candida albicans*. (A) After sexual reproduction the organism is tetraploid and will undergo random chromosome loss until diploid. (B) In normal conditions this occurs successfully, and the organism will only have two copies of each chromosome. (C) Under stressful conditions or due to errors in the replication
719 machinery some chromosomes will only lose one of the extra copies leaving the organism triploid for one chromosome. Image created with BioRender.com
720

721 **1.7.3 - Variants and Large-Scale Mutations**

722 Single nucleotide variants SNVs have been shown to occur at a rate of approximately 0.3% in the
723 standard laboratory strain SC5314 while other strains have showed frequencies ranging from 0.5% in
724 isolates from the same clade and 1.1% in isolates from different clades (*150, 151*). A total of 89% of
725 variants were located within intragenic regions with the capability to alter the protein sequence or
726 expression levels (*152*). Emergent SNPs and indels have also been shown to cluster together within
727 the *C. albicans* genome within 10bp of each other suggesting recombination is at play (*151*). A 2:1
728 transition to transversion (Ti/Tv) ratio has been observed within the reference strain (*153*)
729 highlighting no abnormal ratio compared to other species and a selection for non-deleterious
730 transition mutations (single ring to single ring) that would less likely change an encoded amino acid.

731 **1.7.4 - Population Structure**

732 MLST analysis of *Candida albicans* has identified 18 major and minor clades. These clades showed a
733 significant statistical association with anatomical source and ABC (ATP binding cassette) genotype,
734 emphasising a relationship between clades and virulence factors within *C. albicans*, as well as with
735 geographic location(*154, 155*). The full list of clades and their specific features is shown in table 3.
736 Studies have shown that clade 8 is enriched with isolates sourced from other animals while clade 1
737 does not show any non-primate sourced isolated (*156, 157*). Clade 1 isolates are the most commonly

738 sourced from humans suggesting they are specially adapted for human colonization and infection
739 (*156*) however little has been documented to explain the nature of these adaptations for either
740 genotype or phenotype. All the currently known clades are listed in table 4.

Table 4: List of identified clades within the *Candida albicans* population as well as primary isolation source, primary geographic locations, ABC genotype and known anti-fungal resistances. Adapted from McManus, 2014.

Clade	Origin of Isolates	ABC Genotype	Geographical Enrichment	Reduced Susceptibility
1	Superficial Infections, Vaginal Infections, Oral Tract	A	Worldwide	Fluconazole, 5-Flurocytosine, Terbafine
2	Bloodstream	A	UK	
3	Oral Tract	B	USA	Fluconazole
4	Blood Stream	B/C	Middle East, Africa	Amphotericin B
5	Oral Tract	B/C	Europe	
6	Oral Tract	B	UK	Fluconazole
7	NA	A	South America	
8	Bloodstream, Animals	A/B	Europe	
9	NA	A	Europe	
10	NA	B	Europe	
11	NA	A/C	Europe	
12	NA	B/C	Europe	
13	Vaginal	A	Africa	
14	NA	B	Asia	
15	NA	B/C	Asia	
16	NA	B	Asia	
17	NA	A/B	Asia	
18	Dyspeptic Patients	NA	Asia	

741

742 While approximately 4% of the *C. albicans* genome is heterozygous a range of adverse
743 environmental conditions (antifungal agent exposure, ultraviolet light and oxidative stress) can
744 trigger loss of heterozygosity (LOH) events (*158, 159*). Host-associated isolates are shown to be a
745 much more heterogenous population compared to clonal *in vitro* isolates. This is possibly due to the
746 effects of commensal carriage or due to the wide range of genotoxic stress presented to host-
747 associated isolates, such as environmental changes or interaction with immune cells (*160*). Within an
748 individual a genetically heterogenous *C. albicans* population can be identified with the isolates
749 differing due to LOH events. However, MLST is insufficient to identify this heterogeneity except in
750 occurrences where the LOH event effects one of the MLST loci. MLST is also insufficient in identifying
751 the level of diversity between samples isolated from the same host due to its low discriminatory
752 power. Instead a genome wide analysis is required to fully analyse any potential diversity (*160*).
753 Specifically, this would require substantial resources to sequence and identify, to strain level, the
754 mycobiome. A feat that will, with the advance of sequencing technologies and informatic tools,
755 become far mor accessible in the not-too-distant future.

756 **1.7.5 - Reference sequence**

757 *Candida albicans* SC5314 is the commonly used wild-type control strain from which most of the
758 other common laboratory strains are derived from. The strain was originally isolated from a patient
759 with a generalised *Candida* infection (*161*). The first whole genome assembly of *Candida albicans*
760 SC5314 (Assembly 19) was carried out using shotgun sequencing of the heterozygous diploid
761 genome. This consisted of 412 supercontigs with 266 representing a haploid set, representing 86.5%
762 coverage of the genome and gave an overall haploid genome size of 14.855 megabases.
763 Chromosome size and coverage was determined using the number and distribution of markers on a
764 physical map (*153*).

765 183 supercontigs from assembly 19 were organised into 8 chromosomes using pulse-field gel
766 electrophoresis and hybridization to DNA microarrays to construct assembly 21. Assembly 21 is a
767 haploid assembly with a size of 15.845 megabases in size. In regions where heterozygosity was
768 present in assembly 19 two allelic contigs were assembled (*159*). Centromeres were able to be
769 located in this assembly using sequences identified in Sanyal *et al*. and showed that chromosomes 2
770 and 6 were acrocentric while the other chromosomes were metacentric (*159, 162*).

771 Muzzey *et al* assembled a phased diploid assembly of the SC5314 strain using next-generation
772 sequencing. Variants were identified by mapping reads to assembly 21 and allowing for 3
773 mismatches using Bowtie while INDELS were aligned using a window method and BWA. A phased
774 assembly allowed for better investigation of variants and INDELs as well as seeing allele-specific

775 effects, such as allele-specific mRNA expression (*150*). The current reference sequence has been
776 assembled using short reads and no long read sequencing has been used for the reference genome.

777 **1.7.6 – Genome Assemblies**

778 Most available assemblies on NCBI are haploid with diploid assemblies only available for the
779 reference sequence. The current diploid assemblies held were sequenced using Illumina and used a
780 reference guided assembly method for a chromosome level assembly. Mitochondrial DNA was also
781 sequenced for these assemblies.

782 Variant phasing is an *in silico* approach that allows for identification of alleles on homologous
783 chromosomes, rather than having a single consensus sequence; combining sequences of the two
784 chromosomes into a single representative sequence. An approach that loses this information, while
785 maintaining a haploid genome. Phased genomes can be assembled in a number of ways including
786 sequencing of an individual's parental genomes and observing which of the parent's alleles are
787 inherited, using a reference population's haplotype information and computational methods that
788 extract physical linkage information (*163*). However, by far the most recent approaches use the high-
789 quality short read sequencing data such as Illumina, while combined with long read third generation
790 platforms such as PacBio or Oxford Nanopore (ONP) to yield chromosomal scale phased dipoid
791 representatives of the genome (*164*). These genomes remove the ambiguity that can be generated
792 with the use of IUPAC codes within assemblies. Indeed, a phasing diploid assembly approach
793 remains currently limited in application with ~ 50 entries from long read platforms publicly held
794 within NCBI's SRA. This compares to over 1000 *C. albicans* based Illumina datasets (website accessed
795 Jan 2020).

796 **1.7.7 – Investigation of the Genome**

797 In addition to the genome sequence of an organism, how the genome is shaped, protected and
798 expressed also provides valid information on the function of an organism at any given time or within
799 any given environment. Indeed, epigenetic traits have been postulated to have a role in adaptation
800 and pathogenicity even at the mitochondrial genome level (*165*). Epigenetic modifications are
801 important regulators in altering the phenotype of *Candida albicans*, these can be studied using next-
802 generation sequencing methods to examine base methylation and ChIP-seq to study histone
803 modifications (*166, 167*). A prime example of chromosomal epigenetic regulation in *Candida*
804 *albicans* is regulation of the white-opaque system which involves 8 transcription factors that
805 regulate their own expression and each other's (*168, 169*). Histone modification in *C. albicans* is also
806 an important factor in determining persistence during the host's immune response, with histone

807 acetylation being involved in the response to oxidative stress. ChIP-Seq was used to investigate
808 *Ada2*, a histone acetyltransferase, which was associated with over 200 stress response genes and its
809 loss caused increased sensitivity to oxidative stressors (*170, 171*)

810 Assay for transposase-accessible chromatin during sequencing (ATAC-seq), a method of assessing
811 genome-wide chromatin accessibility using transposases and next-generation sequencing, has been
812 used to investigate chromatin availability and gene expression in response to oxidative stress in *C.*
813 *albicans*. This showed changes to chromatin accessibility in regulatory regions upstream of coding
814 sequences thus changes to transcriptional activity of downstream genes during oxidative stress was
815 observed (*172*). Again, publicly available databases such as the SRA contain predominantly
816 transcriptomic experiments with only few that explore epigenomic methods.

817 **1.8 - Pangenomes**

818 **1.8.1 - Pangenome Concept**

819 The pangenome concept was first described by Tettelin *et al* to describe the genome of a taxa, with
820 genes initially being classified as core genes or accessory genes(*173, 174*) and initially used to species
821 level. Core genes are those present in the majority of members of a pangenome and often encode
822 products that are essential for function, accessory genes are present in a small number of members
823 of the taxa and are genes that encode for supplementary functions that contribute to diversity and
824 confer selective advantages, such as genes involved in antimicrobial resistance (*174*). The
825 combination of these core and accessory genes across the taxa of interest make up the pangenome
826 (*173*). The overall aim of this approach is to better understand genome dynamics of an organism,
827 identifying associations between a gene and a phenotype and how genes may spread within a
828 population.

829 The first reported pangenome was created using 8 *Streptococcus agalactiae* draft genomes (*174*).
830 Since then, the number and quality of bacterial genomes has increased significantly further insights
831 into species level pangenomes. Pangenomes have also been studied at higher levels of taxonomy
832 such as the super kingdom *Eubacteria* pangenome which used 573 bacterial genomes which
833 determined that bacteria have an open pangenome, that bacteria as a whole have extremely large
834 gene pool encoding a large number of proteins (*175*).

835 While prokaryote pangenomes have been well studied, eukaryotic pangenomes are less common
836 due to increased costs for whole genome sequencing and the resources required to analyse
837 eukaryotic genomes (*176*) Eukaryotic genomes show less intraspecies variation due to decreased
838 rates of horizontal gene transfer (*177*). Eukaryotic pangenome studies have shown some diversity

839 however with *Glycine soja* showing an 80:20 core/accessory genome split and the fungi
840 *Zymoseptoria tritici* having an accessory genome accounting for 40% of the total pangenome (*178,*
841 *179*).

842 **1.8.2 - Pangenome Applications**

843 Pangenomes can be used in genome alignment instead of using a single reference sequence. This
844 has many benefits over single reference sequence methods including removal of bias towards highly
845 conserved sequences in the reference and improving mapping results by including known
846 polymorphisms in mapping (*180*).

847 Another benefit of pangenome construction is the ability to reconstruct a phylogeny specific to the
848 group of isolates understudy using core genes only. Traditional phylogeny construction methods
849 have historically used a small number of highly conserved genes for analyses, however, when
850 dealing with more complex population structures this approach can be less accurate (*181*). Further,
851 aligning a pangenome with phylogenetic reconstruction allows for a better understanding of
852 bacterial genome relationships in which there is a high rate of DNA exchange (*181, 182*).

853 Successful application of pangenomic analysis using a multitude of construction techniques has
854 driven this largely prokaryote to be employed in the study of eukaryotes. Whilst early examples have
855 been based on eukaryotes with smaller genomes (such as yeast), recent advances in both
856 sequencing and computational support have led to its use to study plants (*183*) and insects (*184*).
857 Further, recent endeavours have led to the formation of Human pangenome project
858 (https://humanpangenome.org) which aims to create a unified genomic representation of the
859 human genome at a *Homo sapiens* species level employing a pangenome approach. Just like in
860 prokaryotes, this approach to genomics has already enabled better formation of genome references
861 so that they can be used to study human evolution (*185*) and disease (*186, 187*).

862 **1.9 - Recent Postgraduate Research into *Candida albicans***

863 Previous postgraduate research (N Alharbi. (2019)) investigated the prevalence and diversity of
864 *Candida* species isolated from 20 couples of differing fertility status with the overall aim to
865 investigate *Candida* colonisation and its effect on fertility status. The study detected that that mainly
866 *C. albicans* were recovered (by cultural methods) from infertile females in this study. Further most
867 women reporting VVC or RVVC had the same diploid strain type (DST) colonising both reproductive
868 tract and oral cavity. Whilst this research found that the DSTs identified were specific at participant
869 and a couple level, low level diversity across isolates was observed per couple and indeed per host,
870 highlighting potential diversity within the *C. albicans* population at a partnership level. Further

871 phenotypic analyses were conducted to profile anti-fungal resistance and immune responses per
872 host using the reference *C. albicans* SC5314 strain. Cytokine response of participants was explored *in*
873 *vitro* using the SC5314 strain to identify that IL-6 and IL-17 were reduced within the infertile female
874 participants.

875 Despite some very interesting phenotypic observations, sequence-based analysis was limited to an
876 MLST approach. With whole genome sequence reads available for the entire collection of isolates
877 presented within this study, high-resolution genome comparisons have yet to be conducted. In light
878 of this, comparisons of these isolates at a whole genome level may offer further insight into the
879 relationship between *Candida albicans* carriage and fertility status.

880 **1.10 – Genome Wide Association Studies**

881 Genome wide association studies involve the testing of variants across a genome to identify
882 genotype-phenotype associations (V et al., 2019). Variants across the genome can be
883 associated with a range of phenotypes such as drug susceptibility, niche colonisation and
884 disease susceptibility (V et al., 2019). After identification of these associated variants lab-
885 based methods are used to validate these relationships and can provide biomarkers, drug
886 targets and sites for genetic manipulation (E and G, 2020). Population stratification can
887 cause false positive associations in GWAS studies, this can be corrected for using principal
888 component analysis (PCA) (AL et al., 2006). In GWAS PCA models ancestry differences across
889 continuous axes of variation and corrects for false positives arising due to ancestral
890 populations (AL et al., 2006). In highly structured populations where genetic ancestry itself
891 can correct for any population effects can be used to compute any association statistics.
892 More complex models use mixed models where multiple covariate principal components to
893 contribute to the overall statistical power (AL et al., 2010). The power and resolution of
894 GWAS results is largely influenced by linkage disequilibrium. Linkage disequilibrium is the
895 non-random association of alleles at two loci and can be affected by mutation,
896 recombination, genetic drift and other factors (MA et al., 2011).

897 **1.11 - Aims and Hypotheses**

898 Within this project the aim is to expand upon the previous post graduate study exploring the link
899 between *Candida albicans* colonization and reproductive health leveraging higher resolution
900 genome analyses and making comparisons with sequence information held within publicly accessible
901 databases. Broadly, the hypothesis that will be explored within this research is

902 that *Candida albicans* possesses genetic features that are adaptive for
903 colonisation, persistence, and pathogenicity within the reproductive tracts of humans. Further that
904 colonisation of the reproductive tract by *Candida albicans* possessing pathogenic genetic traits may
905 negatively affect fertility.
906 In order to explore the hypothesis outlined above several finer aims have been identified that will
907 allow this exploration.

908 - Aim 1: Extend Phenotypic Analysis of Isolates *in vitro*
909 - Generation of a sensitivity profile of all collected strains exposed to hydrogen
910 peroxide stress and to calcofluor.
911 - Aim 2: Generation of Chromosome Scale Phased Diploid Genome Assemblies
912 - Extraction of high molecular weight DNA (HMW-DNA for the purpose
913 of generating long sequence reads using an Oxford Nanopore minION. Long
914 reads will be combined with short Illumina reads already produced to create
915 hybrid assemblies allowing for finer scaled genomic analyses specifically in for
916 large scale chromosomal rearrangement approaches.
917 - Aim 3: Generation of a *Candida albicans* Pangenome
918 - Pangenome construction using genome assemblies generated here will
919 allow presence/absence analysis of gene level differences within the isolates
920 presented within this research. Further comparisons to database
921 held assemblies of *C. albicans* will also be conducted subsequent to the quality
922 control of database held assemblies.
923 - Aim 4: Identification of variants associated with colonisation of the reproductive tract
924 - Using a genome wide association study (GWAS) approach, isolates documented
925 here along Illumina sequence reads of isolates deposited within
926 NCBI's public SRA database will be analysed to identify if any
927 genetic elements are associated with isolation source/niche, which in this case
928 is anatomical body site.

929 However, research presented within was highly disrupted by the COVID-19 global
930 pandemic. As such, aims 1 and 2 were partially completed and thus results are not presented within.
931 Aim 3 was refined since no hybrid genome assemblies were completed, instead focus was switched
932 to database held genome assemblies. Little impact on aim 4 was observed.

933 **2.0 - Material and Methods**

934 **2.1 - Computational analysis**

935 Computational analysis was carried out using a CLIMB VM with 8 cores and 64Gb of RAM, more
936 intensive operations, such as pangenome construction, were performed using Super Computing
937 Wales (SCW) Sunbird cluster with a single node (40 CPUs with jobs submitted using SLURM).

938 **2.2 - Quality Control of *Candida albicans* database genomes**

939 51 *Candida albicans* assembly accession numbers were retrieved from the NCBI genome database
940 using the wget unix command and the ftp link to each genome assembly in fasta format
941 (Supplementary data table 1). Metadata for each assembly was downloaded using NCBI batch entrez
942 with the BioSample name as query. All methods described from 2.2 to 2.7 used these 51 assembly
943 isolates. An example command for genome assembly download is as follows:

```
wget
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/182/965/GCF_000182965.3_ASM18296v3/GCF_000182965.3_ASM18296
v3_genomic.fna.gz
```

944

945 **2.3 - Meta analysis**

946 Analysis and plotting of genome size, GC content, isolation source and geographical location were
947 carried out using **ggplot2**, **maps** and **scatterpie** packages within R studio using customised
948 markdown scripts. Data displaying the number of isolates present in a geographical location and
949 relative percentages of niche habitation were displayed using pie charts on a world map.

950 **2.4 - Chromosomal dotplots and genomic features**

951 Chromosomal alignments were generated using the DNAdiff command as part of the MUMmer 4
952 package using *Candida albicans* SC5314 as a reference sequence (*188*). A for loop was constructed to
953 analyse and filter the delta files and then to order the query sequence contigs into the optimal order
954 according to the SC5314 chromosome order. Genomic features were also assessed using the Nucmer
955 command. Graphical representations were created using Rstudio and the packages **ggplot2**, **tidyr**,
956 **dplyr**, **knitr**, **magrittr** and **Genomic Rangers**. Example commands are included below:

```
dnadiff Candida_albicans_SC5314.fna Candida_albicans_NCYC_4144.fna -p NCYC_4144
nucmer Candida_albicans_SC5314.fna Candida_albicans_NCYC_4144.fna -p NCYC_4144
```

957

958 **2.5 – Assembly Sequence similarity assessment**

959 Assembly isolate sequence similarity was compared with sourmash (*189*) using the compute and
960 compare commands with k-mers set at 31. 4 non-*albicans candida* (*tropicalis*, *orthopsilosis*,
961 *hispaniensis* and *dubliniensis*) and 3 *Saccharomyces* (2 *cerevisiae* and *paradoxus*) were included in

962 the analysis to act as outgroups. A and *1 Naumovozyma dairenens* sequence was also included as
963 previous literature suggested it may me misclassified and it may be a *Candida albicans* sequence.

```
sourmash compute -k 31 --scaled 1000 *.fastq
sourmash compare *.sig -o Sourmash_reads_matrix_31
sourmash plot Sourmash_reads_matrix_31 --pdf --labels
```

964 Commands are included below:
965
966 **2.6 - Pangenome construction**
967 Pangenome analysis of database *C. albicans* genome assemblies was performed using the Pangloss
968 pipeline(*190*) to perform gene prediction, analysis of sequence similarity and construction of a
969 syntenic pangenome. The config file was altered to give correct pathways for all programmes used
970 and data supplied as well as altering the number of cores used (8 during testing on CLIMB VM, 40
971 when used on the SCW sunbird cluster). Full pangenome construction took approximately 1 week on
972 the CLIMB VM while construction took 68 hours on the SCW sunbird cluster. The SLURM script used
973 to run pangloss is shown below:
974
975

```
python /home/ubuntu/Pangloss-master/Pangloss.py --pred  --ips --plots --karyo
/home/ubuntu/Work/candida_pangloss/pan_config.ini
```

976
977 **2.7 – Annotation of Pangenome Clusters**
978 Core and accessory sequences were functionally annotated using BlastKoala (*191*). Sequences were
979 given KEGG orthology assignments to all sequences determining their molecular function using
980 BLAST and GHOSTX and comparing the results to KEGG genes. Taxonomy id was set as 5476 for
981 *Candida albicans*. The genes identified were used in a gene ontology enrichment analysis, using
982 PANTHER version 14 (*192*), to determine enrichment in biological processes between the core and
983 accessory sequences.
984
985 **2.8 - Obtaining SRA sequences**
986 A metadata evaluation of the Sequence Read Archive (SRA) was carried out to identify *Candida*
987 *albicans* isolated from humans. A total of 320 SRA records hit inclusion criteria and these were batch
988 downloaded using a customised prefetch command from the SRA toolkit. This command is shown
989 below. These reads along with the 48 Swansea isolates were used for variant calling and analysis.
990 The Swansea isolates were obtained from three couples from previous studies, 10 from couple 3, 32
991 from couple 5 and 6 from couple 6. All analyses from this point used all 368 Swansea isolates and
992 SRA sequences.

```
prefetch --option-file sralist.txt
```

2.9 - Variant calling

Reads were aligned to the reference genome using BWA MEM optimised to use all available cores. Produced SAM files were converted into the BAM format using GATK RevertSam command. Read group information was added to the BAM file using the GATK AddOrReplaceReadGroups command then sorted into query name order using GATK SortSam. The BAM and SAM files were merged using GATK MergeBamAlignment to produce a merged BAM file with read group information. Duplicates were marked with GATK MarkDuplicates and then sorted into coordinate order. Nm, Md and Uq tags were calculated using GATK SetNmMdAndUqTags using the coordinate ordered BAM file then SNPs and INDELs were called using BCFtools mpileup and bcftools call. All VCF files were combined using BCFtools merge (*193-196*). The full workflow is shown in Figure 6. All of the commands used in this workflow are shown below.

```
Read alignment
bwa mem -t 1 Candida_albicans_SC5314.fna read1.fastq.gz read2.fastq.gz > sam file.sam
SAM to BAM conversion
gatk RevertSam -I sam file.sam -O bam file.bam;
Assign reads to a single read group
gatk AddOrReplaceReadGroups -I bam file.bam -O bam file_rg.bam -LB bam file -PL ILLUMINA -PU
one -SM bam file;
Sort SAM file by query name
gatk SortSam -I bam filerg.bam -O *bam file*_srt.bam -SO queryname;
Merge alignment data
gatk MergeBamAlignment -O bam filem.bam -UNMAPPED bam filesrt.bam -R
Candida_albicans_SC5314.fna -ALIGNED sam file.sam;
Identify and tag duplicate reads
gatk MarkDuplicates -I bam filem.bam -O bam filemdup.bam -M bam file_mdup.bam.txt;
Sort SAM file by coordinates
gatk SortSam -I bam filemdup.bam -O bam filemdsrt.bam -SO coordinate;
Calculate NM, MD and UQ tags
gatk SetNmMdAndUqTags -I bam filemdsrt.bam -O bam filefix.bam -R Candida_albicans_SC5314.fna -
-CREATE_INDEX true;
Generate VCFs with genotype likelihoods
bcftools mpileup --redo-BAQ -a DP,AD -f Candida_albicans_SC5314.fna bam file_fix.bam -o
outputpileup.mpileup
SNP calling
bcftools call -A -m -Ov -o output.vcf input.mpileup
Merge VCF files
bcftools merge -0 --threads 8 -o combine.vcf -m none -l vcf_list.txt
```

1025
1026
1027
1028
1029

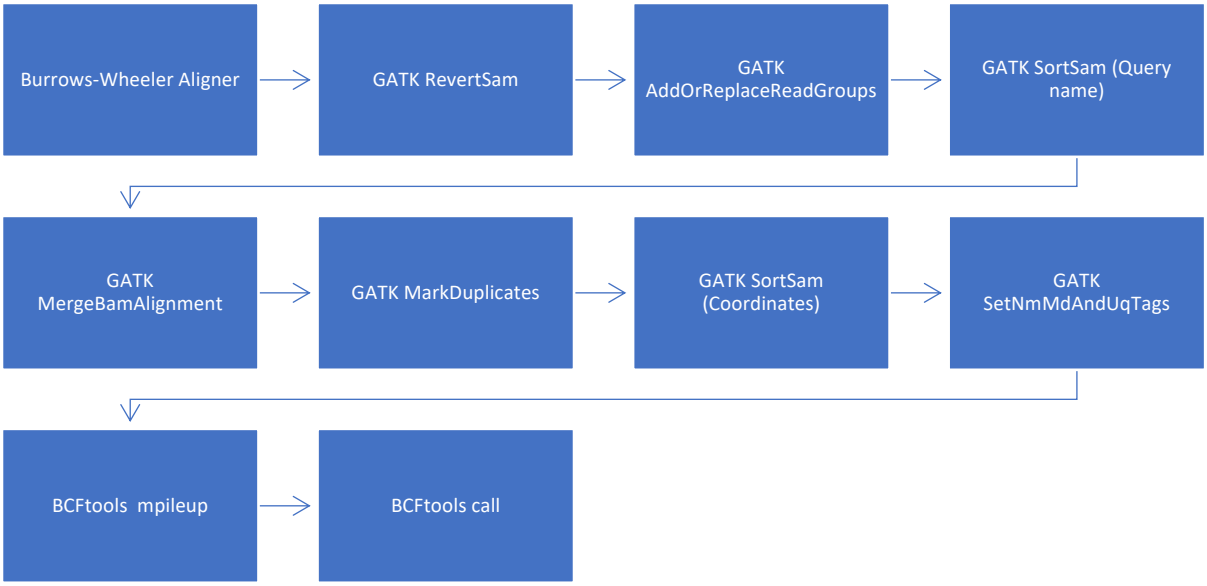


Figure 6 – Customised BWA, GATK and BCFtools workflow used to align SRA and Swansea reads to SC5314 reference genome and then in variant calling.

1030 **2.10 - SNP-based phylogeny construction**

1031 Maximum likelihood phylogenetic trees using the combined VCF file were generated with SNPhylo
1032 (*197*). Minor allele frequency and missing rate were set at 0.1, linkage disequilibrium was at at 0.75.
1033 The command used is shown below:

```
bash /home/ubuntu/SNPhylo-master/snphylo.sh -v combined vcf.vcf -l 0.75 -m 0.1 -M 0.1 -P Merged_reads
```

1034

1035 **2.11 - Functional variant annotation**

1036 Variant annotation and prediction of the effects of these variants was performed using snpEff using
1037 SC5314 as a reference (*198*). A custom database for *Candida albicans* was constructed from the
1038 *Candida albicans* SC5314 reference genome using the *build* command. VCF files for individual
1039 isolates were split into separate chromosomes and each chromosome was functionally annotates.
1040 Default parameters were used for the analysis. The commands used are shown below:

1041

```
java -jar snpEff.jar build -genbank -v SC5314Ref
java -Xmx8G -jar /home/ubuntu/snpEff/snpSift.jar split *vcf file*
java -Xmx8G -jar /home/ubuntu/snpEff/snpEff.jar eff -v -c /home/ubuntu/snpEff/snpEff.config SC5314Ref *vcf file*
```

1042

1043 **2.12 - Genome wide association study**

1044 Whole genome association analysis was carried out with PLINK2 (*199*). PL
1045 INK2 binary files were produced from the combined VCF. Filtering steps included a linkage
1046 disequilibrium window size of at 10 KBases with a r^2 of 0.75. Hardy-Weinberg threshold of $6*10^{-6}$and
1047 a minor allele frequency of 0.1. Phenotype files with isolation source and aneuploidy information
1048 were used in construction of the PLINK2 binary files. Population stratification was accounted for
1049 using a principal component analysis and included in the association analysis. A general linear model
1050 was used. The output file was plotted as a manhattan plot in RStudio using ggplot2 with a
1051 significance threshold of $5 * 10^{-8}$. The commands used are shown below:

```
/home/ubuntu/plink2 --vcf combined vcf.vcf --double-id --make-pgen --allow-extra-chr –out merged_reads --pheno phenotypes.txt --hwe 0.000006 --maf 0.1 --set-all-var-ids @#[ca]$r,$a –indeppairwise 10 0.75 --fa Candida_albicans_SC5314.fasta --new-id-max-allele-len 15 missing
/home/ubuntu/plink2 -pfile merged_reads --pca --allow-extra-chr
/home/ubuntu/plink2 --pfile merged_reads --glm --covar plink2.eigenvec --allow-extra-chr
```

1052

1053 **2.13 - Genome coverage**

1054 Evaluation and basic statistics of alignment data was generated with BamQC, part of the qualimap
1055 tool, using processed bam files from the GATK workflow (*200*). Output files were manually assessed
1056 looking at genome coverage across the isolates. Chromosomes with an average coverage 50%
1057 greater than other chromosomes from the same isolate were determined to be aneuploidy for that
1058 chromosome. The commands used are shown below:

1059

```
bamqc -bam bam file -c -nt 8 -outfile bam file.pdf -oc bam file_coverage
```

1060

1061

1062

1063

1064

1065

1066

1067 **3.1 Quality Control of *Candida albicans* genomes hosted on NCBI**

1068 All results described in this section used the 51 database assemblies described in 2.2.

1069 **3.1.1 Genome Assembly Metrics of Database Sequences**

1070 A total of 51 *Candida albicans* assemblies were obtained from the NCBI genome database (correct at
1071 access date 2019). Genome size of these isolates ranged from 12.4 Mb to 28.6 Mb with a mean
1072 genome size of 15.2 Mb. The GC content ranged from 29.8% to 34.1% with an average of 33.4%. Two
1073 strains were identified with genome assembly sizes which represent a diploid genome, namely
1074 *Candida albicans* SC5314-P0 and *Candida albicans* SC5314-GTH12 (Table 4 and Figure 7A). Only the
1075 SC5314-P0 strain had metadata held within NCBI's BioSample database (accession SAMN08098130)
1076 which highlights this was isolated from human blood in a case of candidemia. Metadata for the
1077 GTH12 strains is unfortunately missing other than it was isolated from a human, with no specific
1078 body site given (BioSample SAMN08098151). What should be noted is that these assemblies are
1079 described as being Illumina only with an assembly method of "reference guided" using the *C.*
1080 *albicans* SC5314 A22-s07-m01-r18 assembly as a reference. With read coverage described as 24X for
1081 the SC5314-P0 strain and 18X for the GTH12 strain, the usefulness and trustworthiness of these
1082 assemblies can be questioned since true telomere to telomere genomes have only started to
1083 become achievable with the use of library preparation techniques such as proximity ligation and the
1084 use of third generation sequencing technologies. Several low GC content isolates were also
1085 identified with GC content below 31.5% and as low as 29.8%. Interestingly, all of which were isolated
1086 from the reproductive tract. Further a mix of host disease state was noted. The UAB040-W3D3 strain
1087 (BioSample: SAMN06854471) and UAB090-W2D7 strain (BioSample: SAMN06854477) were of
1088 vaginal origin with a host disease of vulvovaginal candidiasis, while the UAB012-W3D5, UAB012-
1089 W7D4 strains (BioSamples: SAMN06854254 and SAMN06854404 respectively) were both also
1090 vaginal in isolation source but were from asymptomatic hosts with relation to VVC. The reason for
1091 the lower GC content is currently undocumented. The other 45 isolates all had similar assembly sizes
1092 of around 14 Mb and GC content between 33.5% and 34.1%. Other isolation sources include the
1093 oropharyngeal tract, blood, faeces, and environmental sources. Metadata for all assemblies can be
1094 found in supplementary table 1

Table 5: *Candida albicans* assembly isolates used in initial pangenome construction and metadata information.

Organism Name	Strain	BioSample	BioProject	Assembly	Size(Mb)	GC%	Niche
Candida albicans	12C	SAMN00767974	PRJNA75209	GCA_000773845.1	14.890	34.10	Oral
Candida albicans	19F	SAMN01048008	PRJNA75221	GCA_000775445.1	14.57	33.70	Vaginal
Candida albicans	3153A	SAMN00974104	PRJNA165021	GCA_000447595.1	14.89	33.70	N/A
Candida albicans	A123	SAMN00974110	PRJNA165033	GCA_000447455.1	14.64	33.50	N/A
Candida albicans	A155	SAMN00974111	PRJNA165035	GCA_000447615.1	14.47	33.50	N/A
Candida albicans	A20	SAMN00974106	PRJNA165025	GCA_000447575.1	14.55	33.60	N/A
Candida albicans	A203	SAMN00974113	PRJNA165039	GCA_000447495.1	14.79	33.80	N/A
Candida albicans	A48	SAMN00974107	PRJNA165027	GCA_000447535.1	14.70	33.60	N/A
Candida albicans	A67	SAMN00974108	PRJNA165029	GCA_000447515.1	14.69	33.70	N/A
Candida albicans	A84	SAMN00974112	PRJNA165037	GCA_000447635.1	14.69	33.70	N/A
Candida albicans	A92	SAMN00974109	PRJNA165031	GCA_000447475.1	14.61	33.60	N/A
Candida albicans	ATCC 12031	SAMN04324314	PRJNA305340	GCA_002276455.1	17.07	33.70	Lungs(Bronchitis)
Candida albicans	Ca529L	SAMN02058435	PRJNA200311	GCA_000691765.2	14.67	34.00	Oral mucosa
Candida albicans	Ca6	SAMN03164130	PRJNA120431	GCA_000784695.1	14.72	33.60	N/A
Candida albicans	CHN1	SAMN00974105	PRJNA165023	GCA_000447555.1	14.73	33.60	N/A
Candida albicans	GC75	SAMN00767984	PRJNA75223	GCA_000773735.1	14.70	33.70	Oral
Candida albicans	L26	SAMN01048004	PRJNA75211	GCA_000775455.1	14.52	33.60	Vaginal
Candida albicans	NCYC 4144	SAMN11464299	PRJNA543141	GCA_005890765.1	14.70	33.61	*Quercus petraea*
Candida albicans	NCYC 4144_2	SAMN11464299	PRJNA543257	GCA_005890695.1	12.51	33.50	*Quercus petraea*
Candida albicans	NCYC 4145	SAMN11464300	PRJNA543142	GCA_005890775.1	15.45	33.67	*Quercus petraea*
Candida albicans	NCYC 4145_2	SAMN11464300	PRJNA543275	GCA_005890685.1	13.77	33.50	*Quercus petraea*
Candida albicans	NCYC 4146	SAMN11464301	PRJNA543143	GCA_005890745.1	15.48	33.59	*Quercus robur*
Candida albicans	NCYC 4146_2	SAMN11464301	PRJNA543276	GCA_005890705.1	13.06	33.50	*Quercus robur*
Candida albicans	P34048	SAMN01048010	PRJNA75229	GCA_000775465.1	14.54	33.70	Blood
Candida albicans	P37005	SAMN01048006	PRJNA75217	GCA_000773745.1	14.47	33.80	Oral
Candida albicans	P37037	SAMN01048011	PRJNA75231	GCA_000773825.1	14.48	33.60	Oral

Candida albicans	P37039	SAMN00767975	PRJNA75233	GCA_000784515.1	14.52	33.70	Blood
Candida albicans	P57055	SAMN01048013	PRJNA75239	GCA_000775505.1	14.59	33.70	Blood
Candida albicans	P57072	SAMN00767978	PRJNA75227	GCA_000773805.1	14.51	33.70	Blood
Candida albicans	P60002	SAMN01048007	PRJNA75219	GCA_000784525.1	14.79	34.00	Blood
Candida albicans	P75010	SAMN00769059	PRJNA75235	GCA_000784575.1	14.86	34.10	Blood
Candida albicans	P75016	SAMN01048012	PRJNA75237	GCA_000784595.1	14.68	34.00	Blood
Candida albicans	P75063	SAMN01048014	PRJNA75241	GCA_000775525.1	14.45	33.60	Blood
Candida albicans	P76055	SAMN01048016	PRJNA75243	GCA_000784505.1	14.45	33.70	Blood
Candida albicans	P76067	SAMN01048017	PRJNA75245	GCA_000784495.1	14.62	33.80	Blood
Candida albicans	P78042	SAMN01048015	PRJNA75247	GCA_000784615.1	14.68	33.70	Blood
Candida albicans	P78048	SAMN01048009	PRJNA75225	GCA_000773725.1	14.50	33.70	Blood
Candida albicans	P87	SAMN00767982	PRJNA75215	GCA_000774085.1	14.46	33.60	Oral
Candida albicans	P94015	SAMN01048005	PRJNA75213	GCA_000773755.1	14.74	33.90	Blood
Candida albicans	SC5314	SAMN02953594	PRJNA10701	GCA_000182965.3	14.28	33.48	Blood
Candida albicans	SC5314_2	SAMN01041717	PRJNA191536	GCA_000784655.1	14.70	33.60	N/A
Candida albicans	SC5314_3	SAMN01041717	PRJNA120009	GCA_000784635.1	15.21	34.00	N/A
Candida albicans	SC5314 GTH12	SAMN08098151	PRJNA395439	GCA_002835845.1	28.59	33.57	N/A
Candida albicans	SC5314 P0	SAMN08098130	PRJNA395439	GCA_002837675.1	28.60	33.48	Blood(Candidemia)
Candida albicans	SP CRL 000G1	SAMN09217378	PRJNA471744	GCA_004026255.1	12.56	33.70	Faeces
Candida albicans	TIMM 1768	SAMN09204982	PRJNA471195	GCA_003454735.1	14.43	33.63	Faeces(Candidiasis)
Candida albicans	UAB012 W3D5	SAMN06854254	PRJNA384935	GCA_002259805.1	15.26	29.80	Vaginal(Epithelium)
Candida albicans	WAB012 W7D4	SAMN06854404	PRJNA384935	GCA_002259875.1	15.64	30.00	Vaginal
Candida albicans	UAB040 W3D3	SAMN06854471	PRJNA384935	GCA_002259885.1	14.98	31.50	Vulvovaginal candidiasis
Candida albicans	WAB030 W2D7	SAMN06854477	PRJNA384935	GCA_002259865.1	15.52	30.10	Vulvovaginal candidiasis
Candida albicans	WO 1	SAMN02953609	PRJNA16373	GCA_000149445.2	14.47	33.51	Blood

1095

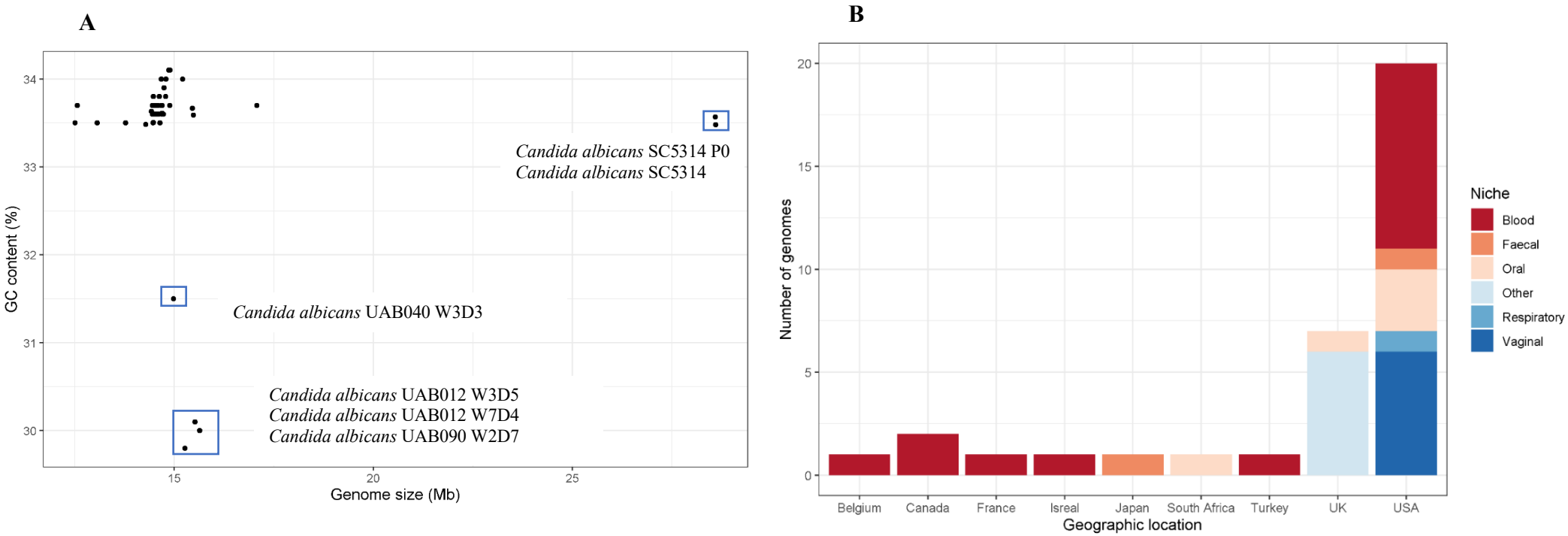


Figure 7 – A. Genome size (Mb) against GC content (%) of GenBank assemblies. Blue boxes indicate assemblies with outlying GC contents and genome sizes. High genome size isolates are diploid assemblies. Low GC content isolates were sourced in the vagina. B, Geographical and anatomical distribution of database genomes. Figures generated using ggplot2

1096 **3.1.2 - Taxonomic Confirmation of *C. albicans* Genome Assemblies**

1097 In order to confirm that all assemblies described within the database were correctly classified, a coarse
1098 scale clustering analysis using Sourmash (*189*) was employed to identify species level taxonomic
1099 assignment. Sourmash signatures were generated for each of the *Candida albicans* genomes in
1100 addition to eight other non-albicans genomes. These included the representative genome for four
1101 other *Candida* species (*C. tropicalis, C. orthopsilosis, C. hispaniensis and C. dubliniensis*) and two
1102 *Saccharomyces* species (2 *cerevisiae* and *paradoxus*). In addition, a *Naumovozyma dairenensis*
1103 genome assembly was included since it was documented within the literature (*201*) as having been
1104 incorrectly classified as *Naumovozyma dairenensis,* when in fact it was likely to belong to the *C.*
1105 *albicans* species. Genome information for these isolates is supplied in table 6.

Table 6: Non-*Candida albicans* assemblies included in sourmash analysis as outgroups.

Organism Name	Strain	BioProject	BioSample	Assembly	Size(Mb)	GC%	Niche
Candida dubliniensis	CD36	PRJEA34697	SAMEA2272258	GCA_000026945.1	14.05	33.10	N/A
Candida hispainiensis	CBS 9996	PRJEB18079	SAMEA4837037	GCA_900535975.1	10.66	41.60	N/A
Candida orthopsilosis	CO 90-125	PRJEA83665	SAMEA2272376	GCA_000315875.1	13.00	37.40	N/A
Candida tropicalis	MYA-3404	PRJNA13675	SAMN02953608	GCA_000006335.3	14.70	33.30	N/A
Saccharomyces cerevisiae	S288C	PRJNA43747	N/A	GCA_000146045.2	11.80	38.20	N/A
Saccharomyces cerevisiae	BY4742	PRJNA429985	SAMN08364553	GCA_003086655.1	12.10	38.20	Laboratory
Saccharomyces paradoxus	UFRJ50816	PRJEB7245	SAMEA4461731	GCA_002079145.1	12.00	38.50	N/A
Naumovozyma dairenensis	CBS 421	PRJEA70961	SAMEA2272418	GCA_000227115.2	13.50	33.80	N/A

1106

1107

In a pairwise comparison of signatures in an all against all manner, all *Candida albicans* assemblies clustered together with an average sequence similarity of 71% and a minimum sequence similarity of 55%. The *Naumovozyma dairenensis* sequence, highlighted by the blue box in Figure 8, clustered with the *Candida albicans* sequences indicating that it has indeed been misclassified as suggested. All other non-*Candida albicans* outgroup sequences clustered as expected, separately from the *Candida albicans* sequences. Of interest, is the fact that within the outgroups, *Saccharomyces paradoxus* appears grouped within the *Candida* sp. assemblies, whilst outside of the scope of the research presented here, it does offer an avenue for further exploration and could highlight additional taxonomic inconsistencies in database held genomes.

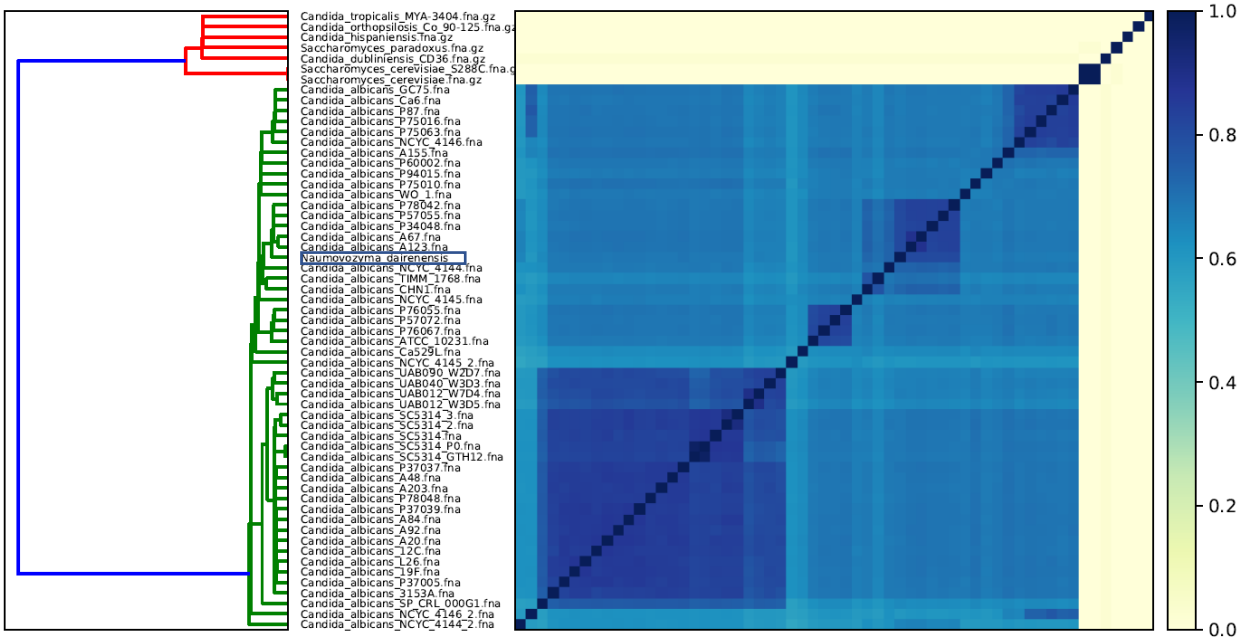
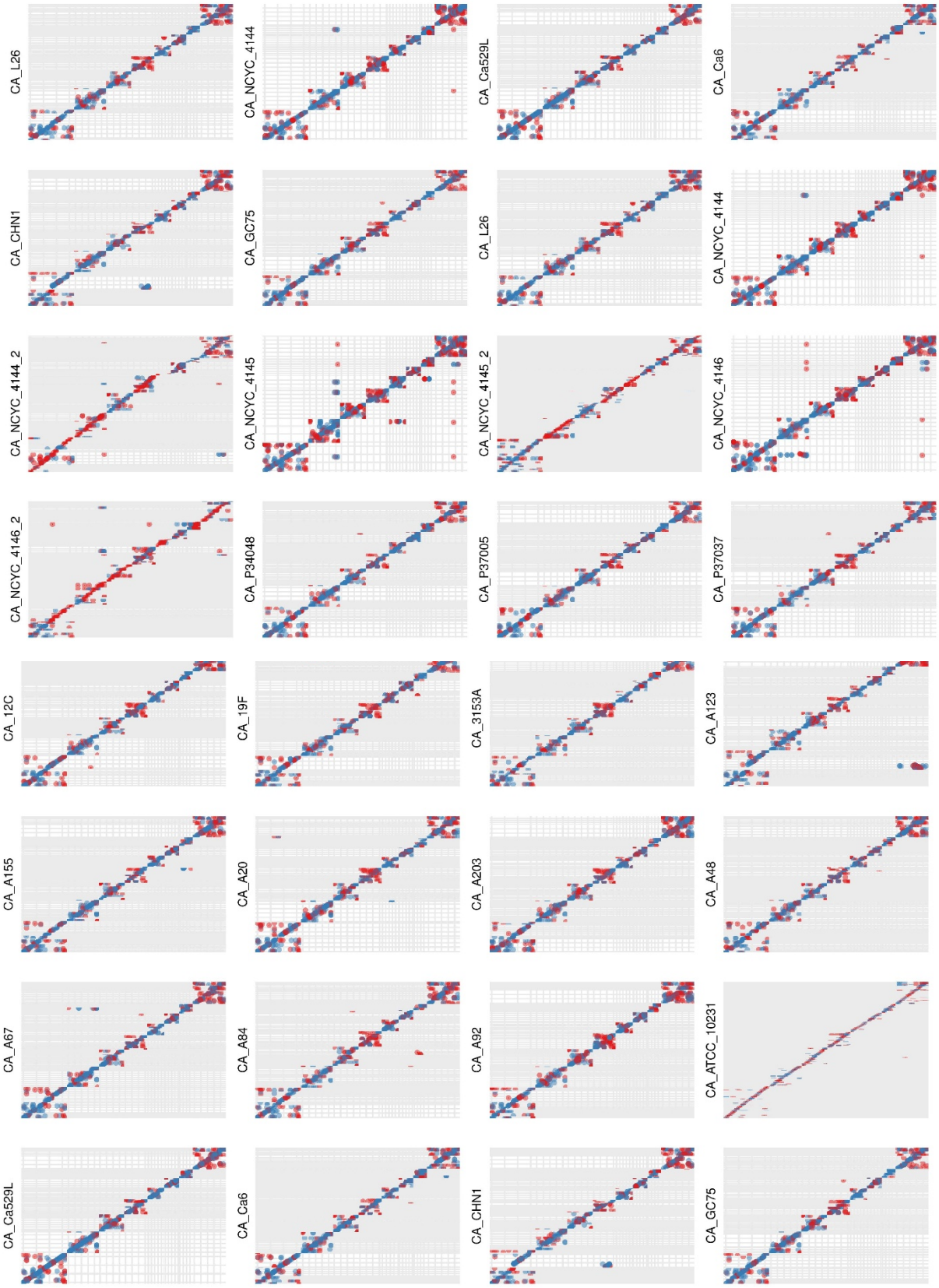


Figure 8: All *C. albicans* assemblies were contained within a single clade, also containing a *Naumovozyma* sequence (green clade). Outgroups remain separate (red clade). Scale indicates sequence similarity. Blue box highlights the *Naumovozyma* sequence

3.1.3 - Genomic Features at Chromosomal Resolution

A comparative analysis of the genome assemblies was performed by comparing each against the *C. albicans* SC5314 reference genome (ASM18296v3) with use of the MUMmer package. The dnadiff wrapper script was used to output alignment statistics including insertions, relocations, and translocations. The nucmer generated delta files were processed to generate dotplots (Figure 9) for each assembly versus the reference for visual inspection. In the case of diploid assemblies two dot plots were presented to highlight each chromosome in the pair. No large-scale insertions or deletions were detected in any isolate although the size of the largest feature was an insertion just under 2000 bp.

1128


CA_L26
CA_NCYC_4144
CA_Ca529L
CA_Ca6
CA_CHN1
CA_GC75
CA_L26
CA_NCYC_4144
CA_NCYC_4144_2
CA_NCYC_4145
CA_NCYC_4145_2
CA_NCYC_4146
CA_NCYC_4146_2
CA_P34048
CA_P37005
CA_P37037
CA_12C
CA_19F
CA_3153A
CA_A123
CA_A155
CA_A20
CA_A203
CA_A48
CA_A67
CA_A84
CA_A92
CA_ATCC_10231
CA_Ca529L
CA_Ca6
CA_CHN1
CA_GC75

1129

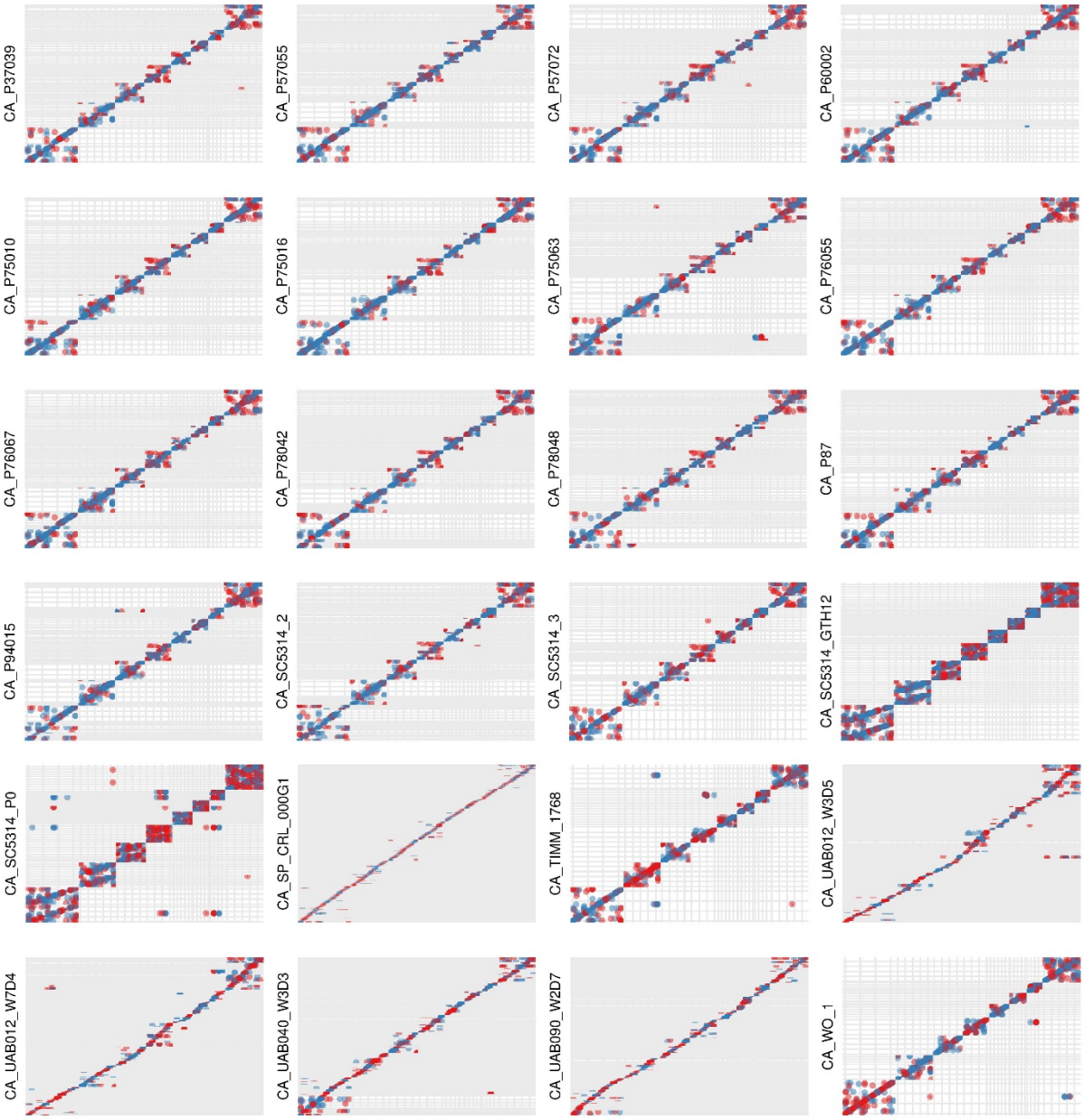


Figure 9: Chromosomal dotplots produced using delta files produced using MUMmer 4 in R using the GGplot2, tidyr, dplyr, knitr, magrittr and Genomic Rangers packages. Blue lines indicate the sequence was matched with the forward strand while red lines indicate the sequence matched with the negative strand.

Quantification of distinct features identified via nucmer was carried out by two means. Firstly, a general quantification and qualification comparison across all isolates for insertions, inversions, relocations and translocations. Secondly to quantify if there is a differential abundance in a genomic feature according to the niche from which the isolate was obtained. The high number of insertions highlighted in Figure 10 were attributed to the two diploid assemblies. Furthermore, these higher numbers of insertions were associated with a larger insertion size (>1500).

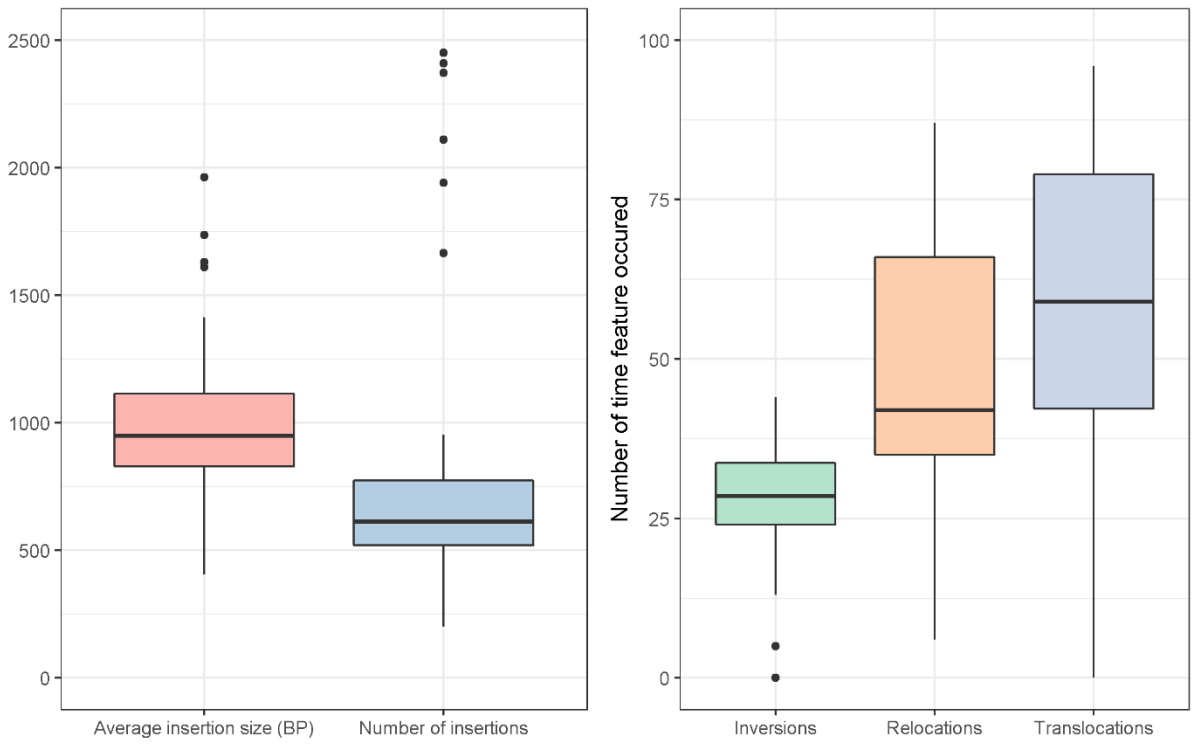


Figure 10 - Genomic features of database assemblies identified using MUMMER 4 presented as whisker dot plots generated using Ggplot2. Black lines inside boxes show the median value, box borders indicate the upper (75th percentile) and lower (25th percentile) quartiles. Whiskers show the maximum (100th percentile) and minimum (0th percentile) data points. Additional dots show outliers.

Pairwise comparisons were made between all isolation sources to identify if significant differences in genomic features were present with a specific emphasis on identifying if vaginal isolates contained altered numbers of genomic features compared to other niches with respect to the reference sequence used. In order to achieve this, a pairwise Wilcox test with corrections for multiple testing highlighted how only three significant differences were detected. These are a difference in numbers of relocations between plant isolates compared to blood isolates (P=0.02). In this instance there are a greater number of relocations in blood isolates. A further differentiation within the plant isolates were the numbers of translocations which were significantly higher in plant isolates than those from a vaginal source and those in which the isolation source was unknown. The genomic feature counts can be found within supplementary table 1, whilst the adjusted P-values are documented in table 7.

Table 7: Adjusted p-values from Wilcox tests of differences between genomic features between assembly isolates from different isolation sources. Adjusted P-values < 0.05 are highlighted in red.

Adjusted P-value of Genomic Features Compared Between Niches							
Groups	**Breakpoints**	**Inversions**	**Relocations**	**Translocations**	**Total SNPs**	**Insertions**	**Insertion average size**
Faeces/Blood	0.983	0.774	0.941	0.645	1.000	1.000	0.945
Oral/Blood	0.989	0.774	0.497	0.568	0.952	1.000	0.989
Oral/Faeces	1.000	0.774	0.714	1.000	0.952	1.000	0.989
Other/Blood	0.278	0.275	0.078	0.103	0.344	0.543	0.152
Other/Faeces	0.648	0.554	0.442	0.419	0.952	0.905	0.459
Other/Oral	0.278	0.774	0.235	0.402	0.740	0.543	0.230
Plant/Blood	0.648	0.774	0.020	0.645	0.249	1.000	0.989
Plant/Faeces	0.964	0.774	0.148	0.402	0.952	1.000	1.000
Plant/Oral	0.648	0.774	0.148	0.419	0.387	1.000	0.989
Plant/Other	0.648	0.349	0.148	0.030	0.065	1.000	0.459
Vaginal/Blood	0.949	0.774	0.450	0.402	0.249	1.000	0.973
Vaginal/Faeces	0.989	0.774	0.714	0.643	0.952	1.000	0.714
Vaginal/Oral	1.000	0.774	0.442	0.661	0.344	1.000	1.000
Vaginal/Other	0.278	0.774	0.148	0.645	0.740	0.543	0.152
Vaginal/Plant	0.739	0.774	0.099	0.037	0.065	0.674	1.000

1174 **3.1.4 Pangenome construction**

1175 To evaluate the total gene content across the collection of isolates, a pangenome approach was used
1176 to generate a dataset of syntenic ortholog clusters with the use of the Pangloss pipeline. Initial
1177 pangenome construction used the full set of 51 genomes described in section 3.1. This is inclusive of
1178 the diploid assemblies. The total pangenome assembly process took 165 hours on an 8 CPUs. This
1179 highlights how computationally intensive this particular analysis process is. The pangenome consisted
1180 of a total of 16,196 unique clusters. This total value is larger than expected and represents
1181 approximately 2.5 times the median protein count across all genomes. Further, the split between core
1182 and accessory genes was 10.99% core and 89.01% accessory, these equate to 1780 orthologous
1183 clusters (core) and 11,416 orthologous clusters respectively as accessory as shown in Figure 11A.
1184 These statistics are at odds with the general expectation of pangenome sizes and do not fit with
1185 previous descriptions of *Candida* specific pangenomes that are described to have an approximate
1186 90:10 split of core to accessory genes (*190*). Of the accessory genes, 5124 were represented as
1187 singleton clusters and 4339 with a cluster size of 2 as shown in Figure 11B. Combined, this accounted
1188 for 9463 (65.64%) of the accessory genome. Further exploration of these singleton and doubletons
1189 found that they were predominantly composed of sequences from the diploid strains (*Candida*
1190 *albicans* SC5314-P0 and *Candida albicans* SC5314-GTH12).

1191 As a result of this initial pangenome construction, it is clear that a mixture of assembly types (haploid
1192 and diploid representations) determines the distribution of the pangenome clusters rather than the

presence or absence and sequence content of the protein sequences themselves. As such, it is likely that an informative pangenome should be constructed from genome assemblies that are equivalent in their ploidy, that being it is built with either all diploid or all haploid assemblies only.

A second pangenome was constructed without the diploid strains. Thus using 49 isolates were taken forward. With this approach, the pangenome contained a total of 7361 ortholog clusters with a split of 64% core (4711 clusters) and 36% accessory (2650 clusters) as seen in Figure 12A. That is 4711 clusters were found in every single isolate. Of the accessory clusters, 1207 were singletons, highlighting the potential genetic diversity within this dataset. Cluster sizes are shown in Figure 12B. The Chao lower bound estimate for the total pangenome size of *Candida albicans* through analysis of this dataset was 12,091 clusters. This value accounts for clusters identified within this study together with an estimation of clusters not yet observed thus highlighting the potential for further sequence and functional diversity within the species. The fact that this estimator is likely to be conservative, i.e. it is more likely to be too small than it is to be too large, highlights that further exploration of *C. albicans* genomes through a pangenome context is warranted.

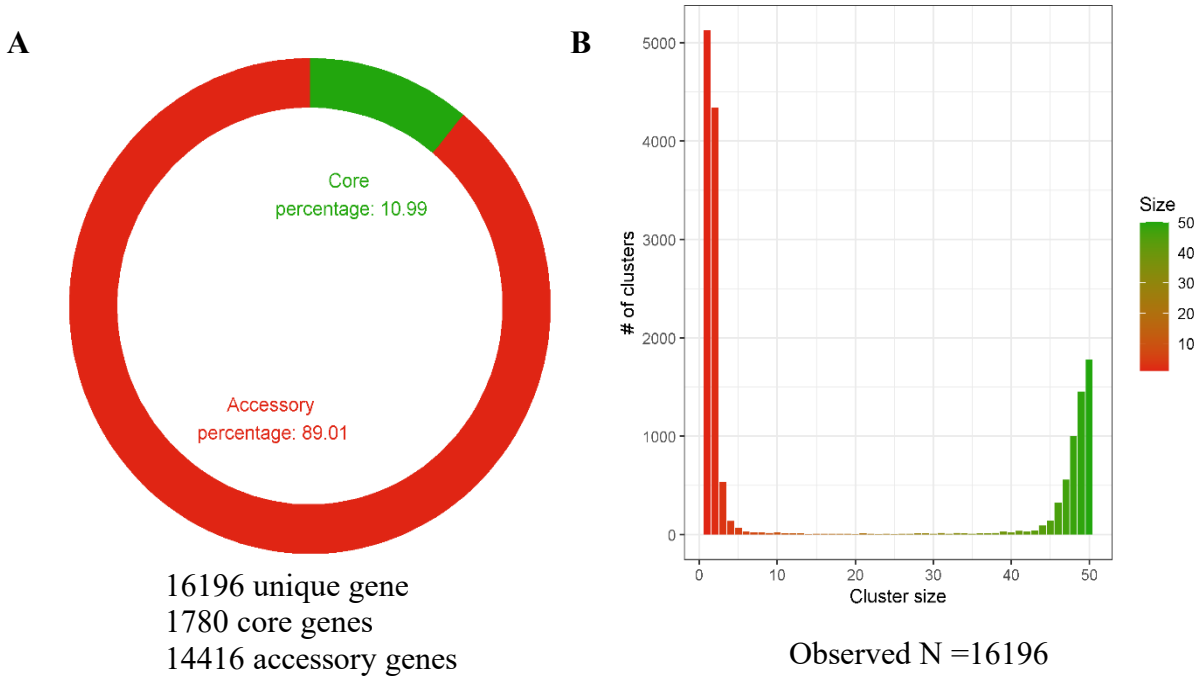


Figure 11 - Number of core and accessory genes within the *Candida albicans* assemblies and size of orthologous clusters as determined by PanOCT as part of the Pangloss pipeline. Core genes are those that are present in 95% of genomes. (A) The percentage split and number of genes present in the core and accessory genomes. (B) The number of clusters in each size group. Plotted in R using ggplot2 and cowplot.

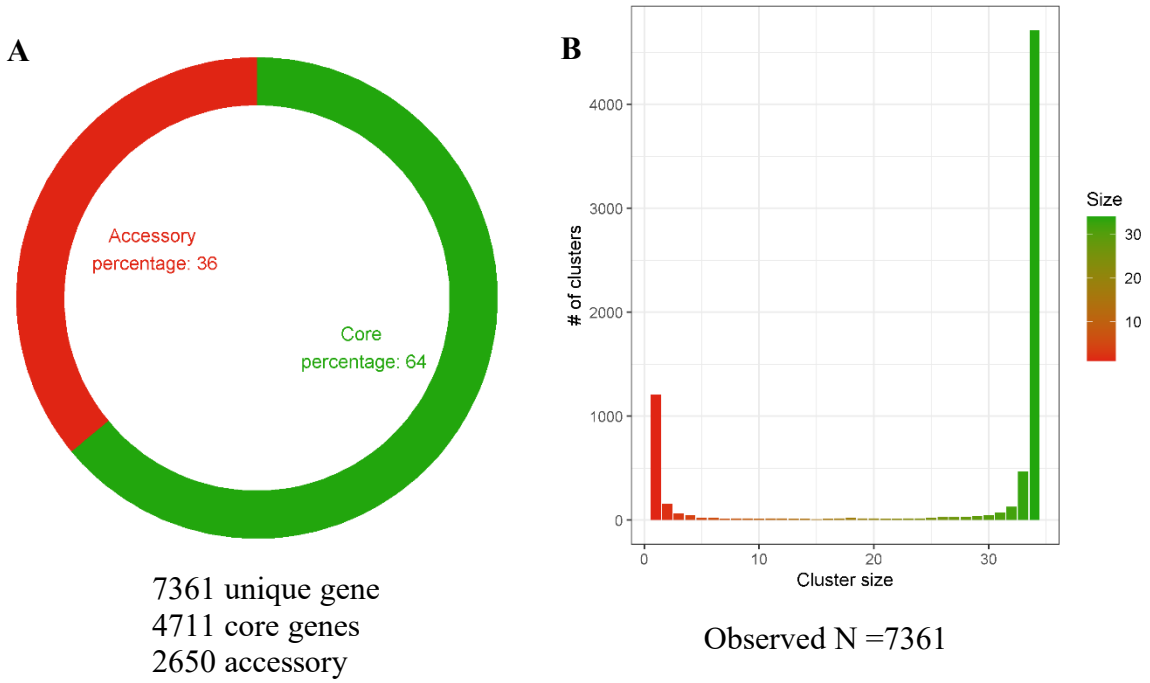


Figure 12 - Number of core and accessory genes within the haploid *Candida albicans* assemblies and size of orthologous clusters as determined by PanOCT as part of the Pangloss pipeline. Core genes are those that are present in 95% of genomes. (A) The percentage split and number of genes present in the core and accessory genomes. (B) The number of clusters in each size group. Plotted in R using ggplot2 and cowplot.

3.1.5

Annotation of the Pangenome

Annotation of the pangenome core and accessory clusters was undertaken to compare the differences in predicted function between core and accessory clusters. The core genome had 2720 matches out of 4711 inputted entries, a percentage of 57.7%. The functional categories representing hits to clusters and the number of clusters part of each category are shown in table 8.

The accessory genome had 910 matches of 2650 entries, a percentage of 34.3%. The functional categories of these clusters and the number of matches in each cluster is shown in table 9.

Table 8: Core genome clusters functional annotations from KoalaBLAST.

Core Genome		
Functional Category	**Number of Matches**	**Percentage of Matches**
Protein families: genetic information processing	664	24.41
Genetic Information Processing	579	21.29
Cellular Processes	195	7.20
Carbohydrate metabolism	180	6.62
Protein families: signalling and cellular processing	148	5.44
Environmental Information Processing	130	4.80
Lipid metabolism	110	4.04
Protein families: metabolism	101	3.71
Amino acid metabolism	93	3.42
Unclassified: metabolism	87	3.20
Energy metabolism	76	2.79
Metabolism of cofactors and vitamins	72	2.65
Organismal systems	57	2.10
Nucleotide metabolism	55	2.02
Human diseases	54	1.99
Glycan biosynthesis and metabolism	52	1.91
Unclassified	24	0.88
Metabolism of terpenoids and polyketides	16	0.59
Metabolism of other amino acids	12	0.44
Xenobiotics degradation and metabolism	7	0.26
Unclassified: signalling and cellular processes	7	0.26
Biosynthesis of other secondary metabolites	1	0.04

1256
1257
1258
1259
1260
1261
1262

Table 9: Accessory genome clusters functional annotations from KoalaBLAST.

Accessory Genome		
Functional Category	**Number of Matches**	**Percentage of Matches**
Protein families: genetic information processing	252	27.69
Genetic Information Processing	173	19.01
Carbohydrate metabolism	64	7.03
Cellular Processes	55	6.04
Protein families: signalling and cellular processing	52	5.71
Environmental Information Processing	47	5.16
Energy metabolism	45	4.95
Protein families: metabolism	41	4.51
Glycan biosynthesis and metabolism	26	2.86
Lipid metabolism	26	2.86
Metabolism of cofactors and vitamins	23	2.53
Organismal systems	21	2.31
Unclassified: metabolism	20	2.20
Amino acid metabolism	19	2.09
Nucleotide metabolism	12	1.32
Metabolism of other amino acids	8	0.88
Unclassified: signalling and cellular processes	8	0.88
Human Diseases	7	0.77
Unclassified: Genetic information processing	6	0.66
Xenobiotics degredation and metabolism	2	0.22
Metabolism of terpernoids and polyketides	2	0.22
Biosynthesis of other secondary metabolites	1	0.11

1263 Gene IDs obtained from KoalaBLAST were used to further differentiate the functional differences of
1264 the clusters between the core and accessory genomes by determining enrichment in biological
1265 processes, molecular function and cellular components. The enrichment results are available in
1266 supplementary file 4.

1267 Of particular interest within the core genome was a 10.31-fold enrichment in within the cytolysis of
1268 symbiont of host cells (GO:0001897) category. Cytolysis of symbiont of host cells is the ability of an
1269 organism to kill a cell of its host organism through cytolysis. In *Candida albicans* this has been

1270 associated with the cytolytic peptide toxin candidalysin (*53*). Orthologs identified in the enrichment
1271 analysis include SHE3, KEX1, ECE1 and KEX2. Within the accessory genome a 13.76-fold enrichment
1272 in dolichol-linked oligosaccharide biosynthetic process specifically (GO: 0006488), an essential
1273 component in protein modification. The orthologs identified were all part of the ALG family and
1274 included ALG5, 6, 7, 9, 11 and ALG13. A 13.38-fold enrichment was also observed in the chromatin
1275 DNA binding category (GO: 0031490) and the orthologs identified were KAE1, GON7, TBF1, SSN6 and
1276 RAP1.

1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301

1302 **3.2 - Genome wide association study of *Candida albicans* with relation to reproductive tract**
1303 **colonization**
1304 All results described below used the 368 SRA and Swansea isolates described in 2.8.
1305 **3.2.1 - Read alignment and variant calling**
1306 In order to explore if any genetic markers are associated with the isolation source and potential
1307 pathogenicity of an isolate, the collection of Swansea isolates presented here were analysed in
1308 combination with sequence reads obtained from NCBI's Sequence Read Archive (SRA). To facilitate
1309 this, a dataset of sequence reads was first obtained from the SRA using the search term "*Candida*
1310 *albicans*", specifically selecting *Candida albicans* as the taxonomy. Only genomic SRAs were selected
1311 that were sequenced from a DNA source, in effect removing transcriptomic datasets (RNA). As a
1312 further filter, only Illumina generated datasets were used in order to fit with the analysis pipeline,
1313 however this had little impact on exclusion of datasets since this platform was represented the vast
1314 majority of SRA hits. Only paired-end sequence reads in FASTQ format were selected. Read sets that
1315 were related to the query "*Candida albicans*" but were not representative of a single isolate genome,
1316 for example metagenomes, were excluded from the analysis despite offering an interesting research
1317 avenue.
1318 320 relevant SRA datasets were obtained from the NCBI SRA database using the SRA toolkit.
1319 Metadata for each BioSample was also retrieved. A full table of all SRA accessions that were
1320 downloaded and analysed can be found as Supplementary Table 2. Within this table are the raw
1321 read metrics. The average number of reads across the samples (on a per sample basis) was 13.1
1322 million. The smallest being 152674 (SRR7704197), and the largest being 50.7 million (SRR6710164).
1323 Average read lengths ranged between 68 and 301 bp. The breakdown of isolation source or niche is
1324 provided below (Table 10). Unfortunately, a large number of human derived isolates did not include
1325 an exact body site.

Table 10: Isolation source and read information from downloaded SRAs. Standard deviations shown with averages

Isolation Source (Niche)	Number of Isolates	Average Read Length	Average Number of Reads (Million)	Average Read Quality
Vaginal	69	107.91 ± 13.74	15.90 ± 6.95	36.00 ± 0.68
GI Tract	3	93.33 ± 0.94	10.50 ± 1.66	36.00 ± 0.83
Sputum	6	119.50 ± 13.19	6.69 ± 2.35	36.00 ± 0.63
Blood	34	109.30 ± 1.49	10.8 ± 4.58	36.00 ± 1.49
Urine	26	94.10 ± 2.65	9.79 ± 2.82	36.00 ± 0.69

1326 The Swansea sampled isolates (Alharbi *et al*-currently unpublished) raw sequence metrics are
1327 presented below (Table 11).

Table 11: Isolation source and read information from Swansea isolates. Standard deviations shown with average

Isolation Source (Niche)	Number of Isolates	Average Read Length	Average Number of Reads (Million)	Average Read Quality
Vaginal	15	84.00 ± 10.04	33.10 ± 9.99	38.00 ± 0.64
Sputum	28	88.00 ± 17.51	19.60 ± 5.47	38.00 ± 0.87
Semen	4	86.00 ± 10.52	21.60 ± 2.17	38.00 ± 0.7

1328

1329 All metadata available for these sequences is available in supplementary table 2. These sequences,
1330 along with 48 Swansea sequences, were aligned to a single reference genome as described within
1331 the methods section (section 2.2.2). Alignments of each sample were merged into a single binary
1332 alignment map (BAM) file for downstream processing. Quality control of the alignments was carried
1333 out using BAMstats to generate overall alignment statistics, this information is available in
1334 supplementary 2. Nucleotide variants were called using BCFtools (*202*), which detail the positions in
1335 the reference that differ within any sample together with the change within the sample that is
1336 observed. Types of changes identified include single nucleotide variants, small insertions and
1337 deletions (INDELS). The variant call format files (VCF) were merged into a multi sample VCF.

1338 **3.2.2 Assessment of aneuploidy**

1339 In order to determine the prevalence of aneuploidy within this dataset, an approach was used to
1340 utilise read coverage information. Coverage data was generated using BamQC and summary files
1341 were manually examined to identify aneuploidy. Any chromosome that showed an approximate 50%
1342 increase in coverage compared to the average coverage across the entire genome was determined
1343 to have trisomy for that chromosome. This increase in sequence reads and thus an increase in
1344 coverage would be evidence that a proportional increase in DNA content for that chromosome was
1345 present at the stage of DNA extraction and is direct evidence of chromosome copy number variation.

1346 Out of the 368 isolates, chromosomal duplication was detected 85 times. However, this was spread
1347 across 69 isolates (18.75 % of isolates) and indicates that some isolates showed multiple cases of
1348 aneuploidy. Aneuploidy, by means of trisomy, was detected in at least 1 isolate for every
1349 chromosome. However, aneuploidy rate was highest for chromosome 7 (29 isolates). Interestingly, it
1350 was also observed that this chromosome also yielded the highest trisomy rate for isolates from the

reproductive tract (present in 6 isolates). This represents the highest aneuploidy count for a single known isolation source within this dataset. On the other hand, for vaginal isolates, no aneuploidy was detected for chromosomes 3, 4, 6 and R. Across all SRAs and Swansea isolates the lowest observed aneuploidy rate was for chromosomes 1 and R. Unexpectedly, the GI-tract isolates showed no aneuploidy at all. Despite these findings, the missing isolation sources on many of the human derived samples make drawing conclusions difficult since this group (missing), contained the highest overall rate. It was the only group to display aneuploidy for chromosomes 3, 4 and 6 and was also the highest for chromosomes 5 and 7. This, of course, clouds judgment on the association of aneuploidy with a body site. Further, it does not provide evidence for the root cause of aneuploidy since treatment data is also missing from the datasets. Stability of aneuploidy detected can also not be determined without culture-based methods.

Full BamQC reports are available at supplementary 5 and full aneuploidy information is present in supplementary 2.

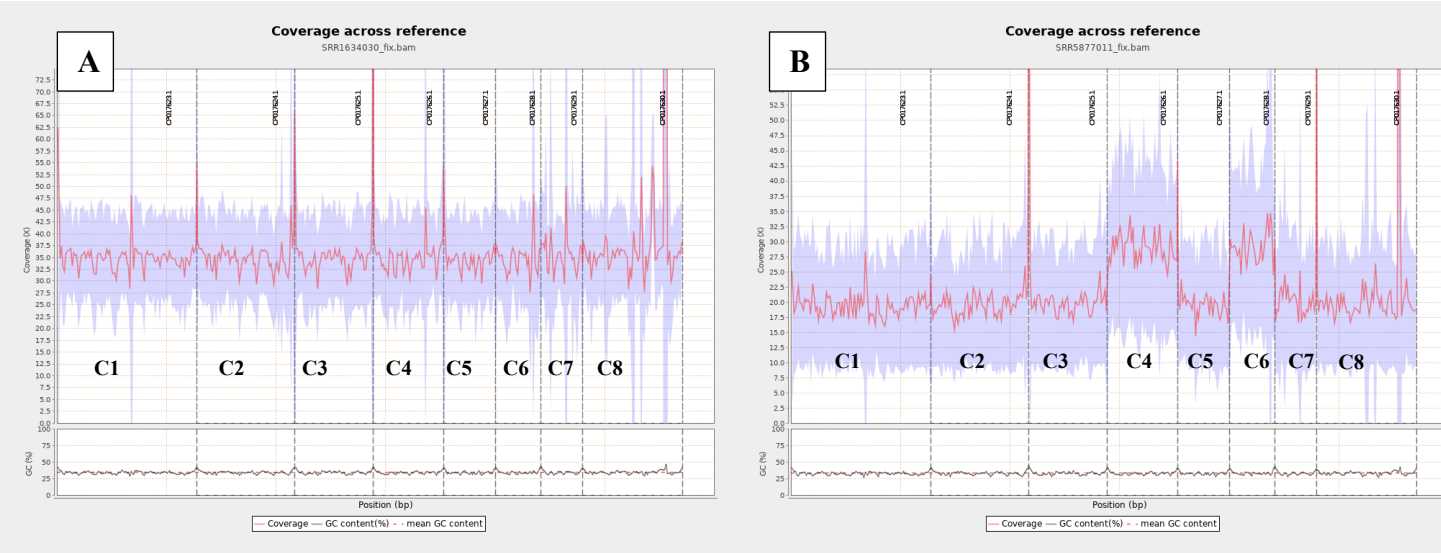


Figure 13: Examples of coverage graphs generated by BamQC. (A) An example of a sequence not showing any aneuploidy. (B) An example of a sequence showing aneuploidy. Chromosomes 4 and 6 show an approximate genome coverage increase of 50% when compared to other chromosomes of the same isolate, indicating that these chromosomes are triploid.

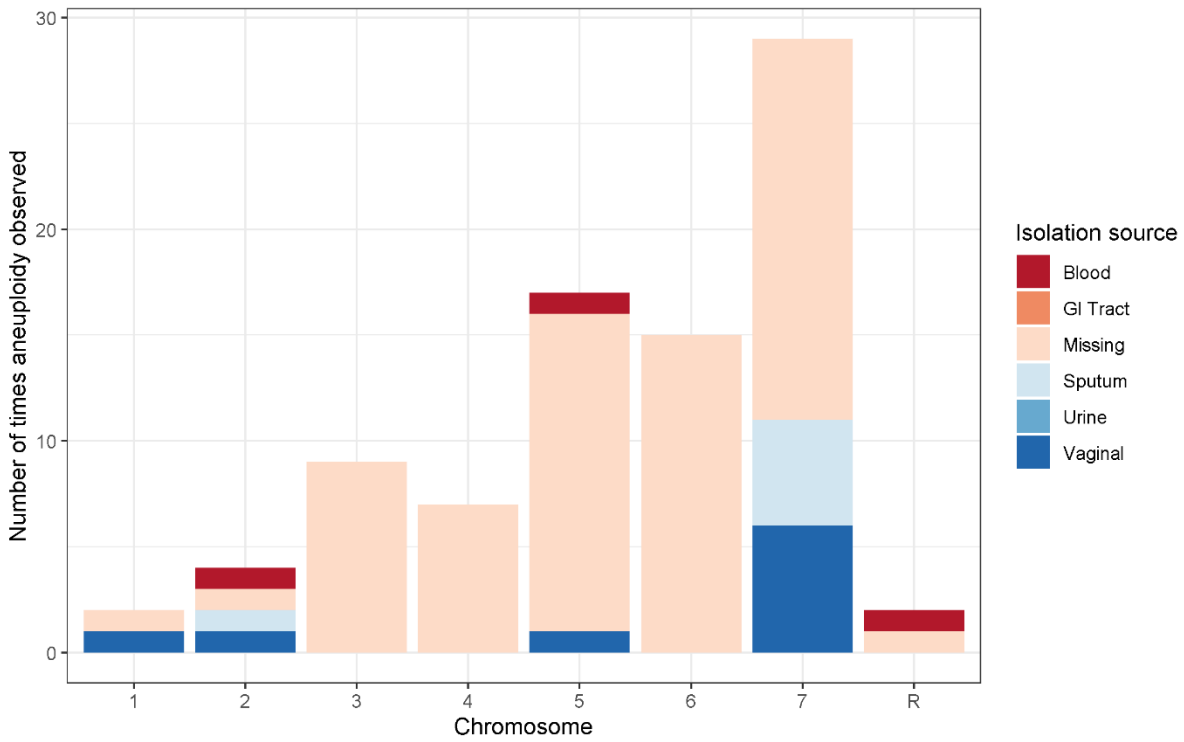
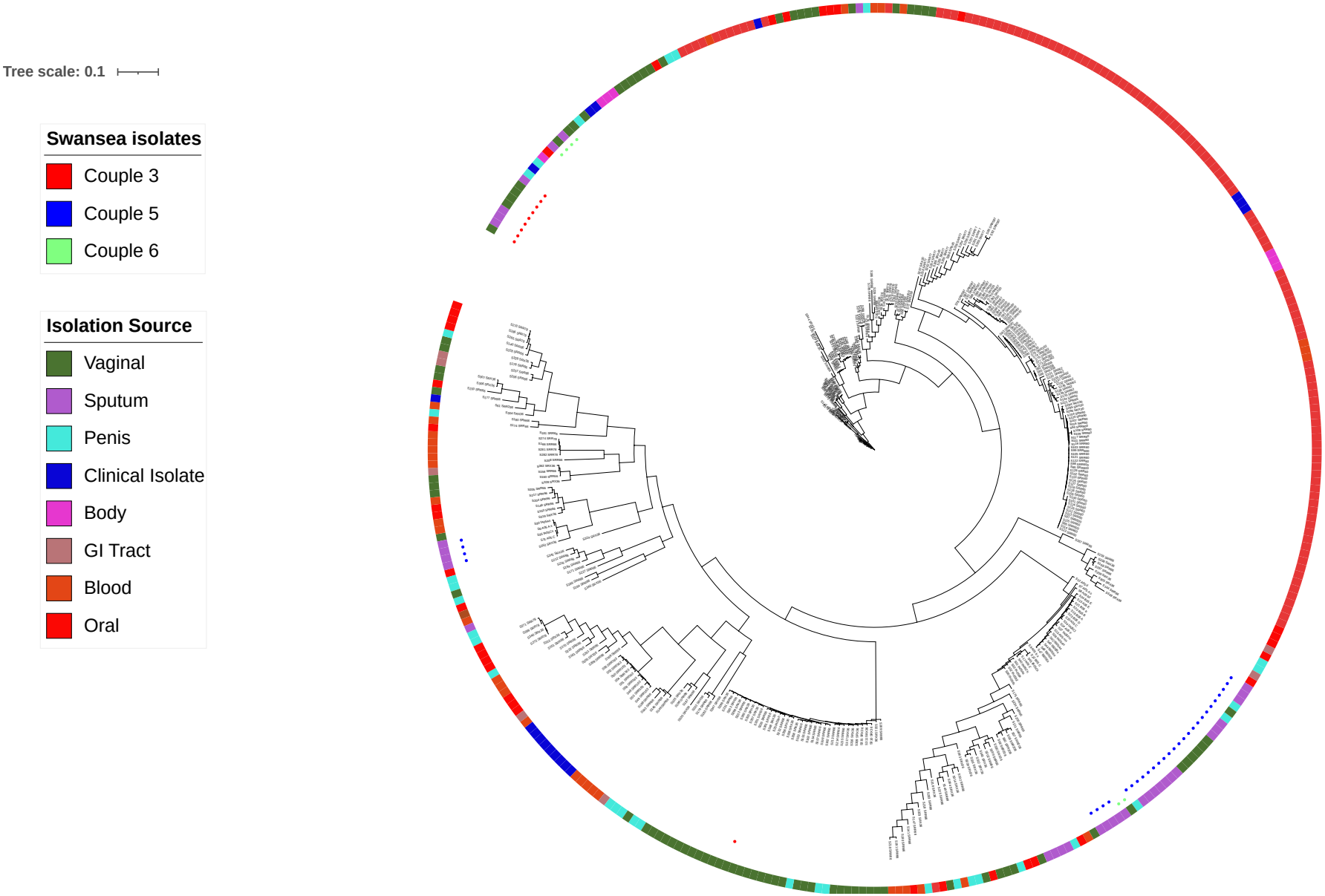


Figure 14 – Summary of the number of times trisomy was observed throughout the Swansea and SRA sequences organised by niche and the chromosome where aneuploidy was detected. Coverage data generated using BamQC. Graph was produced in R using ggplot2.

3.2.3 SNP-based Phylogeny Reconstruction

In order to reconstruct the phylogenetic relationship of all isolates understudy here, an approach was employed which uses shared variant sites across all isolates. A total of 19,050,775 variant positions were detected across this dataset. The smallest number of SNPs was detected in isolate SRR7704197 (15263 SNPs), while the most was detected in isolate SRR6001262 (23934472 SNPs). Of course, this shows only relatedness to the single reference genome. A multi-sample VCF file containing variant information representative for all 368 isolate genome sequences was used in the construction of a Maximum Likelihood SNP phylogenetic tree. After stringent filtering steps (see Methods section 2.2.3), a concatenated sequence of 21,682 high quality variants were used for tree construction using SNPylo applying a maximum likelihood method. The output Newick formatted file was rendered and annotated using ITOL (Figure 15) (*203*). The tree was annotated using the isolation source (niche).



1418

Figure 15: SNP-based maximum likelihood *Candida albicans* phylogeny tree constructed using 21,682 SNPs in SNPhylo and visualised in ITOL.

This showed that sequences from similar sources or the same study were most likely to be closely related to each other. Swansea isolates seemed to share genotypes across body sites and were more closely related to each other than other isolates from the same body source. Swansea isolates from the same couple also seemed to cluster together. There was no obvious clustering based on isolation niche, indicating a good number of contrasting pairs were available to inform the GWAS analysis.

3.2.4 Functional Effects of Identified Variants

Functional effects of variants from all Swansea reads and SRAs were identified. The average mutation rate across all chromosomes for all isolates was 1 in 43, whereby there is a single variant position in 43 bases of the reference genome. Interestingly isolates sourced from sputum had a significantly lower variant rates of when compared to isolates from all other isolation sources as shown in Figure 15.

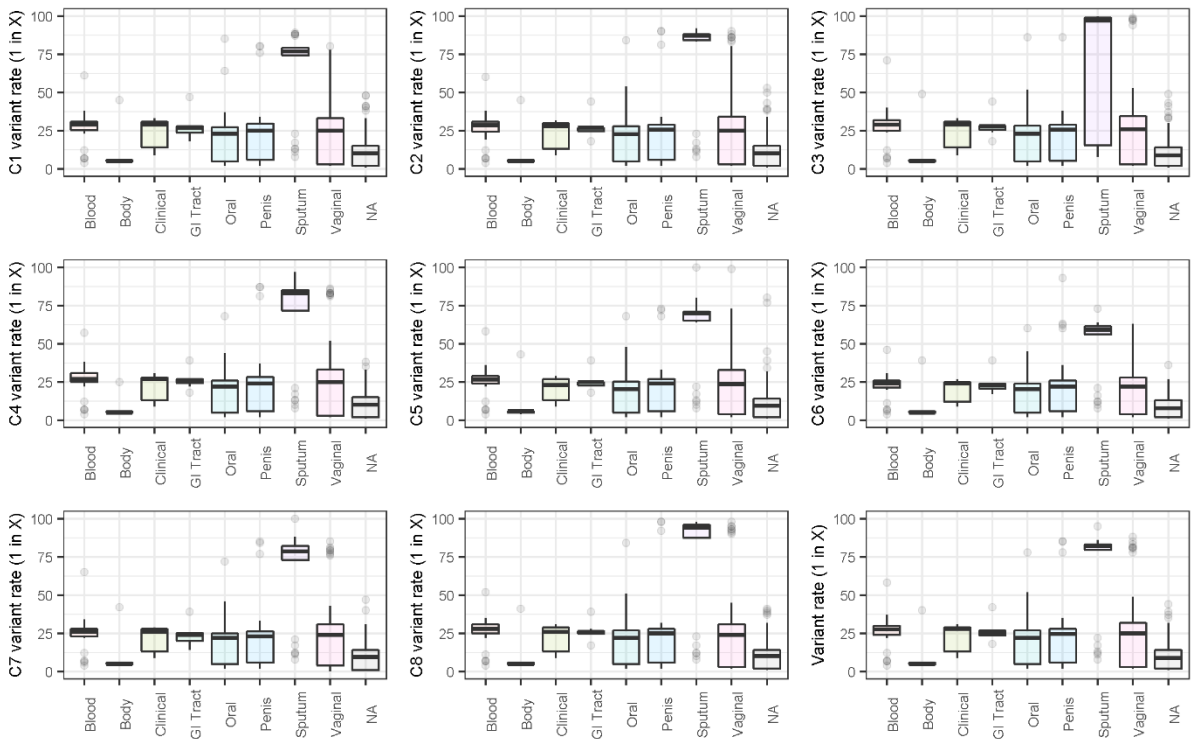


Figure 16 – Rate of variants across all reads, separated by isolation source (X axis) and faceted by chromosome number (Y axis). Variant rate information generated by snpeff and plotted in R using ggplot. Black lines inside boxes show the median value, box borders indicate the upper (75th percentile) and lower (25th percentile) quartiles. Whiskers show the maximum (95th percentile) and minimum (5th percentile) data points. Additional dots show outliers.

1435

1436

1437

Table 12: P-values from wilcox testing comparing isolate variant rates based on isolation source.

Group 1	Group 2	Variant rate P-value
Body	Blood	0.0171
Clinical	Blood	0.365
Clinical	Body	0.021
GI Tract	Blood	0.485
GI Tract	Body	0.043
GI Tract	Clinical	0.913
Oral	Blood	0.013
Oral	Body	0.485
Oral	Clinical	0.306
Oral	GI Tract	0.306
Penis	Blood	0.171
Penis	Body	0.193
Penis	Clinical	0.581
Penis	GI Tract	0.645
Penis	Oral	0.413
Sputum	Blood	0.0000075
Sputum	Body	0.000434
Sputum	Clinical	0.000147
Sputum	GI Tract	0.0157
Sputum	Oral	0.0000004
Sputum	Penis	0.0000199
Vaginal	Blood	0.611
Vaginal	Body	0.611
Vaginal	Clinical	0.933
Vaginal	GI Tract	0.951
Vaginal	Oral	0.193
Vaginal	Penis	0.913

1438

1439 Average transition/transversion (Ts/Tv) ratio across the isolates was 2.12 with isolates sourced from
1440 the sputum showing a significant difference to isolates from other sources, as shown in Figure 16.

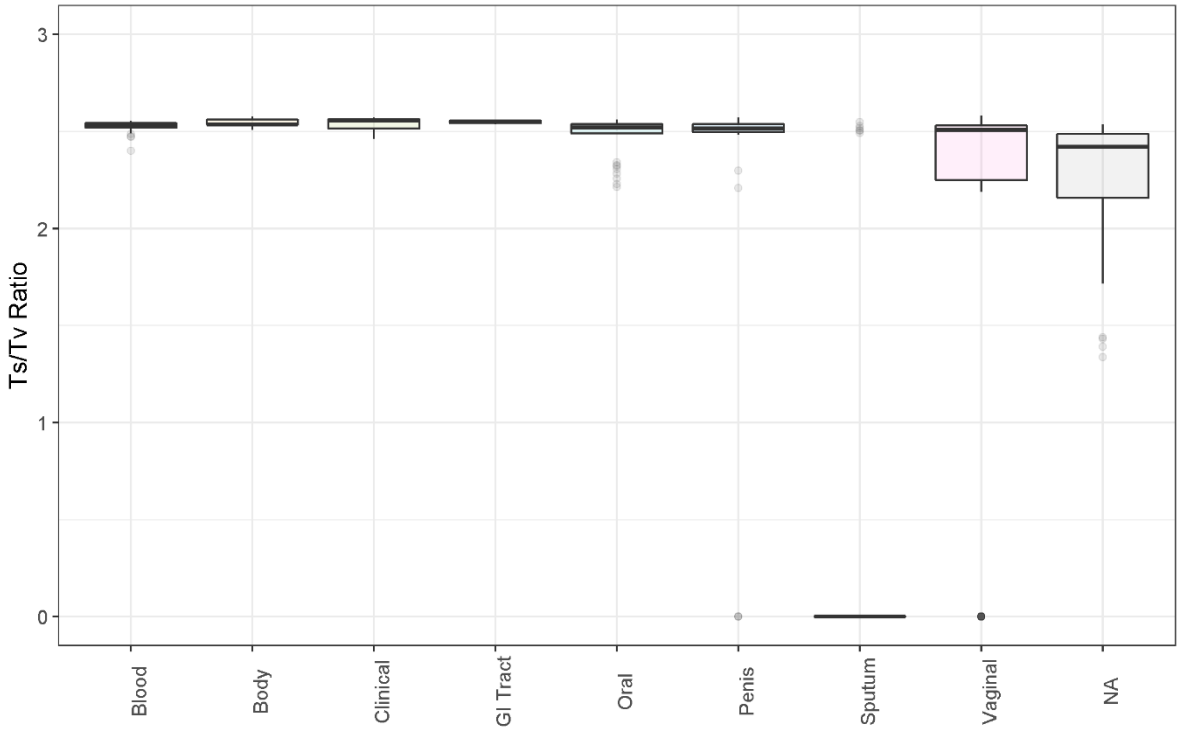


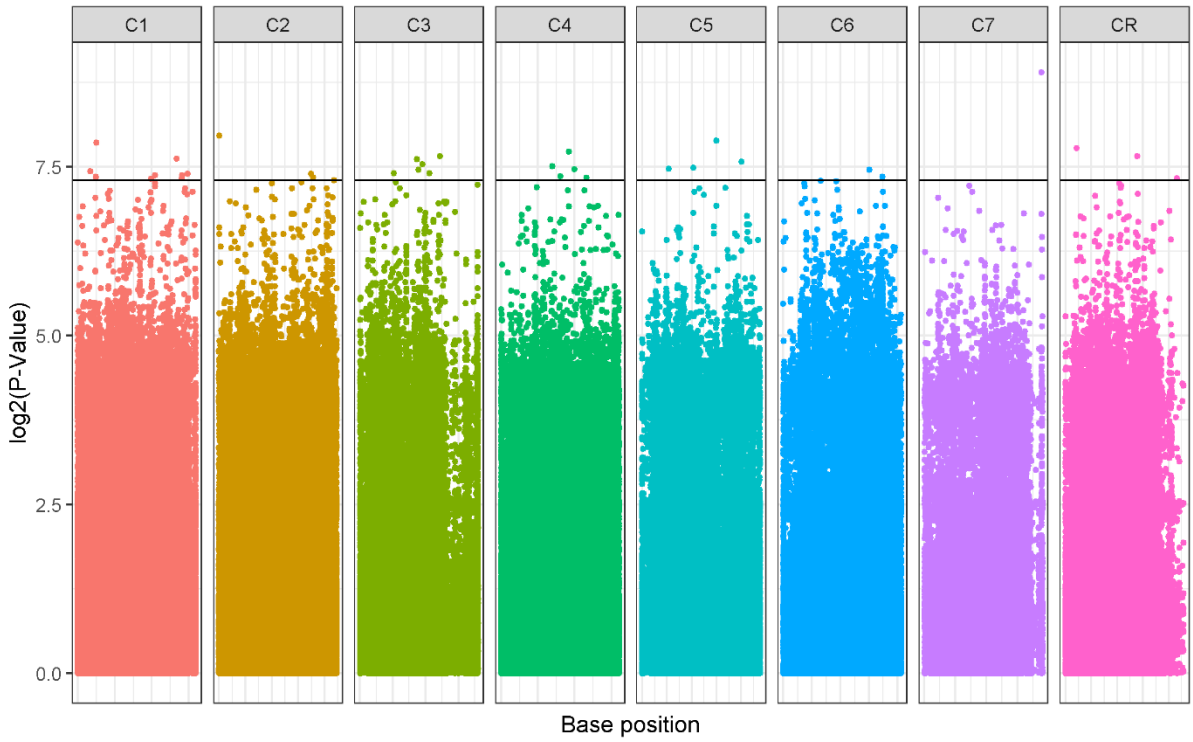
Figure 16– Transition/transversion ratio across all Swansea reads and SRAs, separated by isolation source. Variant rate information generated by SNPeff and plotted in R using ggplot. Black lines inside boxes show the median value, box borders indicate the upper (75th percentile) and lower (25th percentile) quartiles. Whiskers show the maximum (95th percentile) and minimum (5th percentile) data points. Additional dots show outliers.

1441

1442

1443 **3.2.5 - Genome Wide Association Study**

1444 To identify any association between phenotypes and the variants detected through the alignment and
1445 variant calling process, a genome wide association study (GWAS) was carried out. The total 19,050,775
1446 variants from the combined variant count of all samples was entered into a pre-processing or filtering
1447 stage leaving 90,827 variants to be used in the analysis stage. Two separate tests were performed to
1448 identify the presence of an association between the variants and two traits, those being a propensity
1449 to present chromosomal aneuploidy and secondly the isolation source (niche). From these tests, the
1450 only variants determined to have statistically significant association between variant and trait was the
1451 vaginal isolation source. A total of 35 variants had p-values above the threshold log2 P-value. These
1452 were spread across the genome with ten located on chromosome 1, three located on chromosome 2,
1453 six on chromosome 3 and 4 respectively, four on chromosome 5, two on chromosome 6, a single on
1454 chromosome 7 and three on chromosome R (8).



1455

1456 Figure 17 – Manhattan plot of the 90,827 filtered variants p-values when tested against isolation from a vaginal
1457 source. A cut off value of $5*10^{-8}$ was used to determine significance. 35 variants have a p-value greater than that of the cut off value. All p-values were standardized by performing a log2 conversion. Association data was generated using PLINK and plotted in R using ggplot2.

1458

Table 13– Location of all variants significantly associated with isolation from a vaginal source and the gene in which each variant is located.

Chromosome:Base Position	P-value	Gene
1:340102	3.68E-08	ESP1
1:481885	4.52E-08	PSY2
1:487815	4.43E-08	N/A
1:498748	1.39E-08	Arf family GTPase
1:1975851	4.73E-08	poly(A)-specific ribonuclease
1:2083270	4.23E-08	NUP188
1:2672052	2.40E-08	SNT1
1:2802301	4.69E-08	N/A
1:2815637	4.26E-08	N/A
1:2970390	4.04E-08	ssDNA endonuclease
2:1844	1.09E-08	N/A
2:1744236	4.03E-08	YWP1
2:1779845	4.54E-08	TIF4631

3:517086	3.95E-08	ALR1
3:870424	2.46E-08	KRE9
3:888024	3.49E-08	Ctp1p
3:949550	2.92E-08	ADE2
3:1063249	3.96E-08	HGT4
3:1220627	2.22E-08	N/A
4:693436	3.10E-08	N/A
4:803987	4.41E-08	NUP84
4:803991	4.41E-08	NUP84
4:920308	1.89E-08	COX11
4:994178	3.44E-08	N/A
4:1158844	4.57E-08	LYS1
5:270198	3.39E-08	VPS1
5:516168	3.25E-08	CST20
5:750488	1.29E-08	PCL1
5:1005308	2.67E-08	N/A
6:756510	3.48E-08	POX1
6:869278	4.42E-08	ALG2
7:944344	1.26E-09	TLO16
8:209981	1.68E-08	CCE1
8:1391298	2.22E-08	Rab family GTPase
8:2164505	4.66E-08	Cht3p

1459

1460 To follow this analysis, the genes with significantly associated variants identified by GWAS (table 12)
1461 were analysed for enrichment using the gene ontology panther classification system (*192*), querying
1462 the list of genes against the *Candida albicans* specific reference list. Three separate annotation data
1463 sets were used; GO biological process, GO molecular function and GO cellular component.
1464 Enrichment was detected in all annotation data sets and are presented in the tables that follow
1465 (Tables 14-16) which display only FDR $P < 0.05$. Whilst several broader GO hits have been identified,
1466 of interest is the > 100-fold enrichment within the multivesicular body assembly definition
1467 (GO:0036258) together with the 94.3-fold enrichment of protein localisation to endosome
1468 (GO:0036010). Both have roles in the assimilation of nutrients from the environment, a prerequisite

1469 to successful growth and in sustained membrane trafficking. Further to this, enrichment was
1470 detected within the GO cellular component set whereby 28.18-fold enrichment (FDR P-value 3.00E-
1471 03) was observed in the late endosome definition (GO:0005770). Within the molecular function set,
1472 two main categories were found to be enriched, GTPase activity (GO:0003924) showed 26.94-fold
1473 enrichment (FDR p-value 7.52E-10) and GTP binding (GO:0005525) showed 24.49-fold enrichment
1474 (FDR p-value 1.28E-08). These likely to be at least partially due to variants within the Arf family
1475 GTPase and Rab family GTPase genes among others listed within table 12.
1476

Table 14: Gene ontology results from genes with significant variants looking at biological process.

GO biological process	Number of reference genes in category	Number of genes in category input	Number of genes expected in input	Over/Under enrichment of category	Fold enrichment of genes observed	Raw P-Value	False discovery rate
multivesicular body assembly (GO:0036258)	2	2	.01	+	> 100	1.61E-04	1.93E-02
multivesicular body organization (GO:0036257)	3	2	.02	+	> 100	2.67E-04	2.95E-02
protein localization to endosome (GO:0036010)	4	2	.02	+	94.30	3.99E-04	4.09E-02
protein localization to phagophore assembly site (GO:0034497)	14	3	.07	+	40.41	8.62E-05	1.39E-02
protein localization to Golgi apparatus (GO:0034067)	17	3	.09	+	33.28	1.43E-04	1.83E-02
vacuole inheritance (GO:0000011)	17	3	.09	+	33.28	1.43E-04	1.79E-02
small GTPase mediated signal transduction (GO:0007264)	59	9	.31	+	28.77	3.70E-11	2.09E-07
vacuole organization (GO:0007033)	76	7	.40	+	17.37	1.77E-07	3.33E-04
cytosolic transport (GO:0016482)	73	6	.39	+	15.50	2.78E-06	2.24E-03

macroautophagy (GO:0016236)	74	6	.39	+	15.29	3.00E-06	1.54E-03
cellular response to heat (GO:0034605)	71	5	.38	+	13.28	4.18E-05	8.41E-03
retrograde transport, endosome to Golgi (GO:0042147)	60	4	.32	+	12.57	3.24E-04	3.45E-02
response to heat (GO:0009408)	82	5	.43	+	11.50	8.02E-05	1.33E-02
endosomal transport (GO:0016197)	108	6	.57	+	10.48	2.35E-05	5.29E-03
cytoplasm to vacuole transport by the Cvt pathway (GO:0032258)	90	5	.48	+	10.48	1.22E-04	1.76E-02
response to temperature stimulus (GO:0009266)	92	5	.49	+	10.25	1.35E-04	1.85E-02
endocytosis (GO:0006897)	93	5	.49	+	10.14	1.41E-04	1.85E-02
nuclear transport (GO:0051169)	171	8	.91	+	8.82	2.82E-06	1.99E-03
nucleocytoplasmic transport (GO:0006913)	171	8	.91	+	8.82	2.82E-06	1.76E-03
organelle assembly (GO:0070925)	189	8	1.00	+	7.98	5.77E-06	1.81E-03
intracellular signal transduction (GO:0035556)	234	9	1.24	+	7.25	2.86E-06	1.61E-03
protein targeting to vacuole (GO:0006623)	225	8	1.19	+	6.71	1.98E-05	4.66E-03
establishment of protein localization to vacuole (GO:0072666)	231	8	1.22	+	6.53	2.39E-05	5.18E-03
protein localization to vacuole (GO:0072665)	236	8	1.25	+	6.39	2.77E-05	5.79E-03

signal transduction (GO:0007165)	307	10	1.63	+	6.14	3.03E-06	1.32E-03
signaling (GO:0023052)	312	10	1.65	+	6.04	3.50E-06	1.41E-03
regulation of cell cycle (GO:0051726)	251	7	1.33	+	5.26	3.13E-04	3.39E-02
intracellular protein transport (GO:0006886)	514	14	2.73	+	5.14	1.31E-07	3.68E-04
protein transport (GO:0015031)	556	14	2.95	+	4.75	3.40E-07	4.80E-04
establishment of protein localization (GO:0045184)	571	14	3.03	+	4.62	4.70E-07	5.30E-04
protein localization to organelle (GO:0033365)	513	12	2.72	+	4.41	7.00E-06	1.97E-03
establishment of protein localization to organelle (GO:0072594)	446	10	2.36	+	4.23	7.43E-05	1.31E-02
cellular protein localization (GO:0034613)	651	14	3.45	+	4.06	2.26E-06	2.12E-03
cellular macromolecule localization (GO:0070727)	667	14	3.54	+	3.96	3.01E-06	1.41E-03
protein localization (GO:0008104)	685	14	3.63	+	3.85	4.12E-06	1.55E-03
cell communication (GO:0007154)	546	11	2.90	+	3.80	7.60E-05	1.30E-02
vesicle-mediated transport (GO:0016192)	499	10	2.65	+	3.78	1.87E-04	2.16E-02
filamentous growth of a population of unicellular organisms (GO:0044182)	451	9	2.39	+	3.76	4.43E-04	4.46E-02
filamentous growth (GO:0030447)	579	11	3.07	+	3.58	1.29E-04	1.81E-02

nitrogen compound transport (GO:0071705)	740	14	3.92	+	3.57	1.01E-05	2.72E-03
macromolecule localization (GO:0033036)	797	15	4.23	+	3.55	4.30E-06	1.51E-03
growth (GO:0040007)	588	11	3.12	+	3.53	1.47E-04	1.81E-02
cellular component assembly (GO:0022607)	653	11	3.46	+	3.18	3.69E-04	3.86E-02
intracellular transport (GO:0046907)	842	14	4.46	+	3.14	4.42E-05	8.59E-03
establishment of localization in cell (GO:0051649)	874	14	4.63	+	3.02	6.70E-05	1.22E-02
organic substance transport (GO:0071702)	1182	18	6.27	+	2.87	5.25E-06	1.74E-03
cellular response to stimulus (GO:0051716)	1204	18	6.38	+	2.82	6.89E-06	2.04E-03
response to stimulus (GO:0050896)	1424	19	7.55	+	2.52	1.69E-05	4.33E-03
regulation of cellular process (GO:0050794)	1375	18	7.29	+	2.47	4.65E-05	8.73E-03
organelle organization (GO:0006996)	1314	17	6.97	+	2.44	1.08E-04	1.68E-02
regulation of biological process (GO:0050789)	1471	18	7.80	+	2.31	1.19E-04	1.76E-02
cellular component organization (GO:0016043)	1650	19	8.75	+	2.17	1.82E-04	2.14E-02
biological regulation (GO:0065007)	1666	19	8.83	+	2.15	2.01E-04	2.27E-02
transport (GO:0006810)	1764	20	9.35	+	2.14	1.14E-04	1.74E-02

1477

1478

1479

Table 15: Gene ontology results from genes with significant variants looking at molecular function.

GO molecular function complete	Number of reference genes in category	Number of genes in category input	Number of genes expected in input	Over/Under enrichment of category	Fold enrichment of genes observed	Raw P-Value	False discovery rate
GTPase activity (GO:0003924)	76	11	.40	+	27.30	2.78E-13	6.59E-10
GTP binding (GO:0005525)	77	10	.41	+	24.49	1.08E-11	1.28E-08
guanyl ribonucleotide binding (GO:0032561)	78	10	.41	+	24.18	1.22E-11	9.59E-09
guanyl nucleotide binding (GO:0019001)	78	10	.41	+	24.18	1.22E-11	7.19E-09
nucleoside-triphosphatase activity (GO:0017111)	203	11	1.08	+	10.22	5.48E-09	2.59E-06
pyrophosphatase activity (GO:0016462)	231	11	1.22	+	8.98	1.99E-08	7.86E-06
hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides (GO:0016818)	232	11	1.23	+	8.94	2.08E-08	7.03E-06

hydrolase activity, acting on acid anhydrides (GO:0016817)	232	11	1.23	+	8.94	2.08E-08	6.15E-06
purine ribonucleoside triphosphate binding (GO:0035639)	407	12	2.16	+	5.56	6.39E-07	1.68E-04
purine ribonucleotide binding (GO:0032555)	414	12	2.20	+	5.47	7.64E-07	1.81E-04
purine nucleotide binding (GO:0017076)	421	12	2.23	+	5.38	9.10E-07	1.96E-04
ribonucleotide binding (GO:0032553)	437	12	2.32	+	5.18	1.34E-06	2.65E-04
carbohydrate derivative binding (GO:0097367)	442	12	2.34	+	5.12	1.51E-06	2.75E-04
anion binding (GO:0043168)	553	13	2.93	+	4.43	2.32E-06	3.92E-04
nucleotide binding (GO:0000166)	571	13	3.03	+	4.29	3.30E-06	5.21E-04
nucleoside phosphate binding (GO:1901265)	571	13	3.03	+	4.29	3.30E-06	4.89E-04

small molecule binding (GO:0036094)	633	13	3.36	+	3.87	1.02E-05	1.42E-03
hydrolase activity (GO:0016787)	773	13	4.10	+	3.17	8.50E-05	1.06E-02
ion binding (GO:0043167)	909	15	4.82	+	3.11	2.17E-05	2.85E-03

1480

Table 16: Gene ontology results from genes with significant variants looking at cellular component.

GO cellular component complete	Number of reference genes in category	Number of genes in category input	Number of genes expected in input	Over/Under enrichment of category	Fold enrichment of genes observed	Raw P-Value	False discovery rate
late endosome (GO:0005770)	38	5	.20	+	24.81	2.46E-06	2.67E-03
endosome (GO:0005768)	125	6	.66	+	9.05	5.16E-05	1.87E-02
microtubule cytoskeleton (GO:0015630)	107	5	.57	+	8.81	2.65E-04	3.60E-02
cytoplasmic vesicle (GO:0031410)	201	7	1.07	+	6.57	8.16E-05	1.77E-02
intracellular vesicle (GO:0097708)	201	7	1.07	+	6.57	8.16E-05	1.47E-02
vesicle (GO:0031982)	263	8	1.39	+	5.74	5.89E-05	1.60E-02

endomembrane system (GO:0012505)	683	13	3.62	+	3.59	2.31E-05	1.25E-02
cell periphery (GO:0071944)	823	13	4.36	+	2.98	1.62E-04	2.51E-02
cytoplasmic vesicle (GO:0031410)	201	7	1.07	+	6.57	8.16E-05	1.77E-02
intracellular vesicle (GO:0097708)	201	7	1.07	+	6.57	8.16E-05	1.47E-02
vesicle (GO:0031982)	263	8	1.39	+	5.74	5.89E-05	1.60E-02

1502 **Chapter 4 – Discussion**

1503 **4.1 – Quality control of *Candida albicans* database genome assemblies.**

1504 In order to investigate the role that *Candida albicans* plays in reproductive health and fertility, a
1505 genomics-based analysis was conducted leveraging the sequence and associated metadata held
1506 within publicly available databases. Prior to the use of sophisticated techniques for comparative
1507 genomics a thorough quality control procedure should also be applied. To this end, several
1508 exploratory metrics were produced to confirm the reliability of the dataset. Following this a
1509 pangenome was constructed to assess gene presence/absence within the database assemblies.
1510 Good pangenome construction was found to be influenced by consistent assembly type as well as
1511 identifying a number of genes associated with virulence with were enriched in both the core and
1512 accessory genomes.

1513 The GC content and genome size throughout the database assemblies seemed to remain consistent,
1514 with values of approximately 34% (GC content) and 14 Mbases respectively. However, four
1515 assemblies (UAB040-W3D3, UAB090-W2D7, UAB012-W3D5 and UAB012-W7D4) all isolated from
1516 the female reproductive tract showed decreased GC content relative to the other assemblies. Kiktev
1517 *et al* studying altered GC content within *Saccharomyces cerevisiae* has identified a link between
1518 elevated GC content and mutation and recombination rates (*204*). Using an *in vitro* model, they
1519 found that an approximate doubling in GC content within a gene, whilst preserving the protein's
1520 primary sequence, could lead to a seven-fold increase in the rate of mutations compared to the gene
1521 with the lower GC content, with half of these mutations being single-base substitutions caused by
1522 DNA polymerase ζ, as well as deletions and duplications caused by polymerase slippage. The
1523 elevated GC content gene also showed increased rates of recombination relative to lower GC forms
1524 of the gene, this has previously been linked to an increased susceptibility to recombinogenic double-
1525 strand breaks (*205*). This would indicate that up shifts in GC content is one of the driving forces in
1526 evolution because of its effect on mutation and recombination rates. The low GC content within the
1527 *Candida albicans* assemblies mentioned above could be a mechanism to protect against mutation
1528 caused by errors in DNA replication. Kiktev *et al* also theorised that in a diploid organism,
1529 heterozygous genes with a high-GC and low-GC form could counterbalance the high mutation and
1530 recombination rates through gene conversion. When double strand breaks occur in the high GC
1531 content gene it will most likely be replaced by the low GC content form of the gene. Due to the
1532 environment the afore mentioned *Candida albicans* isolates came from, reproductive tract in a case
1533 of vulvovaginal candidiasis, could lead to increased exposure to DNA damage, such as through
1534 exposure to reactive oxygen species, causing more double strand breaks and leading to the

1535 replacement of higher GC content genes with their lower GC content versions. This is however, at
1536 odds with GC-biased gene conversion (gBGC) (*206*) however, the no exploration of the GC content
1537 between genic and intergenic regions was conducted.

1538 **4.2 – *Candida albicans* taxonomic assignments**

1539 All *Candida albicans* assemblies clustered closely and showed high sequence similarity to each other
1540 indicating that they have all been correctly classified as *Candida albicans.* The *Naumovozyma*
1541 *dairenensis* included also clustered within the *Candida albicans* clade indicating that it has been
1542 incorrectly classified, as suggested in the previous literature (*201*), and that it should be relabelled as
1543 a *Candida albicans strain* instead. Also of interest was the location of *Saccharomyces paradoxus*
1544 which, while still showing low sequence similarity, grouped closer to the non-*albicans Candida* than
1545 the two *Saccharomyces cerevisiae* assemblies included in the analysis. This could also suggest
1546 misclassification of the *Saccharomyces paradoxus* assembly, although more robust methods would
1547 need to be used to explore this. However, previous phylogenetic studies in the literature would
1548 suggest that this is not the case and that the taxonomic classification of *S. paradoxus* is correct (*207*).
1549 These findings both highlight how quality control of database sequences is an unfortunate necessity
1550 prior to in depth and computationally intensive analysis techniques. It also highlights limitations to
1551 the use of non-curated databases.

1552 **4.3 – Plant isolated *Candida albicans* assemblies show significant differences in genomic**
1553 **features**

1554 Chromosomal dotplots did not indicate any major chromosomal alterations within the *Candida*
1555 *albicans* assemblies with only the diploid assemblies showing any significant differences due to both
1556 forms of each chromosome being represented on the dotplots. Differences between several types of
1557 genomic features (Insertions, insertion size, deletions, relocations, translocations, breakpoints,
1558 inversions and SNPs) were reported between assemblies of different isolation sources. Diploid
1559 assemblies showed increased numbers of insertions and average insertion size compared to the
1560 reference, as was expected due to the nature of the diploid assemblies. Statistical analyses of the
1561 genomic features detected between niches only showed a significant difference when the
1562 assemblies isolated from plants were included, with these being with relocation with blood isolates
1563 and translocations with the vaginal and other isolates. These differences can potentially be
1564 attributed to the high levels of genetic diversity seen in all three assemblies (*164*), with all three
1565 showing high levels of heterozygosity when compared to common clinical strains. However, all three
1566 strains showed to be genetically closer to clinical strains with similar heterozygosity levels than to
1567 each other suggesting they came from unique sources as opposed to being a contaminant from

1568 humans or wild animals. Bensasson *et al* hypothesised that it is most likely that these levels of
1569 heterozygosity represent the ancestral levels for *Candida albicans* and that clinical strains showing
1570 lower levels have undergone loss-of-heterozygosity events over time. This is further supported by
1571 the high natural fitness of these strains when compared to other clinical strains.

1572

1573 **4.4 – Pangenome construction quality is based on genome assembly type**

1574 When pangenome construction was consistent across all assemblies a much more conventional split
1575 of core and accessory genes was obtained, more in line with previous *Candida albicans* pangenome
1576 construction (*208*). This suggests that consistent assembly type is important in good eukaryotic
1577 pangenome construction. As well as consistent assembly type aneuploidy must also be considered
1578 during pangenome construction as this could potentially skew the core and accessory genome split,
1579 a factor unusually discussed due to the vast majority of pangenome construction reported on
1580 bacterial species. Aneuploidy was detected within the Swansea isolated genomes, however with a
1581 lack of trusted genome assemblies due to the Illumina only sequencing technique, the effect that
1582 this has on pangenome construction has yet to be explored. Theoretically, the pangenome wide
1583 effect would be dependent on genome assembly quality and the nature of aneuploidy (whole
1584 chromosomes vs partial chromosome loss/gain). The objective of pangenome construction is to
1585 essentially understand and explain phenotypes from a genotypic context. However, the presence of
1586 aneuploidy complicates this approach in its traditional sense. A link between core-accessory gene
1587 and chromosomal location has been detected in several pangenomes (*179, 208*) highlighting the
1588 accessory genes localise in regions of chromosomes more likely encounter structural variation.
1589 However, how this is affects the pangenome when aneuploidy is present remains unknown.

1590 Computational resource intensity was also shown during pangenome construction, with pangenome
1591 construction taking three days using 40 CPUs through SCW, while it took over a week using 8 CPUs
1592 on a CLIMB VM. In order to further increase the number of strains added to a pangenome further
1593 processing power would be required to prevent excessive construction time.

1594 The difference between the number of clusters identified in this study and the Chao lower bound
1595 estimate would suggest that a large proportion of sequences are still to be identified within this
1596 pangenome. This number may be skewed by high numbers of singleton and doubleton sequences
1597 due to the presence of highly fragmented genomes, potentially like some of the assemblies used
1598 with high contig numbers (*190*). However, due to the conservative nature of the Chao lower bound it
1599 is still likely there are a high number of genes available to explore within this pangenome that may

1600 become apparent with the addition of more and higher quality genomes. This highlights the
1601 potential for further gene level diversity of *Candida albicans* generated through genetic variayopm
1602 events such as translocations, truncations and aneuploidy of chromosomes together with loss of
1603 heterozygosity (LOH) (*209*).

1604

1605 **4.5 –The pangenome of C. albicans displays enrichment of virulence genes**

1606 Enrichment analysis of core genes identified a 10.31-fold enrichment of genes with the ability to kill
1607 host cells. The orthologs identified were SHE3, KEX1, KEX2 and ECE1. SHE3 is a mRNA binding protein
1608 that is required for normal filamentation and epithelial cell damage (*210*). KEX1 is a
1609 carboxypeptidase involved in the maturation of candidalysin Ece1p (*211*) and KEX2 is a proprotein
1610 convertase involved in hyphal-growth, virulence and maturation of candidalysin Ece1p (*211, 212*).
1611 ECE1 codes for candidalysin, a cytolytic peptide toxin essential for mucosal infection, ECE1 is only
1612 expressed when *Candida albicans* is in the hyphal morphology (*53*). Usually core genes are those
1613 that are necessary for an individual's viability which would suggest that the ability for *Candida*
1614 *albicans* to damage or even kill a host cell even though these genes seem to be primarily associated
1615 with pathogenicity (*213*). This would support theories that *Candida albicans* infections are heavily
1616 influenced by host factors (*14*) given the known commensalism of *Candida albicans*. The pathobiont
1617 concept is affirmed by this, the presence of core genes that regulate and enhance pathogenicity,
1618 such as SHE3, KEX1, KEX2 and ECE1, does not necessarily mean that the organism is acting as a
1619 pathogen given that some isolates within this dataset were not isolated form an infected host.

1620 The accessory genome also showed interesting enrichments for proteins involved in dolichol-linked
1621 oligosaccharide biosynthesis and chromatin DNA binding. Those orthologs involved in dolichol-linked
1622 oligosaccharide biosynthesis were all members of the ALG family, The orthologs identified were all
1623 part of the ALG family and included ALG5, 6, 7, 9, 11 and ALG13. The ALG family are
1624 glycotransferases involved in the synthesis of cell wall mannan through protein glycosylation (*214,*
1625 *215*). While these genes are usually considered to encode an essential function in this pangenome
1626 they were determined to be accessory genes. This could be due to the previously mentioned
1627 fragmented genomes that could have removed these sequences from some assemblies causing
1628 them to be incorrectly classified. This is supported by previous research in which haploid strains with
1629 null mutations in ALG7 and ALG11 were found to be inviable (*216, 217*). Alternatively, it could point
1630 toward splitting of these sequences into multiple orthologous clusters based on sequence content.
1631 Thus, suggesting some form of divergence in sequence could lead to divergence in function.

1632 A limitation in this approach comes from the nature of the genome assembles held within publicly
1633 available databases. It has been described within, how the representation of the genomes as either
1634 true diploids (both sets of chromosomes present) or using the IUPAC codes to indicate
1635 heterozygosity affects the final pangenome metrics. Since many IUPAC containing *C. albicans*
1636 genomes are not phased, they represent a consensus or chimeric representation of the actual
1637 genome. It can be theorised that this allows potential for pseudo- diversity within the resulting
1638 genome assembly and thus pangenome. There may be two approaches that could solve this issue.
1639 The first is to produce fully phased telomere to telomere assemblies (T2T) with use of hybrid
1640 sequencing approaches which utilise both short and long read sequencing platforms to facilitate
1641 phasing of bases (*218-220*) However, this approach requires additional costs and indeed additional
1642 sequencing of isolates currently only sequenced using short read technologies. This also forgoes the
1643 use of currently sequenced strains held within databases. As presented here, a pangenome must be
1644 constructed of equal level assemblies. An alternative approach would be to use a map to
1645 pangenome strategy (*221, 222*) which can utilise a combined traditional approach of genome
1646 assembly with a read alignment approach identifying site of variation at gene level. This offers great
1647 opportunity to study the currently held short sequence reads held within public databases such as
1648 the SRA.

1649 **4.6 – Variant calling and SNP analyses**

1650 To further investigate the relationship between *Candida albicans* and reproductive health a genome
1651 wide association study was performed to identify variants that showed a significant association with
1652 isolation from different body sites. Prior to this 320 SRAs and 48 Swansea isolates were aligned, and
1653 variants called for later use. The quality of these reads were assessed and SNPs were used to
1654 examine aneuploidy of these isolates and the phylogenetic relationship between the isolates. A
1655 GWAS of the SNPs was then carried out to identify any variants showing significant association with
1656 a particular isolation source and the function of genes these variants were discovered in. In total 35
1657 variants showed a significant association with isolation from the female reproductive tract with
1658 genes that have previously been shown to play a role in *Candida albicans* virulence.

1659 **4.7 – Quality of read mapping**

1660 Of the 320 SRAs used the source of the isolate was missing for 182 isolates (56% of the isolates). The
1661 lack of data available caused difficulties later in associating any genes or mutations with a specific
1662 isolation source and may have led to false negatives or false positive during the GWAS section of the
1663 investigation. The most represented niche within the SRAs were the vaginal isolates (69 isolates, 21%
1664 of the total isolates investigates) which may have led to overrepresentation of this group in later

experiments, skewing the results. Of the 320 SRAs used in this study none used were obtained as part of metagenome studies. While no metagenome data was included in this analysis it does offer interesting future research paths. Metagenome data would allow for *Candida albicans* to be investigated *in vivo* there by removing any risk of mutations being caused by culture-based methods and allowing for better investigation of isolates that show poor growth *in vitro*. This would also allow for the relative proportion of these mutations within the population at each isolation source to be quantified.

4.8 – Aneuploidy detection

85 cases of aneuploidy were detected across 69 isolates with the highest numbers of aneuploidy found in chromosomes 5 (17 isolates), 6 (15 isolates) and 7 (29 isolates) with chromosome 7 showing the most instances of aneuploidy in isolates sourced from the reproductive tract (6 isolates). However, because sequencing of Swansea isolates was unable to occur stability of aneuploidy within these isolates was unable to be assessed. Due to the lack of data relating to the host disease state and any treatments applied these phenotypes could not be associated with increased aneuploidy. Aneuploidy within *Candida albicans* has previously been shown to be associated with increased resistance to antifungal agents and utilization of alternate carbon and nitrogen sources (*223*). Aneuploidy of the isochromosome 5L, two copes of the left arm of chromosome 5, has been frequently seen after exposure to fluconazole, which in turn has shown increased fitness in *Candida albicans* when exposed to fluconazole. This is most probably due to the presence of genes (*CYP51*) on chromosome 5 that are targeted by fluconazole and aneuploidy causing an increase in copy number and expression levels (*224, 225*). Chromosome 7 aneuploidy, the most prevalent chromosome with aneuploidy for female reproductive tract located isolates and all isolates in general, has not been shown to give an increase in fitness under stress conditions. So far it has only been associated with increased susceptibility to medium-chain fatty acids (*226*).

4.9 – *Candida albicans* phylogeny construction

Swansea isolates primarily clustered based on the couple they were isolated from and not the anatomical site they were sourced from, however some intra-host diversity was still observed with the Swansea isolates. This is line with what has previously been recorded (*160*). Clustering of isolates based around the couple they were isolated from rather than the individual also suggests the movement and colonization of *Candida albicans* between individuals which has previously been suggested as a potential reason for increases in resistance to antifungals (*227*).

1696 There was also no obvious clustering throughout all isolates included based on isolation source,
1697 providing a good number of contrasting pairs for use in the later GWAS analysis. The lack of
1698 clustering suggests any variants identified in the later GWAS analysis are less likely to be false
1699 positives and that there may be a relationship between significant variants and phenotypes of
1700 interest.

1701 **4.10 – Variant Rates were Higher than Previously Reported Studies**

1702 Mutation rates within the isolates studied had a mean value of 1 mutation per 43 bases. This is an
1703 increase on the previously reported figures that range from 1 in 500 to 1 in 100 (*150, 151*). This
1704 increase could be due to the quality of the database reads used, the Swansea isolates had a mean
1705 mutation rate of 1 in 96 bases, which was more in line with the previously reported figures. Isolates
1706 sourced from sputum showed significantly lower rates of mutations than those sourced from other
1707 body sites. This could potentially be due to being closely related to the reference sequence however
1708 this did not appear to be the case when the phylogenetic tree was studied.

1709 **4.11 – GWAS identified 35 variants significantly associated with isolation from the female**
1710 **reproductive tract**

1711 A 28.18-fold enrichment was observed in the late endosome definition. Trafficking through the
1712 endosome is a significant factor in the efficacy of antifungal agents, which is seen during treatment
1713 of *Candida albicans* with azoles (*228*). Azole treatment inhibits ergosterol biosynthesis and leads to
1714 the accumulation of toxic sterol intermediates that compromise the plasma membrane. Mutants
1715 that have impaired membrane trafficking through the late endosomal prevacuolar compartment
1716 showed significantly better growth when exposed to azoles than wild-type *Candida albicans*. These
1717 mutants also showed improved growth despite the reduction of ergosterol. However, these mutants
1718 were hypersensitive to antifungal agents that impaired other ergosterol synthesis pathways. This
1719 would suggest that common azole antifungal agents would be less effective against these mutants,
1720 necessitating the use of other antifungals. If these mutants are identified in patients suffering from
1721 vulvovaginal candidiasis it would allow for more effective treatment options to be planned, as well
1722 as potentially providing an avenue for removing strains causing recurrent vulvovaginal candidiasis.

1723 GTPase activity and GTP binding were also enriched with the ARF family GTPase and Rab family
1724 GTPase genes likely to be implicated. Hyphal growth is an important factor in *Candida* virulence and
1725 Arf family GTPases have been shown to be important regulators in hyphal growth and subsequent
1726 virulence (*229*). Arf GTPases are key regulators in membrane and protein trafficking to the plasma
1727 membrane (*230*). Loss of function mutations in these genes have been shown to cause impaired

1728 hyphal growth compared to strains with the wild-type gene, leading to decreased virulence in
1729 candidiasis (*229*). With an impaired hyphal morphology it would not be possible for these mutants to
1730 disseminate throughout a host (*1*). ARF GTPases have also been implicated in azole resistance with
1731 loss of function mutants showing an increased susceptibility to fluconazole (*231*). This would also
1732 suggest that if patients with candidiasis are shown to have strains of *Candida albicans* with loss of
1733 function mutations in ARF family genes a more effective treatment regimen can be planned.

1734 To further confirm if the mutations and genes identified in this study are having a phenotypic effect
1735 on *Candida albicans*' ability to colonize and persist within the female reproductive tract further
1736 experimentation is required. Mutagenesis experiments targeting these genes would allow for the
1737 phenotypic effects they are having to be fully explored as well as co-cultures with human cells to see
1738 any alterations to their interactions. No metagenomic data was included in these analyses however
1739 its inclusion would allow for confirmation that any significant mutations identified were due to the
1740 isolation source of the sample and that they have not been caused by the culture-based methods
1741 that were employed. This would further assist in identifying and false positives. Future GWAS
1742 experiments using isolates with more phenotypic data (i.e. growth, antifungal resistance) and using
1743 more isolates from different isolation sources (such as increasing the number of isolates from the GI
1744 tract) would make the results more robust while generating more data that could be useful in
1745 informing clinical practice for *Candida* infections.

1746 Due to the lack of information available relating to the disease state of the host and any treatments
1747 that may have been applied it is not possible to determine if these mutations show an association
1748 with colonisation (ability of a microorganism to occupy a new host niche as a commensal or a
1749 pathogen), persistence (ability of a microbial population to survive exposure to stresses such as
1750 antimicrobial agents), or resistance (ability of a microbial population to actively grow under
1751 sustained exposure to a stress such as antimicrobial agents) of *Candida albicans* within a human
1752 host.

1753 **4.12 - Suggested areas for future study**

1754 The results from this study offer suggestions for areas worthy of further study that can be
1755 segregated into two areas of research; firstly, one which deals with the methodological approaches
1756 to pangenome construction, particularly with regards to the effects of genome assembly type
1757 (haploid vs diploid). Secondly to assess the phenotypic effects of the variants which were identified
1758 as being statistically associated with vaginal colonisation.

1759 Since assembly type appears to alter the metrics of a pangenome, little is understood on how well a
1760 traditional construction technique, using assembled genomes as a starting point, handles datasets of
1761 diploid assemblies. Of course, this would require some in depth analysis comparing both assembly
1762 types and pangenome construction method. As previously mentioned, this requires that the diploid
1763 assemblies are of a high standard (low contig numbers, high completeness) and often necessitates a
1764 hybrid assembly approach. This would require the generation a novel dataset comprising assemblies
1765 generated from both short read and long read technologies of either through a *de novo* or
1766 resequencing approach.

1767 Alternative to this pangenome construction strategy would be a map to pan approach such as one
1768 represented by the Eukaryotic Pangenome Analysis Toolkit (EUPAN) (*221*). This strategy utilised both
1769 an assembly based pangenome construction, followed by aligning short sequence reads to a
1770 reference genome or pangenome and using coverage information to determine presence or absence
1771 of the gene along with identifying genotype of the aligned reads. This offers the benefit of not
1772 requiring resequencing approaches, but can take advantage of existing short read datasets, such as
1773 those held within NCBI's SRA. A large-scale map to pangenome approach for Candida albicans has
1774 yet to have been documented and offers an interesting complement to traditional GWAS
1775 approaches through the inclusion of presence absence variation (PAV) analysis or gene-PAV-based
1776 genome-wide association studies to correlate or associate gene presence with phenotype.

1777 To confirm the significance of the mutations identified in the GWAS study with successful
1778 colonization of the female reproductive tract further investigation is required. Mutagenesis
1779 experiments to investigate the effects of loss of function mutations in the genes identified will allow
1780 for the identification of phenotypic changes associated with these mutations. *In vitro* studies can be
1781 used to study changes to expression within these mutants, through RNA-sequencing to investigate
1782 the transcriptome. Phenotype studies can also be carried out to find any changes to the mutants
1783 ability to resist antifungal agents, changes in ability to grow on different mediums, biofilm formation
1784 and resistance to other stresses, such as hydrogen peroxide to represent reactive oxygen species
1785 generated by the immune system. Co-cultures with human cells can also be used to investigate
1786 differences in how these mutants interact with human cells, previous co-cultures have been able to
1787 study changes to *Candida albicans* biofilm formations, and the differences in cytokines expressed by
1788 human cells (*232*). *In vivo* studies can also be exploited to investigate the effect these mutations
1789 might have while interacting with mammalian cells without exposing *Candida albicans* to laboratory
1790 conditions that might affect gene expression, aneuploidy, or selection of mutants. Metagenome
1791 experiments can be used to analyse any changes to the genome of *Candida albicans in vivo* as well
1792 as using RNA-sequencing to investigate which genes are being actively expressed. Metagenome

experiments also allow for the identification of any strains that may be present that cannot be cultured under laboratory conditions. This can be used to confirm that the mutations identified are associated with isolation from the female reproductive tract and not from growth using culture-based methods.

1827 **References**

1828 1. N. SM, G. BA, W. JN, Candida albicans cell-type switching and functional plasticity in the
1829 mammalian host. *Nature reviews. Microbiology* **15**, (2017).
1830 2. J. M. Achkar, B. C. Fries, Candida Infections of the Genitourinary Tract. *Clin Microbiol Rev* **23**,
1831 253-273 (2010).
1832 3. W. JM, J. DJ, Carbohydrate digestibility and metabolic effects. *The Journal of nutrition* **137**,
1833 (2007).
1834 4. G. MJ, K. PW, T. SC, Glucosamine synthetase activity of the colonic mucosa in ulcerative
1835 colitis and Crohn's disease. *Gut* **18**, (1977).
1836 5. P. K, C. C, N. SM, Passage through the mammalian gut triggers a phenotypic switch that
1837 promotes Candida albicans commensalism. *Nature genetics* **45**, (2013).
1838 6. C. C, P. K, F. SD, T. BB, N. SM, An iron homeostasis regulatory circuit with reciprocal roles in
1839 Candida albicans commensalism and pathogenesis. *Cell host & microbe* **10**, (2011).
1840 7. L. SR *et al.*, In Candida albicans, white-opaque switchers are homozygous for mating type.
1841 *Genetics* **162**, (2002).
1842 8. S. B *et al.*, "White-opaque transition": a second high-frequency switching system in Candida
1843 albicans. *Journal of bacteriology* **169**, (1987).
1844 9. S. C, H. M, W. M, G. M, M. J, White-opaque switching of Candida albicans allows immune
1845 evasion in an environment-dependent fashion. *Eukaryotic cell* **12**, (2013).
1846 10. L. CY *et al.*, Metabolic specialization associated with phenotypic switching in
1847 Candidaalbicans. *Proceedings of the National Academy of Sciences of the United States of*
1848 *America* **99**, (2002).
1849 11. K. C *et al.*, Misexpression of the opaque-phase-specific gene PEP1 (SAP1) in the white phase
1850 of Candida albicans confers increased virulence in a mouse model of cutaneous infection.
1851 *Infection and immunity* **67**, (1999).
1852 12. G. J, W. D, L. SR, S. DR, Release of a potent polymorphonuclear leukocyte chemoattractant is
1853 regulated by white-opaque switching in Candida albicans. *Infection and immunity* **72**,
1854 (2004).
1855 13. S. NV *et al.*, Candida albicans White-Opaque Switching Influences Virulence but Not Mating
1856 during Oropharyngeal Candidiasis. *Infection and immunity* **86**, (2018).
1857 14. B. A. McManus, D. C. Coleman, Molecular epidemiology, phylogeny and evolution of Candida
1858 albicans. *Infect Genet Evol* **21**, 166-178 (2014).

1859 15. M. A. Hickman, C. Paulson, A. Dudley, J. Berman, Parasexual Ploidy Reduction Drives
1860 Population Heterogeneity Through Random and Transient Aneuploidy in Candida albicans.
1861 *Genetics* **200**, 781-794 (2015).
1862 16. B. J, H. L, Does stress induce (para)sex? Implications for Candida albicans evolution. *Trends in*
1863 *genetics : TIG* **28**, (2012).
1864 17. H. RA, N. MC, Fungal interactions with the human host: exploring the spectrum of symbiosis.
1865 *Current opinion in microbiology* **40**, (2017).
1866 18. J. D. Sobel, Recurrent vulvovaginal candidiasis. *Am J Obstet Gynecol* **214**, 15-21 (2016).
1867 19. v. d. V. FL *et al.*, STAT1 mutations in autosomal dominant chronic mucocutaneous
1868 candidiasis. *The New England journal of medicine* **365**, (2011).
1869 20. P. A *et al.*, Chronic mucocutaneous candidiasis in humans with inborn errors of interleukin-
1870 17 immunity. *Science (New York, N.Y.)* **332**, (2011).
1871 21. L. SK, K. CJ, K. DJ, K. JH, Immune cells in the female reproductive tract. *Immune network* **15**,
1872 (2015).
1873 22. M. B, F. LJ, R. J, Vaginal microbiome: rethinking health and disease. *Annual review of*
1874 *microbiology* **66**, (2012).
1875 23. M. D *et al.*, Genital tract infection and associated factors affect the reproductive outcome in
1876 fertile females and females undergoing in vitro fertilization. *Biomedical reports* **10**, (2019).
1877 24. F. U *et al.*, Bacterial vaginosis--a microbiological and immunological enigma. *APMIS : acta*
1878 *pathologica, microbiologica, et immunologica Scandinavica* **113**, (2005).
1879 25. I. M, G. AB, The epidemiology, pathogenesis, and diagnosis of vulvovaginal candidosis: a
1880 mycological perspective. *Critical reviews in microbiology* **37**, (2011).
1881 26. S.-M. L *et al.*, Estrogen Receptor-Alpha (ESR1) Governs the Lower Female Reproductive Tract
1882 Vulnerability to Candida albicans. *Frontiers in immunology* **9**, (2018).
1883 27. G. M, Secreted mucosal antimicrobials in the female reproductive tract that are important to
1884 consider for HIV prevention. *American journal of reproductive immunology (New York, N.Y. :*
1885 *1989)* **71**, (2014).
1886 28. A. M, F. I, L. R, F. J, B. IM, Frequency of bacteria, Candida and malassezia species in
1887 balanoposthitis. *Acta dermato-venereologica* **88**, (2008).
1888 29. C. M. WD, D. P. SS, V. PA, Semen as virus reservoir? *Journal of assisted reproduction and*
1889 *genetics* **33**, (2016).
1890 30. E. X. Castrillón-Duque, J. Puerta Suárez, W. D. Cardona Maya, Yeast and Fertility: Effects of In
1891 Vitro Activity of Candida spp. on Sperm Quality. *J Reprod Infertil* **19**, 49-55 (2018).

1892 31. A. IA, I. V, I. M, Superficial fungal infections of the male genitalia: a review. *Critical reviews in*
1893 *microbiology* **37**, (2011).
1894 32. W. GJ, T. GS, M. VK, Fungal infections of the genitourinary system: manifestations, diagnosis,
1895 and treatment. *The Urologic clinics of North America* **26**, (1999).
1896 33. M. R, Microbiota of male genital tract: impact on the health of man and his partner.
1897 *Pharmacological research* **69**, (2013).
1898 34. d. B. G *et al.*, The role of short-chain fatty acids in the interplay between diet, gut
1899 microbiota, and host energy metabolism. *Journal of lipid research* **54**, (2013).
1900 35. N. JM, V. EF, Modulation of intestinal barrier by intestinal microbiota: pathological and
1901 therapeutic implications. *Pharmacological research* **69**, (2013).
1902 36. B. AJ, S. V, Interactions between the microbiota and pathogenic bacteria in the gut. *Nature*
1903 **535**, (2016).
1904 37. T. E, J. N, Introduction to the human gut microbiota. *The Biochemical journal* **474**, (2017).
1905 38. K. CA, Inflammation and gastrointestinal Candida colonization. *Current opinion in*
1906 *microbiology* **14**, (2011).
1907 39. R. A, D. D, P. JV, W. M, K. CA, Adaptations of Candida albicans for growth in the mammalian
1908 intestinal tract. *Eukaryotic cell* **9**, (2010).
1909 40. H.-A. HE, S. MJ, Fungi in the healthy human gastrointestinal tract. *Virulence* **8**, (2017).
1910 41. R. JA, K. CA, On Commensalism of Candida. *Journal of fungi (Basel, Switzerland)* **6**, (2020).
1911 42. I. DC *et al.*, Defective trained immunity in patients with STAT-1-dependent chronic
1912 mucocutaneaous candidiasis. *Clinical and experimental immunology* **181**, (2015).
1913 43. M. L *et al.*, Pre-colonization with the commensal fungus Candida albicans reduces murine
1914 susceptibility to Clostridium difficile infection. *Gut microbes* **9**, (2018).
1915 44. H. LL, The ALS gene family of Candida albicans. *Trends in microbiology* **9**, (2001).
1916 45. Z. K *et al.*, In vivo transcript profiling of Candida albicans identifies a gene essential for
1917 interepithelial dissemination. *Cellular microbiology* **9**, (2007).
1918 46. Z. X, O. SH, Y. KM, H. LL, Analysis of the Candida albicans Als2p and Als4p adhesins suggests
1919 the potential for compensatory function within the Als family. *Microbiology (Reading,*
1920 *England)* **151**, (2005).
1921 47. Z. X, O. SH, H. LL, Deletion of ALS5, ALS6 or ALS7 increases adhesion of Candida albicans to
1922 human vascular endothelial and buccal epithelial cells. *Medical mycology* **45**, (2007).
1923 48. S. P, B. E, A. CM, Essential role of the Candida albicans transglutaminase substrate, hyphal
1924 wall protein 1, in lethal oroesophageal candidiasis in immunodeficient mice. *The Journal of*
1925 *infectious diseases* **185**, (2002).

1926 49. L.-M. E *et al.*, Candida albicans Inhibits Pseudomonas aeruginosa Virulence through
1927 Suppression of Pyochelin and Pyoverdine Biosynthesis. *PLoS pathogens* **11**, (2015).
1928 50. M. FL, W. D, H. B, Candida albicans pathogenicity mechanisms. *Virulence* **4**, (2013).
1929 51. R. JP, M. DL, Adaptive immune responses to Candida albicans infection. *Virulence* **6**, (2015).
1930 52. C. J. Nobile, A. D. Johnson, Candida albicans Biofilms and Human Disease. *Annu Rev*
1931 *Microbiol* **69**, 71-92 (2015).
1932 53. M. DL *et al.*, Candidalysin is a fungal peptide toxin critical for mucosal infection. *Nature* **532**,
1933 (2016).
1934 54. N. A *et al.*, Transcription profiling of Candida albicans cells undergoing the yeast-to-hyphal
1935 transition. *Molecular biology of the cell* **13**, (2002).
1936 55. K. D, J. AD, Induction of the Candida albicans filamentous growth program by relief of
1937 transcriptional repression: a genome-wide analysis. *Molecular biology of the cell* **16**, (2005).
1938 56. C. PL *et al.*, Expression levels of a filament-specific transcriptional regulator are sufficient to
1939 determine Candida albicans morphology and virulence. *Proceedings of the National*
1940 *Academy of Sciences of the United States of America* **106**, (2009).
1941 57. M. DL *et al.*, A biphasic innate immune MAPK response discriminates between the yeast and
1942 hyphal forms of Candida albicans in epithelial cells. *Cell host & microbe* **8**, (2010).
1943 58. P. BM *et al.*, Fungal morphogenetic pathways are required for the hallmark inflammatory
1944 response during Candida albicans vaginitis. *Infection and immunity* **82**, (2014).
1945 59. W. B *et al.*, Candida albicans-epithelial interactions: dissecting the roles of active
1946 penetration, induced endocytosis and host factors on the infection process. *PloS one* **7**,
1947 (2012).
1948 60. D. C. P *et al.*, Surgical pathology and the diagnosis of invasive visceral yeast infection: two
1949 case reports and literature review. *World journal of emergency surgery : WJES* **8**, (2013).
1950 61. M. AM *et al.*, NRG1 represses yeast-hypha morphogenesis and hypha-specific gene
1951 expression in Candida albicans. *The EMBO journal* **20**, (2001).
1952 62. R. G, M. E, J. B, W. C, L.-R. J, Our current understanding of fungal biofilms. *Critical reviews in*
1953 *microbiology* **35**, (2009).
1954 63. G. Ramage, S. P. Saville, D. P. Thomas, J. L. López-Ribot, Candida Biofilms: an Update.
1955 *Eukaryot Cell* **4**, 633-638 (2005).
1956 64. C. JW, S. PS, G. EP, Bacterial biofilms: a common cause of persistent infections. *Science (New*
1957 *York, N.Y.)* **284**, (1999).
1958 65. C. LE, The evolution of fungal drug resistance: modulating the trajectory from genotype to
1959 phenotype. *Nature reviews. Microbiology* **6**, (2008).

1960 66. A. JB, Evolution of antifungal-drug resistance: mechanisms and pathogen fitness. *Nature*
1961 *reviews. Microbiology* **3**, (2005).
1962 67. N. J *et al.*, Putative role of beta-1,3 glucans in Candida albicans biofilm resistance.
1963 *Antimicrobial agents and chemotherapy* **51**, (2007).
1964 68. Z. W, F. SG, Interactions of Candida albicans with epithelial cells. *Cellular microbiology* **12**,
1965 (2010).
1966 69. P. H *et al.*, Role of the fungal Ras-protein kinase A pathway in governing epithelial cell
1967 interactions during oropharyngeal candidiasis. *Cellular microbiology* **7**, (2005).
1968 70. D. F *et al.*, Cellular interactions of Candida albicans with human oral epithelial cells and
1969 enterocytes. *Cellular microbiology* **12**, (2010).
1970 71. V. CC, K. H, N. CJ, M. AP, D.-B. A, Mucosal tissue invasion by Candida albicans is associated
1971 with E-cadherin degradation, mediated by transcription factor Rim101p and protease Sap5p.
1972 *Infection and immunity* **75**, (2007).
1973 72. I. ID, L. I, Fungal dysbiosis: immunity and interactions at mucosal barriers. *Nature reviews.*
1974 *Immunology* **17**, (2017).
1975 73. F. Z *et al.*, Human beta-defensins: differential activity against candidal species and regulation
1976 by Candida albicans. *Journal of dental research* **84**, (2005).
1977 74. S. F *et al.*, Flagellin stimulation of intestinal epithelial cells triggers CCL20-mediated
1978 migration of dendritic cells. *Proceedings of the National Academy of Sciences of the United*
1979 *States of America* **98**, (2001).
1980 75. N. MG, M. L, Innate immune mechanisms for recognition and uptake of Candida species.
1981 *Trends in immunology* **31**, (2010).
1982 76. P. D, B. G, C. C, P. S, V. A, Candida albicans mannoprotein influences the biological function
1983 of dendritic cells. *Cellular microbiology* **8**, (2006).
1984 77. B. GD *et al.*, Dectin-1 is a major beta-glucan receptor on macrophages. *The Journal of*
1985 *experimental medicine* **196**, (2002).
1986 78. N. SL, H. A, Candida albicans is phagocytosed, killed, and processed for antigen presentation
1987 by human dendritic cells. *Infection and immunity* **69**, (2001).
1988 79. I. BZ *et al.*, Skin-resident murine dendritic cell subsets promote distinct and opposing
1989 antigen-specific T helper cell responses. *Immunity* **35**, (2011).
1990 80. B. DW, S. AG, M. HL, Growth inhibition of Candida albicans hyphae by CD8+ lymphocytes.
1991 *Journal of immunology (Baltimore, Md. : 1950)* **154**, (1995).
1992 81. d. R. L, L. D, J. P, Immunopathogenesis of oropharyngeal candidiasis in human
1993 immunodeficiency virus infection. *Clinical microbiology reviews* **17**, (2004).

82. V. B. M, W. C, Fertility and infertility: Definition and epidemiology. *Clinical biochemistry* **62**, (2018).

83. O. W, C. I, D. S, S. G, D. P, Infertility and the provision of infertility medical services in developing countries. *Human reproduction update* **14**, (2008).

84. M. MN, F. SR, B. T, V. S, S. GA, National, regional, and global trends in infertility prevalence since 1990: a systematic analysis of 277 health surveys. *PLoS medicine* **9**, (2012).

85. A. A, M. A, H. A, C. MR, A unique view on male infertility around the globe. *Reproductive biology and endocrinology : RB&E* **13**, (2015).

86. N. RD, International disparities in access to infertility services. *Fertility and sterility* **85**, (2006).

87. I. MC, P. P, Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century. *Human reproduction update* **21**, (2015).

88. G. C *et al.*, Definition and prevalence of subfertility and infertility. *Human reproduction (Oxford, England)* **20**, (2005).

89. Y. JA *et al.*, Women and Their Microbes: The Unexpected Friendship. *Trends in microbiology* **26**, (2018).

90. S. RM, A. AM, M. AM, M. ASh, Bacterial vaginosis and infertility: cause or association? *European journal of obstetrics, gynecology, and reproductive biology* **167**, (2013).

91. B. G, S. BG, S. R, R. SV, K. A, Comparative Study on the Vaginal Flora and Incidence of Asymptomatic Vaginosis among Healthy Women and in Women with Infertility Problems of Reproductive Age. *Journal of clinical and diagnostic research : JCDR* **11**, (2017).

92. M. I *et al.*, Evidence that the endometrial microbiota has an effect on implantation success or failure. *American journal of obstetrics and gynecology* **215**, (2016).

93. H. Vander, V. Prabha, Evaluation of fertility outcome as a consequence of intravaginal inoculation with sperm-impairing micro-organisms in a mouse model. *J Med Microbiol* **64**, 344-347 (2015).

94. D. T *et al.*, Influence of Escherichia coli on motility parameters of human spermatozoa in vitro. *International journal of andrology* **19**, (1996).

95. D. T *et al.*, Escherichia coli-induced alterations of human spermatozoa. An electron microscopy analysis. *International journal of andrology* **23**, (2000).

96. F. M, S.-K. A, J. P, K. M, K. M, Bacteria trigger oxygen radical release and sperm lipid peroxidation in in vitro model of semen inflammation. *Fertility and sterility* **88**, (2007).

2026 97. N. Burrello *et al.*, Candida albicans experimental infection: effects on human sperm motility,
2027 mitochondrial membrane potential and apoptosis. *Reprod Biomed Online* **18**, 496-501
2028 (2009).
2029 98. S. O. Onemu, I. N. Ibeh, Studies on the significance of positive bacterial semen cultures in
2030 male fertility in Nigeria. *Int J Fertil Womens Med* **46**, 210-214 (2001).
2031 99. Y. H. Tian, J. W. Xiong, L. Hu, D. H. Huang, C. L. Xiong, Candida albicans and filtrates interfere
2032 with human spermatozoal motility and alter the ultrastructure of spermatozoa: an in vitro
2033 study. *Int J Androl* **30**, 421-429 (2007).
2034 100. C. Rennemeier, T. Frambach, F. Hennicke, J. Dietl, P. Staib, Microbial Quorum-Sensing
2035 Molecules Induce Acrosome Loss and Cell Death in Human Spermatozoa ▿ . *Infect Immun* **77**,
2036 4990-4997 (2009).
2037 101. C. ER *et al.*, Morphology-function relationships and repeatability in the sperm of Passer
2038 sparrows. *Journal of morphology* **276**, (2015).
2039 102. S. JD, Vulvovaginal candidosis. *Lancet (London, England)* **369**, (2007).
2040 103. D. DW, K. M, S. JD, R.-R. R, Global burden of recurrent vulvovaginal candidiasis: a systematic
2041 review. *The Lancet. Infectious diseases* **18**, (2018).
2042 104. @nhsuk. (@nhsuk, 2020).
2043 105. v. S. J, Y. MH, Vulvovaginitis: screening for and management of trichomoniasis, vulvovaginal
2044 candidiasis, and bacterial vaginosis. *Journal of obstetrics and gynaecology Canada : JOGC =*
2045 *Journal d'obstetrique et gynecologie du Canada : JOGC* **37**, (2015).
2046 106. D. GGG *et al.*, Role of Molecular Biology in Diagnosis and Characterization of Vulvo-Vaginitis
2047 in Clinical Practice. *Gynecologic and obstetric investigation* **82**, (2017).
2048 107. G. AM, F. B, Risk factors for vulvovaginal candidiasis: a case-control study among university
2049 students. *Epidemiology (Cambridge, Mass.)* **7**, (1996).
2050 108. C. W, F. B, S. JD, Association of recurrent vaginal candidiasis and secretory ABO and Lewis
2051 phenotype. *The Journal of infectious diseases* **176**, (1997).
2052 109. R. DC *et al.*, Gene polymorphisms in pattern recognition receptors and susceptibility to
2053 idiopathic recurrent vulvovaginal candidiasis. *Frontiers in microbiology* **5**, (2014).
2054 110. S. JD *et al.*, Maintenance fluconazole therapy for recurrent vulvovaginal candidiasis. *The New*
2055 *England journal of medicine* **351**, (2004).
2056 111. D. G *et al.*, Individualized decreasing-dose maintenance fluconazole regimen for recurrent
2057 vulvovaginal candidiasis (ReCiDiF trial). *American journal of obstetrics and gynecology* **199**,
2058 (2008).

2059 112. D. GG, M. I, B. G, P. S, Self-elimination of risk factors for recurrent vaginal candidosis.
2060 *Mycoses* **54**, (2011).
2061 113. K. T. Potts, The Chemistry of 1,2,4-Triazoles. *Chem. Rev.* **61**, 87-127 (1961).
2062 114. R. K *et al.*, Discovery of fluconazole, a novel antifungal agent. *Reviews of infectious diseases*
2063 **12 Suppl 3**, (1990).
2064 115. L.-F. C, Triazole antifungal agents in invasive fungal infections: a comparative review. *Drugs*
2065 **71**, (2011).
2066 116. P. LR, G. S, H. M, Triazole antifungals: a review. *Drugs of today (Barcelona, Spain : 1998)* **51**,
2067 (2015).
2068 117. Z. J *et al.*, The Fungal CYP51s: Their Functions, Structures, Related Drug Resistance, and
2069 Inhibitors. *Frontiers in microbiology* **10**, (2019).
2070 118. F. DR, P. AC, Profile of isavuconazole and its potential in the treatment of severe invasive
2071 fungal infections. *Infection and drug resistance* **6**, (2013).
2072 119. G.-R. R, C.-E. M, M. E, Triazole Resistance in Aspergillus Species: An Emerging Problem. *Drugs*
2073 **77**, (2017).
2074 120. F. MG, B. M, B. P, An improved model of the Aspergillus fumigatus CYP51A protein.
2075 *Antimicrobial agents and chemotherapy* **55**, (2011).
2076 121. W. AG *et al.*, In Vitro Biochemical Study of CYP51-Mediated Azole Resistance in Aspergillus
2077 fumigatus. *Antimicrobial agents and chemotherapy* **59**, (2015).
2078 122. M. E *et al.*, A new Aspergillus fumigatus resistance mechanism conferring in vitro cross-
2079 resistance to azole antifungals involves a combination of cyp51A alterations. *Antimicrobial*
2080 *agents and chemotherapy* **51**, (2007).
2081 123. v. d. L. JW *et al.*, Aspergillosis due to voriconazole highly resistant Aspergillus fumigatus and
2082 recovery of genetically related resistant isolates from domiciles. *Clinical infectious diseases :*
2083 *an official publication of the Infectious Diseases Society of America* **57**, (2013).
2084 124. C. AT, K. M, I. F, B. J, S. D, TAC1, transcriptional activator of CDR genes, is a new transcription
2085 factor involved in the regulation of Candida albicans ABC transporters CDR1 and CDR2.
2086 *Eukaryotic cell* **3**, (2004).
2087 125. L. A, K. AF, K. O, Amphotericin B. *Applied microbiology and biotechnology* **68**, (2005).
2088 126. M.-A. AC, S. L, Z. O, It only takes one to do many jobs: Amphotericin B as antifungal and
2089 immunomodulatory drug. *Frontiers in microbiology* **3**, (2012).
2090 127. G. KC *et al.*, Amphotericin primarily kills yeast by simply binding ergosterol. *Proceedings of*
2091 *the National Academy of Sciences of the United States of America* **109**, (2012).

2092 128. H.-P. A *et al.*, Multiple functions of sterols in yeast endocytosis. *Molecular biology of the cell*
2093 **13**, (2002).
2094 129. P. DS, A. TM, B. MD, A post-PKS oxidation of the amphotericin B skeleton predicted to be
2095 critical for channel formation is not required for potent antifungal activity. *Journal of the*
2096 *American Chemical Society* **129**, (2007).
2097 130. P. AJ, S. I, R. M, Apoptosis induced by environmental stresses and amphotericin B in Candida
2098 albicans. *Proceedings of the National Academy of Sciences of the United States of America*
2099 **100**, (2003).
2100 131. O. K, R. VB, B. JF, B. RA, K. VE, Amphotericin B protects cis-parinaric acid against peroxyl
2101 radical-induced oxidation: amphotericin B as an antioxidant. *Antimicrobial agents and*
2102 *chemotherapy* **41**, (1997).
2103 132. S.-A. ML, B. J, M. G, Amphotericin B-induced oxidative damage and killing of Candida
2104 albicans. *The Journal of infectious diseases* **154**, (1986).
2105 133. V. P *et al.*, Reduced susceptibility to polyenes associated with a missense mutation in the
2106 ERG6 gene in a clinical isolate of Candida glabrata with pseudohyphal growth. *Antimicrobial*
2107 *agents and chemotherapy* **51**, (2007).
2108 134. K. PD, S. PA, M. RL, N. RD, T. BJ, A small subpopulation of blastospores in candida albicans
2109 biofilms exhibit resistance to amphotericin B associated with differential regulation of
2110 ergosterol and beta-1,6-glucan pathway genes. *Antimicrobial agents and chemotherapy* **50**,
2111 (2006).
2112 135. S. D, I. F, P. T, F. D, B. J, Candida albicans mutations in the ergosterol biosynthetic pathway
2113 and resistance to several antifungal agents. *Antimicrobial agents and chemotherapy* **47**,
2114 (2003).
2115 136. A. DM, N. C, P. J, Fluconazole at subinhibitory concentrations induces the oxidative- and
2116 nitrosative-responsive genes TRR1, GRE2 and YHB1, and enhances the resistance of Candida
2117 albicans to phagocytes. *The Journal of antimicrobial chemotherapy* **65**, (2010).
2118 137. B. KS *et al.*, Genome-wide expression profiling reveals genes associated with amphotericin B
2119 and fluconazole resistance in experimentally induced antifungal resistant isolates of Candida
2120 albicans. *The Journal of antimicrobial chemotherapy* **54**, (2004).
2121 138. H. WW, T. L, D. DW, A. MJ, Molecular mechanisms of primary resistance to flucytosine in
2122 Candida albicans. *Antimicrobial agents and chemotherapy* **48**, (2004).
2123 139. V. A, G. HJ, D. J, Flucytosine: a review of its pharmacology, clinical indications,
2124 pharmacokinetics, toxicity and drug interactions. *The Journal of antimicrobial chemotherapy*
2125 **46**, (2000).

2126 140. P. A, 5-Fluorocytosine--current status with special references to mode of action and drug
2127 resistance. *Contrib Microbiol Immunol* **4**, (1977).
2128 141. F. M, K. D, Isolation and characterization of fluoropyrimidine-resistant mutants in two
2129 Candida species. *Annals of the New York Academy of Sciences* **544**, (1988).
2130 142. F. P, W. TJ, Evolving role of flucytosine in immunocompromised patients: new insights into
2131 safety, pharmacokinetics, and antifungal therapy. *Clinical infectious diseases : an official*
2132 *publication of the Infectious Diseases Society of America* **15**, (1992).
2133 143. H. Chibana, J. L. Beckerman, P. T. Magee, Fine-resolution physical mapping of genomic
2134 diversity in Candida albicans. *Genome Res* **10**, 1865-1877 (2000).
2135 144. R. E, Chromosome instability in Candida albicans. *FEMS yeast research* **7**, (2007).
2136 145. T. A. Defosse *et al.*, [Yeasts from the CTG clade (Candida clade): Biology, impact in human
2137 health, and biotechnological applications]. *J Mycol Med* **28**, 257-268 (2018).
2138 146. K. Bouchonville, A. Forche, K. E. Tang, A. Selmecki, J. Berman, Aneuploid chromosomes are
2139 highly unstable during DNA transformation of Candida albicans. *Eukaryot Cell* **8**, 1554-1566
2140 (2009).
2141 147. M. Legrand *et al.*, Homozygosity at the MTL locus in clinical strains of Candida albicans:
2142 karyotypic rearrangements and tetraploid formation. *Mol Microbiol* **52**, 1451-1462 (2004).
2143 148. A. Kravets *et al.*, Widespread Occurrence of Dosage Compensation in Candida albicans. *PLoS*
2144 *One* **5**, (2010).
2145 149. C. Tucker *et al.*, Transcriptional Regulation on Aneuploid Chromosomes in Diverse Candida
2146 albicans Mutants. *Sci Rep* **8**, (2018).
2147 150. M. D, S. K, W. JS, S. G, Assembly of a phased diploid Candida albicans genome facilitates
2148 allele-specific measurements and provides a simple model for repeat and indel structure.
2149 *Genome biology* **14**, (2013).
2150 151. J. M. Wang, R. J. Bennett, M. Z. Anderson, K. Nielsen, The Genome of the Human Pathogen
2151 Candida albicans Is Shaped by Mutation and Cryptic Sexual Recombination. (2018).
2152 152. F. A, M. PT, M. BB, M. G, Genome-wide single-nucleotide polymorphism map for Candida
2153 albicans. *Eukaryotic cell* **3**, (2004).
2154 153. T. Jones *et al.*, The diploid genome sequence of Candida albicans. *Proceedings of the*
2155 *National Academy of Sciences of the United States of America* **101**, (2004).
2156 154. F. C. Odds *et al.*, Molecular Phylogenetics of Candida albicans. *Eukaryot Cell* **6**, 1041-1052
2157 (2007).

2158 155. J. H. Shin *et al.*, Genetic diversity among Korean Candida albicans bloodstream isolates:
2159 assessment by multilocus sequence typing and restriction endonuclease analysis of genomic
2160 DNA by use of BssHII. *J Clin Microbiol* **49**, 2572-2577 (2011).
2161 156. J. MD, B. ME, d. E. C, O. FC, Multilocus sequence typing of Candida albicans isolates from
2162 animals. *Research in microbiology* **159**, (2008).
2163 157. W. L *et al.*, Molecular phylogenetic analysis of a geographically and temporally matched set
2164 of Candida albicans isolates from humans and nonmigratory wildlife in central Illinois.
2165 *Eukaryotic cell* **7**, (2008).
2166 158. A. Forche, P. T. Magee, A. Selmecki, J. Berman, G. May, Evolution in Candida albicans
2167 Populations During a Single Passage Through a Mouse Host. *Genetics* **182**, 799-811 (2009).
2168 159. M. van het Hoog *et al.*, Assembly of the Candida albicans genome into sixteen supercontigs
2169 aligned on the eight chromosomes. *Genome Biol* **8**, R52 (2007).
2170 160. E. Sitterlé *et al.*, Within-Host Genomic Diversity of Candida albicans in Healthy Carriers. *Sci*
2171 *Rep* **9**, (2019).
2172 161. O. FC, B. AJ, G. NA, Candida albicans genome sequence: a platform for genomics in the
2173 absence of genetics. *Genome biology* **5**, (2004).
2174 162. S. K, B. M, C. J, Centromeric DNA sequences in the pathogenic yeast Candida albicans are all
2175 different and unique. *Proceedings of the National Academy of Sciences of the United States*
2176 *of America* **101**, (2004).
2177 163. C. Y, C. AP, K. E, T. A, S. NJ, Comparison of phasing strategies for whole human genomes.
2178 *PLoS genetics* **14**, (2018).
2179 164. H. JAP, D. GB, B. CM, B. D, Phased Diploid Genome Assemblies for Three Strains of Candida
2180 albicans from Oak Trees. *G3 (Bethesda, Md.)* **9**, (2019).
2181 165. B. TF, B. DCF, B. MRS, Evidence for Mitochondrial Genome Methylation in the Yeast Candida
2182 albicans: A Potential Novel Epigenetic Mechanism Affecting Adaptation and Pathogenicity?
2183 *Frontiers in genetics* **9**, (2018).
2184 166. B. C, N. SM, O. NK, H. SB, Chromatin proteomics and epigenetic regulatory circuits. *Expert*
2185 *review of proteomics* **5**, (2008).
2186 167. B. A *et al.*, High-resolution profiling of histone methylations in the human genome. *Cell* **129**,
2187 (2007).
2188 168. H. AD *et al.*, Structure of the transcriptional network controlling white-opaque switching in
2189 Candida albicans. *Molecular microbiology* **90**, (2013).
2190 169. F. C *et al.*, Epigenetic cell fate in Candida albicans is controlled by transcription factor
2191 condensates acting at super-enhancer-like elements. *Nature microbiology* **5**, (2020).

2192 170. K. J, P. S, L. JS, Epigenetic Control of Oxidative Stresses by Histone Acetyltransferases in
2193 Candida albicans. *Journal of microbiology and biotechnology* **28**, (2018).
2194 171. S. A *et al.*, Genome-wide mapping of the coactivator Ada2p yields insight into the functional
2195 roles of SAGA/ADA complex in Candida albicans. *Molecular biology of the cell* **20**, (2009).
2196 172. J. S, T. M, M. T, K. K, ATAC-Seq Identifies Chromatin Landscapes Linked to the Regulation of
2197 Oxidative Stress in the Human Fungal Pathogen Candida albicans. *Journal of fungi (Basel,*
2198 *Switzerland)* **6**, (2020).
2199 173. T. H *et al.*, Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae:
2200 implications for the microbial "pan-genome". *Proceedings of the National Academy of*
2201 *Sciences of the United States of America* **102**, (2005).
2202 174. M. D, D. C, T. H, M. V, R. R, The microbial pan-genome. *Current opinion in genetics &*
2203 *development* **15**, (2005).
2204 175. L. P, G. JP, Estimating the size of the bacterial pan-genome. *Trends in genetics : TIG* **25**,
2205 (2009).
2206 176. G. AA, B. J, E. D, Towards plant pangenomics. *Plant biotechnology journal* **14**, (2016).
2207 177. K. PJ, P. JD, Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* **9**,
2208 (2008).
2209 178. L. YH *et al.*, De novo assembly of soybean wild relatives for pan-genome analysis of diversity
2210 and agronomic traits. *Nature biotechnology* **32**, (2014).
2211 179. P. C, H. FE, C. D, Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the
2212 structural basis of a highly plastic eukaryotic genome. *BMC biology* **16**, (2018).
2213 180. S. K *et al.*, Simultaneous alignment of short reads against multiple genomes. *Genome biology*
2214 **10**, (2009).
2215 181. V. G, M. D, R. DR, T. H, Ten years of pan-genome analyses. *Current opinion in microbiology*
2216 **23**, (2015).
2217 182. K. V, G. L, D. N, O. CA, The net of life: reconstructing the microbial phylogenetic network.
2218 *Genome research* **15**, (2005).
2219 183. D. C. R, Q. Y, O. S, H. MB, H. CN, How the pan-genome is changing crop genomics and
2220 improvement. *Genome biology* **22**, (2021).
2221 184. G.-F. Richard, *Eukaryotic Pangenomes*. (Springer, 2020).
2222 185. M. Y *et al.*, A high-quality bonobo genome refines the analysis of hominid evolution. *Nature*
2223 **594**, (2021).
2224 186. A. A *et al.*, A global reference for human genetic variation. *Nature* **526**, (2015).

2225 187. G. DF *et al.*, Large-scale whole-genome sequencing of the Icelandic population. *Nature*
2226 *genetics* **47**, (2015).
2227 188. G. Marcais *et al.*, MUMmer4: A fast and versatile genome alignment system. *PLoS Comput*
2228 *Biol* **14**, e1005944 (2018).
2229 189. C. T. I. Brown, Luiz., sourmash: a library for MinHash sketching of DNA. *Journal of Open*
2230 *Source Software* **1**, 27 (2016).
2231 190. C. G. P. McCarthy, D. A. Fitzpatrick, Pangloss: A Tool for Pan-Genome Analysis of Microbial
2232 Eukaryotes. *Genes (Basel)* **10**, (2019).
2233 191. K. M, S. Y, M. K, BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of
2234 Genome and Metagenome Sequences. *Journal of molecular biology* **428**, (2016).
2235 192. M. H, M. A, E. D, H. X, T. PD, PANTHER version 14: more genomes, a new PANTHER GO-slim
2236 and improvements in enrichment analysis tools. *Nucleic acids research* **47**, (2019).
2237 193. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.
2238 (2013).
2239 194. A. McKenna *et al.*, The Genome Analysis Toolkit: a MapReduce framework for analyzing
2240 next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
2241 195. M. A. DePristo *et al.*, A framework for variation discovery and genotyping using next-
2242 generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
2243 196. G. A. Van der Auwera *et al.*, From FastQ data to high confidence variant calls: the Genome
2244 Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.11-33 (2013).
2245 197. T. H. Lee, H. Guo, X. Wang, C. Kim, A. H. Paterson, SNPhylo: a pipeline to construct a
2246 phylogenetic tree from huge SNP data. *BMC Genomics* **15**, 162 (2014).
2247 198. P. Cingolani *et al.*, A program for annotating and predicting the effects of single nucleotide
2248 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2;
2249 iso-3. *Fly (Austin)* **6**, 80-92 (2012).
2250 199. S. Purcell *et al.*, PLINK: a tool set for whole-genome association and population-based
2251 linkage analyses. *Am J Hum Genet* **81**, 559-575 (2007).
2252 200. K. Okonechnikov, A. Conesa, F. García-Alcalde, Qualimap 2: advanced multi-sample quality
2253 control for high-throughput sequencing data. *Bioinformatics* **32**, 292-294 (2016).
2254 201. A. A. Stavrou, V. Mixão, T. Boekhout, T. Gabaldón, Misidentification of genome assemblies in
2255 public databases: The case of Naumovozyma dairenensis and proposal of a protocol to
2256 correct misidentifications. *Yeast* **35**, 425-429 (2018).

2257 202. L. H, A statistical framework for SNP calling, mutation discovery, association mapping and
2258 population genetical parameter estimation from sequencing data. *Bioinformatics (Oxford,*
2259 *England)* **27**, (2011).
2260 203. L. I, B. P, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and
2261 annotation. *Nucleic acids research* **49**, (2021).
2262 204. K. DA, S. Z, L. KS, P. TD, GC content elevates mutation and recombination rates in the yeast
2263 Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United*
2264 *States of America* **115**, (2018).
2265 205. P. TD, Meiotic recombination hot spots and cold spots. *Nature reviews. Genetics* **2**, (2001).
2266 206. L. Y, M. D, D. L, GC-biased gene conversion in yeast is specifically associated with crossovers:
2267 molecular mechanisms and evolutionary significance. *Molecular biology and evolution* **30**,
2268 (2013).
2269 207. Y. JX *et al.*, Contrasting evolutionary genome dynamics between domesticated and wild
2270 yeasts. *Nature genetics* **49**, (2017).
2271 208. C. G. P. McCarthy, D. A. Fitzpatrick, Pan-genome analyses of model fungal species. *Microb*
2272 *Genom* **5**, (2019).
2273 209. H. MP *et al.*, Genetic and phenotypic intra-species variation in Candida albicans. *Genome*
2274 *research* **25**, (2015).
2275 210. E. SL, N. SM, S. NV, F. SG, J. AD, An RNA transport system in Candida albicans regulates
2276 hyphal morphology and invasive growth. *PLoS genetics* **5**, (2009).
2277 211. N. G *et al.*, Inactivation of Kex2p diminishes the virulence of Candida albicans. *The Journal of*
2278 *biological chemistry* **278**, (2003).
2279 212. R. JP *et al.*, Processing of Candida albicans Ece1p Is Critical for Candidalysin Maturation and
2280 Fungal Virulence. *mBio* **9**, (2018).
2281 213. N. JR, G. SL, H. B, Candidalysin: discovery and function in Candida albicans infections. *Current*
2282 *opinion in microbiology* **52**, (2019).
2283 214. C. J. Nobile *et al.*, A recently evolved transcriptional network controls biofilm development in
2284 Candida albicans. *Cell* **148**, 126-138 (2012).
2285 215. X. D *et al.*, Genome-wide fitness test and mechanism-of-action studies of inhibitory
2286 compounds in Candida albicans. *PLoS pathogens* **3**, (2007).
2287 216. O. J *et al.*, Gene annotation and drug target discovery in Candida albicans with a tagged
2288 transposon mutant collection. *PLoS pathogens* **6**, (2010).
2289 217. S. ES *et al.*, Gene Essentiality Analyzed by In Vivo Transposon Mutagenesis and Machine
2290 Learning in a Stable Haploid Isolate of Candida albicans. *mBio* **9**, (2018).

2291 218. P. S, H. H, I. SA, P. A, S. K, Utilization of Hybrid Assembly Approach to Determine the Genome
2292 of an Opportunistic Pathogenic Fungus, Candida albicans TIMM 1768. *Genome biology and*
2293 *evolution* **10**, (2018).
2294 219. D. G. A, B.-A. E, O. S, S. MF, Efficient hybrid de novo assembly of human genomes with
2295 WENGAN. *Nature biotechnology* **39**, (2021).
2296 220. C. Z, E. DL, M. J, Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial
2297 Pathogens Using Oxford Nanopore Sequencing. *International journal of molecular sciences*
2298 **21**, (2020).
2299 221. H. Z *et al.*, EUPAN enables pan-genome studies of a large number of eukaryotic genomes.
2300 *Bioinformatics (Oxford, England)* **33**, (2017).
2301 222. L. H, F. X, C. C, The design and construction of reference pangenome graphs with minigraph.
2302 *Genome biology* **21**, (2020).
2303 223. R. EP, H. DH, S. F, Variation in assimilating functions occurs in spontaneous Candida albicans
2304 mutants having chromosomal alterations. *Microbiology (Reading, England)* **143 (Pt 5)**,
2305 (1997).
2306 224. S. AM, D. K, C. LE, A. JB, B. J, Acquisition of aneuploidy provides increased fitness during the
2307 evolution of antifungal drug resistance. *PLoS genetics* **5**, (2009).
2308 225. A. Selmecki, A. Forche, J. Berman, Aneuploidy and Isochromosome Formation in Drug-
2309 Resistant Candida albicans. *Science* **313**, 367-370 (2006).
2310 226. M. Q, O. M, I. E, B. G, Susceptibility to Medium-Chain Fatty Acids Is Associated with Trisomy
2311 of Chromosome 7 in Candida albicans. *mSphere* **4**, (2019).
2312 227. D. F *et al.*, Oral transmission of Candida albicans between partners in HIV-infected couples
2313 could contribute to dissemination of fluconazole-resistant isolates. *AIDS (London, England)*
2314 **11**, (1997).
2315 228. L.-T. A, K. ME, E. KE, J. BS, P. GE, Trafficking through the late endosome significantly impacts
2316 Candida albicans tolerance of the azole antifungals. *Antimicrobial agents and chemotherapy*
2317 **59**, (2015).
2318 229. L. H *et al.*, Role of Arf GTPases in fungal morphogenesis and virulence. *PLoS pathogens* **13**,
2319 (2017).
2320 230. S. N, Coordination of intracellular transport steps by GTPases. *Seminars in cell &*
2321 *developmental biology* **22**, (2011).
2322 231. E. E *et al.*, Reverse genetics in Candida albicans predicts ARF cycling is essential for drug
2323 resistance and virulence. *PLoS pathogens* **6**, (2010).

232. C. J, M. TS, I. Y, M. PK, G. MA, Interaction of Candida albicans with adherent human peripheral blood mononuclear cells increases C. albicans biofilm formation and results in differential expression of pro- and anti-inflammatory cytokines. *Infection and immunity* **75**, (2007).

Supplementary Files

Supplementary files are stored at https://swanseauniversity-my.sharepoint.com/:f:/g/personal/907243_swansea_ac_uk/EmNO6hpR53VAjyTJgzO1xSUB73iAmcyzwvcYvvVP6ETHgw?e=a6pZ1A