CMEE Project

Detecting differences between strains of Cryptococcus neoformans through a new tool in detecting ploidy

Oliver Tarrant

Imperial College London, Department of Biological Sciences

MRes Computational Methods in Ecology and Evolution

Word count:

Supervised by Dr. Matteo Fumagalli

m.fumagalli@imperial.ac.uk

1 Abstract

2 Introduction

Cryptococcus neoformans is an opportunistic fungi that is responsible for up to 30% of AIDs related deaths through infections of Crytococcus Meningitis (CM) (**Vanhove2016**). Part of the reason for its success as a pathogen is its ability to evolve quickly within its hosts (**Rhodes2017**). This micro-evolution seems to occur through multiple pathways with one of the most important being through the production of copy number variation (CNV). CNV is the presence of extra or missing copies of genes within a genome (**Joao2015**). The occurrence of CNV of an entire chromosome is referred to as aneuploidy or chromosomal copy number variation (CCNV). In C.neoformans, CCNV of certain areas of the genome increase the concentration of drug resistant genes promoting more virulent stands of CM (**Rhodes2017**).

It is believed that CNV arise predominantly through misalignments during meiotic recombination and methods such as breakage-fusion-bridge cycles or non-homologous end joining (**Hastings2010**).

A recent paper on C.neoformans (**Rhodes2017**) focused on 17 HIV infected individuals from sub-Saharan Africa. Each individual had contracted CM and after initial treatment had been reinfected. The study concluded that 15 of the individuals had suffered a relapse of their original infection whilst for the remaining two, one had been initially infected by a mixture of strains whilst the infection in the other patient had developed a hyper-mutator state and thus formed a new strand. Those infections that caused relapses within the patients showed high levels of CCNV as a pathway for within host evolution to form a drug resistant isolate.

Research within this paper was restricted by the uncertainty surrounding the genotypes of the isolates. In particular .....INCLUDE RESTRICTIONS....

The aim of this study was predominately to utilise the new techniques developed to accurately detect the cases of CCNV within the dataset of c.neofromans isolates. With the resulting information, the gentic diversity of the isolates was studied providing a more realistic estimation from those derived with the original assumptions. By using the actual ploidy levels present it has also been possible to investigate the effect of including this extra level of complexity into the phylogenetic analysis of the isolates. After improving the assumptions on the phylogenetic clustering and genetic diversity more power was available to study the genes under selection within the isolates thus providing a better idea at the affects of microeveolution caused by drug pressures for the C.neoformans fungi. All together these tests were performed to quantify and qualify the differences between the separate lineages of the C.neoformans fungus and to understand the role of CCNV within these differences.

3 Methods

3.1 Data wrangling

The data is first filtered for monomorphic sites. A monomorphic site (one where only one allele can be found) is not informative of the ploidy as it would be impossible to distinguish between the likelihoods of genotypes. Thus to remove excess noise, the data is filtered so that remaining bases all have a major allele frequency of less than 0.8. Thus the remaining data are the single nucleotide polymorphisms (SNPs).

To save on computational time the triallelic and tetrallelic sites are also filtered out. The proportions of the third and fourth most common allele at the site are calculated, if greater than 0.1 then the site is also removed. Thus genotype likelihoods are calculated for these sites assuming they are biallelic. Evidently the most likely genotype at a site was triallelic then it would imply the individual had a ploidy level > 2 (and > 3 if tetrallelic). Whilst being a limitation of this method, it is likely that the added information from including these sites would be outweighed by the extra computational time required.

3.2 Genotype likelihoods

Denote by $O_{i,j}$, all the read information observed for individual i at base position j. Thus $O_{i,j}$ consists of a list of two element components, the first of which is the nucleotide reads observed, denoted $o_{i,j}$, and the second are the corresponding phred quality scores denoted $q_{i,j}$. The phred quality scores can be converted into the probability that a base has been incorrectly called by using the relationship shown in equation 1.

$$bP = 10^{-Q/10} \tag{1}$$

The genotype likelihood at base position j for an individual i is the likelihood of observing genotype $G_{i,j}$ at that base. The log-likelihood of the data given each genotype can be observed by using a baysian approach taken from (**Nielsen2011**) as outlined below:

$$ln[P(O_{i,j}|G_{i,j}, y_{i,j})] = \sum_{r=1}^{R} ln[\sum_{k=1}^{y_{i,j}} \frac{1}{y_{i,j}} P(o_{i,j,r}|g_{i,j,k}, q_{i,j,r}, y_{i,j})] \tag{2}$$

Where

$$P(o_{i,j,r}|g_{i,j,k}, q_{i,j,r}, y_{i,j}) = \begin{cases} 1 - \epsilon_{i,j,r}, & \textit{if } o_{i,j,r} = g_{i,j,k} \\ \frac{\epsilon_{i,j,r}}{3} & \textit{otherwise} \end{cases} \tag{3}$$

R is the total number of reads and $y_{i,j}$ is the ploidy level for individual i at base j. $g_{i,j,k}$ represents the k'th value of the genotype $G_{i,j}$.

3.3 Ploidy likelihood

The overall likelihood that a sample has ploidy level y_i can be calculated as shown in equation 4 by summing across the likelihoods of each genotype belonging to that ploidy at each base and then multiplying across all the bases.

$$ln[P(O_i|y_i)] = \sum_{j=1}^{N} ln[\sum_{G_{i,j} \in S_{y_i}} P(O_{i,j}|G_{i,j}, y_{i,j})P(G_{i,j}|y_{i,j})] \tag{4}$$

Here S_{y_i} is the set of genotypes that are possible for ploidy y_i, N the number of all bases in the supercontig. The value of $P(G_{i,j}|y_{i,j})$ is approximated assuming that the alleles at each base are in Hardy-Weinberg equilibrium (HWE). A major and minor allele are assigned by calculating the genotype likelihoods of each allele being haploid from $O_{i,j}$ (see 2). The allele with greatest likelihood is then classed as the major allele at that base and the second most likely the minor allele. The allele frequencies are then estimated by weighting the read available by their quality score as follows:

$$\bar{P} = P(Major\ allele\ at\ base\ j) = \frac{Sum\ of\ phred\ scores\ for\ major\ alleles\ at\ base\ j}{Sum\ of\ phred\ scores\ for\ major\ and\ minor\ alleles\ at\ base\ j}$$

$$\bar{Q} = P(Minor\ allele\ at\ base\ j) = \frac{Sum\ of\ phred\ scores\ for\ minor\ alleles\ at\ base\ j}{Sum\ of\ phred\ scores\ for\ major\ and\ minor\ alleles\ at\ base\ j} = 1 - \bar{P}$$

Thus we can apply HWE assumptions and calculate:

$$P(G_{i,j} = n_1 \times major, n_2 \times minor|y_{i,j}) = \binom{n_1+n_2}{n_1}\bar{P}^{n_1}\bar{Q}^{n_2} \tag{5}$$

To further increase the accuracy of this calculation, the co-efficient of inbreeding, F, has been included (**Vieira2013**). F is an estimate for the proportion of the population that is inbred (**Howard2017**). Thus now the frequency of $G_{i,j}$ can be expressed as a sum of terms representing the probability that the individual is not inbred with genotype $G_{i,j}$ and the term that the individual is inbred with genotype $G_{i,j}$ (6). Explicitly it has been assumed that the population is established enough that an equilibrium in the genotypes has been reached. At this stage all inbred individuals are assumed to be homozygous.

$$P(G_{i,j} = n_1 \times major, n_2 \times minor|y_{i,j}) = \begin{cases} (1-F)\binom{n_1+n_2}{n_1}\bar{P}^{n_1}\bar{Q}^{n_2} + F(\bar{P}) & if\ homozygous\ for\ major \\ (1-F)\binom{n_1+n_2}{n_1}\bar{P}^{n_1}\bar{Q}^{n_2} + F(\bar{Q}) & if\ homozygous\ for\ minor \\ (1-F)\binom{n_1+n_2}{n_1}\bar{P}^{n_1}\bar{Q}^{n_2} & if\ heterozygous \end{cases} \tag{6}$$

Note heterozygous case has no additional term for inbred individuals as probability of getting an heterozygous individual from an inbred homozygous individual is 0.

Under the assumption that the ploidy level is constant across a supercontig and that it is the same across all samples, the log ploidy likelihoods are summed over the samples. Summing over all samples would give an overall likelihood of the data given each ploidy level. Choosing the maximum of these values provides a value for the most likely ploidy of the supercontig. Thus the maximum likelihood estimation (MLE) of the ploidy level over the supercontig is:

$$Y_{MLE} = y\ where\ y\ satisfies\ Sup_y \sum_i ln[P(O_i|y)] \tag{7}$$

To provide a confidence value for the ploidy level, bootstrap analysis has been performed (**Davison1997**). Rather than summing the the entire supercontig together, it is broken into segments of 5 bi-allilic site windows. This allows the segments to be sampled at random with replacement so provide a level of confidence in the inferred ploidy. The method employed in this paper takes a random sample of 100 of these segments and sums the log likelihoods of each ploidy. The most likely ploidy is recorded and the process repeated until there are 100 ploidy levels, each inferred from 100 random segments. The distribution of the resulting inferred ploidy levels can then be used to quantify the confidence in the ploidy values.

3.4 Detecting aneuploidy

Maximum likelihood estimation (MLE) is used to infer aneuploidy within a set of sample. By considering the entire data-set, an MLE for the ploidy level of each sample is inferred as above by maximising the likelihood of the data given the ploidy level of the sample. The data is then sampled at random with replacement from all samples to retrieve an independent sample of bases. Using this independent sample of bases, a likelihood is calculated for the overall sample using the following hypohtesis:

$$H_0 : \textit{Each sample has ploidy level equal to the MLE ploidy level for the overall dataset}$$

$$H_1 : \textit{Each sample has it's own MLE ploidy level}$$

The resulting values of MLE_{H_0} and MLE_{H_1} can then be compared, in particular by studyingthe value of $\Delta = MLE_{H_1} - MLE_{H_0}$. By analysing the Δ it appears that this analysis is not suitable for a likelihood ratio test as the distribution of $\Lambda = 2\Delta$ does not follow the χ^2 distribution and thus simulations are needed to understand this statistic (**Pinheiro2000**).

Genomes were simulated for samples of 17-23 individuals with mean haploid read depths of 5,10,15,20,50,100 with a base ploidy levels of haploid, diploid and triploid and with either 0,1,2,3 or 4 samples with aneuploidy of haploid, diploid, triploid, tetraploid or pentaploid. Each scenario has 10 simulations producing a total of 3600 sets of samples. Using python's sklearn (**Sklearn2011**), 75% of the simulations have been used to fit a logistic regression classifier which is fitted on the observable independent variables of: inferred MLE ploidy under H_0, mean haploid read depth, δ and the number of samples. Where $\delta = \frac{\Delta}{\textit{Number of bases in the sample}}$ (the average deviation from H_0 caused by each base) is used to compensate for the fact that Δ will increase with the number of bases in the sample. All values used in the regression are standardised so that all variables are on the same scale and the impact of one variable is not missed due to it's scale being dominated by another. Interaction terms and polynomial terms up to degree 2 were used in the regression to provide the optimal fit.

Applied to the test dataset, the following results are obtained from using the classifiers predict function:

Table 1: AIC values for distributions fitted to genomes of uniform ploidy

		True Aneuploidy	
		Yes	No
Predicted Aneuploidy	Yes	698	7
	No	6	189

With this classifier, it is possible to predict whether or not the sample contains aneuploidy or not. Each prediction returns a probability that there is aneuploidy in a sample or not. A desired level of accuracy is chosen and whilst the probability of aneuploidy is greater than this value, the sample most likely to have aneuploidy (largest contribution to Δ) is removed from the set of samples and the remaining samples are tested again for aneuploidy. The inferred ploidy is also updated for the sub sample in case an individual with aneuploidy was skewing the results for the inferred ploidy. This iterative process is repeated until the probability of aneuploidy in the remaining samples falls below the chosen accuracy level. At this point it can be assumed that the remaining samples all have the ploidy level of that inferred for the remaining sub-sample and ploidy of the individuals with aneuploidy can be inferred as their most likely ploidy given their genotypes as shown in equation 4.

3.5 Generalising to include triallelic and tetrallelic sites

The method above is limited to the study of biallelic sites. If triallelic and tetrallelic sites were included the framework would remain the same with a few additional calculations.

Now assume there are m possible alleles at base j, $A_1, A_2, ..., A_m$ with frequencies $P_1, P_2, ..., P_m$. Thus as before the frequency P_k can be approximated as:

$$\bar{P_k} = \frac{\sum_{o_{i,j,r}=A_k} 1 - bP_{o_{i,j,r}}}{\sum_{r=1}^{R} 1 - bP_{o_{i,j,r}}} \tag{8}$$

Where $bP_{o_{i,j,r}}$ is the probability of the read $o_{i,j,r}$ being incorrectly called and $\sum_k \bar{P_k} = 1$

The generalised versions of the genotype likelihoods would be:

$$P(G_{i,j} = n_1 \times A_1, n_2 \times A_2..., n_m \times A_m | y_{i,j}) = \binom{n_1 + n_2 + ... + n_m}{n_1, n_2, ..., n_m} \prod_{k=1}^{m} \bar{P_k}^{n_k} \tag{9}$$

Where

$$\binom{n_1 + n_2 + ... + n_m}{n_1, n_2, ..., n_m} = \frac{(n_1 + n_2 + ... + n_m)!}{n_1! n_2! ... n_m!}$$

Again with the inclusion of inbreeding, the genotype likelihood is:

$$P(G_{i,j} = n_1 \times A_1, ..., n_m \times A_m | y_{i,j}) = \begin{cases} (1-F)\binom{n_1+n_2+...+n_m}{n_1,n_2,...,n_m} \prod_{k=1}^{m} \bar{P_k}^{n_k} + F(\bar{P_k}) & \textit{if homozygous for } A_k \\ (1-F)\binom{n_1+n_2+...+n_m}{n_1,n_2,...,n_m} \prod_{k=1}^{m} \bar{P_k}^{n_k} & \textit{if heterozygous} \end{cases} \tag{10}$$

4 Results

4.1 Initial checks for aneuploidy in the Cryptococcus dataset

By fitting a mixture of negative binomial distributions to the depths of each sample from the Cryptococcus dataset it is possible to visualise which samples are likely to have aneuploidy and how many different ploidy levels they are likely to contain (**Tarrant2018**).

Initial tests suggest that is aneuploidy at one level in samples CCTP50 and IRN-R21, and at 2 levels in CCTP-d257 and CCTP50-d409.

Other samples show variation in the read depths suggesting the possibility of CNV or even aneuploidy but not significant enough to be picked up by the fitted models.

4.2 Testing on simulated data

Data has been simulated to mimic the cryptococcus neoformans dataset. This dataset contains 35 samples with haploid genomes which display some levels of diploid and triploid aneuploidy (**Rhodes2017**). The samples come from 17 pairs (1 being a triplet) of before and after treatment samples. Chromosome 12 showed the most variation with 4 of the 17 pairs exhibiting aneuploidy.

5 Discussion

6 Conclusion

7 Acknowledgements